
Master in Management + Analytics Thesis

# A Machine Learning Approach for Prediction of Corrugated Cases Prices

by

**Ignacio Mera**

Advisor: **Magdalena Cornejo**

June 2021

# A Machine Learning Approach for Prediction of Corrugated Cases Prices

## Master in Management + Analytics Thesis

**Ignacio Mera**

## Abstract

Even though machine learning is widely spread among different industries, its application in the fast-moving consumer goods (FMCG) business is not a common practice and even today it remains in its early stages. Moreover, to our knowledge, there has never been a systematic approach to predict packaging materials costs in this kind of markets using machine learning algorithms, from the buyer's perspective. On the other hand, the FMCG business is a highly competitive environment, in which profitability depends not only upon sales, but also upon keeping healthy product margins. This means not only setting the right prices that consumers are willing to pay, but also getting the lowest possible costs in the supply chain. Cases usually represent between 15% and 25% of the total packaging cost, being a material with functional requirements that usually does not add value to the consumer. Therefore, it is of high importance to maintain low cases prices to achieve competitivity. In this work we propose a machine learning approach for prediction of prices of a corrugated cases portfolio of a big FMCG firm in LATAM, to understand if real prices are higher or lower than what is suggested by the model. In this way, anomalies in the dataset will be unveiled, which might become opportunities for further negotiations and costs reductions.

# Un Enfoque de Aprendizaje Automático para Predicción de Precios de Cajas Corrugadas

## Tesis de Maestría en Análisis y Gestión de Negocios

**Ignacio Mera**

## Resumen

Si bien el aprendizaje automático es una práctica ampliamente extendida entre diferentes industrias, su aplicación en el mercado del consumo masivo no es una práctica común y todavía se encuentra en sus etapas iniciales. Adicionalmente, para nuestro conocimiento, nunca hubo un enfoque sistemático para predecir costos de materiales de empaque en este tipo de mercados con algoritmos de aprendizaje automático, desde la perspectiva de la firma compradora. Por otro lado, la industria del consumo masivo es un ambiente altamente competitivo, en el que la rentabilidad no solo depende de las ventas, sino también de mantener márgenes sanos en los productos. Esto significa no solo asignar los precios correctos que los consumidores están dispuestos a pagar, sino también conseguir los menores costos posibles en la cadena de suministros. Las cajas usualmente representan entre el 15% y el 25% del costo total de empaque, siendo un material con requerimientos funcionales pero que generalmente no agrega valor al consumidor. Es por esto que es de alta importancia mantener precios bajos de cajas para lograr la competitividad. En este trabajo proponemos un enfoque de aprendizaje automático para la predicción de precios del portafolio de cajas corrugadas de una importante firma de consumo masivo en LATAM, con el objetivo de entender si los precios reales son más altos o más bajos que los sugeridos por el modelo. De esta forma, se descubrirán anomalías en el *dataset*, que podrán luego convertirse en oportunidades para futuras negociaciones y reducciones de costos.

*To my father, Mario.*

# Contents

# 1. Introduction

The fast-moving consumer goods (FMCG) business is a highly competitive environment, in which profitability depends not only upon sales, but also upon keeping healthy product margins. This means not only setting the right prices that consumers are willing to pay, but also getting the lowest possible costs in the supply chain.

The supply chain cost is mainly composed by the operating cost (production, logistics, etc.), raw material costs and packaging costs. It is essential for the business to keep these figures low to be a competitive player in the market, as lower costs derive in two possible favorable scenarios:

- Lowering prices while keeping healthy product margins, enhancing in this way sales and market share.
- Keeping current prices and incrementing product margins, enhancing in this way net revenues which can then be reinvested in the company's brands (for example, in marketing campaigns or in R&D).

In this work we will merely focus on packaging costs, more specifically on corrugated cases costs. Among the different materials that are needed for manufacturing a product to be commercialize in the market, cases are a common element, present in absolutely all SKUs. Cases are a secondary packaging, being generally not perceived by consumers. Even though these materials are generally not relevant for consumers, they are of great interest for the business, as they usually represent between 15% and 25% of the total packaging cost. Therefore, it is mandatory for all FMCG firms to get competitive cases prices.

## 1.1. Background

### 1.1.1. Cases Cost Structure

Given a specific case, with its specific technical features, a commonly accepted formula for a corrugated case is [1]:

$$Price = Fixed\ Costs\ (FC) + Variable\ Costs\ (VC) + Margin\ (M) \qquad (1)$$

Fixed costs, which can be 30-40% of the total manufacturing cost, include items which are insensitive to production rates, such as:

- Interest on loans.
- Depreciation.
- Insurance.
- Administration overhead.
- R&D.

Variable costs include items which are sensitive to production rates, such as:

- Raw materials.
- Labor.
- Maintenance.
- Energy and Freight.

It is important to notice that given formula (1), for the same case specification, it is highly expected for a FMCG firm to get different prices from different suppliers.

On the one hand, fixed costs are directly related to the financial aspects of the firm, to the size of its structure and to economies of scale. In the first place, firms that are able to get cheaper financing will be able to set more competitive prices, as they will need less capital to pay debt. In the second place, firms with bigger structures will need higher prices to get higher margins, in order to sustain the high structure cost. Finally, firms with higher levels of production will be able to distribute the fixed costs between more units, reducing then the cost per unit and being able to charge lower prices for their products [2].

On the other hand, variable costs will highly depend on economies of scale. Higher levels of production will make it possible for a firm to increment the volumes of raw materials purchases, getting then lower prices per unit. Also, as set up times will be reduced, production efficiency will increase, making it possible to optimize labor and energy, among others.

Furthermore, all these aspects will highly vary between different countries. Interest rates, labor, energy costs and freight costs will depend upon the local competitivity, the country's economy and its politics. In addition, raw materials will depend upon paper availability in the local market and the robustness of the local recycling system.

Finally, it is worth noticing that this analysis is done for a specific case with some specific technical features defined by a particular specification. It is expected for a FMCG firm to buy a portfolio of cases, as each product manufactured needs a specific case to be packed. The final price of each case will be highly affected by the specification requested, as the aspects mentioned before may vary as different technical features are needed. Therefore, each supplier will produce and sell a wide range of cases at a wide range of prices.

### 1.1.2. Machine Learning

Learning, as intelligence, covers such a broad range of processes that it is difficult to precisely define it. However, as regards machines, we might say that a machine learns whenever it changes its structure, program, or data in such a manner that its expected future performance improves. For example, when the performance of a speech-recognition machine improves after hearing several samples of a person's speech, we feel quite justified in that case to say that the machine has learned. Machine learning usually refers to changes in systems that perform tasks associated with *artificial intelligence (AI)*. Such tasks involve recognition, diagnosis, planning, robot control, prediction, etc. The changes might be either enhancements to already performing systems or *ab initio* synthesis of new systems [3]. In this work, we will solely focus on the development of new systems for predictions, and we will refer to machine learning as the development of models that, being given certain amount of robust data, are able to learn and predict an output for new data with a certain degree of confidence.

There are two types of learning within machine learning's boundaries. On the one hand, supervised learning which involves building a model for predicting or estimating an output based on one or more inputs. Problems of this nature occurs in fields as diverse as business, medicine, astrophysics, and public policy. On the other hand, unsupervised learning, which involves one or more inputs but no supervising output; nevertheless, it is possible to learn relationships and structures from such data [4]. In this work we will focus on supervise

learning, as our problem involves, given certain technical and commercial inputs for a specific corrugated case, predicting its price as a supervised output.

Machine learning supervised algorithms have proved extremely useful in forecasting. Efficiency of supervised learning in forecasting tasks has been reflected several times [5]. Moreover, there are plenty of industries that have taken advantage of machine learning algorithms to predict prices, such as finance, airline, housing and hotel industry. For example, stocks markets have been deeply disrupted by the use of machine learning models, which enables analysts to understand whether a stock is over or undervalued, and which are the probabilities for that stock's price to go up or down within the next day's [6]. However, surprisingly, machine learning is not a widespread practice in the FMCG business. There have been some intents to use machine learning for diverse tasks as demand prediction [7] or to optimize digital marketing campaigns, but it is no common practice to use it for prediction of materials costs in the supply chain. Hence, machine learning still holds a great potential to disrupt this industry and to become an important competitive advantage.

## 1.2. Justification

As discussed before, prices of corrugated cases may have high variation between different suppliers and different countries. Furthermore, FMCG firms usually buy a portfolio of different cases, as each product to be pack needs a different case with a unique specification, with specific technical features, hence with different prices. In addition, technical features may affect the final price in different ways among different countries or suppliers, depending on local capabilities or supplier's production structure. Due to this enormous complexity, a lot of effort is usually invested by *Procurement* teams to understand the drivers of corrugated cases prices in each negotiation, to pay the correct price for the correct product.

Even though machine learning is widely spread among different industries, achieving efficient results which enables firms to optimize different business aspects, its application in the FMCG business is not a common practice and even today remains in its early stages. Moreover, to our knowledge, there has never been a systematic approach to predict packaging materials costs in this kind of markets with machine learning algorithms, from the buyer's perspective. Therefore, our proposal is to develop a predicting tool which will enable corrugated cases buyers to take data driven decisions during new negotiations and to understand further opportunities within current negotiated prices. In addition, it will help to understand which are the main drivers of cases prices for different suppliers and to understand whether a supplier is correctly being chosen for a given product.

## 1.3. Objective

Concretely, the objective of this project is to develop a machine learning model that, by taking advantage of the full database of the cases portfolio of a big FMCG firm in LATAM, is capable of predicting the price of a corrugated case, given some technical and commercial features. This will be very helpful for the business, as it will allow the buyers of the respective firm to:

- Understand which is the correct price that a supplier should be asking in a negotiation regarding a new case specification. This information will enhance negotiation power of the buyers, making it possible to achieve lower prices.
- Make a revision of all portfolio's prices and find inconsistencies between reality and the model's predictions. This will impulse negotiations of current prices, finding saving opportunities.
- Understand if it is convenient for a current case, given certain technical and commercial features, to be bought from another supplier.

Finally, it will also be our objective to understand which are the main technical and commercial drivers in cases prices, to influence these features as possible in a competitive direction.

## 1.4. Size of the Opportunity

Finally, it is worth mentioning the size of the market that will be analyzed, to understand the order of magnitude that the opportunities might represent. 406 M of units are purchased each year in LATAM by the firm being analyzed, representing a total expenditure of € 71.8M. Figure 1.a shows the quantity of cases that are annually purchased by country. In addition, figure 1.b shows the total amount of money each country annually spends on these materials.

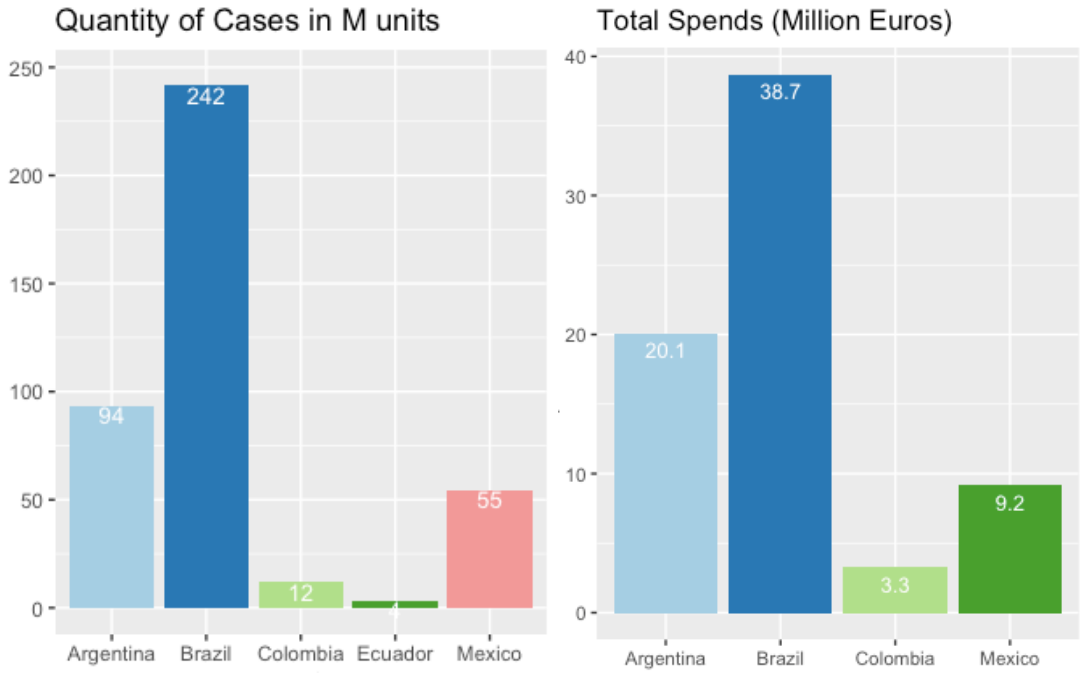


*Figure 1: a) Quantity of cases purchased by country; b) Money expenditure in cases by country.*

# 2. Methods and Procedures

## 2.1. Data

### 2.1.1. Dataset Description

The data used for the development of this project has been provided by the key corrugated cases suppliers of an international FMCG firm in LATAM, under a *non-disclosure agreement*. The dataset contains all the relevant technical and commercial information of each SKU the firm is currently buying. In particular, the dataset contains the following information.

**General Information:**

1. *country*: Argentina, Brazil, Mexico, Colombia and Ecuador.

2. *supplier*: for confidential compliance, the name of each supplier will be modified to a generic form (Supplier 1, Supplier 2, etc).

3. *plant*: name of the plant in which each SKU is delivered.

4. *material_id*: internal code of the given SKU.

5. *material_description*: description of the given SKU.

6. *category*: category for which the given SKU is bought. Deos, Hair Care, Household Care, Skin Cleansing, Skin Care, Laundry, Ice Cream, Savoury, Spreads and Dressings, Oral Care.

**Technical Information:**

7. *case_type*: in this project, we will include three types of cases:

- **Regular Cases.**
- **Shelf Ready Cases:** these are similar to regular cases, but with punched holes to easily remove the front part. Typically, being all other parameters equal, these cases are more expensive than regular cases, as they need a better-quality paper to achieve the same resistance.
- **Wrap Around Cases:** typically used for Deos category. These cases do not provide any resistance. Being all other parameters equal, they are typically more competitive than regular cases.

a) b) c)

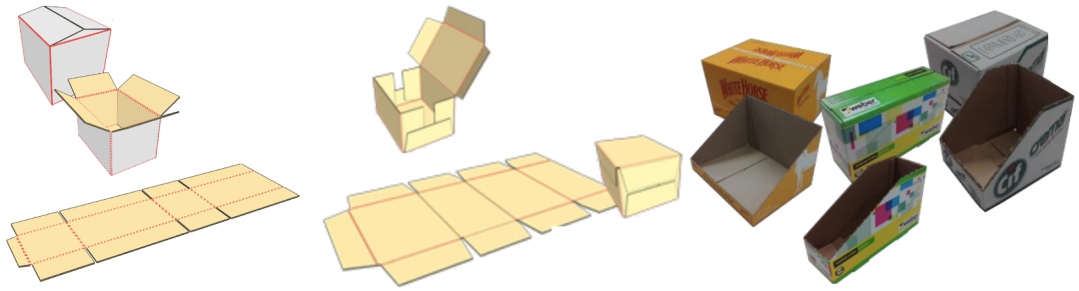

*Figure 2: a) Regular case; b) Wrap around; c) Shelf ready.*

8. *colors:* quantity of colors.

9. *printing_technology*: the technology of the printing process is specified. Flexography or Offset.

10. *flute*: the flute type determines the thickness of a case, as described in figure 3.

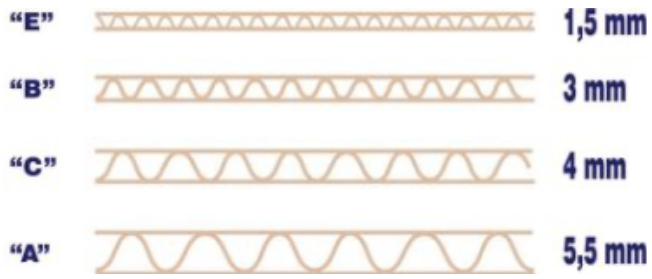


*Figure 3: Flute type and thickness.*

Different flute types can be combined, in other to achieve a better resistance in a case.

11. *white*: Yes or No. Determines if the liner to be used is whiteboard, which generally gives a premium aspect (e.g. CIF case in figure 2.c), incrementing the price of a given case.

12. *recycled*: % of recycled material used to manufacture a given SKU. It is expected to get worse mechanical properties as the percentage of recyclable material used increases.

*13. height* (in mm).

*14. width* (in mm).

*15. length* (in mm).

16. *weight* (in grams).

17. *board_area:* this parameter represents the total area of corrugated board that is needed for manufacturing a given case, in $m^2$. A good estimation for this property is given by equation (2), which will be used in this project.

$$Area = 2 * (Height * Width + Width * Height * Lenght + 2 * Width * Length) \quad (2)$$

18. *top_load:* this parameter represents the weight that a given case can support without collapsing, in kgf; in other words, the resistance of the case. It is one of the most relevant technical features to be analyzed for regular and shelf ready cases, as this resistance determines the possible palletization of the packed products. We expect that, as top load increases, so does the price of the respective case.

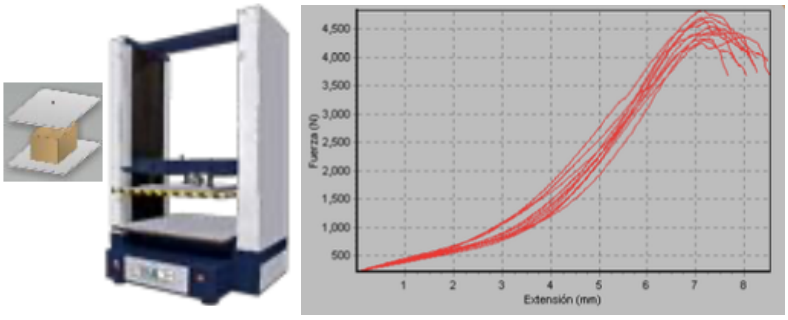


*Figure 4: Top load measurement.*

**Commercial Information:**

19. *volume*: quantity of SKUs bought during 2020. It is a good estimation of the quantity that is expected to be bought during 2021.

20. *moq*: minimum order quantity. This value represents the minimum quantity of material that can be requested for one purchase order.

21. *local_price:* price of the case in the local currency.

22. *price_eur*: price converted to euros, by a reference foreign exchange given by the firm.

23. *price_board_area:* it is a measure of the price paid, in €, for 1 $m^2$ of corrugated material, for a given SKU. Mathematically, it is the result of *price_eur / board_area.*

### 2.1.2. Dataset Size

In terms of columns, hence of variables, as described in the previous section the dataset contains 23. In terms of the quantity of rows, the entire dataset contains 1007 observations.

It is important to notice that, in general, machine learning projects use large datasets to fit their respective models. A dataset of 1007 observations is small in comparison with what is usually used in the industry. Hence, this is our main limitation. In order to have accurate results regardless of this limitation, special validation techniques will be used, which will be later described.

## 2.2. Machine Learning Techniques

### 2.2.1. Supervised Learning

Suppose that we observe a quantitative response Y and *p* different predictors $X_1$, $X_2$, …, $X_p$. We assume that there is some relationship between Y and X = ($X_1$, $X_2$, …, $X_p$), which can be written in the very general form:

$$Y = f(X) + \epsilon \tag{3}$$

Here, $f$ is some fixed but unknown function of $X_1$, …, $X_p$, and $\epsilon$ is a random error term which is independent of X and has mean zero. In this formulation, $f$ represents the systematic information that X provides about Y. $X_1$, …, $X_p$ are input variables while Y is an output variable. For the first, we will use different names as *predictors*, *independent variables*, *features* or, in sometimes, just *variables*. For the last, we will refer to *response*, *target* or *dependent variable*, as its value is not independent but rather depends upon the values of $X_1$, …, $X_p$. [8]

There are two main reasons for which we may wish to estimate $f$: *prediction* and *inference*.

#### *2.2.1.1. Prediction*

In many situations, a set of inputs X are readily available, but the output Y cannot be easily obtained. In this setting, since the error term averages to zero, we can predict Y using:

$$\hat{Y} = \hat{f}(X) \tag{4}$$

where $\hat{f}$ represents our estimation for $\hat{f}$ and $\hat{Y}$ represents the resulting prediction for Y. In this setting, $\hat{f}$ is often treated as a black box, in the sense that one is not typically concerned with the exact form of $\hat{f}$ provided that it yields accurate predictions for Y. [8]

#### *2.2.1.2. Inference*

We are often interested in understanding the way that Y is affected as $X_1$, …, $X_p$ change. In this situation we wish to estimate $f$ but our goal is not necessarily to make predictions for Y. Instead, we want to understand the relationship between X and Y or, more specifically, to understand how Y changes as a function of $X_1$, …, $X_p$. Now $f$ cannot be treated as a black box, because we need to know its exact form. In this setting, one may be interested in answering the following questions:

- Which predictors are associated with the response? It is expected that just some variables make a strong impact in the dependent variable.

- What is the relationship between the response and each predictor? Some predictors might have a positive correlation with the dependent variable, while others might have a negative correlation. Also, magnitudes will probably vary within a set of different predictors.

- Can the relationship between Y and each predictor be adequately summarized using a linear equation, or is the relationship more complicated? It is commonly desired for the correlation to be accurately explained by a linear function, as this provides an easy understanding of the impact of each predictor. However, reality is often not that simple, and more flexible models are required. [8]

### 2.2.2. Trade-off Between Prediction Accuracy and Model Interpretability

Among the different machine learning models that can be used for prediction or inference, some are less flexible, or more restrictive, in the sense that they can produce just a relatively small range of shapes to estimate $f$. For example, linear regression is a relatively inflexible approach, because it can only generate linear functions.

One might reasonably ask why we would ever choose to use a more restrictive method instead of a very flexible approach. There are several reasons for which we might prefer a more restrictive model. If we are mainly interested in inference, then restrictive models are much more interpretable. For instance, when inference is the goal, the linear model may be a good choice since it will be relatively easy to understand the relationship between Y and $X_1$, …, $X_p$. In contrast, very flexible approaches, such as the bagging methods that will be further discussed, might lead to such complicated estimates of $f$ that it is difficult to understand how any individual predictor is associated with the response. [8]

### 2.2.3. Regression vs Classification Problems

Variables can be characterized as either quantitative or qualitative (also known as categorical). Quantitative variables take on numerical values. In contrast, qualitative variables take on values in one of *K* different classes or categories. We tend to refer to problems with a quantitative response as regression problems, while those involving a qualitative response are often referred to as classification problems. [8]

### 2.2.4. Describing our Business Problem

In our problem, *price_board_area* will be the response variable, while the other variables will be the predictors. The objective of predicting *price_board_area* rather than just *price_eur* is to get a fair measure of price, independent of the quantity of material used for manufacturing a specific case. In this way, we will predict how technical and commercial features impact on the specific price of the material, rather than in a particular SKU.

Notice that our dependent variable is a quantitative one. Hence, we are in a regression problem. However, among the different predictors there are both qualitative and quantitative variables. In table 1 each variable is set in the corresponding classification.

| Qualitative Predictors | Quantitative Predictors |
|---|---|
| *country* | *colors* |
| *supplier* | *recycled* |
| *plant* | *height* |
| *material_id* | *width* |
| *material_description* | *length* |
| *category* | *weight* |
| *case_type* | *board_area* |
| *printing_technology* | *top_load* |
| *flute* | *volume* |
| *white* | *moq* |

*Table 1: Classification of predictors between qualitative and quantitative variables.*

It will be of our interest to develop not only a prediction analysis, but also an inference one. In the first case, our objective will be to predict, in terms of the commercial and technical features of a given SKU, which is the price that should be charged by our supplier. With this information we will find negotiation opportunities and understand which are the potential savings associated to these opportunities. For this analysis, more flexible models, such a *Boosting* and *Bagging* will be used, as it will be of no importance the interpretability of the model, but rather the accuracy of the predictions. In the second case, our objective will merely be to understand which are the key features that affect the final price of a given SKU. This information will be highly valuable, as it will give visibility on how a price could be reduced by slightly modifying these relevant features. For this analysis, linear methods will be used, as they will provide the simplicity and interpretability required.

## 2.3. Assessing Accuracy

### 2.3.1. Mean Square Error (MSE)

In order to evaluate the performance of a statistical learning method on a given dataset, we need some way to measure how well its predictions actually match the observed data. That is, we need to quantify the extent to which the predicted response value for a given observation is close to the true response value for that observation. In the regression setting, the most commonly used measure is the Mean Squared Error (MSE), given by:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2 \tag{5}$$

where $\hat{f}(x_i)$ is the prediction for the i[th] observation. The MSE will be small if the predicted responses are very close to the true responses and will be large if, for some observations, the predicted and true responses differ substantially. [8]

### 2.3.2. Mean Absolute Percentage Error (MAPE)

Even though MSE is a robust metric for assessing the accuracy of a model, it is unfortunately not intuitive for communicating results to high level stakeholders. Furthermore, it is scale depending, as a 10% error in a high value prediction will impact stronger in the MSE than a 10% error in a low value prediction. Therefore, in this project, we will not only focus on

measuring MSE, but will also focus on measuring more robust metrics such as the Mean Absolute Percentage Error (MAPE):

$$MAPE = \frac{1}{n}\sum_{i=1}^{n} 100 * \left|\frac{y_i - \hat{f}(x_i)}{y_i}\right| \quad (6)$$

MAPE is a percentage-based measure, not dependent on scale [9], thus more appropriate for our business problem than MSE. However, it presents two main disadvantages that will be further discussed.

### 2.3.3. Median Absolute Percentage Error (MdAPE)

One key disadvantage of MAPE is having an extremely skewed distribution when any value of $y_i$ is close to 0 [10]. Therefore, we will not only analyze MAPE but will also calculate the Median Absolute Percentage Error (MdAPE):

$$MdAPE = Median\left(100 * \left|\frac{y_i - \hat{f}(x_i)}{y_i}\right|\right) \qquad 1 \leq i \leq n \quad (7)$$

MdAPE is less sensitive to outliers than MAPE, as values close to 0 will be commonly excluded from the calculus, therefore it is more robust. However, it still exhibits a technical disadvantage, which is putting a heavier penalty on positive errors than in negative ones [10]. This observation led to the use of the so-called "symmetric" measures.

### 2.3.4. Symmetric Mean Absolute Percentage Error (sMAPE) & Symmetric Median Absolute Percentage Error (sMdAPE)

Even though MdAPE is a robust metric not very sensitive to outliers, it still holds the problem of asymmetry. Therefore, Makridakis introduced in 1993 the following two metrics [11]:

$$sMAPE = \frac{1}{n}\sum_{i=1}^{n} 200 * \left|\frac{y_i - \hat{f}(x_i)}{y_i + \hat{f}(x_i)}\right| \quad (8)$$

$$sMdAPE = Median\left(200 * \left|\frac{y_i - \hat{f}(x_i)}{y_i + \hat{f}(x_i)}\right|\right) \qquad 1 \leq i \leq n \quad (9)$$

This two metrics will also be calculated in our business problem to gain more information regarding the accuracy of the models. In particular, the sMdAPE metric given by equation (9) will be taken in high consideration and will be used as an objective criterion to decide whether a model is more accurate than other.

### 2.3.5. Training and Testing Sets

When choosing models, it is common practice to separate the available data into two portions, training and test data, where the training data is used to estimate any parameters of the model and the test data is used to evaluate its accuracy. Because the test data is not used

in determining the model, it should provide a reliable indication of how well the model is likely to perform on new data. [9]


*Figure 5: Division of dataset in training data and test data.*

The size of the test set is typically about 20% of the total sample, although this value depends on how big the sample is. The following points should be noted:

- A perfect fit with the training data can always be obtained by using a model with enough flexibility.
- A model which fits the training data well will not necessarily perform well on new data.

From now on, we will refer as *training error* to the accuracy metrics calculated with the training data, and *test error* to the accuracy metrics calculated with the test data. In addition, we will refer to the method of splitting the dataset into training data and testing data, with the purpose of measuring accuracy in the second, as the *validation set method*.

In this project, as well as in any machine learning development, the performance of our models will be mainly evaluated with the test error, as this represents how the model will perform on new unknown observations.

### 2.3.6. The Bias-Variance Trade Off

It is possible to show that the expected MSE for a new value $x_0$ can always be decomposed into the sum of three fundamental quantities: the variance of $\hat{f}(x_0)$, the squared bias of $\hat{f}(x_0)$ and the variance of the error term $\epsilon$. That is:

$$E\left(y_0 - \hat{f}(x_0)\right)^2 = Var\left(\hat{f}(x_0)\right) + \left[Bias(\hat{f}(x_0))\right]^2 + Var(\epsilon) \qquad (10)$$

Equations 7 tells us that to minimize the expected error we need to select a machine learning model that simultaneously achieves low variance and low bias.

Variance refers to the amount by which $\hat{f}$ would change if we estimated it using a different dataset. On the other hand, bias refers to the error that is introduced by approximating a real-life problem, which may be extremely complicated, by a much simpler model. As a general rule, as we use more flexible methods, the variance will increase, and the bias will decrease. The relative rate of change of these two quantities determines whether the test MSE (and other accuracy metrics) increases or decreases. [12]

Figure 6.a shows, for a new observation, the typical U-shape obtained for the MSE. As the flexibility increases, so does the variance, resulting in a higher MSE. This is called overfitting and is highly undesirable, especially when using very flexible models as bagging or boosting.

On the other side, when the flexibility is low, the MSE is high due to a high bias. This is called underfitting and is typically observed in linear models.

Figure 6.b shows the behavior of both the training error and the testing error, as the flexibility of the model increases. It can be observed that, as this happens, the training data fits better the model, resulting in a lower MSE. However, this is not true for the testing data. As discussed before, there is an inflection point at which the variance of $\hat{f}$ starts to increase faster than the reduction of the bias, resulting in a lower MSE due to overfitting.

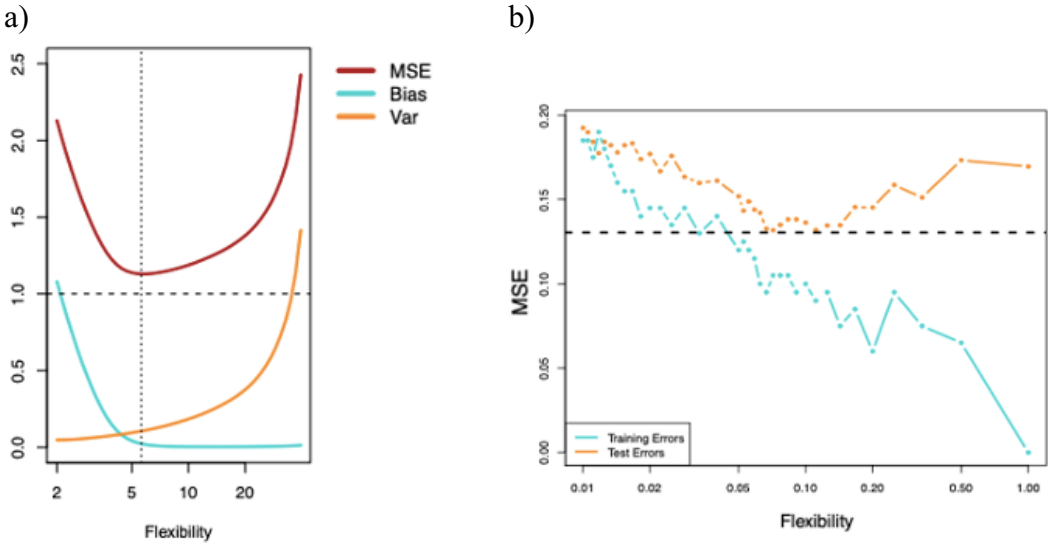


*Figure 6: a) Bias-Variance Trade-Off for a new observation;*
*b) Training Error and Testing Error vs Flexibility.*

It is important to mention that the Bias-Variance Trade Off is not only present when calculating the MSE, but the intuition is also valid for the other metrics presented in this work.

## 2.4. Cross Validation

The test error is the average error that results from using a machine learning method to predict the response on a new observation — that is, a measurement that was not used in training the method. Given a dataset, the use of a particular machine learning method is warranted if it results in a low test error. The test error can be easily calculated if a designated test set is available. Unfortunately, this is not always the case.

In the absence of a very large, designated test set that can be used to directly estimate the test error rate, several techniques can be used to estimate this quantity using the available training data. For this project, we will focus on the Leave-One-Out Cross-Validation (LOOCV) method.

### 2.4.1. Leave-One-Out Cross-Validation (LOOCV)

Like the validation set approach, LOOCV involves splitting the set of observations into two parts. However, instead of creating two subsets of comparable size, a single observation ($x_1$, $y_1$) is used for the validation set, and the remaining observations $\{(x_2, y_2), \ldots, (x_n, y_n)\}$ make

up the training set. The machine learning algorithm is fitted on the $n - 1$ training observations, and a prediction $\hat{y}_1$ is made for the excluded observation, using its value $x_1$. Since $(x_1, y_1)$ was not used in the fitting process, $MSE_1 = (y_1 - \hat{y}_1)^2$ provides an approximately unbiased estimate for the test error. But even though $MSE_1$ is unbiased for the test error, it is a poor estimate because it is highly variable, since it is based upon a single observation $(x_1, y_1)$.

We can repeat the procedure by selecting $(x_2, y_2)$ for the validation data, training the machine learning algorithm on the $n - 1$ observations $\{(x_1, y_1), (x_3, y_3), \ldots, (x_n, y_n)\}$, and computing $MSE_2 = (y_2 - \hat{y}_2)^2$. Repeating this approach $n$ times produces $n$ squared errors, $MSE_1$, …, $MSE_n$. The LOOCV estimate for the test MSE is the average of these $n$ test errors estimates:

$$CV_{(n)} = \frac{1}{n}\sum_{i=1}^{n} MSE_i \tag{11}$$

A schematic of the LOOCV approach is illustrated in figure 7.

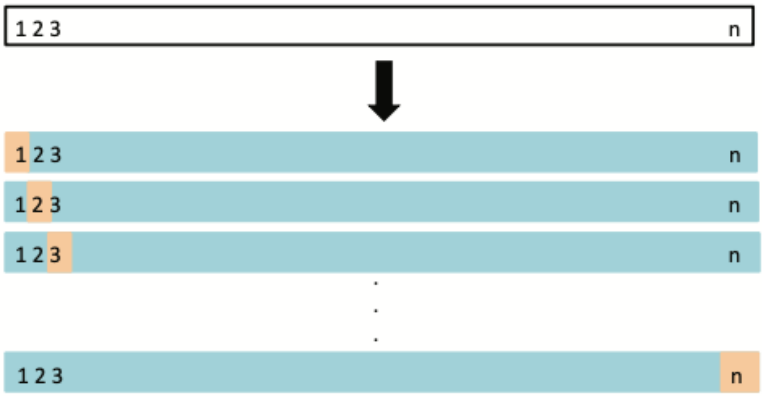


*Figure 7: Schematic display of LOOCV. A set of n data points is repeatedly split into a training set (shown in blue) containing all but one observation, and a validation set that contains only that observation (shown in beige). The first training set contains all but observation 1, the second training set contains all but observation 2, and so forth.*

Even though the LOOCV approach was explained using the MSE accuracy metric, it is extensive to the rest of the metrics presented in this project. The formulas are the same as the ones presented in section 2.3, but taking in consideration that, for the calculus of each $\hat{f}(x_i)$, a particular i-model with the rest of the observations must be trained.

### 2.4.2. LOOCV vs Validation Set Approach

LOOCV has a couple of major advantages over the validation set approach. First, it has far less bias. In LOOCV, we repeatedly fit the machine learning algorithm using training sets that contain $n - 1$ observations, almost as many as are in the entire dataset. This differs from the testing set approach, in which the training set is typically around 80% the size of the original dataset. Consequently, the LOOCV approach tends not to overestimate the test error rate as much as the testing set approach does. Second, in contrast to the validation approach, which will yield different results when applied repeatedly due to randomness in

the training/validation set splits, performing LOOCV multiple times will always yield the same result. There is no randomness in the training/validation set splits. [13]

Even though the LOOCV approach has many advantages over the validation set approach, it is not commonly used because of a main disadvantage: time. As it might be intuited, this method consumes a lot of time and computational work, as instead of fitting one model, $n - 1$ models must be fitted. Hence, given a large dataset, using this LOOCV method becomes unpractical. In addition, the validation set performs good when a large test set can be supplied, therefore in such cases this method is not recommended, as the accuracy benefit is low in comparison with the time penalty.

#### 2.4.3. LOOCV for our Business Problem

As it was mentioned in section 2.1.2, our dataset contains only 1007 observations. This number is low in comparison with the datasets that are usually used to fit machine learning algorithms and represents the main limitation of this project. Therefore, it was decided not to use the validation set approach for measuring test error, but rather using the LOOCV.

The LOOCV approach will give us a more appropriate test error than a validation set approach, as the bias will be far lower. Furthermore, as the dataset is not large, the time penalty for using this method will not be significant.

### 2.5. Machine Learning Models & Algorithms

#### 2.5.1. Linear Regression

Linear regression is a very simple approach for supervised learning, useful for predicting a quantitative response. Linear regression has been around for a long time and is the topic of innumerable textbooks. Though it may seem somewhat dull compared to some of the more modern machine learning approaches, linear regression is still a useful and widely used statistical learning method for inference, due to its high interpretability and easy communication. Being said that, linear regression is an excellent starting point for understanding which key features are involved in an unexplored business problem, and their potential positive or negative impact on a dependent variable.

If we analyze the linear regression from the Bias-Variance Trade Off perspective, it is a method with very low variance but very high bias. This means that, if we splitted the dataset in two randomly, we would probably get the same result for both datasets. However, both results will probably be different than the real value that we are looking for.

##### *2.5.1.1. Linear Regression with a Single Regressor*

Simple linear regression lives up to its name: it is a very straightforward approach for predicting a quantitative response Y on the basis of a single predictor variable X. It assumes that there is approximately a linear relationship between X and Y. Mathematically, we can write this linear relationship as follows:

$$Y \approx \beta_0 + \beta_1 X \tag{12}$$

In equation (12) $\beta_0$ and $\beta_1$ are two unknown constants that represent the intercept and slope terms in the linear model. Together, $\beta_0$ and $\beta_1$ are known as the model coefficients or parameters. Once we have used our training data to produce estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ for the model coefficients, we can predict future values of Y on the basis of a particular value of X, by computing:

$$\hat{y} \approx \hat{\beta}_0 + \hat{\beta}_1 x \tag{13}$$

where $\hat{y}$ indicates a prediction of Y on the basis of X = x.

Our goal is to obtain coefficients estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ such that the linear model fits the available data well, that is: $\hat{y}_i \approx \hat{\beta}_0 + \hat{\beta}_1 x_i$ for *i = 1, …, n*. In other words, we want to find an intercept $\hat{\beta}_0$ and a slope $\hat{\beta}_1$ such that the resulting line is as close as possible to the *n* data points. There are several ways of measuring closeness. However, by far the most common approach involves minimizing the least squares criterion. [14]

Let $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ be the prediction for Y based on the i$^{th}$ value of X. Then $e_i = y_i - \hat{y}_i$ represents the i$^{th}$ residual, that is to say, the difference between the i$^{th}$ observed response value and the i$^{th}$ response value that is predicted by our linear model. We define the residual sum of squares (RSS) as follows:

$$RSS = {e_1}^2 + {e_2}^2 + \cdots + {e_n}^2 \tag{14}$$

Or equivalent:

$$RSS = \left(y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)^2 + \left(y_2 - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)^2 + \cdots + \left(y_n - \hat{\beta}_0 - \hat{\beta}_1 x_n\right)^2 \tag{15}$$

#### 2.5.1.2. Multiple Linear Regression

Simple linear regression is a useful approach for predicting a response on the basis of a single predictor variable. However, in practice we often have more than one predictor. Linear regression can be extended so that it can directly accommodate multiple predictors. We can do this by giving each predictor a separate slope coefficient in a single model. In general, suppose that we have *p* distinct predictors. Then the multiple linear regression model takes the following form:

$$Y \approx \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p \tag{16}$$

where $X_j$ represents the j$^{th}$ predictor and $\beta_j$ quantifies the association between that variable and the response. We interpret $\beta_j$ as the average effect on Y of a one unit increase in $X_j$, holding all other predictors fixed.

As was the case in the simple linear regression setting, the regression coefficients $\beta_0, \beta_1, \ldots, \beta_p$ are unknown and must be estimated. Given estimates $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$, we can make predictions using the following formula:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i,1} + \hat{\beta}_2 x_{i,2} + \cdots + \hat{\beta}_p x_{i,p} \tag{17}$$

For the coefficient's estimation, the same least squares approach used for linear regression can be used [14]. We choose $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$ to minimize the sum of squared residuals:

$$RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i,1} - \hat{\beta}_2 x_{i,2} - \cdots - \beta_p x_{i,p}\right)^2 \quad (18)$$

#### 2.5.1.3. Dummy Predictor Variables

Up to this point, we have considered the independent variables $X_1, \ldots, X_p$ to be numerical. However, this is not always the case. For example, in our project we have categorical variables such as *country* and *supplier*, which will probably have a meaningful impact on the target variable *price_board_area*, and therefore we would like to include in our models.

To do this, we will proceed with a *Dummy Coding* technique, which consists in creating, for each categorical variable, *m – 1* new variables, being *m* the number of categories that the corresponding categorical variable can take [15]. For example, being the possible values for the variable *country:* Argentina, Mexico and Brazil; the Dummy Coding technique will consist in creating two new variables *country_argentina* ($D_A$) and *country_mexico* ($D_M$), with the following regressors associated: $\beta_A$ and $\beta_M$. Having *q* numerical variables, in this particular case, equation (16) would then be modified as follows:

$$Y \approx \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_q X_q + \beta_A D_A + \beta_M D_M \quad (19)$$

$$D_A = \begin{matrix} 1 \; if \; the \; observation \; is \; from \; Argentina \\ 0 \; if \; not \end{matrix} \quad (20)$$

$$D_M = \begin{matrix} 1 \; if \; the \; observation \; is \; from \; Mexico. \\ 0 \; if \; not \end{matrix} \quad (21)$$

In this way, $\beta_A$ and $\beta_M$ will represent how much (or less) a case will cost in Argentina and in Mexico respectively, compared to the benchmark Brazil. It is important to mention that, as this technique requires to create *m – 1* new variables for each possible value a categorical variable can take, we will always need to arbitrarily determine a benchmark value for each categorical variable. In this case, the country Brazil.

Being calculated the estimators $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p, \hat{\beta}_A$ and $\hat{\beta}_M$ for the respective regressors, predictions can be made with the following formula:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i,1} + \hat{\beta}_2 x_{i,2} + \cdots + \hat{\beta}_p x_{i,p} + \hat{\beta}_A D_A + \hat{\beta}_M D_M \quad (22)$$

This procedure can be applied to all the categorical variables within our business problem, including *m – 1* new binary variables for each categorical variable.

#### 2.5.1.4. Interpretation of Regressors

Given an estimated parameter $\hat{\beta}_1$, it can be estimated that a $\Delta X_1$ change will produce a $\hat{\beta}_1 \Delta X_1$ change in the dependent variable Y. For example, if the variable $X_1$ is increased in one unit, the variable Y will increase (or decrease) according to the magnitude and sign of $\hat{\beta}_1$, all else

being equal (i.e. holding all other variables fixed at their observed values). Therefore, the value of $\hat{\beta}_1$ can be interpreted as how the independent variable $X_1$ affects the dependent variable Y. If $\hat{\beta}_1$ is greater than 0, $X_1$ will have a positive impact on Y. In an analogous way, if $\hat{\beta}_1$ is lower than 0, $X_1$ will have a negative impact on the dependent variable Y. In addition, as bigger is the absolute value of $\hat{\beta}_1$, bigger will be the positive or negative impact that the associated variable will have on Y.

However, knowing the value of $\hat{\beta}_1$ is not enough for assessing the influence of the corresponding independent variable on the dependent variable. It is also important to understand the distribution of $\hat{\beta}_1$ and which is the confidence of the calculated value. For this we will use confidence intervals and hypothesis tests, reporting in each case the *p-values* obtained.

### 2.5.2. Log-Log Regression Models and Elasticity

In our business problem, it will be important not only to understand the marginal impact of a particular variable in the dependent variable *price_board_area*, but also the percentual impact. For example, we could ask ourselves the following questions:

- How less will the same case cost if we increased the volume by 5%?
- How less will a case cost if we reduced the top load by 10%?

Even though an answer in $\$/m^2$ can be helpful to answer these questions, it is much more interpretable and easier to communicate to stakeholders a percentual value. Therefore, we will use log-log regression models. Logarithms convert changes in variables into percentage changes. Hence, regression specifications that use natural logarithms allow regression models to estimate percentage relationships. [16]

The following relation, with multiple regressors, can be established:

$$\ln(Y) \approx \beta_0 + \beta_1 \ln(X_1) + \cdots + \beta_p \ln(X_p) \tag{23}$$

In the log-log regression model, a 1% change in $X_i$ is associated with a $\beta_i$% change in Y. In other terms, $\beta_i$ represents the elasticity of Y to the variable $X_i$, and can be defined as follows:

$$\beta_i = \frac{\Delta Y}{\Delta X_i} \tag{24}$$

As was the case in the linear regression setting, the regression coefficients $\beta_0, \beta_1, \ldots, \beta_p$ are unknown and must be estimated. Given the estimates $\hat{\beta}_0$, $\hat{\beta}_1$, …, $\hat{\beta}_p$, we can make predictions using the following formula:

$$\ln(\hat{y}_i) = \hat{\beta}_0 + \hat{\beta}_1 \ln(x_{i,1}) + \cdots + \hat{\beta}_p \ln(x_{i,p}) \tag{25}$$

It is important to mention that equation (23) is only valid when the $X_1$, …, $X_p$ independent variables are numerical. For example, when assessing elasticity to *top_load* or *volume*, as given in the example questions. However, when there are categorical independent variables, such as *country* and *supplier*, a different approach must be implemented.

As described in section 2.5.1.3, when leading with categorical variables the Dummy Coding technique results appropriate to incorporate them in the corresponding regression models. Therefore, given $X_1$, …, $X_q$ numerical independent variables and $D_{q+1}$, …, $D_p$ dummy variables created after the possible values the respective categorical variables can take (without counting the benchmarks), we can establish the following relationship:

$$\ln(Y) \approx \beta_0 + \beta_1 \ln(X_1) + \cdots + \beta_q \ln(X_q) + \beta_{q+1} D_{q+1} + \cdots + \beta_p D_p \quad (26)$$

Hence, when estimated the parameters $\beta_0, \beta_1, \ldots, \beta_q, \beta_{q+1}, \ldots, \beta_p$, we can make predictions according to the following equation:

$$\ln(\hat{y}_i) = \hat{\beta}_0 + \hat{\beta}_1 \ln(x_{i,1}) + \cdots + \hat{\beta}_q \ln(x_{i,q}) + \hat{\beta}_{q+1} D_{q+1} + \cdots + \hat{\beta}_p D_p \quad (27)$$

For instance, if $D_{q+1}$ represents the dummy variable *country_argentina*, associated with the categorical variable *country,* the regressor $\beta_{q+1}$ will represent how much, in percentual terms, the same case will cost in Argentina versus an arbitrarily chosen benchmark.

### 2.5.3. Tree-Based Methods

Even though linear and log-log regression models are easy to interpret, hence, are appropriate for inference analysis, their power of prediction is low compared to more sophisticated machine learning methods. Therefore, for prediction analysis, we will focus on tree-based methods.

#### *2.5.3.1. Decision Trees*

*Decision Trees* is a machine learning method with a low prediction power. However, it is the basis for much powerful algorithms, such as *Random Forest* and *XGBoost*, which will be further explored. Therefore, a short introduction is required.

As well as in linear methods there was an error function that needed to be minimize (RSS, corresponding to equations (15) and (18)), the main objective of Decision Trees will also be to minimize a defined error function. In this case, the predictor space defined by the set of possible values for $X_1$, …, $X_p$, will be divided into *J* distinct and non-overlapping regions $R_1$, …, $R_J$. Then, for every observation that falls into the region $R_j$, we make the same prediction, which is simply the mean of the response values for the training observations in $R_j$ [17]. Mathematically speaking, the predictor space will be divided in order to minimize the following RSS value:

$$RSS = \sum_{j=1}^{J} \sum_{i \in R_j} \left(y_i - \hat{y}_{R_j}\right)^2 \quad (28)$$

where $\hat{y}_{R_j}$ is the mean response for the training observations within the $j^{th}$ box.

Unfortunately, it is computationally infeasible to consider every possible partition of the feature space into *J* boxes. For this reason, a top-down, greedy approach is taken, which is known as recursive binary splitting. The approach is top-down because it begins at the top of the tree (at which point all observations belong to a single region) and then successively

splits the predictor space; each split is indicated via two new branches further down on the tree. It is greedy because at each step of the tree-building process, the best split is made at that particular step, rather than looking ahead and picking a split that will lead to a better tree in some future step.

To perform recursive binary splitting, we first select the predictor $X_j$ and the cutpoint $s$ such that splitting the predictor space into the regions $R_1(j,s) = \{X|Xj < s\}$ and $R_2(j,s) = \{X|Xj > s\}$ leads to the greatest possible reduction in RSS. The new RSS value will then be:

$$RSS = \sum_{i:\, x_i \in R_1(j,s)} \left(y_i - \hat{y}_{R_1}\right)^2 + \sum_{i:\, x_i \in R_2(j,s)} \left(y_i - \hat{y}_{R_2}\right)^2 \qquad (29)$$

This process continues until a stopping criterion is reached.

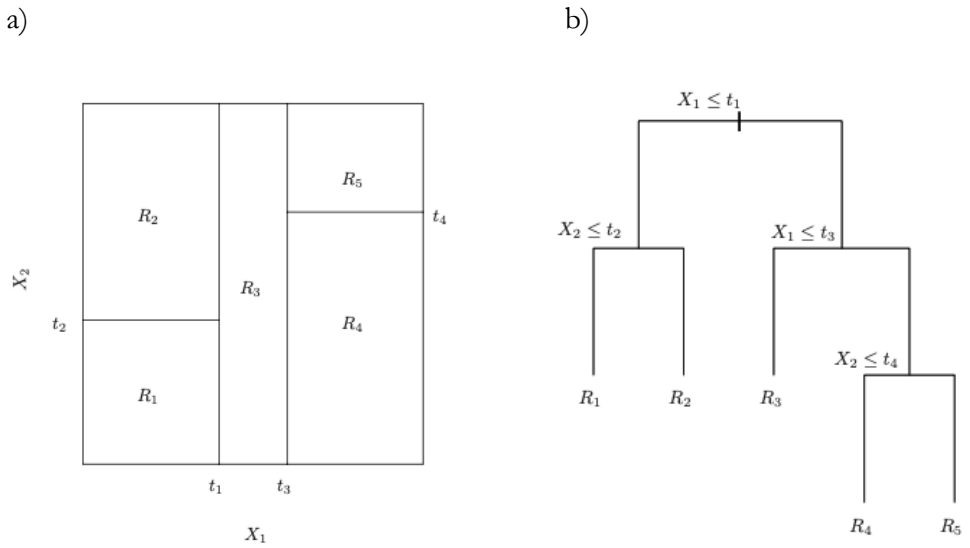


*Figure 8: Example of a decision tree. a) Partition of a two-dimensional predictor space; b) Tree representation of the corresponding partition.*

We will refer to the starting point of the tree as the *root*, and to the final predictor spaces $R_1$, …, $R_J$, as *terminal nodes* or *leaves*. The points along the tree where the predictor space is splitted are referred to as *internal nodes*.

#### 2.5.3.2. Stopping Criterions for Decision Trees

A very deep tree might lead to overfitting, which is highly undesirable. There are many ways to prevent this. In this project we will mainly use the following stopping criterions, as they are the most common in the industry [18]:

i) **Minsplit:** the minimum number of observations that must exist in a node for a split to be attempted.
ii) **Minbucket**: the minimum number of observations in any leaf for a split to be attemped.
iii) **Maxdepth**: maximum depth of any terminal node of the final tree, with the root node counted as depth 0.
iv) **Complexity parameter ($\alpha$)**: the error function is modified as:

$$RSS = \sum_{m=1}^{|T|} \sum_{x_i \in R_j} \left(y_i - \hat{y}_{R_j}\right)^2 + \alpha|T| \tag{30}$$

|T| indicates the number of terminal nodes of the tree T. The tuning parameter $\alpha$ controls a trade-off between the subtree's complexity and its fit to the training data. As $\alpha$ increases, there is a price to pay for having a tree with many terminal nodes, and, therefore, the quantity RSS defined in equation (30) will tend to be minimized for a smaller subtree.

#### *2.5.3.3. Bagging and Bootstrap*

Even though Decision Trees is a method with low bias, it suffers from high variance (at the contrary of linear regression models). This means that if we splitted the training data into two parts at random and fitted a decision tree to both halves, the results that we would get would probably be significantly different. In contrast, a procedure with low variance will yield similar results if applied repeatedly to distinct datasets. [19]

Given a set of *n* independent observations $z_1, ..., z_n$, each with variance $\sigma$, the variance of the mean $\bar{Z}$ of the observations is given by $\sigma^2/n$. In other words, averaging a set of observations reduces variance. Therefore, a natural way to reduce the variance and hence increase the prediction accuracy of a statistical learning method is to take many training sets from the population, build a separate prediction model using each training set, and average the resulting predictions. In other words, we could calculate $\hat{f}^1(x), \hat{f}^2(x), \ldots, \hat{f}^B(x)$ using *B* separate training sets, and average them in order to obtain a single low-variance statistical learning model, given by:

$$\hat{f}_{avg}(x) = \frac{1}{B}\sum_{b=1}^{B} \hat{f}^b(x) \tag{31}$$

However, this is not practical because we generally do not have access to multiple training sets. Instead, we can bootstrap. *Bootstrap* is a resampling method approach that allows us to use a computer to emulate the process of obtaining new sample sets $Z^{*1}, Z^{*2}, \ldots, Z^{*B}$ from a unique dataset $Z$. It consists in randomly selecting observations from the original dataset, with reposition, to build the *B* bootstrap datasets $Z^{*1}, Z^{*2}, \ldots, Z^{*B}$ [20]. An illustrative example for this method can be observed in figure 9.

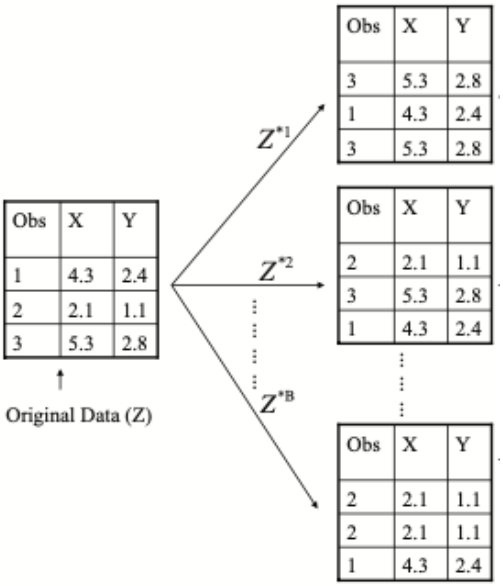


*Figure 9: Bootstrap example. Three bootstrap datasets, $Z^{*1}$, $Z^{*2}$ and $Z^{*B}$ are created from $Z$, by randomly selecting 3 samples with reposition in each case.*

Bagging mainly consists of generating *B* bootstraps datasets from an original dataset and training our method on the b[th] bootstrapped training set to get $\hat{f}^{*b}(x)$. Finally, all the predictions are averaged by:

$$\hat{f}_{bag}(x) = \frac{1}{B}\sum_{b=1}^{B}\hat{f}^{*b}(x) \tag{32}$$

Bagging is particularly useful when working with deep decision trees. Each individual tree has high variance but low bias. Averaging these *B* trees reduces the variance. Bagging has been demonstrated to give impressive improvements in accuracy by combining hundreds or even thousands of trees into a single procedure.

#### *2.5.3.4. Random Forest*

Bagged trees predictions are highly correlated, as each individual tree uses the same strong predictors; hence, all the trees will look quite similar to each other. Unfortunately, averaging many highly correlated quantities does not lead to a large reduction in variance as averaging many uncorrelated quantities. In particular, this means that bagging will not lead to a substantial reduction in variance over a single tree in this setting. [19]

Random Forest overcomes this problem by way of a small tweak that decorrelates the trees. As in bagging, we build a number of decision trees on bootstrapped training samples. But when building these decision trees, each time a split in a tree is considered, a random sample of *m* predictors is chosen from the full set of *p* predictors to fit the model. In this way, trees are decorrelated, thereby making the average of the resulting trees less variable and hence more reliable. Algorithm 1 describes the method for developing Random Forest.

**Algorithm** *Random Forest for Regression or Classification.*

1. For $b = 1$ to $B$:
   (a) Draw a bootstrap sample $\mathbf{Z}^*$ of size $N$ from the training data.
   (b) Grow a random-forest tree $T_b$ to the bootstrapped data, by recursively repeating the following steps for each terminal node of the tree, until the minimum node size $n_{min}$ is reached.
      i. Select $m$ variables at random from the $p$ variables.
      ii. Pick the best variable/split-point among the $m$.
      iii. Split the node into two daughter nodes.
2. Output the ensemble of trees $\{T_b\}_1^B$.

To make a prediction at a new point $x$:

*Regression:* $\hat{f}_{\mathrm{rf}}^B(x) = \frac{1}{B}\sum_{b=1}^{B} T_b(x)$.

*Algorithm 1: Random Forest. Source: [21].*

#### 2.5.3.5. Boosting

Bagging involves creating multiple copies of the original training dataset using the bootstrap technique, fitting a separate decision tree to each copy, and then combining all the trees in order to create a single predictive model. Notably, each tree is built on a bootstrap dataset, independent of the other trees. Boosting works in a similar way, except that the trees are grown sequentially: each tree is grown using information from previously grown trees. Boosting does not involve bootstrap sampling; instead, each tree is fitted on a modified version of the original dataset.

Like bagging, boosting involves combining a large number of decision trees $\hat{f}^1(x)$, $\hat{f}^2(x)$, …, $\hat{f}^B(x)$. Unlike fitting a single large decision tree to the data, which amounts to fitting the data hard and potentially overfitting, the boosting approach instead learns slowly. Given the current model, we fit a decision tree to the residuals from the model. That is, we fit a tree using the current residuals, rather than the outcome Y, as the response. We then add this new decision tree into the fitted function to update the residuals. By fitting small trees to the residuals, we slowly improve $\hat{f}$ in areas where it does not perform well. The shrinkage parameter λ slows the process down even further, allowing more and different shaped trees to attack the residuals. In boosting, unlike bagging, the construction of each tree depends strongly on the trees that have already been grown. The method to perform boosting is slightly more difficult than the method for bagging and Random Forest; it is described in algorithm 2.

**Algorithm** *Boosting for Regression Trees*

1. Set $\hat{f}(x) = 0$ and $r_i = y_i$ for all $i$ in the training set.
2. For $b = 1, 2, \ldots, B$, repeat:
   (a) Fit a tree $\hat{f}^b$ with $d$ splits ($d+1$ terminal nodes) to the training data $(X, r)$.
   (b) Update $\hat{f}$ by adding in a shrunken version of the new tree:
   $$\hat{f}(x) \leftarrow \hat{f}(x) + \lambda \hat{f}^b(x).$$
   (c) Update the residuals,
   $$r_i \leftarrow r_i - \lambda \hat{f}^b(x_i).$$
3. Output the boosted model,
   $$\hat{f}(x) = \sum_{b=1}^{B} \lambda \hat{f}^b(x).$$

*Algorithm 2: Boosting regression trees. Source:* [19].

## 2.6. Hyperparameters

Every model has regularization parameters that are not learned automatically in the model training process. Rather, they must be tuned on the problem at hand and given to the training algorithm. These parameters are called *hyperparameters* [15] and will be of high importance when tuning tree-based models.

### 2.6.1. Hyperparameters for Tree-Based Models

#### *2.6.1.1. Hyperparameters for Random Forest*

In Random Forest, the quantity of trees used for predicting does not produce overfitting. The following hyperparameters will be used for this technique [22]:

- **ntree**: number of trees grown.
- **mtry**: number of variables randomly sampled at each split.
- **sampsize**: size(s) of sample(s) to draw in each tree.
- **nodesize**: minimum size of terminal nodes for each tree.
- **maxnodes**: maximum number of terminal nodes trees in the forest can have.

#### *2.6.1.2. Hyperparameters for Boosting*

Different from Random Forest, the quantity of trees used for boosting might lead to overfitting. In particular, the following hyperparameters will be explored [23].

Hyperparameters which make the model more flexible, hence increasing the probability of overfitting:

- **nround**: number of trees.
- **max_depth**: max deepness that the trees can reach.
- **eta (λ):** shrinkage parameter.
- **colsample_bytree**: quantity of variables to randomly select in each tree.
- **subsample**: size(s) of sample(s) to draw in each tree.

Hyperparameters which make the model less flexible, hence reducing the probability of overfitting:

- **gamma**: minimum error reduction for making a split.
- **min_child_weight**: minimum size of terminal nodes.

### 2.6.2. Hyperparameters Search

#### *2.6.2.1. Grid Search*

A grid search consists in defining a grid of possible hyperparameters and using an algorithm to test the performance of the corresponding model trained with each of the possible hyperparameters set [15].

#### *2.6.2.2. Random Search*

Different to grid search, in which the possible hyperparameter are pre-defined, in a random search the hyperparameters are chosen randomly *n* times. After the model is trained and tested with the *n* sets of randomly chosen hyperparameters, we choose the ones that delivered the best performance.

In this project, we will rather use random search than grid search, as it is more robust to the presence of irrelevant hyperparameters. Also, it can accelerate the search process [24].

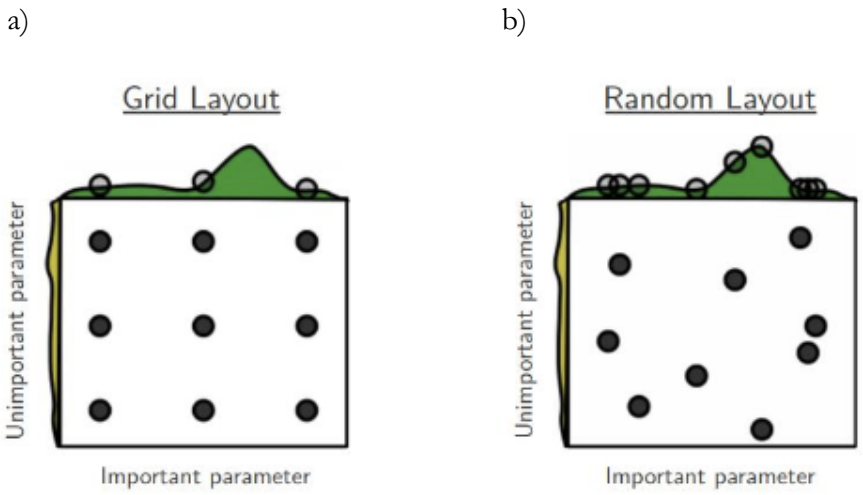


*Figure 10: Schematic view of: a) Grid layout; b) Random layout.*

# 3. Packaging Prices Prediction: State of the Art

Even though cases represent between 15% and 25% of the total packaging cost in the FMCG industry, there has never been a systematic approach to predict these materials costs in this kind of markets, from the buyer's perspective. Therefore, we will analyze in this section the state of the art regarding packaging costs evaluation in the entire supply chain, and existing efforts to use machine learning techniques for prediction of packaging costs, from the manufacturer's perspective.

## 3.1. Packaging Costs in the Supply Chain

There have been some efforts to evaluate the total packaging costs along the supply chain. Regattieri et al. (2012) developed an integral model which includes all costs related to packaging purchases and elaborations [25]. This model is described below:

$$\begin{aligned} C_{TOT} = {} & C_{ENG} + C_{ORD} + C_{PUR} + C_{RENT} + C_{EXT\,TRAN} + C_{REC} + C_{COND} + C_{INT\,TRAN} \\ & + C_{STOCK} + C_{PICK} + C_{INT\,TRAN1} + C_{MAN} + C_{REV1} + C_{INT\,TRAN2} \\ & + C_{STOCK1} + C_{PICK1} + C_{INT\,TRAN3} + C_{REV2} + C_{RE-USE} + C_{DISP} - R_{SUB} \\ & - R_{UDC} \end{aligned} \quad (33)$$

$C_{ENG}$ - Cost of Engineering: Cost for studying each type of packaging and for making prototypes. It includes the labor costs of engineering the product.

$C_{ORD}$ - Cost of Purchase Order: Cost for managing the internal purchase orders if the manufacturer produces the packaging internally; otherwise, it represents the purchase orders for buying and/or renting packaging from suppliers. It includes the labor costs for making the order.

$C_{RENT}$ - Cost of Rent: Cost to rent packages.

$C_{EXT\,TRAN}$ - Cost of External Transport: Cost for transporting raw materials and/or packages from the supplier to the manufacturer: it comprises labor costs, depreciation of vehicles (e.g. truck), cost of the distance travelled.

$C_{REC}$ - Cost of Receiving: Cost for receiving raw materials and/or packages. It includes the labor costs and depreciation of vehicles (e.g. truck, forklift) used to unload products.

$C_{COND}$ - Cost of Conditioning: Cost for sorting raw materials and/or packages before storing them in the warehouse. It includes the labor costs and depreciation of mechanical devices (if used), for example for unpacking and re-packing products.

$C_{INT\,TRAN}$ - Cost of Internal Transport: Cost for transporting raw materials and/or packages from the manufacturer's receiving area to the warehouse. It includes the labor costs, depreciation of vehicles (e.g. forklift), cost of the distance travelled.

$C_{STOCK}$ - Cost of Stocking: Cost for storing raw materials and/or packages in the warehouse. It includes the labor costs and the cost of the space for storing the packages.

$C_{PICK}$ - Cost of Picking: Cost for picking raw materials from the warehouse for producing the packages. It includes the labor costs and depreciation of vehicles (e.g. forklift) for picking the products.

$C_{INT\ TRAN1}$ - Cost of Internal Transport 1: Cost for transporting raw materials from the warehouse to the manufacturing area to produce the packages. It includes the labor costs, depreciation of vehicles (e.g. forklift), cost of the distance travelled.

$C_{MAN}$ - Cost of Packages Manufacturing: Cost for producing packages internally; it includes the labor costs, depreciation of production plants and utilities (e.g. electricity, water, gas, etc.).

$C_{REV1}$ - Cost of Internal Reverse Logistics 1: Cost of transport for bringing the raw materials not used during manufacturing back to the warehouse.

$C_{INT\ TRAN2}$ - Cost of Internal Transport 1: Cost for transporting the packages produced by the company from the production area to the warehouse. It includes the labor costs, depreciation of vehicles (e.g. forklift), cost of the distance travelled.

$C_{STOCK1}$ - Cost of Stocking 1: Cost for stocking packages produced internally by the company. It includes the labor costs and cost of the space for storing the packages.

$C_{PICK1}$ - Cost of Picking 1: Cost for picking packages (produced/bought/rented) from the warehouse. It includes the labor costs and depreciation of vehicles (e.g. forklift) for picking the packages.

$C_{INT\ TRAN3}$ - Cost of Internal Transport 3: Cost for transporting packages from the warehouse to the manufacturing area. It includes the labor costs, depreciation of vehicles (e.g. forklift), cost of the distance travelled.

$C_{REV2}$ - Cost of Internal Reverse Logistics 2: Cost of transport for bringing packages not used during the manufacturing of finished products back to the warehouse.

$C_{RE-USE}$ - Cost of Re-Use: Cost of re-using packaging after the delivery of finished products to the customer.

$C_{DISP}$ - Cost of Disposal: Cost of disposing damaged packages during the manufacturing stage. It comprises the cost of disposal, the cost of transporting damaged packages from the company to the landfill (labor costs, depreciation of vehicles used (e.g. truck), cost of the distance travelled).

$R_{SUB}$ – Gain from Sub-Product: This parameter identifies the possible gain obtained from the disposal of damaged products.

$R_{UDC}$ – Gain from Direct Sale of Pallet: This parameter identifies the possible gain obtained from the sale of tertiary packaging to the final customer.

Understanding each of the components included in equation (33) allows a firm to have a better control over the total packaging cost and to optimize it. In particular, our work

focusses on obtaining deep knowledge and optimize the direct cost of the packaging materials purchases, hence affecting the Cost of Purchase Order ($C_{ORD}$).

## 3.2. Packaging Costs from the Manufacturer's Perspective

Even though there has never been a systematic approach to predict corrugates prices from a buyer's perspective, there have been some efforts to do so from the manufacturer's side. Zang and Fuh (1997) developed a neural network algorithm to predict the price that a firm should charge for their corrugated cases products [26]. Their model has the following features as predicting inputs:

**Cost-related features based on materials**

- Type of corrugated board construction.
- Quality of material.
- Height of fluting.

**Cost-related features based on printing plates**

- Availability of printing features.
- Material of printing plates.
- Types of cut for the rubber plates.
- Size of the printing plates.

**Cost-related features based on printing features**

- Number of colors.
- Printed area.
- Effects of color trapping.
- Availability of printing inks.

**Cost-related features based on production requirement**

- Lot size

It is interesting to notice that most of the parameters included in this model may be related to parameters considered in our own model. Table 2 shows these similarities.

| Zang & Fuh Model Features | Present Model Features |
|---|---|
| *Type of corrugated board construction* | *case_type* |
| *Quality of material* | *top-load and recycled* |
| *Height of fluting* | *flute* |
| *Size of the carton* | *board_area* |
| *Availability of printing features* | - |
| *Material of printing plates* | *printing_technology* |
| *Types of cut for the rubber plate* | - |
| *Size of the printing plates* | *length, width and height* |
| *Number of colors* | *colors* |
| *Effects of color trapping* | - |
| *Availability of printing inks* | - |
| *Lot size* | *volume and moq* |

*Table 2: Comparison between Zang and Fuh's model features and our own model.*

Naturally, manufacturers have more information regarding the technical features that might affect the price of a corrugated case, therefore not all parameters present in Zang & Fuh model can be replicated in our own model. However, having the buyer's perspective implies different advantages that will be further explored in the next section.

### 3.3. Contribution of Current Work to Existing Bibliography

Even though efforts for prediction of corrugated cases with machine learning techniques have already taken place, they are all from the manufacturer's perspective. This means that only the products produced by one firm are being analyzed. In the present work, not only products from one supplier are being analyzed, but products from many suppliers from different countries. This allows us to:

- Make an integral analysis of the features that influence the final price of corrugated cases, not being biased by the particular cost structure of a unique supplier.
- Assess how the variables *country* and *supplier* affect the final output of the model. In other terms, understand the prices differences between different countries and different suppliers.
- Assess the different impact that the technical and commercial features have on the final price, by country. For example, we do not expect the variable *top_load* to have the same impact in Argentina than in Brazil.
- Expand the database to train the model. The work of Zang & Fuh is based on a 40 observations database, using 20 observations for training and 20 observations for testing. In the present work, we use a dataset of 1007 observations.

Finally, when comparing to Zang & Fuh's model, their validation method is not robust and might lead to biased conclusions (validation set method). In the present work, we will use the LOOCV to assess the performance of the model, and we will calculate more robust metrics than percentual error, such as MdAPE and sMdAPE.

# 4. Data Exploration: Descriptive Analysis

## 4.1. Data Quality: NA's Analysis

When analyzing NA's in the dataset, we find that only the variables *top_load* and *weight* present this type of data.

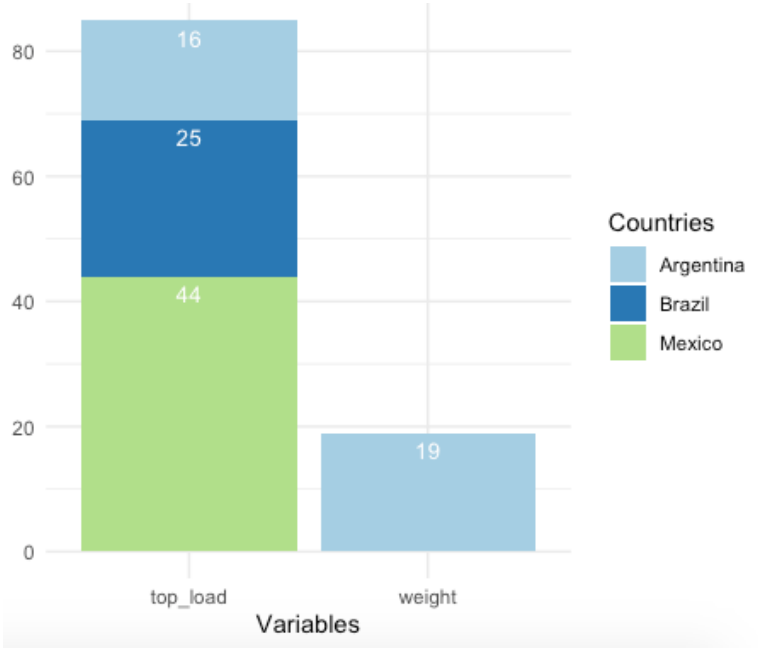


*Figure 11: NA's in the dataset.*

These values are of no surprise. In the case of the variable *weight*, these 19 NA correspond to Supplier 5, which does not inform *weight* in its specifications.

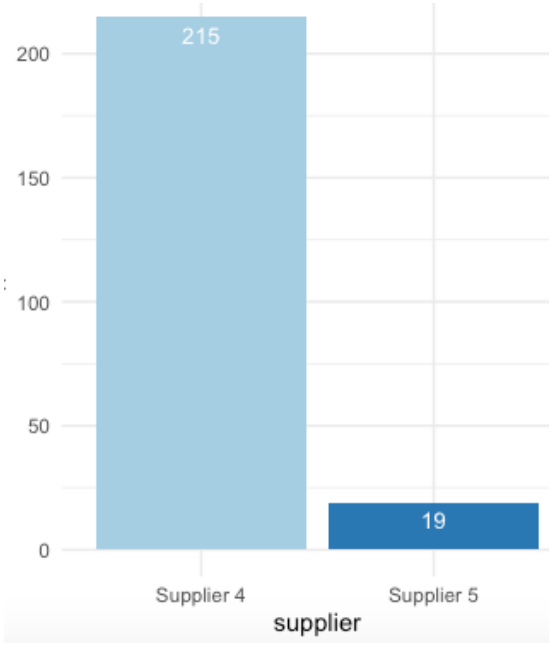


*Figure 12: Quantity of observations by supplier in Argentina.*

Regarding the 85 NA's in the variable *top_load*, these correspond mostly to the flute E cases. Flute E cases are not supposed to provide any resistance, as they are utilized to pack self-supporting products. These means, products whose own structures provide the required resistance for palletization. As it can be observed in figure 13, most of the flute E cases correspond to the wrap around cases, usually used for packing deodorants. Deodorants are light products with resistant aluminum packaging; hence, they do not need further resistance provided by cases.

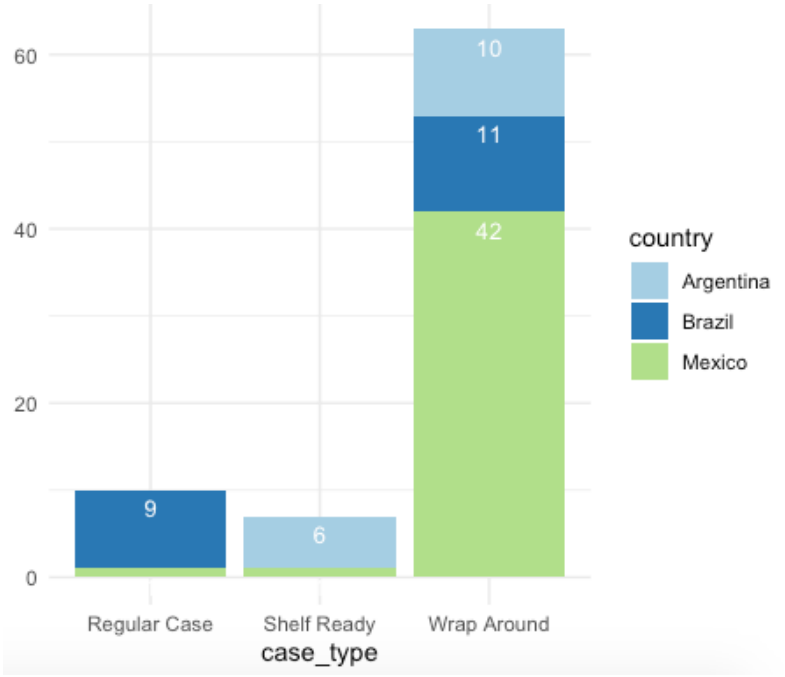


*Figure 13: case_type for flute E cases, by country.*

When looking into detail, it can be noticed that in figure 11 there are 25 NA's in Brazil, whereas in figure 13, only 20 appear. This is because Brazil has 5 wrap around cases which are not flute E, but flute B (shown in figure 14).

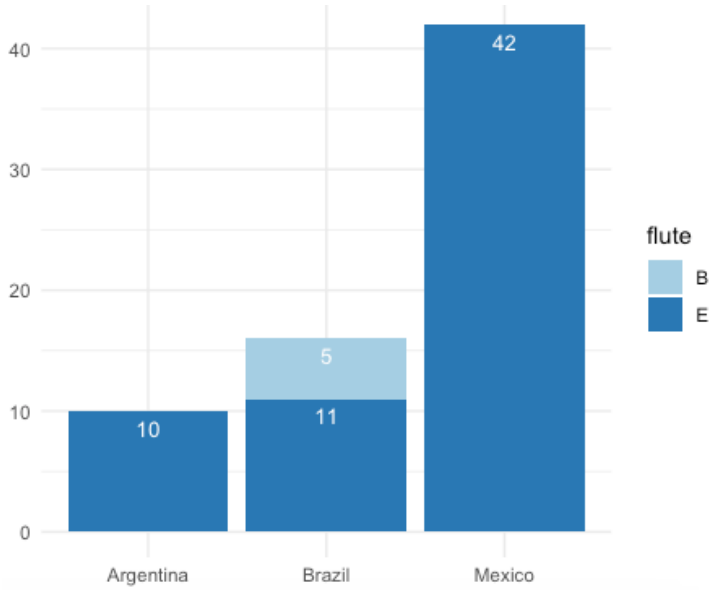


*Figure 14: Wrap around flute type, by country.*

This is an interesting insight, as wrap around cases are usually manufactured in flute E which tends to be more competitive than flute B. Further research should be made to understand if there is a technical requirement for using flute B and, if not, how can the price be reduced by changing to flute E.

## 4.2. General Analysis of Independent Variables

In this section, the distribution of the key variables will be analyzed by country. Before doing this, we find necessary to understand how many observations per country are being considered. This is shown in figure 15.

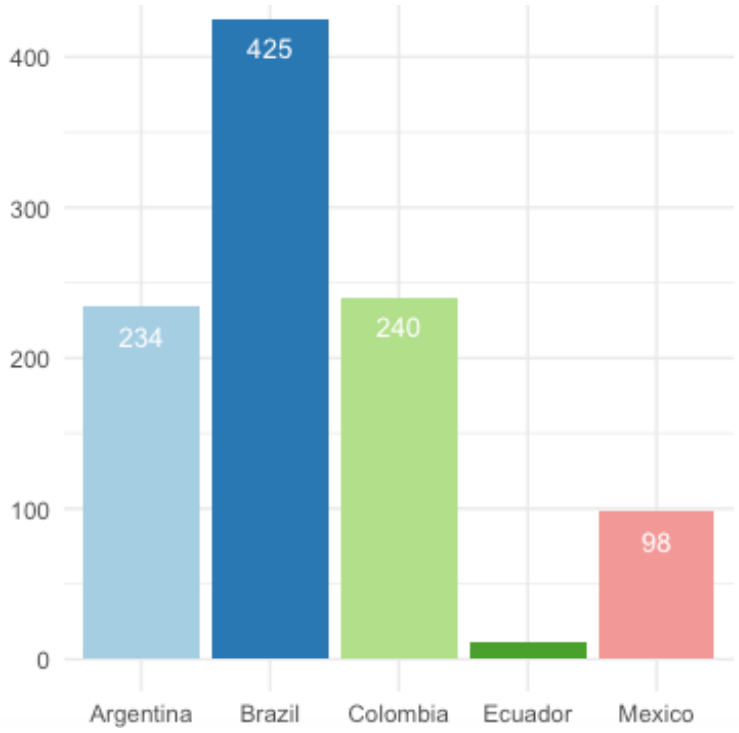


*Figure 15: Quantity of observations per country.*

Brazil has the greatest quantity of observations, followed by Colombia and Argentina. Mexico has fewer observations than these countries, but still representative for a statistical analysis, while Ecuador is almost not represented in the dataset. When doing regressions, it is possible that the value for the dummy regressor *country_ecuador* might be highly biased.

### 4.2.1. Quantitative Variables

For outlier's detection in this section, Rosner's test will be applied [27].

#### *4.2.1.1. Price*

Figures 16 and 17 show the distribution of the variable *price* for each country.

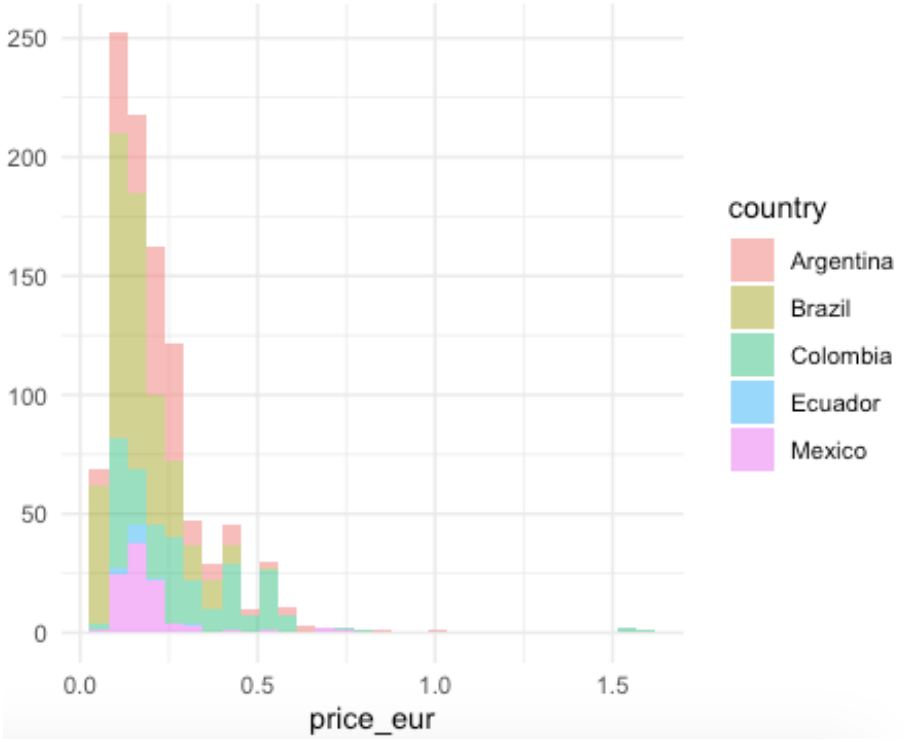


*Figure 16: Histogram of the variable price, by country.*

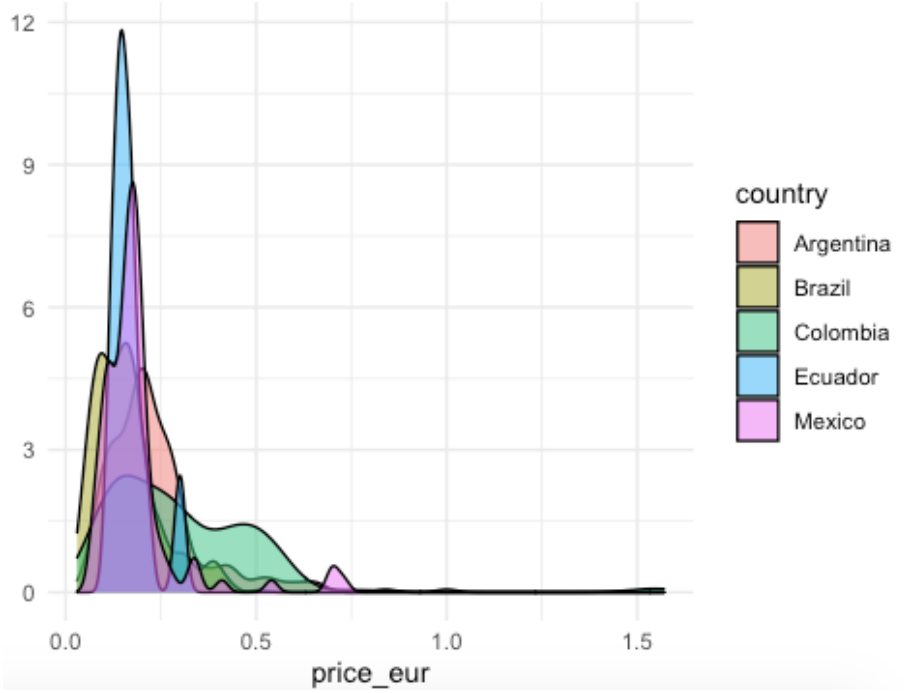


*Figure 17: Density plot of the variable price, by country.*

It is interesting to notice that Mexico and Ecuador have the most centered distributions, while Colombia has the greatest variability. Brazil has the lowest prices while Colombia has the highest.

Rosner's test detects 10 outliers in this distribution. Figure 18 shows that most of the outliers correspond to both Colombia and Mexico, with 5 and 3 respectively, followed by Argentina, with 2. Brazil does not have outliers in *price* distribution.

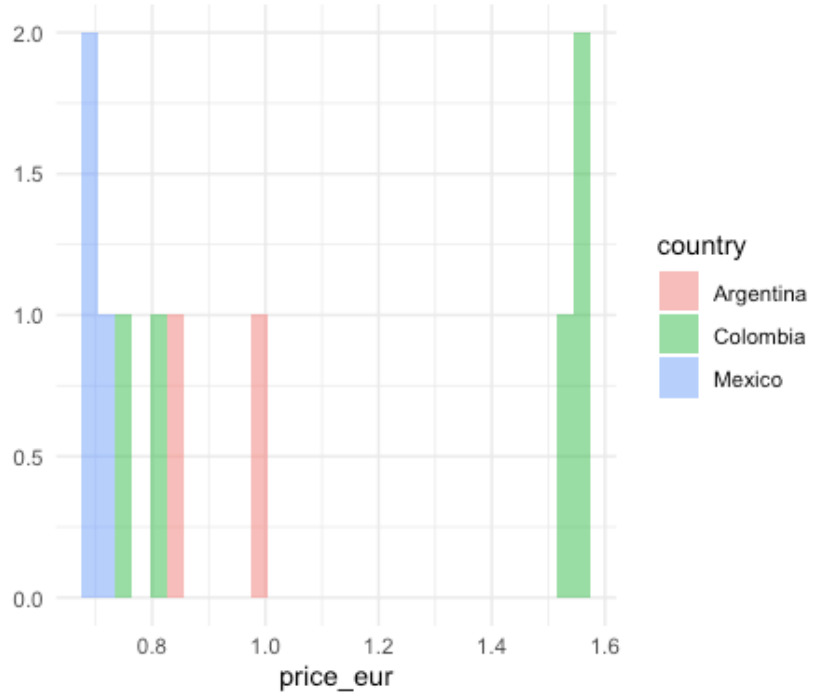


*Figure 18: Outliers of the variable price, by country.*

#### 4.2.1.2. Price/Board Area

Figures 19 and 20 show the distribution of the variable *price_board_area*.

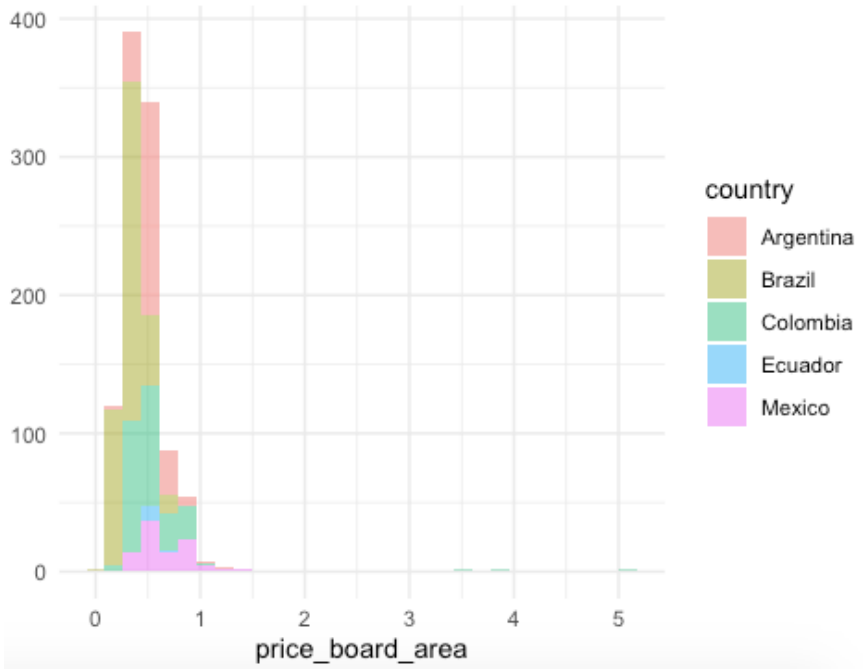


*Figure 19: Histogram of the variable price_board_area, by country.*

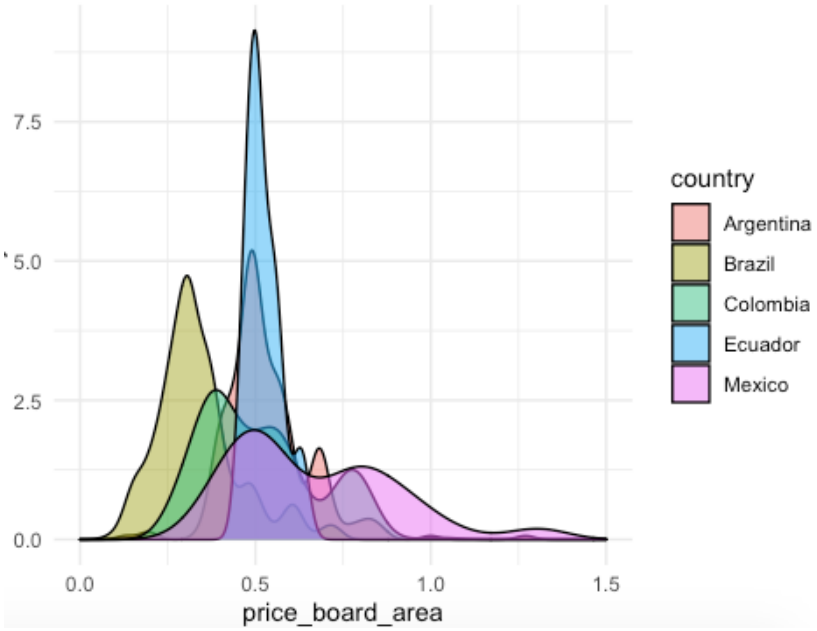


*Figure 20: Density Plot of the variable price_board_area, by country.*

We find more variability than before. When analyzing *price*, the distribution was mostly between 0 and 0.5, whereas now it is mostly between 0 and 1.

Brazil has the lowest prices per board area unit, followed by Colombia. This is different from what was observed when analyzing *price*, where Colombia has the highest prices. Another important difference is that Colombia does not have a wide distribution as before, but Mexico does.

When analyzing outliers, Rosner's test indicates that there are 8. These are shown in figure 21; it can be observed that most are from Mexico and Colombia, 4 and 3 respectively, followed by Argentina, with 1 outlier. Brazil has no outliers for this variable.

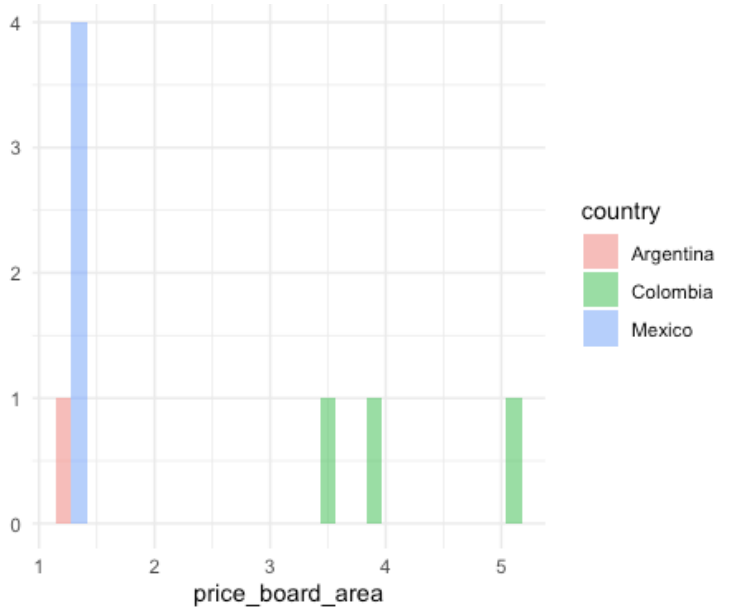


*Figure 21: Outliers of the variable price_board_area, by country.*

#### 4.2.1.3. Top Load

Figures 22 and 23 show the distribution of the key technical variable *top_load*.

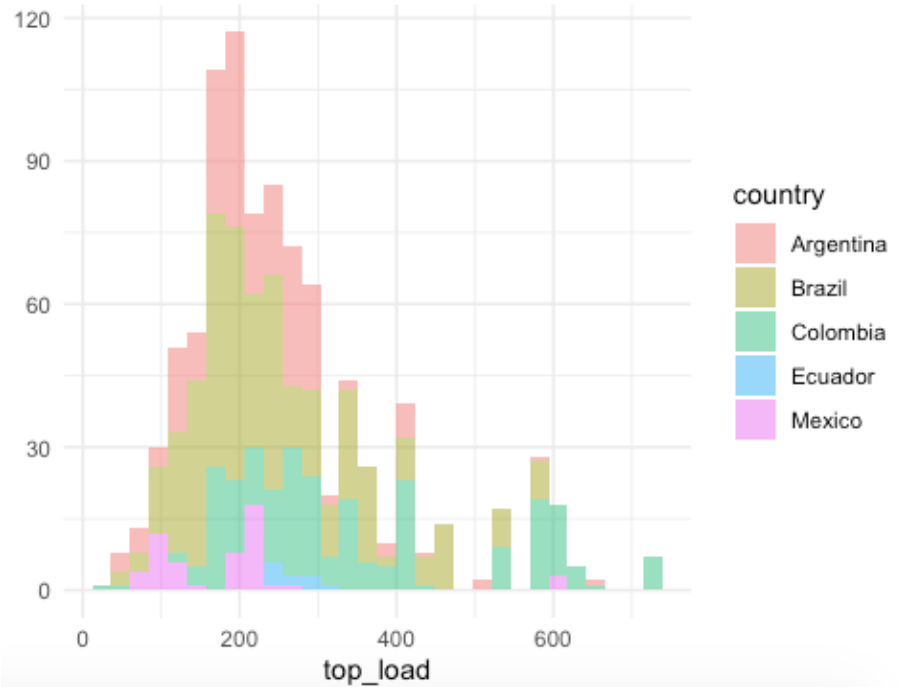


*Figure 22: Histogram of the variable top_load, by country.*

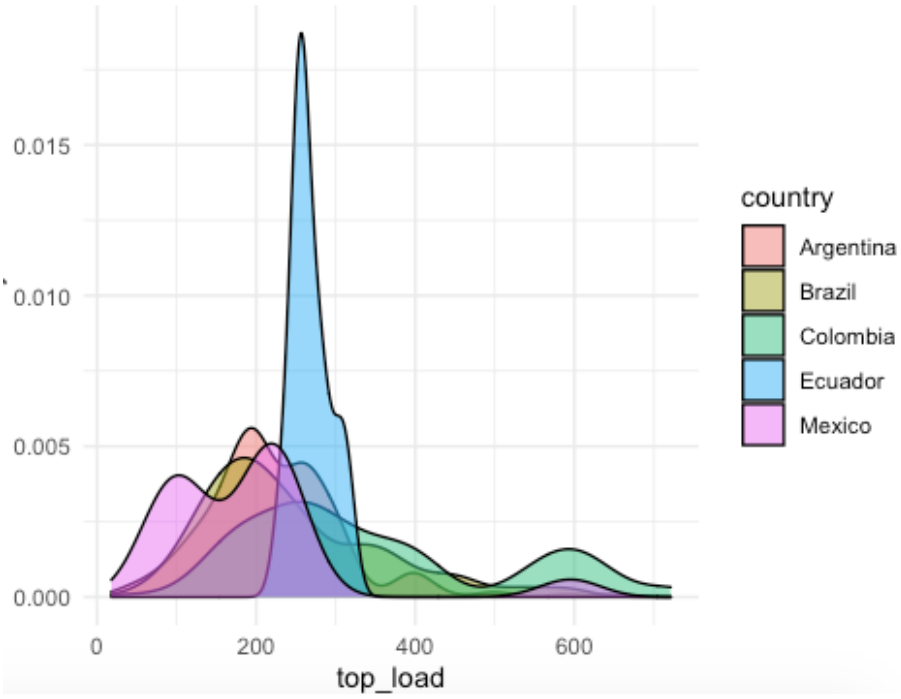


*Figure 23: Density Plot of the variable top_load, by country.*

The distribution of the variable *top_load* for Ecuador is very centered, which probably has to do with the low quantity of observations for this country. It can be noticed that Mexico, Brazil, and Argentina have similar *top_load* values, while Colombia has a wider distribution,

with higher values. When observing this plot and comparing with *price_board_area* distribution (figure 20), we can conclude that lower prices from Brazil and higher prices for Mexico are not related to higher top load necessities in these countries. Furthermore, it is quite interesting to notice that Colombia has bigger top load necessities, but this does not imply higher price per board area unit.

Regarding outliers, Rosner's test indicates that there are no outliers for this variable.

#### 4.2.1.4. Volume

Distribution of variable *volume* goes between 181 and 11,275,275. However, more than 90% of the volume data is under 1,000,000; thus, in figure 24 we only present these values.

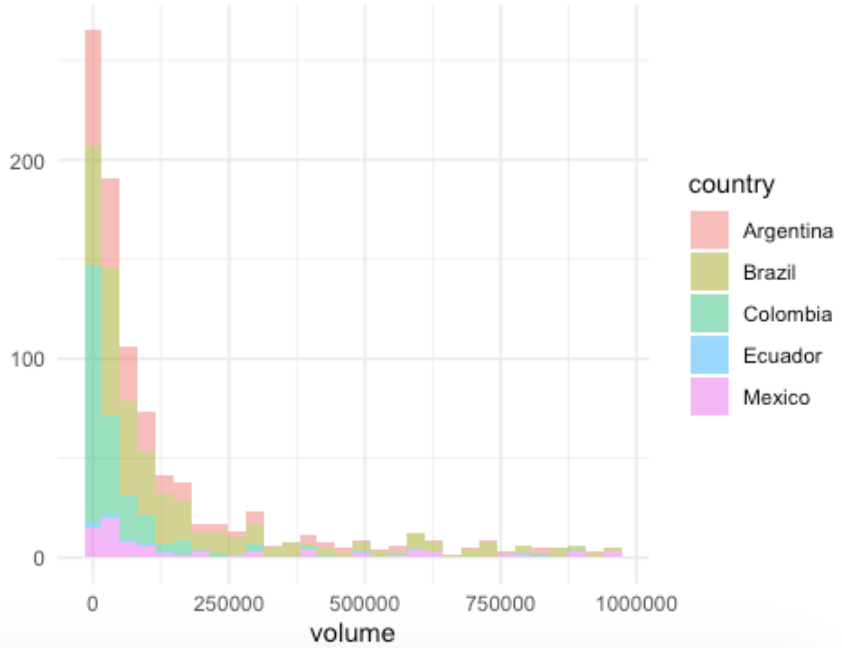


*Figure 24: Histogram of the variable volume, under 1,000,000, by country.*

Volume distribution is similar for all countries, concentrating most of the values between 181 and 125,000, and rapidly descending for higher values.

Rosner's test detected 98 outliers. These are shown in figure 25, where it can be observed that most values are from Brazil, followed by Argentina, Mexico, Colombia and finally Ecuador.

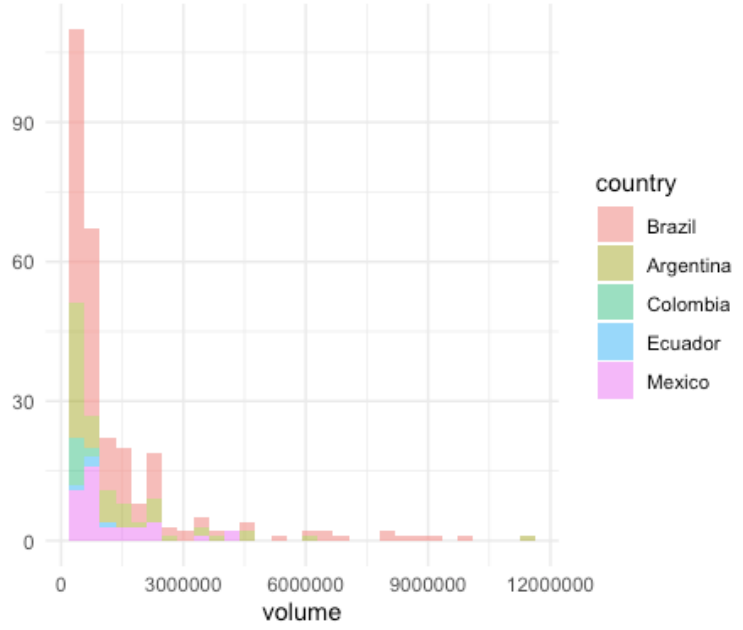


*Figure 25: Outliers of the variable volume, by country.*

#### 4.2.1.5. MOQ (Minimum Order Quantity)

Distribution of *moq* goes between 800 and 500,000. However, more than 93% of the volume data is under 27,000; thus, in figure 26 we present these values.

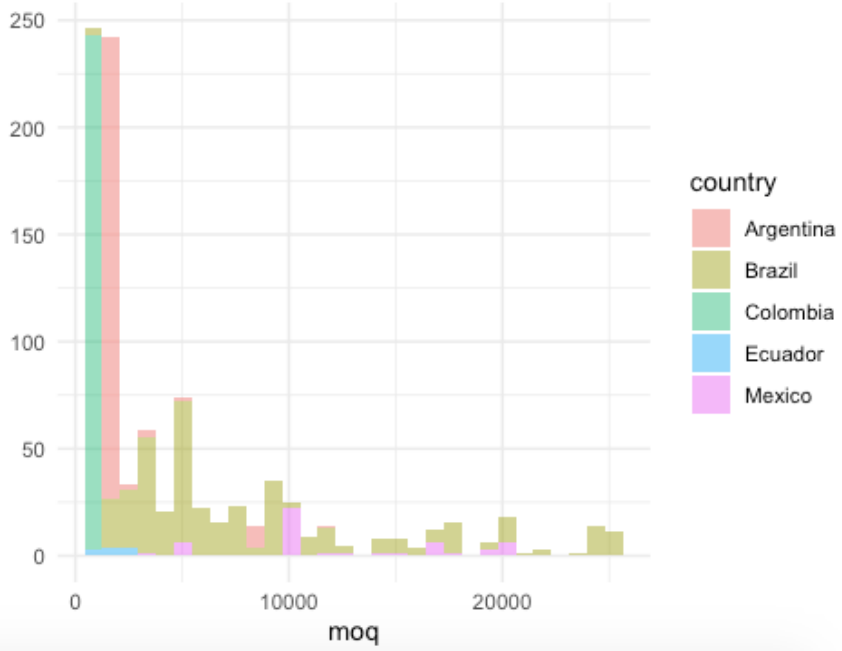


*Figure 26: Histogram of the variable MOQ, under 27,000, by country.*

Brazil and Mexico have wide distributions, while Argentina and Colombia present centered ones. In the case of Argentina, this is expected, as the main supplier (Supplier 4, shown in figure 12) works with a *Just in Time* [28] methodology, delivering for every SKU 2,000 cases

per order. A similar situation occurs in Colombia, where the biggest 2 suppliers (of a total of 3) work with this methodology, delivering MOQs of 1,000 per SKU.

Rosner's test detected 67 outliers. These are displayed in figure 27. It can be observed that most of these values correspond to Mexico, followed by Brazil.

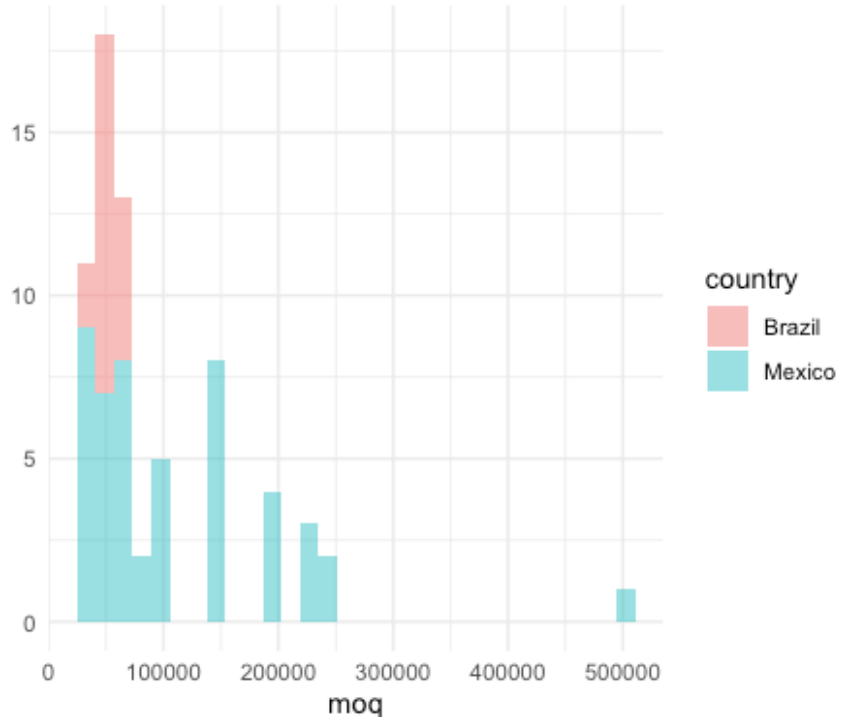


*Figure 27: Outliers of the variable MOQ, by country.*

#### 4.2.1.6. Volume/MOQ Ratio

As MOQ increases, it is common that price is reduced because of economies of scale, therefore there is a motivation for increasing MOQs. However, MOQs can be increased up to a certain limit, given by volume of production. If volume is low and MOQ is high, it is possible that a huge amount of stock is generated, which might hinder operation and increment hidden costs. Therefore, it is mandatory to analyze the ratio between volume and MOQ. If this ratio is very high, there might be an opportunity to increment MOQ and probably reduce prices.

The volume/MOQ ratio was calculated for each observation. In addition, Rosner's test was performed, finding 87 outliers, being the threshold a ratio value equal to 83. Figure 28 shows this ratio for Brazil, Mexico, and Ecuador. As it was mentioned before, Argentina and Colombia MOQs are highly standardized, so a cost benefit by reducing MOQ is not possible in these countries. Figure 28.a shows the ratios under 83, whereas outliers are shown in figure 28.b (values over 83).

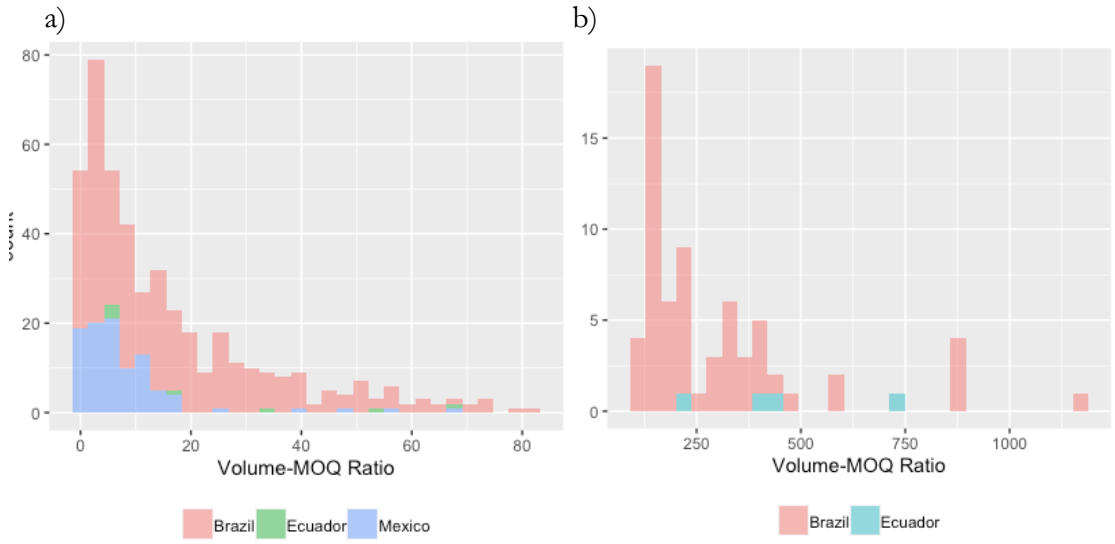


*Figure 28: Histogram of Volume-MOQ Ratio. a) Values under 83; b) Values over 83 (outliers).*

It can be observed that Brazil has a big quantity of SKUs with a volume/MOQ ratio over 83, which might represent a MOQ increase opportunity. In addition, there are further opportunities for Ecuador.

#### 4.2.1.7. Printing Technology and Colors

All observations of the dataset present the value Flexography for the variable *printing_technology*. Hence, given that there is no variability, this variable will be removed from the analysis. Regarding the variable *colors*, the distribution by country is shown in figure 29.

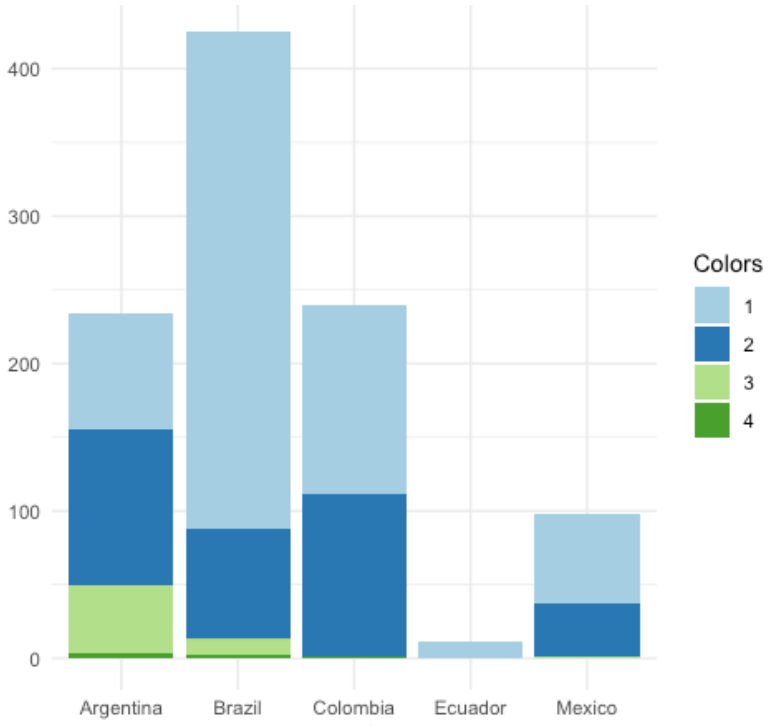


*Figure 29: Distribution of country's observations, by colors.*

Argentina presents a high quantity of 2 and 3 colors SKUs, which might be incrementing the average value of *price_board_area* for this country. This might be related to the high quantity of shelf ready cases that this country has. Regarding Colombia and Mexico, even though they do not have 3 or 4 colors SKUs, they have a big proportion of 2 colors SKUs, which is not efficient from a cost perspective.

Figure 30 shows the distribution of *colors* by *case_type* in Argentina.

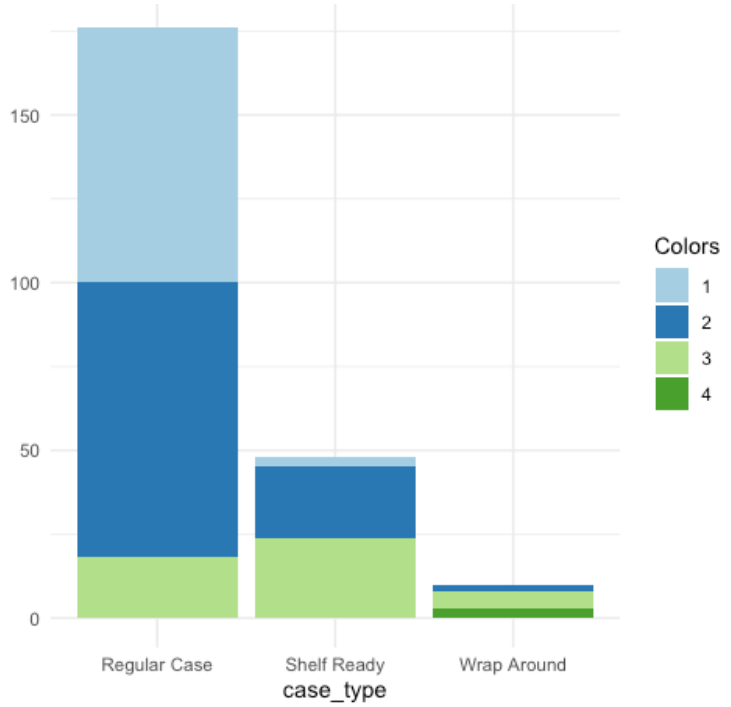


*Figure 30: Distribution of colors by case_type, in Argentina.*

Even though shelf ready cases have mostly 2 and 3 colors, there is still a big proportion of 2 and 3 colors SKUs within regular cases portfolio. Therefore, there might still be an opportunity to reduce quantity of colors, hence, also prices.

### 4.2.2. Qualitative Variables

#### *4.2.2.1. Category*

Figure 31 shows the distribution of each country's observations, by the variable *category*.

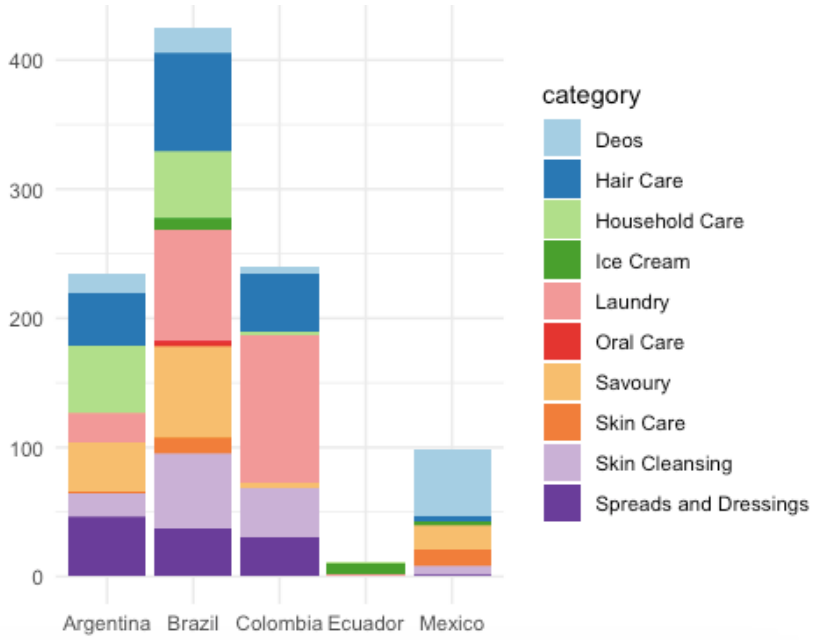


*Figure 31: Distribution of country's observations, by category.*

Most of Deos observations belong to Mexico, so it is expected that most wrap around cases will also belong to this country. Almost all observations from Ecuador belong to Ice Cream Category. Regarding Colombia, it can be observed than more than half of its observations belong to Laundry category. This makes sense, as in figure 23 it was observed that this country has the highest top loads. It is expected that Laundry cases will need more top load than other categories, as the products from this category are usually packed in heavy stand-up pouches [29], which do not provide any resistance as they are not self-supporting (different to bottles and cans).

#### 4.2.2.2. Type of Case

Figure 32 shows the distribution of each country's observations, by the variable *case_type*.

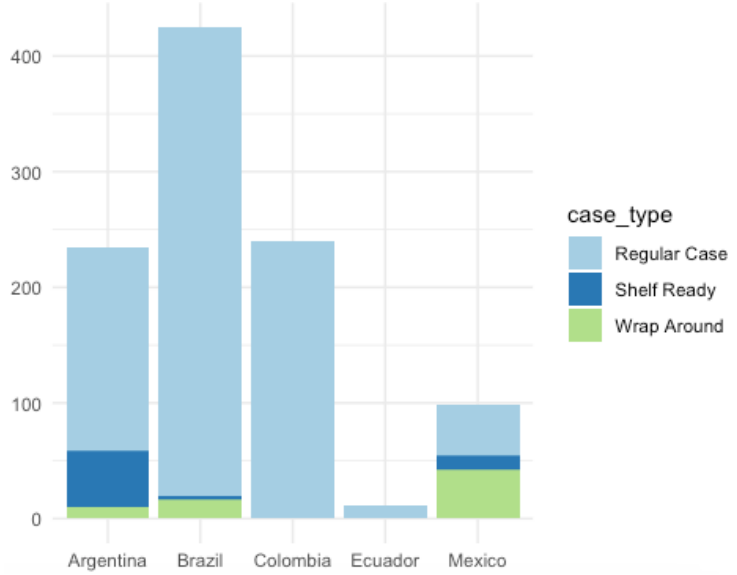


*Figure 32: Distribution of country's observations, by case_type.*

As was expected, most of wrap around cases belong to Mexico. It is interesting to notice that Argentina, different from other countries, has a high quantity of shelf ready cases, which are usually more expensive than regular cases. When estimating regressions, it will be of high importance to include the variable *case_type* to avoid having biased estimators of both Argentina and Mexico dummy variables regressors.

#### 4.2.2.3. Flute Type

Figure 33 shows the distribution of the variable *flute* for each country's observations.

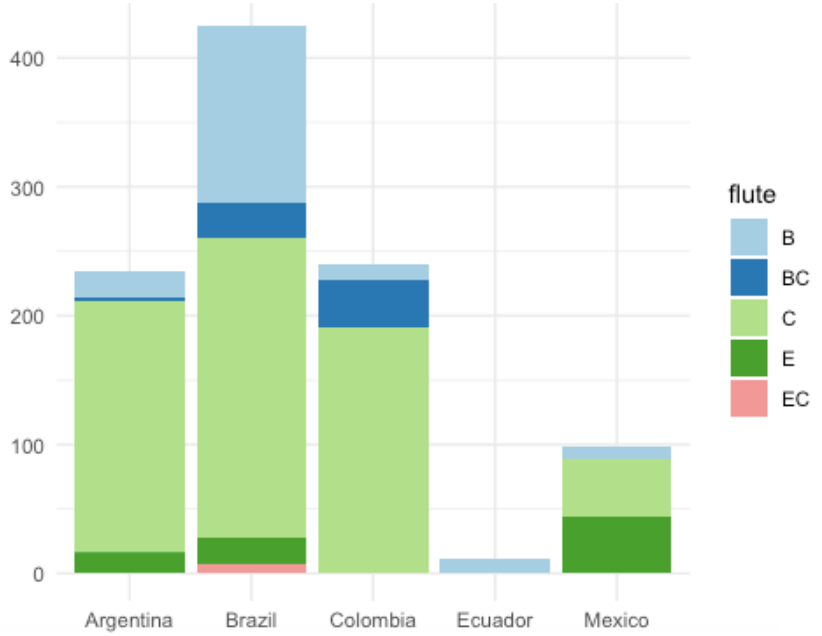


*Figure 33: Distribution of country's observations, by flute.*

Argentina and Colombia have most of their SKUs in flute C. This is of no surprise, as flute C is the biggest corrugate commodity in the market. When analyzing Mexico, we observe a big quantity of flute E observations. Again, this is expected, as it was found that most of Mexico's SKUs are wrap around cases, which we know are mostly manufactured in flute E (figure 14). It is interesting to notice that Brazil has a high quantity of flute B cases, which might be lowering its average *price_board_area* value, as this flute has a lower thickness than flute C.

It is of great interest to analyze for Brazil the distribution of the variable *flute* by *category*. This is shown in figure 34.

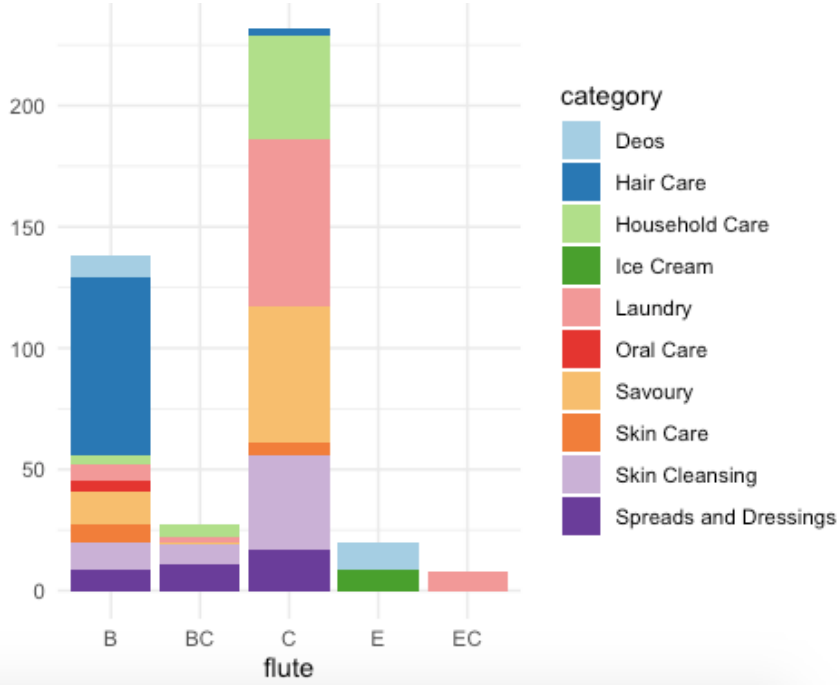


*Figure 34: Distribution of flute, by category.*

Most of Hair Care's SKUs are Flute B cases, which might be reducing the average *price_board_area* of this category, in comparison with other important categories as Laundry, Savoury and Household Care. Regarding flute E cases, they all belong to Deos or Ice Cream categories. Finally, all Oral Care's SKUs are manufactured with flute B.

#### 4.2.2.4. Whiteboard

Figure 35 shows the distribution of the variable *white* for each country's observations.

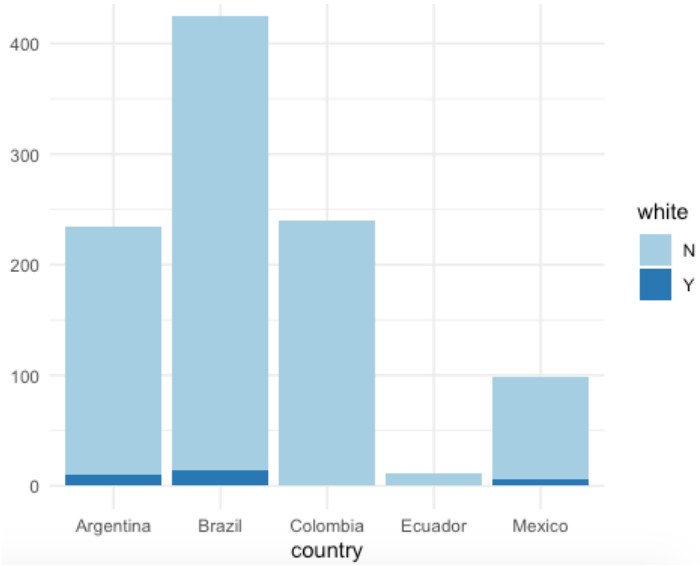


*Figure 35: Distribution of country's observations, by white.*

There are not many SKUs manufactured with whiteboard. This is expected, as whiteboard is a special feature used to give a case a more premium aspect, which is usually not necessary

in the consumer goods industry. The few observations belong to Argentina (10; 4.27%), Brazil (14; 3.29%) and Mexico (6; 6.12%). It is interesting to understand for which type of case, whiteboard is being used. This can be observed in figure 36.

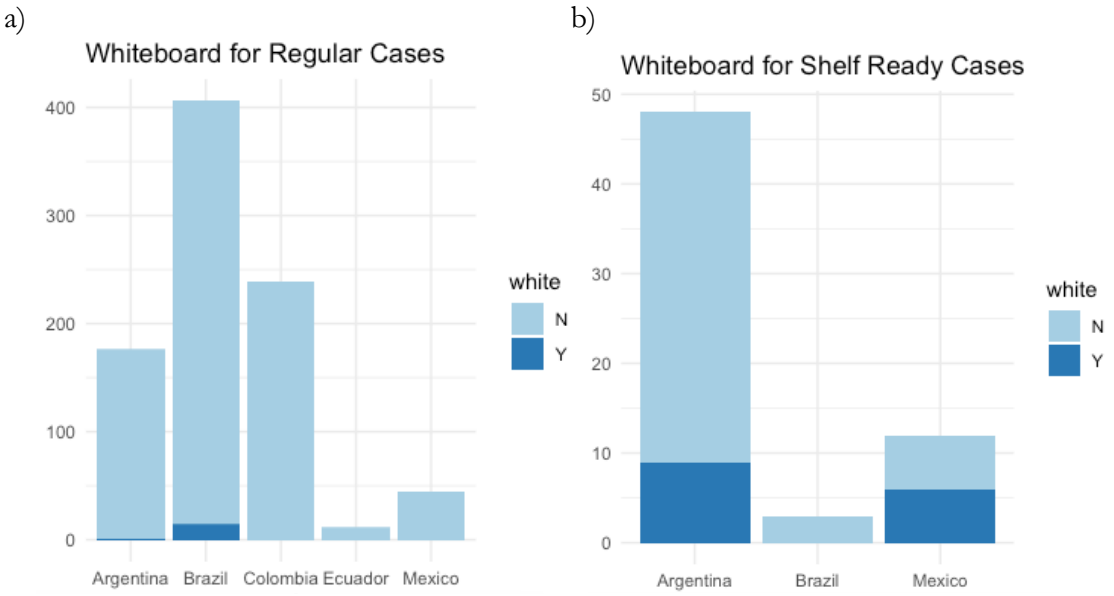


*Figure 36: Distribution of country's observations, by white, splitting by case_type. a) Regular cases; b) Shelf ready cases.*

For Argentina and Mexico, most of the cases manufactured with whiteboard are shelf ready cases. This makes sense, as shelf ready cases will be perceived by the customers, hence, they need a premium aspect. However, we notice that cases manufactured with whiteboard for Brazil are mostly regular cases. Therefore, there might be an opportunity to remove this special feature, reducing then the cost. Finally, it was not observed whiteboard in wrap around cases.

## 4.3. Influence of Independent Variables on Target Variable

The objective of this section is to understand the relation between the main independent variables previously described, with the target variable *price_board_area*, by country. In this way, we aim to get rapid insights about which will be the sign and magnitude of the parameters that will be estimated in our further models. In addition, it will be investigated the quantity of suppliers per country, the quantity of observations for each supplier, and the impact of the associated dummy variables on *price_board_area*.

Ecuador will not be included in the general analysis, because it only has 11 observations. However, some comments will be made at the end of the section. Outliers detected in figure 21 will be removed to avoid biasing the analysis and to achieve the necessary data quality.

### 4.3.1. Price/Board Area vs Suppliers and Type of Case

Figure 37 shows the quantity of observations per country, splitted by their respective suppliers.

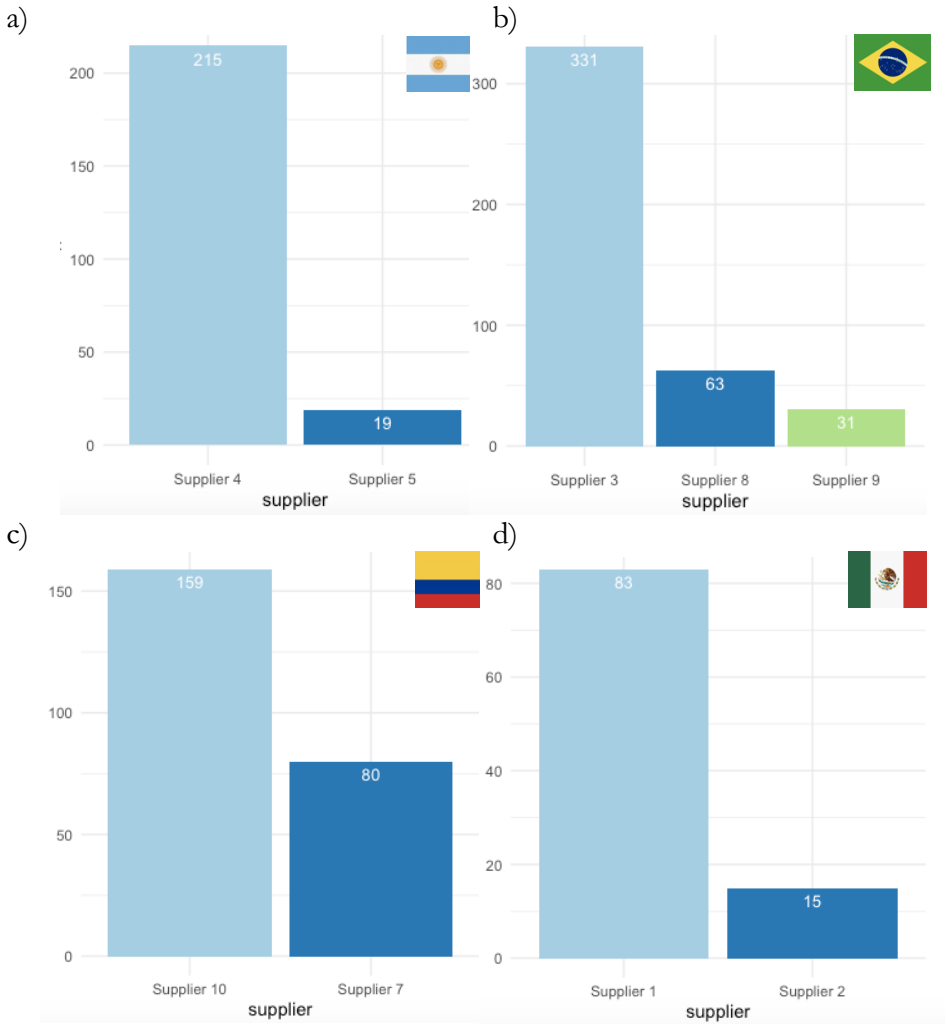


*Figure 37: Quantity of observations by supplier, for each country.*
*a) Argentina; b) Brazil; c) Colombia; d) Mexico.*

It can be observed that each country has a key supplier, that concentrates most of the respective SKUs. This gives us an insight of how concentrated the corrugated market is in LATAM.

In addition, figure 38 shows the relation between *supplier* and *price_board_area*, for each *case_type*.

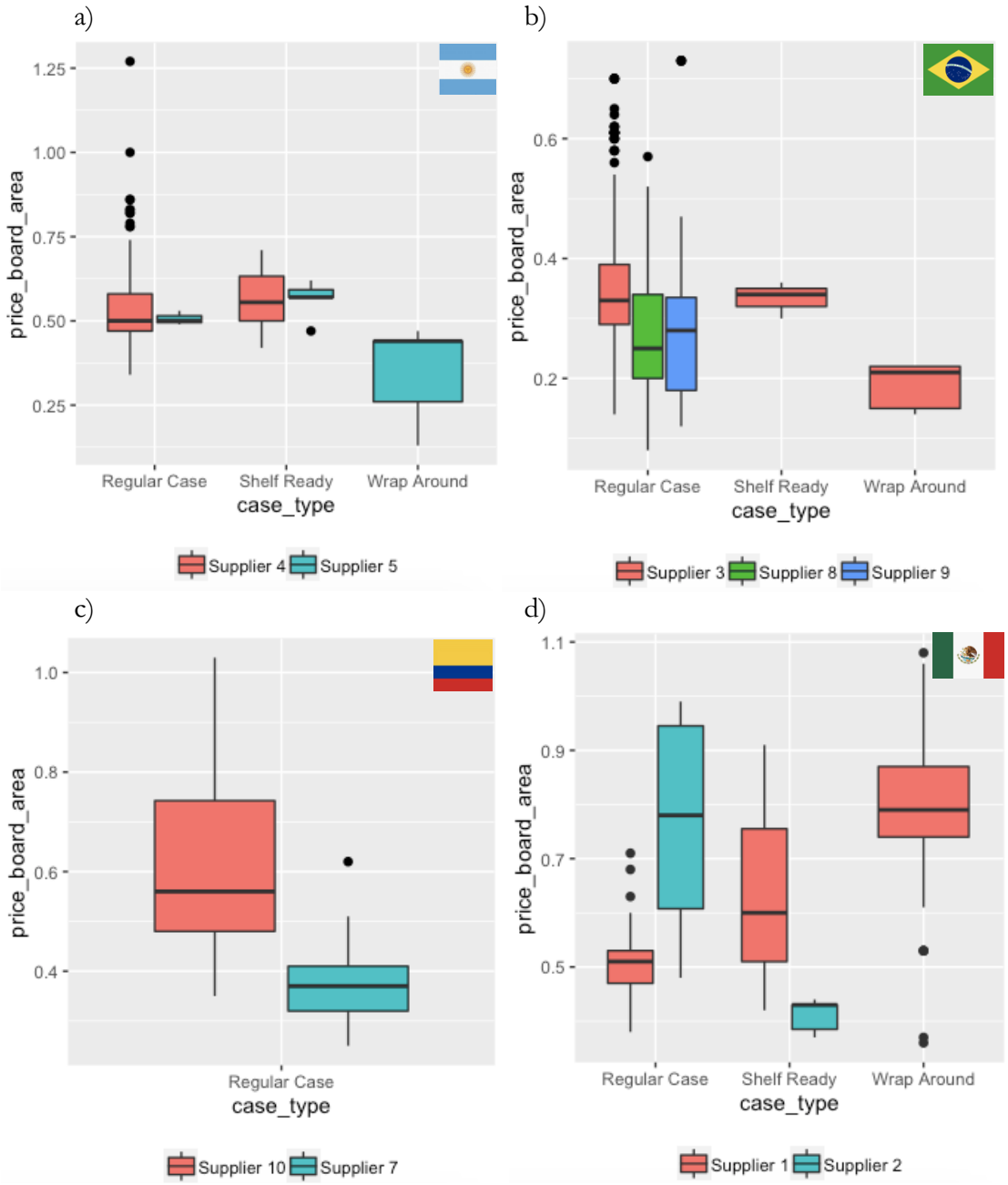


*Figure 38: price_board_area vs case_type, by supplier. a) Argentina; b) Brazil; c) Colombia; d) Mexico.*

- Argentina: Supplier 5 is more expensive than Supplier 4. Also, shelf ready cases are more expensive than regular cases, while wrap around cases are more competitive.
- Brazil: Supplier 3 is more expensive than Suppliers 8 and 9 for regular cases. Regarding the type of case, we observe that regular cases and shelf ready cases have similar prices, while wrap around cases are more competitive, as occurs in Argentina. Shelf ready and wrap around cases are only provided by Supplier 3.
- Colombia: We observe that Supplier 7 is more competitive than Supplier 10. There are only regular cases in this country's portfolio.

- Mexico: There is not a clear relation between *price_board_area* and *case_type*. For Supplier 1 shelf ready cases are more expensive than regular cases, while for Supplier 2 the opposite is valid. Also, when analyzing competitivity of suppliers, we observe that for regular cases Supplier 2 is more expensive than Supplier 1 while for shelf ready cases this relation is the opposite. Finally, a surprising insight is that wrap around cases are more expensive than the other type of cases. This is not true for other countries and was not expected.

### 4.3.2. Price/Board Area vs Volume

Figure 39 shows the relation between *volume* and *price_board_area*, for each *case_type*.

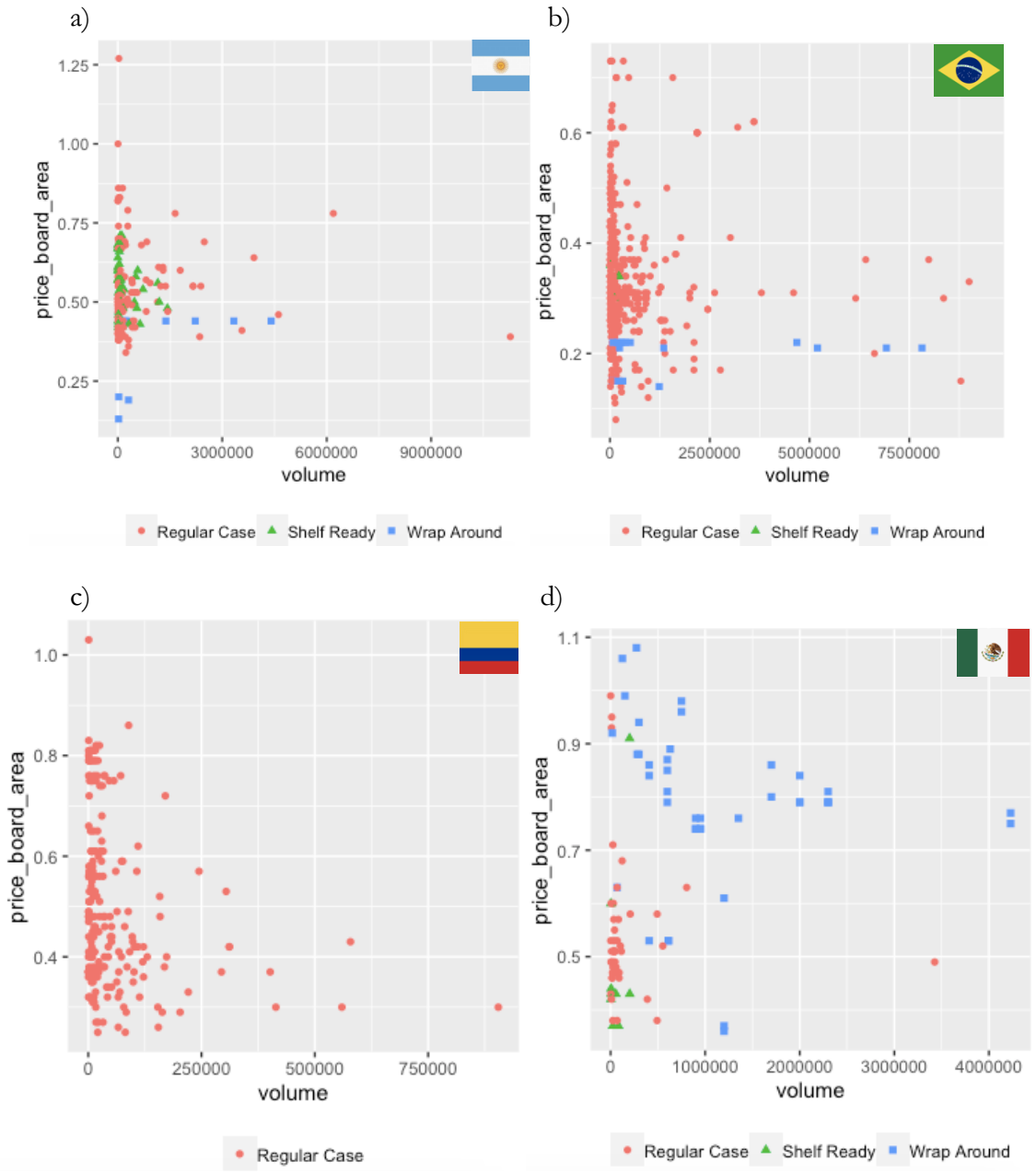


*Figure 39: price_board_area vs volume, by case_type. a) Argentina; b) Brazil; c) Colombia; d) Mexico.*

It can be observed that there is not a clear correlation between *volume* and *price_board_area,* for any country.

### 4.3.3. Price/Board Area vs MOQ

As was mentioned before, Argentina and Colombia work with a *Just in Time* service, so MOQs are mostly equal to 2,000 and 1,000 respectively, having this variable almost no variability. Hence, in figure 40 we only plot *price_board_area* vs *moq* for Brazil and Mexico.

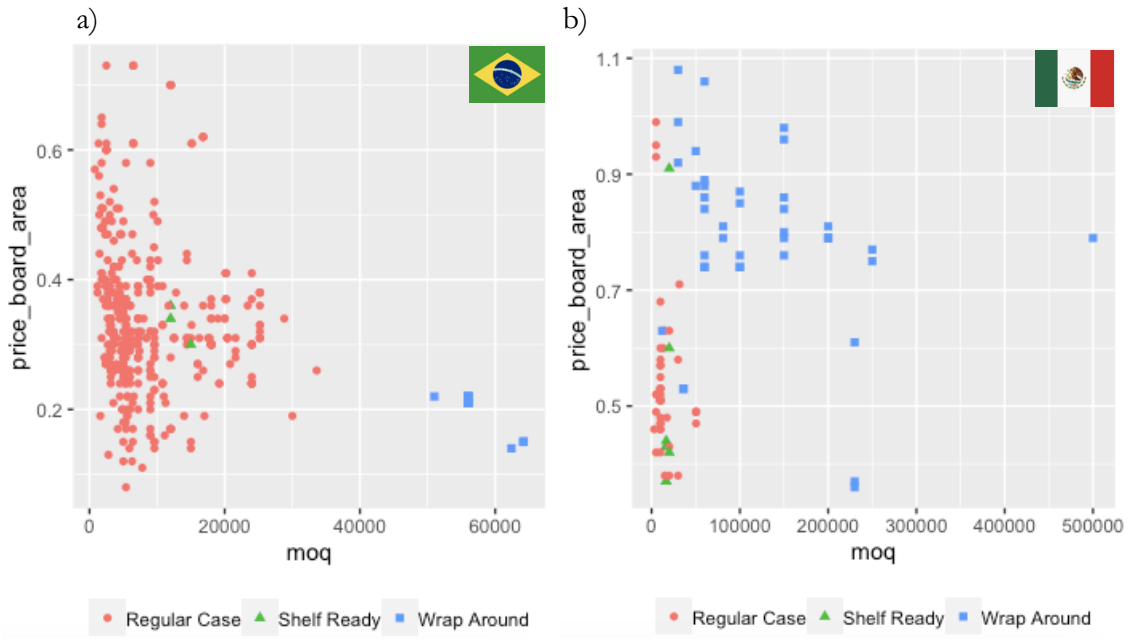


*Figure 40: price_board_area vs moq, by case_type. a) Brazil; b) Mexico.*

For Mexico, there is not a clear correlation between *price_board_area* and *moq*. However, for Brazil this is not true, as we find a negative correlation for wrap around cases of - 0.94 ± 0.10 and - 0.14 ± 0.10 for regular cases[1]. We do not inform a correlation value for shelf ready cases, due to the low quantity of available observations.

### 4.3.4. Price/Board Area vs Top Load

As was mentioned before, for wrap around cases *top_load* is not a relevant variable, so it is usually not specified. For the following analysis, it will also be investigated the relation between *price_board_area* and *top_load* only for shelf ready and regular cases. Figure 41 shows the relation between *top_load* and *price_board_area*, for each *case_type*. In addition, the correlation between these variables can be observed in table 3.

[1] In this occasion, as well as in the rest of this project, when calculating correlation, we will use Pearson method [30].

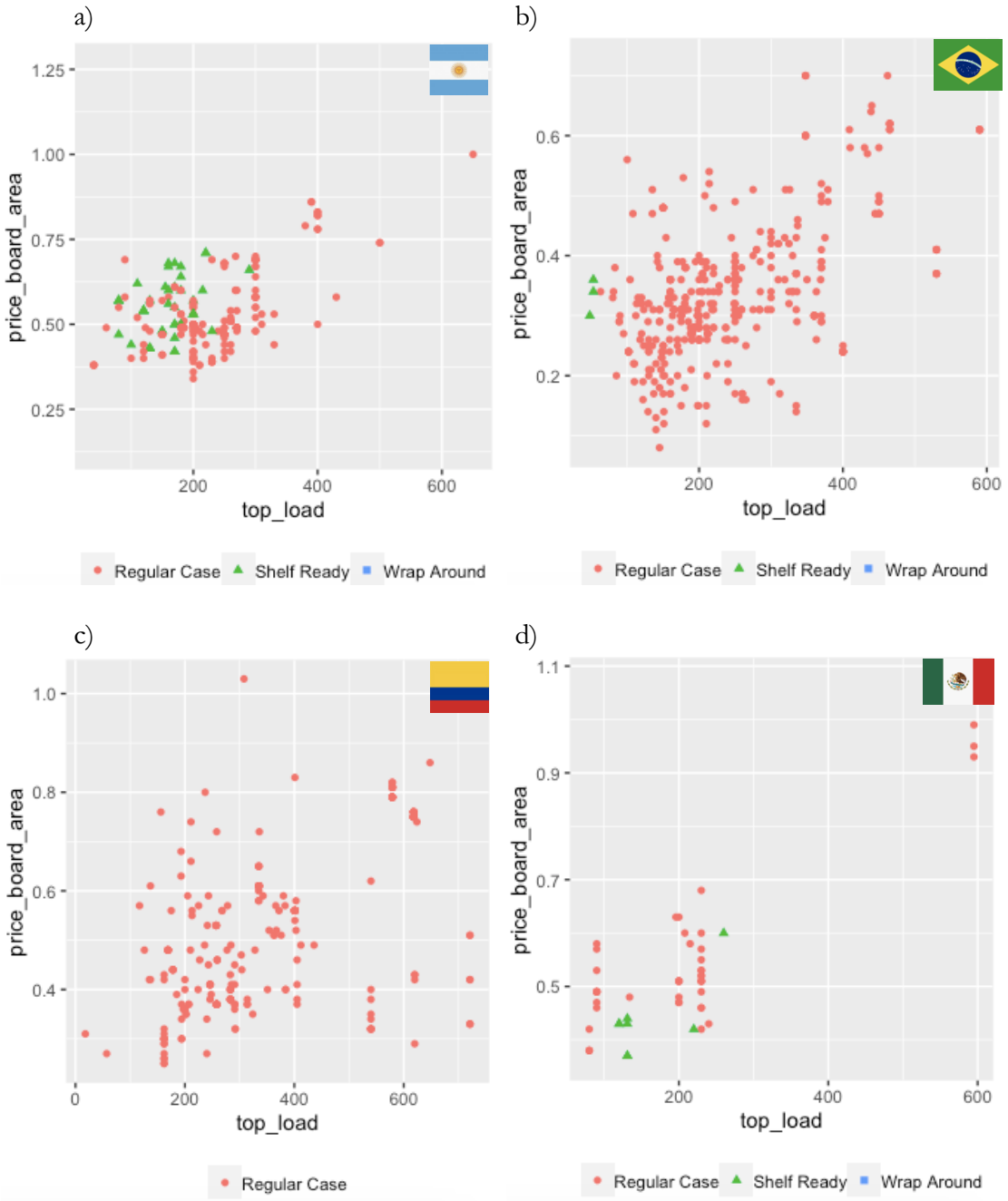


*Figure 41: price_board_area vs top_load, by case_type. a) Argentina; b) Brazil; c) Colombia; d) Mexico.*

| | Regular Cases | Shelf Ready |
|---|---|---|
| **Argentina** | 0.66 ± 0.09 | 0.32 ± 0.30 |
| **Brazil** | 0.59 ± 0.07 | - |
| **Colombia** | 0.50 ± 0.10 | - |
| **Mexico** | 0.84 ± 0.07 | 0.74 ± 0.21 |

*Table 3: Correlation of price_board_area and top_load for each country, by case_type.*

It can be noticed that in almost all cases there is a high positive correlation between these two variables. In Brazil we do not inform a correlation for shelf ready cases due to the low

quantity of observations, whereas in Colombia this relation cannot be analyzed for this type of case, as this country only exhibits regular cases.

### 4.3.5. Price/Board Area vs Colors

Figure 42 shows the relation between *price_board_area* and *colors*, for each *case_type*.

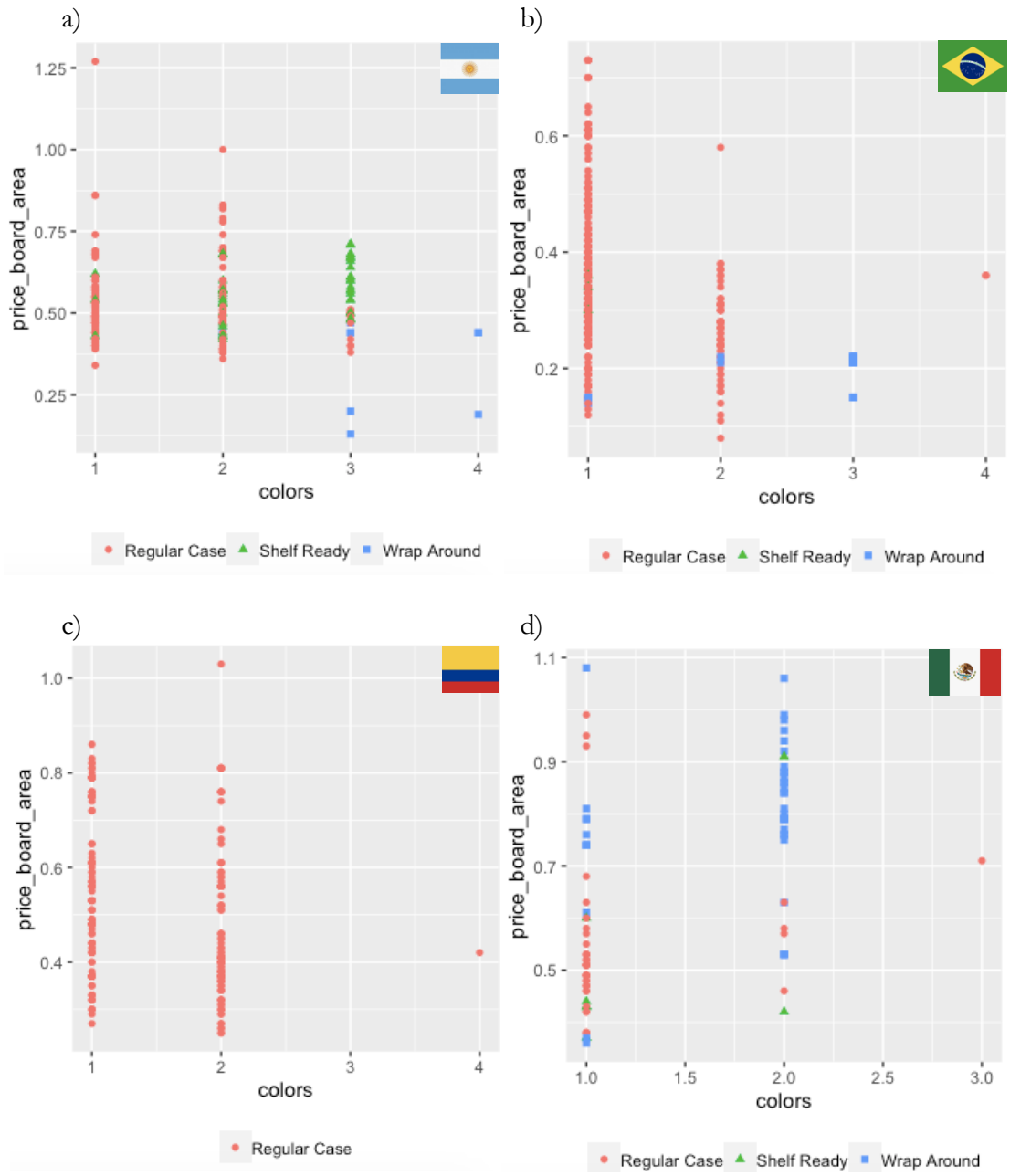


*Figure 42: price_board_area vs colors, by case_type. a) Argentina; b) Brazil; c) Colombia; d) Mexico.*

There is not a clear correlation between *colors* and *price_board_area* for Argentina, Colombia and Mexico. In Brazil we find a negative correlation between these variables, which is not expected as price should increase as the quantity of colors increases. We analyze whether there is a negative correlation between *colors* and *top_load* that can explain this unexpected negative correlation. In figure 43 we plot these variables.

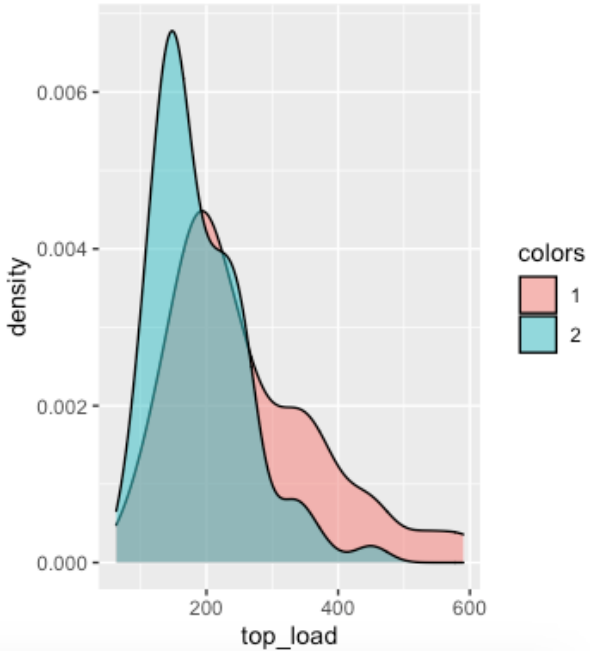


*Figure 43: Density Plot for top_load, by colors, for Brazil.*

It can be observed that generally cases with 1 color have higher values of *top_load* than cases with 2 colors, being cases with higher top loads usually more expensive. This might be the reason why apparently there is a negative correlation between *price_board_area* and *colors*. Therefore, we expect this relation to be non-significant when developing regression models.

### 4.3.6. Price/Board Area vs Flute

Figure 44 shows the relation between *price_board_area* and *flute*, for each *case_type*.

- Argentina: Flute B and C cases have similar prices. Flute E cases have the lowest prices per board area unit, while Flute BC cases have the highest prices. This is expected, as these cases have the lowest and highest thickness respectively.
- Brazil: Again, flute B and C cases have similar prices, while flute BC and E have the highest and lowest prices respectively. In addition, in Brazil there are cases with flute EC, which have prices above flute B and flute C cases, but below flute BC cases. This is also explained by the thickness associated with each flute.
- Colombia: Different to other countries, flute C cases are more expensive than flute B cases. Again, the most expensive cases are the flute BC ones.
- Mexico: Flute B and C have similar prices. On the contrary of other countries and what is expected, flute E cases have the highest *price_board_value* values. This is aligned with what was seen in figure 38.d, where we observed that wrap around cases are surprisingly expensive in this country.

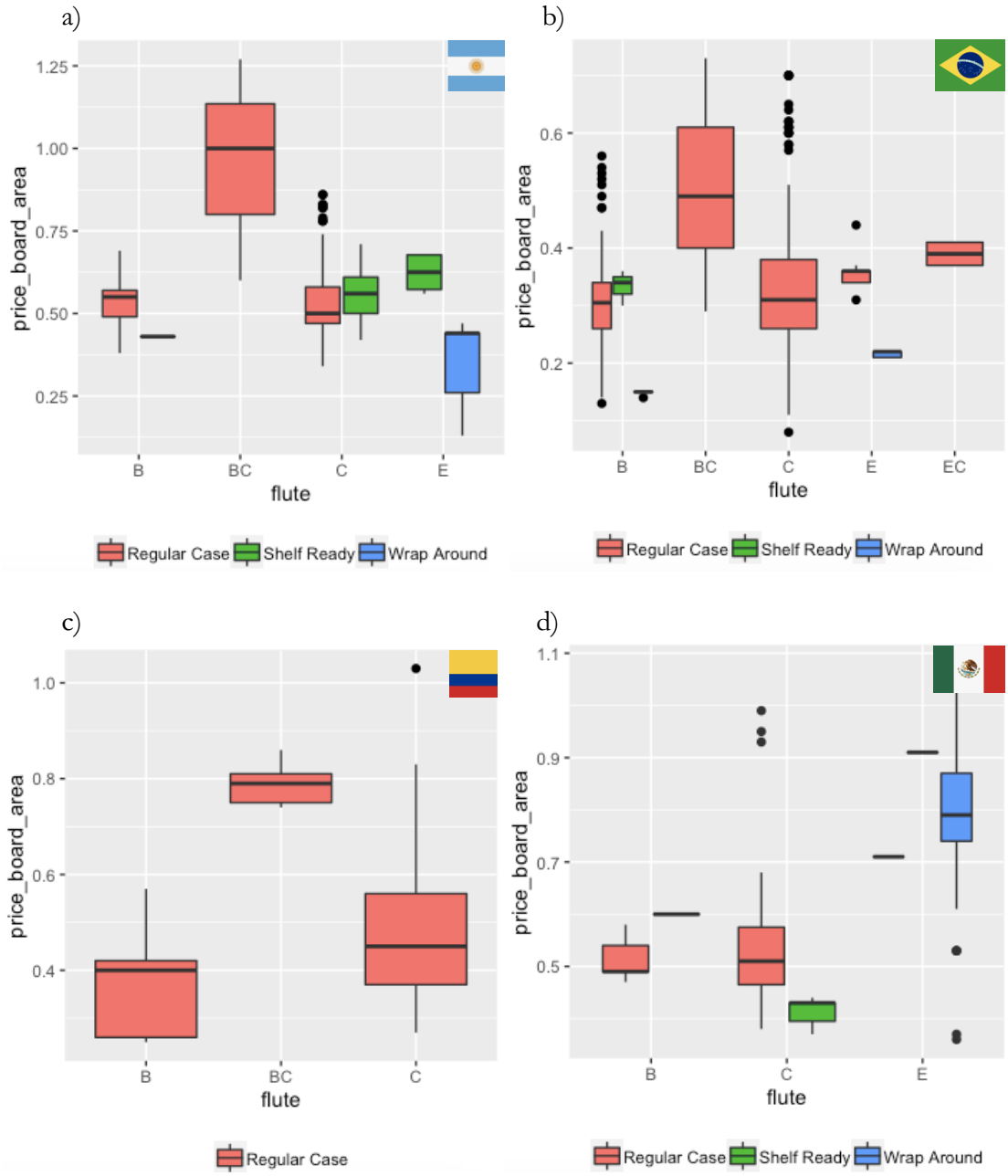


*Figure 44: price_board_area vs flute, by case_type. a) Argentina; b) Brazil; c) Colombia; d) Mexico.*

### 4.3.7. Price/Board Area vs Category

Figure 45 shows the relation between *price_board_area* and *category*, for each *case_type*.

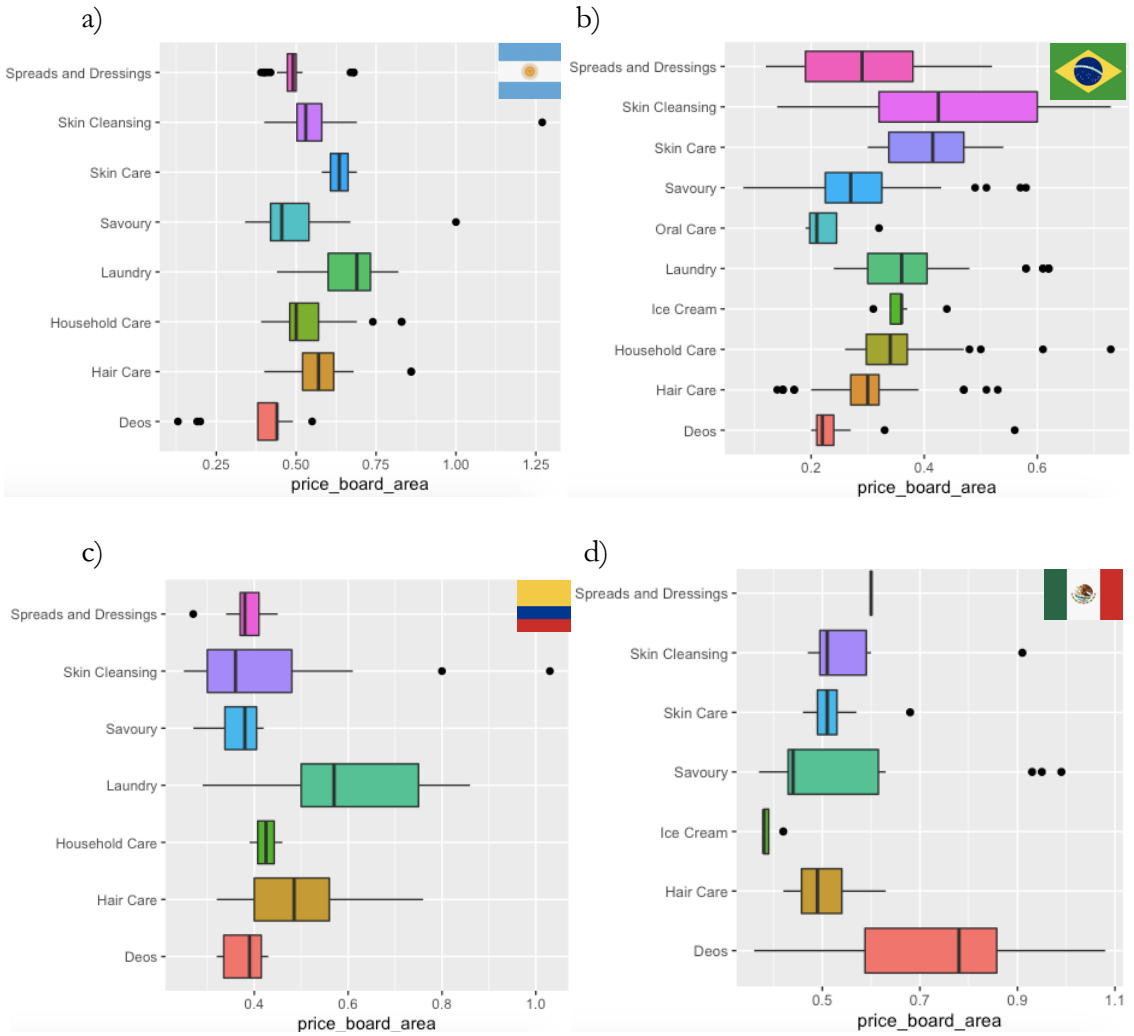


*Figure 45: price_board_area vs category. a) Argentina; b) Brazil; c) Colombia; d) Mexico.*

- Argentina: Deos has the lowest prices, mainly due to the high proportion of wrap around cases and the low top loads in other types of cases. Laundry has the highest prices, due to the necessity of high top loads due to heavy products with non-self-supporting packaging (stand up pouches). Figure 46.a shows this relation.
- Brazil: The categories with the highest prices per board area unit are Skin Cleansing and Skin Care, while the lowest prices correspond to Oral Care and Deos. In figure 46.b we can observe that Oral Care and Deos have effectively the lowest top loads, which explains the correlation with lower prices. However, we do not find a robust explanation for the high prices of Skin Cleansing and Skin Care categories.
- Colombia: As is expected, we find the highest prices in Laundry category, which requires high top loads (figure 46.c) and the lowest prices in Deos category, which has a big quantity of wrap around cases.
- Mexico: We find the highest prices in Deos category. This is explained because Deos category has a high quantity of wrap around SKUs, and we have already observed that, for this country, this type of cases presents very high prices.

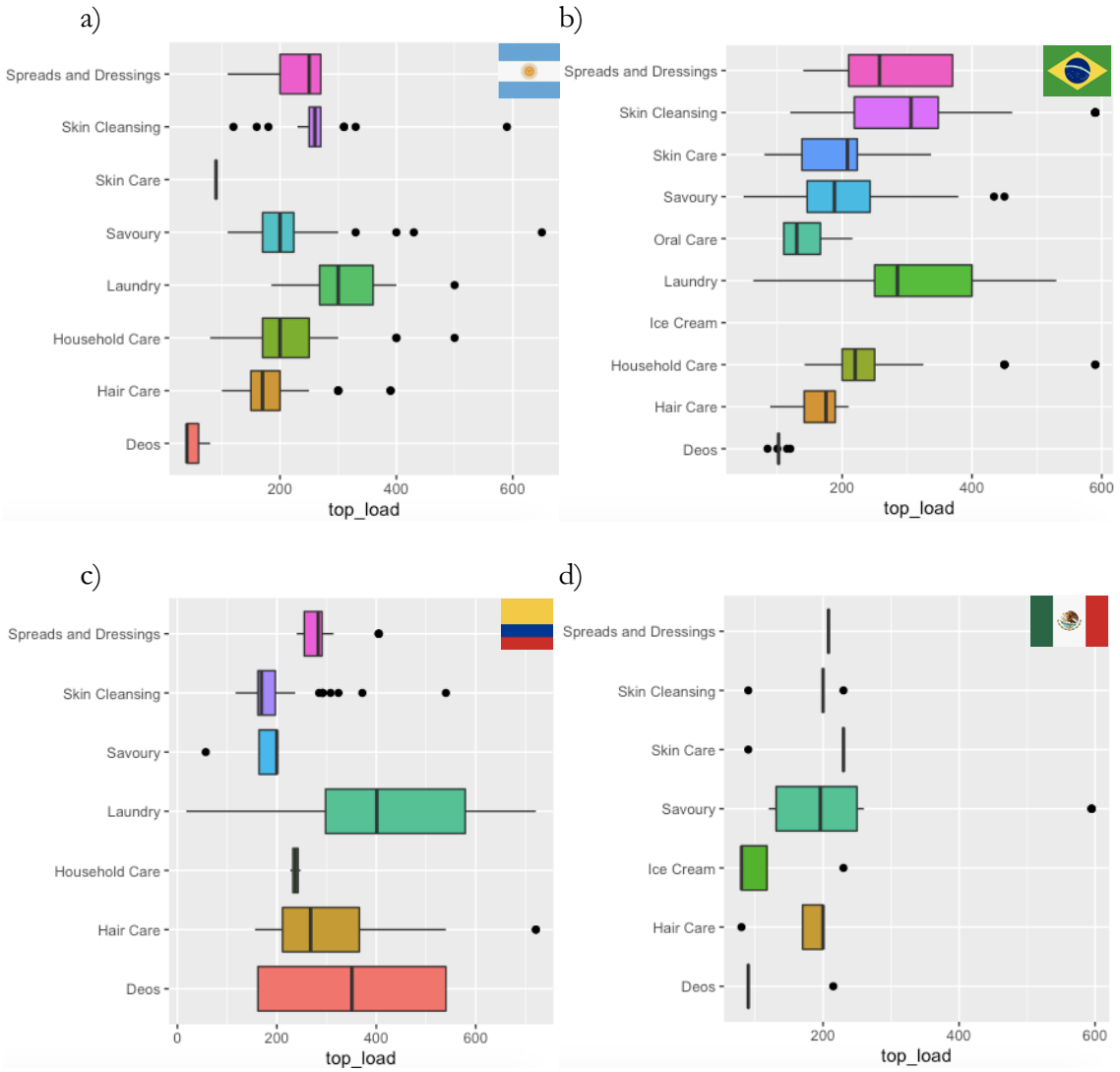


*Figure 46: top_load vs category. a) Argentina; b) Brazil; c) Colombia; d) Mexico.*

### 4.3.8. Ecuador

Ecuador has only 11 observations; thus, we do not find useful to make the same analysis that was performed for other countries. However, it is interesting to remark the main characteristics of these observations:

- Supplier: They all belong to Supplier 6.
- Type of Case: All SKUs are regular cases.
- Category: 8 from the 11 observations belong to Ice Cream category.
- Colors: All SKUs have 1 color.
- Flute: All SKUs are manufactured with flute B.
- Whiteboard: None of the SKUs are manufactured with whiteboard.

When analyzing *volume* and *moq*, we do not observe a correlation with *price_board_area*. However, this is not true when analyzing *top_load*, as shown in figure 47.

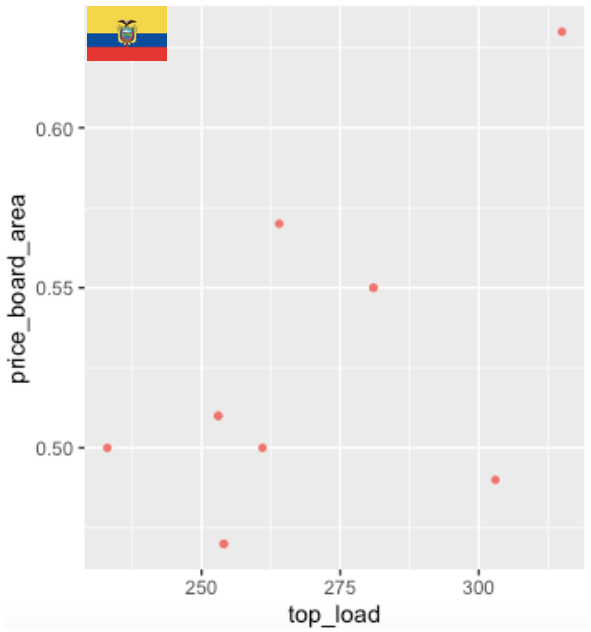


*Figure 47: price_board_area vs top_load for Ecuador.*

For this variable we find a positive correlation of 0.62 ± 0.59. The huge variability of the confidence interval is due to the low quantity of observations used for the statistical analysis.

### 4.3.9. Summary

In this section we analyzed the influence of the different independent variables on the target variable *price_board_area*. Below we sum up the main insights discovered:

1. Argentina, Colombia and Mexico have 2 suppliers, Brazil has 3 and Ecuador has just 1. In all cases, more than 75% of the observations correspond to one supplier, which makes us think of rather concentrated markets. In Argentina, the competitivity of suppliers is similar. However, this is not true for other countries, where there are clear price differences between the different suppliers.
2. Variables such as *volume* and *moq* are generally not relevant. For the first one, we did not find any impact on the target variable. For *moq*, we found a small negative correlation in Brazil, being the impact insignificant for the other countries.
3. Variable *top_load* always presents a positive correlation with the target variable. This was expected, as we know that this technical feature is one of the most relevant when defining a specification.
4. Shelf ready cases are always more expensive than regular cases. Regarding wrap around cases, they are usually more competitive than regular cases. This was observed in Argentina and Brazil, but not in Mexico, where surprisingly wrap around cases present the highest prices.
5. Colombia and Ecuador only have regular cases in their portfolios. On the contrary, the other countries also have shelf ready and wrap around cases.
6. The variable *colors* has a negative impact on *price_board_area* for Brazil. We believe that this result is due to a negative correlation between *top_load* and *colors*. This impact was not observed for the other countries.

7. Regarding the variable *flute*, it was observed that the most expensive one is BC, while the most competitive one is E. This it logical, as these flutes have the highest and lowest thickness respectively. Flute B and C usually have the same prices, except in Colombia, where we observed that flute C is more expensive than flute B.
8. Regarding the variable *category*, we usually observe that Deos is the most competitive category, whereas Laundry is the most expensive. This makes sense, as Deos' portfolio is mainly composed by wrap around – flute E cases, while Laundry cases usually require more top load than other categories. However, this general rule has two big exceptions: we observed high *price_board_area* values for Deos category in Mexico, and in Brazil we observed average *price_board_area* values for Laundry and high values for Skin Care and Skin Cleansing.

# 5. Regression Models

For linear models, the variable to be regressed will be *price_board_area_1000,* which is equal to 1000 * *price_board_area*. In this way, the values that will be obtained for the estimated parameters will be easier to interpret. For log-log models, this will not be necessary, as estimated parameters will represent the elasticity, that is, the percentual impact on *price_board_area*.

Regarding the methodology of analysis, we will start with a generic approach, in which the objective will be to understand the influence of each country on the target variable. Later, we will continue with more specific analysis, country by country, in which we will measure the impact of each variable in each case. Finally, we will make a comparison of the differences found for the key parameters' estimators in each country.

## 5.1. Feature Engineering

Before starting with regression models, some modifications are needed in the dataset. Firstly, the variable *flute* will be replaced by the variable *thickness*, according to:

- Flute E = Thickness 1.5.
- Flute B = Thickness 3.0.
- Flute C = Thickness 4.0.
- Flute BE = Thickness 5.5.
- Flute BC = Thickness 7.0.

These variables are highly correlated, so no information will be lost when dropping *flute*. A numerical variable as *thickness* will allow us to enhance our analysis.

Secondly, for dummy variables, the following values will be established as baselines:

- *country*: Brazil.
- *white*: N.
- *case_type*: Regular Case.
- *supplier*: Supplier 4 for Argentina, Supplier 3 for Brazil, Supplier 10 for Colombia and Supplier 1 for Mexico. These are the suppliers with more observations in each country.

Finally, it is important to mention that the variable *category* will not be included in the regressions to avoid high correlations with other important variables, such as *top_load*.

## 5.2. General Analysis

The main purpose of the following models is to understand the influence of the different countries on *price_board_area*. Hence, to avoid biased estimators for these dummy regressors, we remove the variable *supplier* which is highly correlated with the variable *country*.

In addition, the dataset will be splitted in two, one part containing the shelf ready and regular cases observations, and the other containing the wrap around SKUs. This division is motivated because wrap around cases have NA's for the variable *top_load*; therefore, if the

dataset was not splitted, we would be compromised between the two following undesirable scenarios:

- Not including *top_load* as a regressor, so that wrap around observations can be included in the model. This is highly inconvenient, as *top_load* is a key regressor for shelf ready and regular cases. For example, this would bias Colombia's cases which have higher *top_load* values, leading to an overestimation of the associated country dummy estimator.
- Excluding wrap around cases from the regressions, which is also highly inconvenient. Wrap around cases are very expensive in Mexico, hence, doing this would impact on an underestimation of the associated country dummy estimator.

### 5.2.1. Linear Regression Model 1: Shelf Ready and Regular Cases

Table 4 shows the results for the estimated parameters of this model.

***Linear Regression Model 1: Shelf Ready & Regular Cases***

| **Target Variable** | | |
|---|---|---|
| *price_board_area_1000* | | |
| **Numerical Variables** | | |
| **Regressor** | | **Parameter** |
| *volume* | | 0.00 |
| *moq* | | 0.00 |
| *top_load* | | 0.28 ** |
| *colors* | | -56.03 *** |
| *thickness* | | 39.24 ** |
| *Intercept* | | 201.09 *** |
| **Categorical Variables** | | |
| **Variable** | **Regressor** | **Parameter** |
| *white* | Y | -39.17 |
| *case_type* | Shelf Ready | 101.40 ** |
| *country* | Argentina | 206.71 *** |
| | Colombia | 168.35 *** |
| | Ecuador | 188.33 * |
| | Mexico | 231.33 *** |
| **Statistics** | | |
| R | | 0.2174 |
| Adjusted R | | 0.2079 |
| | | *Significance levels: ***: 0.001 ; **: 0.01; *: 0.05* |

*Table 4: Estimated parameters of Linear Regression Model 1 (Shelf Ready and Regular Cases).*

In general, results are aligned with what was observed in the dataset exploration. Variables *volume* and *moq* do not have a significant effect on *price_board_area_1000*. Variables *top_load* and *thickness* have positive impacts on the target variable, while the variable *colors* has a negative impact. Regarding the variable *type_of_case*, we observe that shelf ready cases are more expensive than regular cases.

Regarding the variable *country*, whose analysis is the objective of this section, we can observe that Brazil has notably lower prices. Following Brazil are Colombia, Ecuador, Argentina and

finally Mexico. Figure 48 shows the values of country estimated parameters, with their respective 95% confidence intervals.

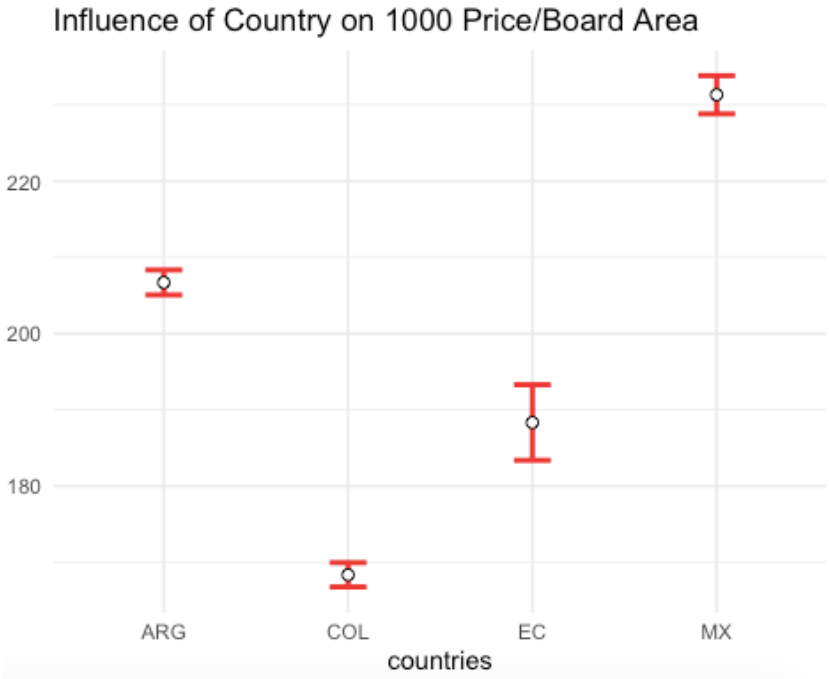


*Figure 48: Country parameters estimators for Linear Regression Model 1 (Shelf Ready and Regular Cases).*

### 5.2.2. Linear Regression Model 2: Wrap Around Cases

Table 5 shows the results for the estimated parameters of this model.

***Linear Regression Model 2: Wrap Around***

| **Target Variable** | | |
|---|---|---|
| *price_board_area_1000* | | |
| **Numerical Variables** | | |
| **Regressor** | | **Parameter** |
| *volume* | | 0.00 |
| *moq* | | 0.00 |
| *colors* | | 43.65 |
| *thickness* | | -13.20 |
| *Intercept* | | 121.89 |
| **Categorical Variables** | | |
| **Variable** | **Regressor** | **Parameter** |
| *country* | Argentina | 124.42 * |
| | Mexico | 655.44 *** |
| **Statistics** | | |
| R | | 0.7860 |
| Adjusted R | | 0.7649 |

*Significance levels: ***: 0.001 ; **: 0.01; *: 0.05*

*Table 5: Parameters of Linear Regression Model 2 (Wrap Around).*

It is of no surprise that Mexico's wrap around cases are more expensive than Brazil's ones. Argentina's wrap around cases are more competitive than Mexico's ones, but more expensive than Brazil's. In figure 49, we show the values of the estimated parameters with their corresponding 95% confidence intervals.

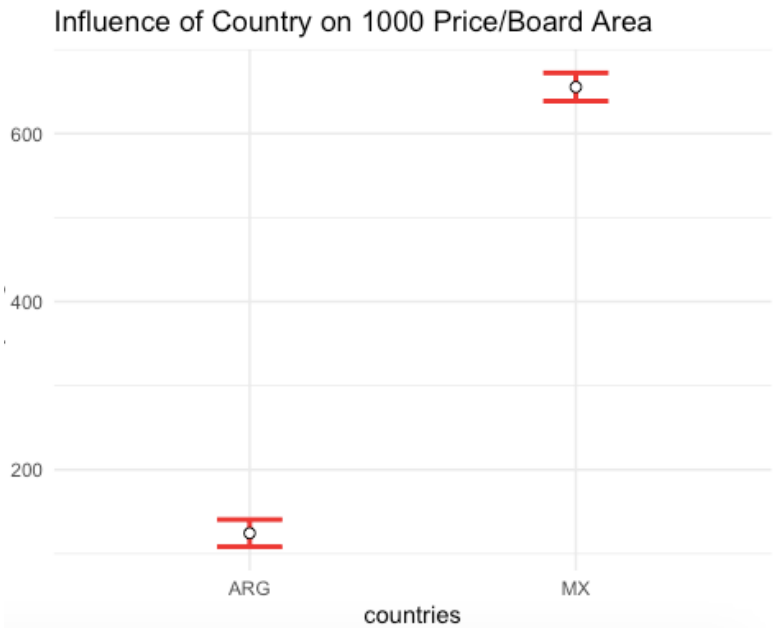


*Figure 49: Country estimated parameters for Linear Regression Model 2 (Wrap Around).*

### 5.2.3. Log-Log Regression Model 1: Shelf Ready and Regular Cases

Even though linear models give us good insights about the influence of the country on the price per board area unit, the percentual impact is much more interpretable. Therefore, we use log-log regression models, regressing *log(price_board_area)*. In table 6 we show the results for the estimated parameters of this model.

| Log-Log Regression Model 1: Shelf Ready & Regular Cases | | |
|---|---|---|
| **Target Variable** | | |
| *log (price_board_area)* | | |
| **Numerical Variables** | | |
| **Regressor** | | **Parameter** |
| *log (volume)* | | -0.02 ** |
| *log (moq)* | | -0.07 *** |
| *log (top_load)* | | 0.21 *** |
| *log (colors)* | | -0.19 *** |
| *log (thickness)* | | 0.29 *** |
| *Intercept* | | 5.04 *** |
| **Categorical Variables** | | |
| **Variable** | **Regressor** | **Parameter** |
| *white* | Y | -0.06 |
| *case_type* | Shelf Ready | 0.21 *** |
| *country* | Argentina | 0.45 *** |
| | Colombia | 0.22 *** |
| | Ecuador | 0.39 *** |
| | Mexico | 0.55 *** |
| **Statistics** | | |
| R | | 0.5092 |
| Adjusted R | | 0.5032 |
| | | *Significance levels: ***: 0.001 ; **: 0.01; *: 0.05* |

*Table 6: Estimated parameters of Log-Log Regression Model 1 (Shelf Ready and Regular Cases).*

It can be observed that a case, being compared with Brazil, is 45% more expensive in Argentina, 22% more expensive in Colombia, 39% more expensive in Ecuador and 55% more expensive in Mexico. These results are represented, with their corresponding 95% confidence intervals, in figure 50.

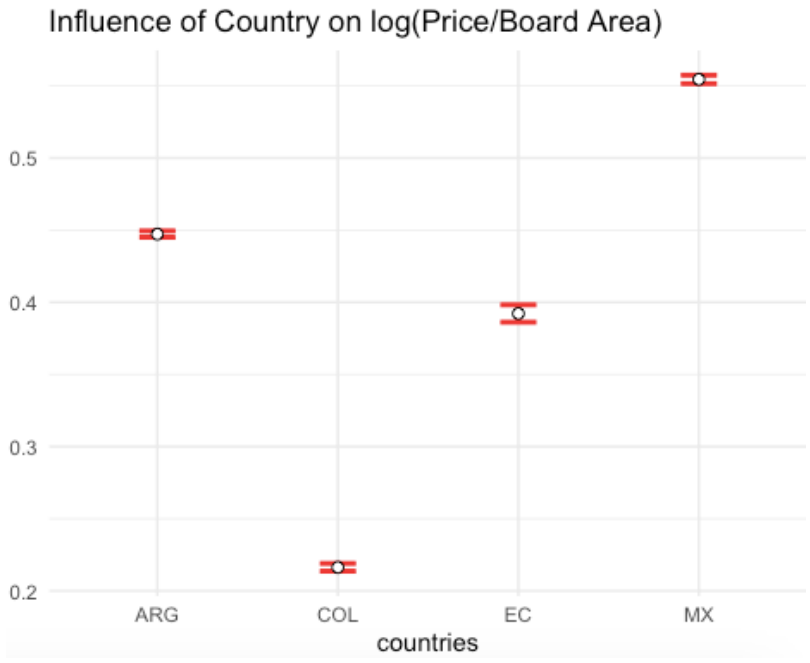


*Figure 50: Country estimated parameters for Log-Log Regression Model 1 (Shelf Ready and Regular Cases).*

### 5.2.4. Log-Log Regression Model 2: Wrap Around

In the same way, we run the log-log regression model for wrap around cases. The results can be observed in table 7.

**Log-Log Regression Model 2: Wrap Around**

| Target Variable | | |
|---|---|---|
| *log (price_board_area)* | | |
| **Numerical Variables** | | |
| **Regressor** | | **Parameter** |
| *log (volume)* | | 0.03 |
| *log (moq)* | | -0.04 |
| *log (colors)* | | 0.16 |
| *log (thickness)* | | -0.38 |
| *Intercept* | | 5.47 *** |
| **Categorical Variables** | | |
| **Variable** | **Regressor** | **Parameter** |
| *country* | Argentina | 0.37 * |
| | Mexico | 1.39 *** |
| **Statistics** | | |
| R | | 0.8527 |
| Adjusted R | | 0.8382 |
| | | *Significance levels: ***: 0.001 ; **: 0.01; *: 0.05* |

*Table 7: Estimated parameters of Log-Log Regression Model 2 (Wrap Around).*

It can be observed that the same wrap around case costs 37% more in Argentina and 139% more in Mexico, when comparing to Brazil. This result was expected, as this behavior was observed during the exploratory analysis. Figure 51 shows the estimated parameters with their respective 95% confidence intervals.

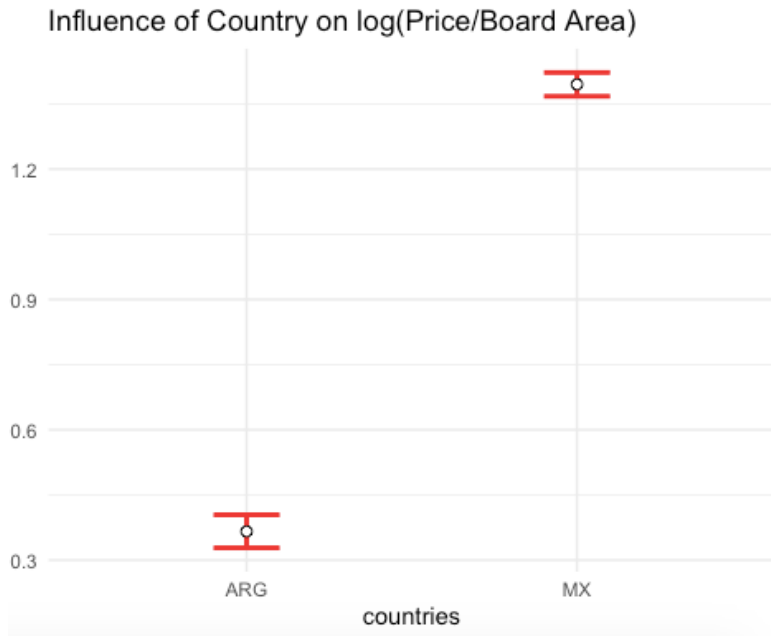


*Figure 51: Country estimated parameters for Log-Log Regression Model 2 (Wrap Around Cases).*

#### 5.2.5. Performance of Regression Models

Even though in this project the main objective of regression models is inference and not prediction, we will use the LOOCV method to calculate different performance measures on linear and log-log regression models. These metrics will be later compared with the metrics obtained for more flexible models, such as Random Forest and XGBoost. As interpretability is not strictly desired in this exercise, we will use all the available variables in the dataset to get the best possible predictions.

Table 8 shows the performance metrics calculated for the 4 regression models, regressing *price_board_area*.

| | MSE | MAPE | MdAPE | sMAPE | sMdAPE |
|---|---|---|---|---|---|
| **Linear Model 1** | 0.050 | 17.91 | 13.24 | 17.07 | 12.84 |
| **Linear Model 2** | 0.017 | 18.14 | 9.63 | 18.67 | 9.19 |
| **Log-Log Model 1** | 0.053 | 14.81 | 11.02 | 14.67 | 10.91 |
| **Log-Log Model 2** | 0.020 | 13.96 | 5.45 | 13.33 | 5.31 |

*Table 8: Performance metrics for regression models.*

It can be observed a slightly better performance for log-log regression models than linear regression models. In addition, the performance is better when predicting wrap around cases prices than shelf ready and regular cases prices. However, the values obtained are very similar for all models, and we expect to get a significant improvement in performance when using more flexible models such as Random Forest and XGBoost.

### 5.3. Country by Country Analysis

For this section, we will remove wrap around observations from the dataset. To include these observations, we would need to remove the variable *top_load*, which we know is highly correlated with *price_board_area* for shelf ready and regular cases. On the other hand, when analyzing country by country, if we splitted the dataset in 2 as was done in the previous section, we would get very small wrap around sub-datasets, from which we would not be able to get significant non-biased insights.

#### 5.3.1. Argentina

Argentina has 1 outlier, with a *price_board_area* of 1.27. This value will be filtered from regressions, so that getting biased estimators is avoided.

##### *5.3.1.1. Linear Regression Model 3*

Table 9 shows the results for the estimated parameters of this model, regressing *price_board_area_1000*.

**Linear Regression Model 3: Argentina**

| Target Variable | | |
|---|---|---|
| *price_board_area_1000* | | |
| **Numerical Variables** | | |
| **Regressor** | | **Parameter** |
| *volume* | | 0.00 |
| *moq* | | -0.01 |
| *top_load* | | 0.91 *** |
| *colors* | | -15.64 |
| *thickness* | | 11.39 |
| *Intercept* | | 310.38 *** |
| **Categorical Variables** | | |
| **Variable** | **Regressor** | **Parameter** |
| *White* | Y | 13.65 |
| *Case Type* | Shelf Ready | 95.98 *** |
| *Supplier* | Supplier 5 | 39.85 |
| **Statistics** | | |
| R | | 0.4170 |
| Adjusted R | | 0.3947 |

*Significance levels: ***: 0.001 ; **: 0.01; *: 0.05*

*Table 9: Estimated parameters of Linear Regression Model 3 (Argentina).*

As was expected, *volume*, *moq*, *colors* and *white* do not have a significant impact on the target variable, whereas *top_load* has a positive impact. Also, shelf ready cases are more expensive than regular cases. Variable *thickness* does not have a meaningful impact as well, which is logical as it was observed in the exploratory analysis that flute B and C cases have similar prices, there are few observations from flute BC cases, and flute E cases were removed from the dataset when removing wrap around cases. Finally, as it was expected, Supplier 5 does not exhibit a different competitivity than Supplier 4.

#### 5.3.1.2. Log-Log Regression Model 3

Table 10 shows the results for the estimated parameters of this model, regressing *log (price_board_area)*.

**Log-Log Regression Model 3: Argentina**

| Target Variable | |
|---|---|
| *log (price_board_area)* | |
| **Numerical Variables** | |
| **Regressor** | **Parameter** |
| *log (volume)* | 0.00 |
| *log (moq)* | -0.09 |
| *log (top_load)* | 0.25 *** |
| *log (colors)* | -0.05 |
| *log (thickness)* | 0.03 |
| *Intercept* | -1.34 *** |

| Categorical Variables | | |
|---|---|---|
| **Variable** | **Regressor** | **Parameter** |
| *White* | Y | 0.01 |
| *Case Type* | Shelf Ready | 0.14 *** |
| *Supplier* | Supplier 5 | 0.15 |

| Statistics | |
|---|---|
| R | 0.2536 |
| Adjusted R | 0.2251 |

*Significance levels: ***: 0.001 ; **: 0.01; *: 0.05*

*Table 10: Estimated parameters of Log-Log Regression Model 3 (Argentina).*

The only variables that have a statistically significant impact are *case_type* (for shelf ready cases) and *top_load*. For the first variable, a shelf ready case is 14% more expensive than a regular case, being all other parameters equal. For *top_load*, an increment of 1% results in an increase of *price_board_area* of 0.25%.

### 5.3.2. Brazil

#### *5.3.2.1. Linear Regression Model 4*

Table 11 shows the results for the estimated parameters of this model, regressing *price_board_area_1000*.

**Linear Regression Model 4: Brazil**

| Target Variable | | |
|---|---|---|
| *price_board_area_1000* | | |
| **Numerical Variables** | | |
| **Regressor** | | **Parameter** |
| *volume* | | 0.00 |
| *moq* | | -0.01 *** |
| *top_load* | | 0.73 *** |
| *colors* | | -25.85 |
| *thickness* | | -9.35 |
| *Intercept* | | 282.29 *** |
| **Categorical Variables** | | |
| **Variable** | **Regressor** | **Parameter** |
| *White* | Y | 25.48 |
| *Case Type* | Shelf Ready | 123.23 * |
| *Supplier* | Supplier 8 | -62.90 *** |
| | Supplier 9 | -89.26 *** |
| **Statistics** | | |
| R | | 0.4947 |
| Adjusted R | | 0.4830 |

*Significance levels: ***: 0.001 ; **: 0.01; *: 0.05*

*Table 11: Estimated parameters of Linear Regression Model 4 (Brazil).*

As was expected, *top_load* has a strong positive impact on *price_board_area*. Also, shelf ready cases are more expensive than regular cases. Supplier 9 is the most competitive, followed by Supplier 8 and finally Supplier 3. Regarding *moq*, as was shown in the exploratory analysis, this variable has a negative significant effect on the target variable. However, this influence is very small, with no practical implications.

#### 5.3.2.2. Log-Log Regression Model 4

Table 12 shows the results for the estimated parameters of this model, regressing *log (price_board_area)*.

**Log-Log Regression Model 4: Brazil**

| Target Variable | | |
|---|---|---|
| *log (price_board_area)* | | |
| **Numerical Variables** | | |
| **Regressor** | **Parameter** | |
| *log (volume)* | -0.02 | ** |
| *log (moq)* | -0.09 | *** |
| *log (top_load)* | 0.38 | *** |
| *log (colors)* | -0.17 | |
| *log (thickness)* | 0.07 | |
| *Intercept* | -2.16 | *** |
| **Categorical Variables** | | |
| **Variable** | **Regressor** | **Parameter** |
| *White* | Y | 0.11 |
| *Case Type* | Shelf Ready | 0.58 *** |
| *Supplier* | Supplier 8 | -0.25 *** |
| | Supplier 9 | -0.32 *** |
| **Statistics** | | |
| R | | 0.4524 |
| Adjusted R | | 0.4398 |

*Significance levels: ***: 0.001 ; **: 0.01; *: 0.05*

*Table 12: Estimated parameters for Log-Log Regression Model 4 (Brazil).*

From results presented in table 12, we understand that a 1% increase in *top_load* increments *price_board_area* by 0.38%, and a 1% increase in *moq* reduces *price_board_area* by 0.09%. Regarding *volume*, is has a significant impact on the target variable, but very small for practical implications. In addition, we discovered that shelf ready cases are 58% more expensive than regular cases, being all other parameters equal. Finally, Supplier 8 is 25% more competitive than Supplier 3, while Supplier 9 is 32% more competitive than this baseline supplier.

In the exploratory analysis, we observed that Supplier 8 and Supplier 9 only provide regular cases, and that both suppliers were more competitive than Supplier 3. However, it could not be identified a *price_board_area* differentiation between them, whereas the current regression is telling us that Supplier 9 is more competitive than Supplier 8. This can be explained by analyzing the respective top loads of each supplier's portfolio. Supplier 9 provides cases with more top load than Supplier 8, at similar prices; thus, the estimated parameter associated with Supplier 9 is lower than the estimated parameter associated with Supplier 8. This *top_load* difference by supplier can be observed in figure 52.

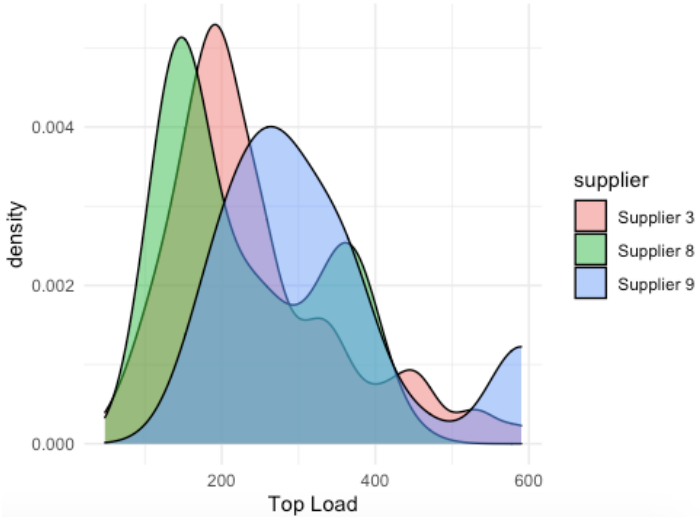


*Figure 52: Density plot of the variable top_load, by supplier, for Brazil.*

### 5.3.3. Colombia

Colombia only has regular cases and uses no whiteboard. Hence, these variables will not be included in the regressions. Furthermore, all Colombia's cases have *moq* equal to 1,000, so this variable will also be excluded from the analysis. Finally, Colombia has 3 outliers in terms of *price_board_area*, with the values 5.17, 3.90 and 3.90. These values will be filtered from regressions, in order to avoid getting biased estimators.

#### *5.3.3.1. Linear Regression Model 5*

Table 13 shows the results for the estimated parameters of this model, regressing *price_board_area_1000*.

**Linear Regression Model 5: Colombia**

| Target Variable | | |
|---|---|---|
| *price_board_area_1000* | | |
| **Numerical Variables** | | |
| **Regressor** | **Parameter** | |
| *volume* | 0.00 | |
| *top_load* | 0.17 | *** |
| *colors* | 10.37 | |
| *thickness* | 65.66 | *** |
| *Intercept* | 206.52 | *** |
| **Categorical Variables** | | |
| **Variable** | **Regressor** | **Parameter** |
| *supplier* | Supplier 7 | -164.83 *** |
| **Statistics** | | |
| R | | 0.7298 |
| Adjusted R | | 0.7239 |

*Significance levels: ***: 0.001 ; **: 0.01; *: 0.05*

*Table 13: Estimated parameters of Linear Regression Model 5 (Colombia).*

From table 13 it can be concluded that *top_load* and *thickness* have positive impacts on *price_board_area* and that Supplier 7 is more competitive than baseline Supplier 10.

#### 5.3.3.2. Log-Log Regression Model 5

Table 14 shows the results for the estimated parameters of this model, regressing *log (price_board_area)*.

**Log-Log Regression Model 5: Colombia**

| Target Variable | | | |
|---|---|---|---|
| *log (price_board_area)* | | | |
| **Numerical Variables** | | | |
| **Regressor** | | **Parameter** | |
| *log (volume)* | | -0.01 | |
| *log (top_load)* | | 0.12 | *** |
| *log (colors)* | | 0.03 | |
| *log (thickness)* | | 0.54 | *** |
| *Intercept* | | -2.03 | *** |
| **Categorical Variables** | | | |
| **Variable** | **Regressor** | **Parameter** | |
| *supplier* | Supplier 7 | -0.36 | *** |
| **Statistics** | | | |
| R | | 0.7188 | |
| Adjusted R | | 0.7127 | |

*Significance levels: ***: 0.001 ; **: 0.01; *: 0.05*

*Table 14: Estimated parameters of Log-Log Regression Model 5 (Colombia).*

An increment of 1% in *top_load* has an impact of 0.12% in *price_board_area*. In addition, incrementing *thickness* in 1% increases *price_board_area* in 0.54%. This means that changing from flute B to C (3mm to 4mm) would represent a 13.15% increase in the target variable, or that changing from B to BC (3 mm to 7 mm) would represent a 72% increase. Regarding suppliers, Supplier 7 is 36% more competitive than Supplier 5.

### 5.3.4. Mexico

Mexico has 4 shelf ready outliers, with *price_board_area* values of 1.32, 1.32, 1.30 and 1.28. These values will be filtered from regressions, in order to avoid getting biased estimators.

#### 5.3.4.1. Linear Regression Model 6

Table 15 shows the results for the estimated parameters of this model, regressing *price_board_area_1000*.

**Linear Regression Model 6: Mexico**

| Target Variable | | |
|---|---|---|
| *price_board_area_1000* | | |
| **Numerical Variables** | | |
| **Regressor** | | **Parameter** |
| *volume* | | 0.00 |
| *top_load* | | 0.84 *** |
| *colors* | | 0.00 |
| *thickness* | | 56.10 |
| *Intercept* | | 269.07 *** |
| **Categorical Variables** | | |
| **Variable** | **Regressor** | **Parameter** |
| *case_type* | Shelf Ready | -73.21 |
| *supplier* | Supplier 2 | 98.32 ** |
| **Statistics** | | |
| R | | 0.8029 |
| Adjusted R | | 0.7708 |

*Significance levels: ***: 0.001 ; **: 0.01; *: 0.05*

*Table 15: Estimated parameters for Linear Regression Model 6 (Mexico).*

The variable *top_load* has a high positive impact on *price_board_area*. Regarding suppliers, Supplier 1 is more competitive than Supplier 2. Other variables have no significant impact on the target variable.

#### 5.3.4.2. Log-Log Regression Model 6

Table 16 shows the results for the estimated parameters of this model, regressing *log (price_board_area)*.

**Log-Log Regression Model 6: Mexico**

| Target Variable | | |
|---|---|---|
| *log (price_board_area)* | | |
| **Numerical Variables** | | |
| **Regressor** | | **Parameter** |
| *log (volume)* | | -0.01 |
| *log (top_load)* | | 0.21 *** |
| *log (colors)* | | 0.02 |
| *log (thickness)* | | 0.14 |
| *Intercept* | | -1.86 ** |
| **Categorical Variables** | | |
| **Variable** | **Regressor** | **Parameter** |
| *case_type* | Shelf Ready | -0.14 |
| *supplier* | Supplier 2 | 0.26 *** |
| **Statistics** | | |
| R | | 0.6645 |
| Adjusted R | | 0.6099 |

*Significance levels: ***: 0.001 ; **: 0.01; *: 0.05*

*Table 16: Estimated parameters for Log-Log Regression Model 6 (Mexico).*

A 1% increment in *top_load* has a 0.21% increment in *price_board_area*. Regarding suppliers competitivity, we can observe that Supplier 2 is 26% more expensive than Supplier 1. As well as in *Linear Regression Model 6*, other variables have no significant impact on *price_board_area*.

### 5.3.5. Ecuador

As was mentioned in the exploratory analysis, Ecuador has only 1 supplier, regular cases, SKUs with 1 color, 1 flute type and no whiteboard manufactured cases. Hence, these variables will not me included in the regressions.

#### *5.3.5.1. Linear Regression Model 7*

Table 17 shows the results for the estimated parameters of this model, regressing *price_board_area_1000*.

**Linear Regression Model 7: Ecuador**

| Target Variable | |
|---|---|
| *price_board_area_1000* | |
| **Numerical Variables** | |
| **Regressor** | **Parameter** |
| *volume* | 0.00 |
| *moq* | 0.06 |
| *top_load* | 2.16 * |
| *Intercept* | -164.27 |
| **Statistics** | |
| R | 0.6198 |
| Adjusted R | 0.4568 |

*Significance levels: ***: 0.001 ; **: 0.01; *: 0.05*

*Table 17: Estimated parameters for Linear Regression Model 7 (Ecuador).*

The only variable that has a significant impact on *price_board_area* is *top_load*. There is a strong positive correlation between both variables.

#### *5.3.5.2. Log-Log Regression Model 7*

Table 18 shows the results for the estimated parameters of this model, regressing *log (price_board_area)*.

| Log-Log Regression Model 7: Ecuador | |
|---|---|
| **Target Variable** | |
| *log (price_board_area)* | |
| **Numerical Variables** | |
| **Regressor** | **Parameter** |
| *log (volume)* | 0.00 |
| *log (moq)* | 0.18 |
| *log (top_load)* | 1.12 * |
| *Intercept* | -8.22 * |
| **Statistics** | |
| R | 0.5842 |
| Adjusted R | 0.4061 |
| *Significance levels: ***: 0.001 ; **: 0.01; *: 0.05* | |

*Table 18: Estimated parameters of Log-Log Regression Model 7 (Ecuador).*

If *top_load* increases 1% in an SKU from Ecuador, the impact on *price_board_area* is of 1.12%. This value is higher than the ones observed for other countries.

### 5.3.6. Country Comparisson

Features *colors* and *white* do not have a significant impact on *price_board_area* for any country. Features *volume* and *moq* only have a significant impact for Brazil and *thickness* only do for Colombia. Therefore, these variables will not be analyzed between countries.

On the contrary, we will analyze *top_load*, which has a significant impact for every country, *case_type: Shelf Ready*, which has a significant impact for Argentina and Brazil, and suppliers competitivity between all countries except Ecuador (which only has 1 supplier).

#### *5.3.6.1. Top Load*

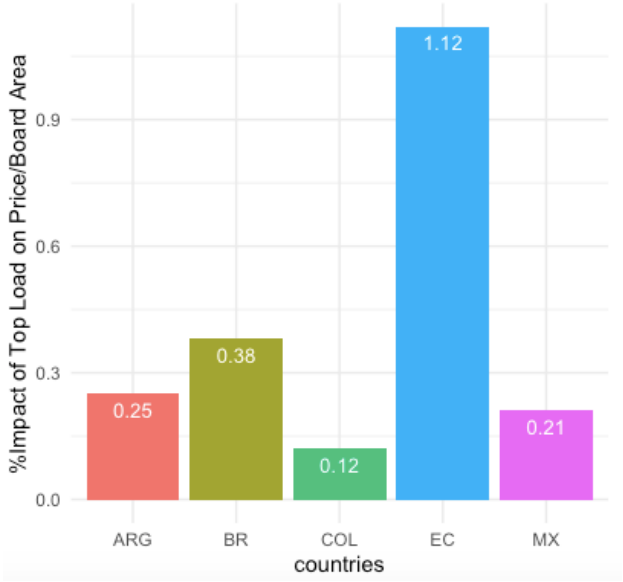


*Figure 53: % Impact of top_load on price_board_area, by country.*

Figure 53 shows that the country with most sensibility to *top_load* is Ecuador, having a 1.12% impact on *price_board_area* for each increase of 1%. Following Ecuador is Brazil with 0.38%, then Argentina with 0.25%, then Mexico with 0.21% and finally Colombia with 0.12%.

#### 5.3.6.2. Shelf Ready Cases vs Regular Cases

Figure 54 shows that shelf ready cases are 58% more expensive than regular cases in Brazil, while this number is reduced to 14% when analyzing Argentina.

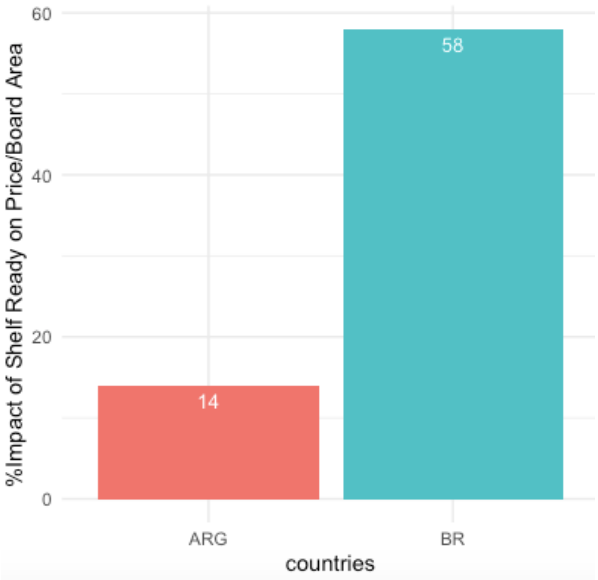


*Figure 54: % Impact of case_type: Shelf Ready on price_board_area, by country.*

#### 5.3.6.3. Suppliers

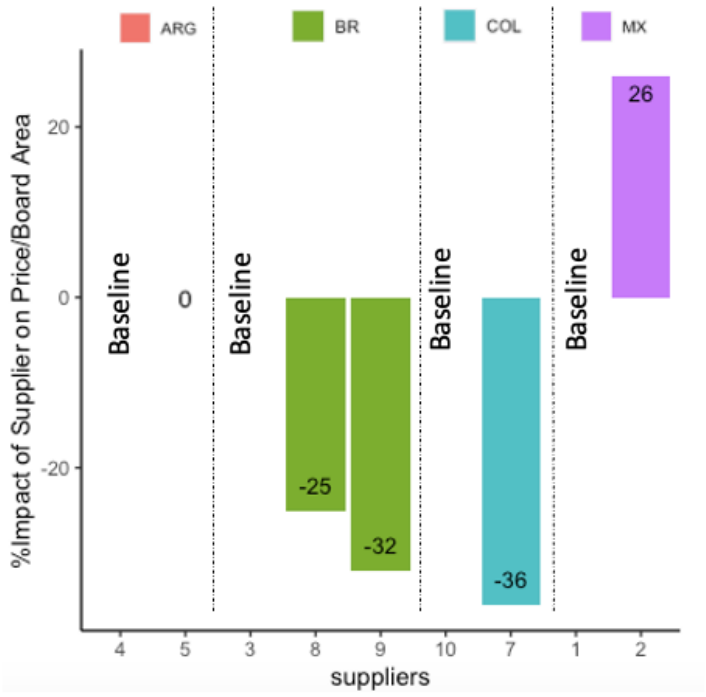


*Figure 55: % Impact of supplier on price_board_area, by country.*

Figure 55 shows that in Argentina there is no significant difference between suppliers' competitivity, but that this is not true for the other countries. Brazil has 3 suppliers, being Supplier 8 and Supplier 9 more competitive than the baseline Supplier 3, 25% and 32% respectively. Colombia has 2 suppliers, being Supplier 7 36% more competitive than baseline Supplier 10. Finally, Mexico has 2 suppliers, being Supplier 2 26% more expensive than baseline Supplier 1.

## 5.4. Regression Models Limitations

As was expected for regression models, the predictive power is poor. This can be observed in the values of $R^2$, which are mostly low. Table 19 shows the values of $R^2$ and adjusted $R^2$ for each linear regression model developed, while table 20 shows these values for the log-log regression models.

| Linear Regression Model | $R^2$ | Adjusted $R^2$ |
|---|---|---|
| **Model 1: Shelf Ready & Regular Cases** | 0.22 | 0.21 |
| **Model 2: Wrap Around** | 0.79 | 0.76 |
| **Model 3: Argentina** | 0.42 | 0.39 |
| **Model 4: Brazil** | 0.49 | 0.48 |
| **Model 5: Colombia** | 0.73 | 0.72 |
| **Model 6: Mexico** | 0.80 | 0.77 |
| **Model 7: Ecuador** | 0.62 | 0.46 |

*Table 19: $R^2$ and adjusted $R^2$ values for linear regression models.*

| Log-Log Regression Model | $R^2$ | Adjusted $R^2$ |
|---|---|---|
| **Model 1: Shelf Ready & Regular Cases** | 0.51 | 0.50 |
| **Model 2: Wrap Around** | 0.85 | 0.84 |
| **Model 3: Argentina** | 0.25 | 0.23 |
| **Model 4: Brazil** | 0.45 | 0.44 |
| **Model 5: Colombia** | 0.72 | 0.71 |
| **Model 6: Mexico** | 0.66 | 0.61 |
| **Model 7: Ecuador** | 0.58 | 0.41 |

*Table 20: $R^2$ and adjusted $R^2$ values for log-log regression models.*

The $R^2$ metric gives us an insight about how well the inputs given as predictors explain the variability of the target variable. Therefore, a low value of this metric implies that the target variable is not correctly explained by the linear effects of the predictive variables, and it may be better explained by non-linear relationships or by other additional predictive variables. For this reason, it can be stated that regression models will not perform well on predictions for this specific business problem.

Nevertheless, despite limitations of regression models for predictions, their contribution for inference is relevant. In the first 2 models it was possible to get statistically significant estimators for country dummy parameters, while in the models 3 to 7, it was possible to get statistically significant estimators for *top_load* and *suppliers* parameters. This information gives us strong insights of how these variables contribute to the value of the target variable, and how they will affect predictions when developing more flexible models such as Random Forest and XGBoost.

# 6. Predictive Models

In this section we will develop predictive models to understand which should be the price to be charged for each SKU. For this objective we will implement Random Forest and XGBoost models, as these flexible models usually deliver good results for prediction analysis. We will firstly deliver baseline models, with hyperparameters chosen by default. Then, we will make a random grid search to find the optimum hyperparameters for each model.

As we proceed with regression models, the dataset will be splitted into a Shelf Ready-Regular Cases sub-dataset and a Wrap Around sub-dataset. Hence, two models will be developed for each sub-section.

## 6.1. Random Forest

### 6.1.1. Baseline Models

For the baseline models we chose the following values for hyperparameters:

- **ntree** = 1000
- **mtry** = 5
- **nodesize** = 5

All values except **nround** = 1000 were set by default by the algorithm. In particular, the value of **mtry** is equal to $p/3$, where $p$ is the quantity of variables involved [22], being 15 in our case. Other hyperparameters as **maxnodes** and **sampsize** were not modified. In the first case, because we want trees to be pruned by **nodesize**; in the second case, because we do not have a huge amount of data, hence we want to work with all available information.

Table 21 shows the main accuracy measures calculated for these models, being *Random Forest Baseline 1* associated with shelf ready and regular cases, while *Random Forest Baseline 2* is associated with wrap around cases.

| | MSE | MAPE | MdAPE | sMAPE | sMdAPE |
|---|---|---|---|---|---|
| **Random Forest Baseline 1** | 0.701 | 223.72 | 197.48 | 97.52 | 99.37 |
| **Random Forest Baseline 2** | 0.525 | 220.86 | 68.84 | 79.43 | 51.21 |

*Table 21: Performance metrics for Random Forest baseline models.*

Surprisingly, the performance of Random Forest baseline models is worse than the ones obtained for linear and log-log regressions (table 8). We will try to revert this result in the following sub-section, adjusting hyperparameters.

### 6.1.2. Hyperparameters Optimization

#### *6.1.2.1. Random Forest Model 1: Shelf Ready and Regular Cases*

We search the optimum hyperparameters for this model, by choosing 40 random combinations from a range of possible values. This range can be observed in table 22.

| Hyperparameter | MIN | MAX |
|---|---|---|
| ntree | 1000 | 2000 |
| mtry | 3 | 7 |
| nodesize | 3 | 7 |

*Table 22: Range of possible optimum hyperparameters for Random Forest Model 1.*

This window of possible values was chosen by setting the default values of **mtry** and **nodesize** ± 2 and choosing high values for **ntree** (Random Forest has no risk of overfitting by incrementing the quantity of trees).

In table 23 we present the 10 best combinations of hyperparameters obtained, sorted descending by sMdAPE. As it was mentioned in the introduction, the criterion for deciding whether a model is better than other is the sMdAPE metric, which reflects the median of the symmetric percentual errors. Hence, table 23 shows the possible combinations, from best accuracy to worst accuracy.

| ntree | mtry | nodesize | MSE | MAPE | MdAPE | sMAPE | sMdAPE |
|---|---|---|---|---|---|---|---|
| 1992 | 7 | 4.30 | 0.052 | 10.21 | 4.90 | 9.51 | 4.92 |
| 1995 | 7 | 5.39 | 0.052 | 10.31 | 5.01 | 9.61 | 4.93 |
| 1984 | 7 | 5.64 | 0.052 | 10.30 | 5.00 | 9.58 | 4.94 |
| 1406 | 7 | 5.59 | 0.052 | 10.36 | 5.01 | 9.64 | 4.99 |
| 1050 | 6 | 4.33 | 0.052 | 10.29 | 4.94 | 9.56 | 5.02 |
| 1815 | 7 | 4.30 | 0.052 | 10.33 | 5.06 | 9.59 | 5.07 |
| 1204 | 7 | 5.21 | 0.052 | 10.29 | 5.17 | 9.61 | 5.08 |
| 1062 | 5 | 3.06 | 0.053 | 10.27 | 5.13 | 9.51 | 5.09 |
| 1643 | 7 | 6.63 | 0.051 | 10.35 | 5.17 | 9.65 | 5.10 |
| 1599 | 5 | 4.25 | 0.052 | 10.34 | 5.16 | 9.60 | 5.11 |

*Table 23: Best hyperparameters obtained for Random Forest Model 1.*

As it can be observed, the performance of the algorithm was greatly enhanced by the modification of the hyperparameters, when comparing with *Random Forest Baseline Model 1*. In particular, with **ntree** = 1992, **mtry** = 7 and **nodesize** = 4.30, the best sMdAPE was obtained. Therefore, these values are set as definitive hyperparameters for *Random Forest Model 1*, getting a sMdAPE of 4.92, which is better than the ones obtained for linear and log-log regression models (12.84 and 10.91 respectively).

#### 6.1.2.2. Random Forest Model 2: Wrap Around Cases

As well as we proceed for shelf ready and regular cases, we search the optimum hyperparameters for this model, by choosing 40 random combinations from a range of possible values. The window chosen is the same as the one chosen for shelf ready and regular cases, which can be observed in table 22.

In table 24 we present the best 10 combinations of hyperparameters obtained, sorted descending by sMdAPE.

| ntree | mtry | nodesize | MSE | MAPE | MdAPE | sMAPE | sMdAPE |
|---|---|---|---|---|---|---|---|
| 1066 | 7 | 5.24 | 0.009 | 10.04 | 3.87 | 9.20 | 3.88 |
| 1644 | 5 | 5.01 | 0.009 | 10.19 | 4.21 | 9.34 | 4.13 |
| 1991 | 4 | 3.29 | 0.009 | 9.88 | 4.35 | 9.10 | 4.34 |
| 1361 | 6 | 3.72 | 0.009 | 9.45 | 4.35 | 8.70 | 4.35 |
| 1539 | 7 | 3.48 | 0.009 | 9.41 | 4.27 | 8.64 | 4.37 |
| 1975 | 7 | 3.07 | 0.009 | 9.47 | 4.46 | 8.68 | 4.46 |
| 1733 | 5 | 5.19 | 0.009 | 10.24 | 4.47 | 9.37 | 4.47 |
| 1951 | 7 | 5.02 | 0.009 | 9.95 | 4.67 | 9.13 | 4.56 |
| 1399 | 7 | 5.78 | 0.009 | 9.81 | 4.63 | 9.00 | 4.60 |
| 1297 | 4 | 5.87 | 0.009 | 10.79 | 4.64 | 9.81 | 4.64 |

*Table 24: Best hyperparameters obtained for Random Forest Model 2.*

Table 24 shows that the performance of the model was greatly enhanced compared to the one of the *Random Forest Baseline Model 2*. In particular, the best sMdAPE was obtained for **ntree** = 1066, **mtry** = 7 and **nodesize** = 5.24. Therefore, we set these values as definitive hyperparameters for *Random Forest Model 2*, getting a sMdAPE of 3.88, which is better than the ones obtained for linear and log-log regression models (9.19 and 5.31 respectively).

### 6.1.3. Random Forest Features Importance

Random Forest allows to measure the importance of each variable, measuring the total decrease in node impurities when splitting the dataset by the corresponding feature, averaged over all trees. For regression, the node impurity is measured by the residual sum of squares [31].

#### *6.1.3.1. Random Forest Model 1: Shelf Ready and Regular Cases*

Figure 56 shows the features importance for *Random Forest Model 1*.

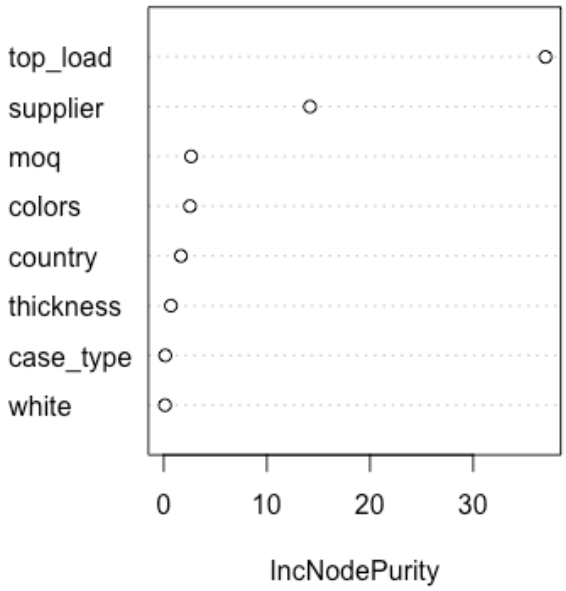


*Figure 56: Feature importance for Random Forest Model 1 (Shelf Ready and Regular Cases).*

It can be observed that *top_load* is the most relevant feature for Random Forest's decision rules to decrease the residual sum of squares. The second most important feature is *supplier*. This is aligned with what was observed in linear and log-log regressions.

#### 6.1.3.2. Random Forest Model 2: Wrap Around Cases

In the same way it was measured the feature importance of *Random Forest Model 1*, figure 57 displays the feature importance of *Random Forest Model 2*.

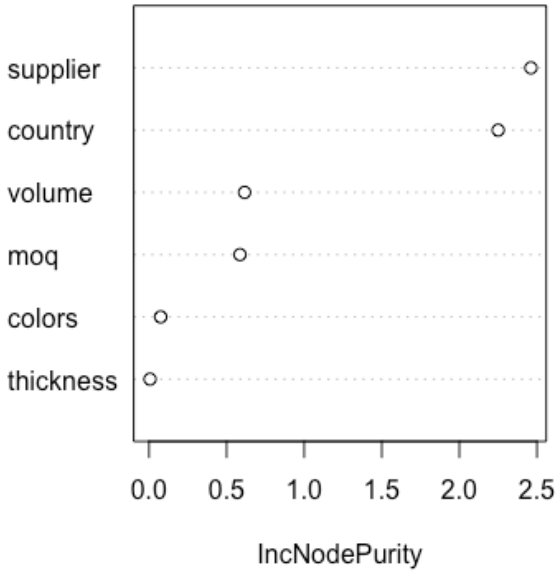


*Figure 57: Feature importance for Random Forest Model 2 (Wrap Around).*

Features *supplier* and *country* are the most relevant variables for impurity decrease in this model, followed by *volume* and *moq*.

## 6.2. XGBoost

### 6.2.1. Baseline Models

For the baseline models we chose the following values for hyperparameters:

- **nround** = 1000
- **max_depth** = 6
- **eta** = 0.3
- **gamma** = 0
- **colsample_bytree** = 1

All values except **nround** = 1000 were set by default by the algorithm. We will not use **min_child_weight** in the models, as we want the trees to be pruned by **max_depth**. As well as in Random Forest, we will not use the hyperparameter **subsample**, as we want to have full usage of the available data.

Table 25 shows the result for these models, being *XGBoost Baseline 1* associated with shelf ready and regular cases, while *XGBoost Baseline 2* is associated with wrap around cases.

| | MSE | MAPE | MdAPE | sMAPE | sMdAPE |
|---|---|---|---|---|---|
| **XGBoost Baseline 1** | 0.058 | 10.44 | 3.94 | 9.57 | 3.95 |
| **XGBoost Baseline 2** | 0.008 | 6.29 | 2.39 | 6.21 | 2.39 |

*Table 25: Performance metrics for XGBoost baseline models.*

As was expected, these models effectively show better performances than linear and log-log regression models (table 8). Furthermore, performances are also better than the ones obtained for Random Forest optimized models. In the following sub-section, we will try to further enhance these results.

### 6.2.2. Hyperparameters Optimization

#### *6.2.2.1. XGBoost Model 1: Shelf Ready and Regular Cases*

We start the search of the optimum hyperparameters choosing 40 random combinations, from a wide range of possible values. This window of values can be observed in table 26.

| Hyperparameter | MIN | MAX |
|---|---|---|
| **nrounds** | 100 | 2000 |
| **max_depth** | 4 | 8 |
| **eta** | 0.20 | 0.40 |
| **gamma** | 0.000 | 0.200 |
| **colsample_bytree** | 0.80 | 1.00 |

*Table 26: Range of possible optimum hyperparameters for XGBoost Model 1.*

In table 27 we present the 10 best combinations of hyperparameters obtained, sorted descending by sMdAPE.

| nround | max_depth | eta | gamma | colsample_bytree | MSE | MAPE | MdAPE | sMAPE | sMdAPE |
|---|---|---|---|---|---|---|---|---|---|
| 919 | 4 | 0.34 | 0.001 | 0.99 | 0.062 | 11.22 | 4.96 | 10.27 | 5.03 |
| 1917 | 5 | 0.21 | 0.002 | 0.85 | 0.061 | 11.39 | 5.37 | 10.31 | 5.34 |
| 906 | 8 | 0.34 | 0.044 | 0.88 | 0.066 | 12.70 | 6.47 | 11.25 | 6.50 |
| 873 | 7 | 0.31 | 0.029 | 0.92 | 0.057 | 12.26 | 6.78 | 11.22 | 6.70 |
| 360 | 5 | 0.23 | 0.031 | 0.95 | 0.054 | 12.50 | 7.10 | 11.42 | 7.09 |
| 1240 | 4 | 0.32 | 0.022 | 0.98 | 0.060 | 12.60 | 7.18 | 11.53 | 7.23 |
| 214 | 6 | 0.31 | 0.047 | 0.89 | 0.053 | 12.53 | 7.34 | 11.60 | 7.29 |
| 769 | 4 | 0.25 | 0.037 | 0.91 | 0.060 | 13.09 | 7.53 | 11.94 | 7.55 |
| 972 | 6 | 0.37 | 0.060 | 0.95 | 0.060 | 12.85 | 7.54 | 11.65 | 7.55 |
| 1908 | 5 | 0.22 | 0.061 | 0.87 | 0.055 | 12.99 | 7.60 | 12.03 | 7.58 |

*Table 27: Best hyperparameters obtained for XGBoost Model 1, iterating over a range of possible values.*

Unfortunately, none of the hyperparameters combinations chosen delivered better results than the one obtained for *XGBoost Baseline 1*, which uses hyperparameters set by default. In a second attempt to enhance the performance of the model, we set the values of

**max_depth**, **eta**, **gamma** and **colsample_bytree** by default, and we iterate over different values of **nround**. The results are presented in table 28.

| nround | max_depth | eta | gamma | colsample_bytree | MSE | MAPE | MdAPE | sMAPE | sMdAPE |
|---|---|---|---|---|---|---|---|---|---|
| 300 | 6 | 0.30 | 0.000 | 1.00 | 0.058 | 10.44 | 3.88 | 9.57 | 3.95 |
| 500 | 6 | 0.30 | 0.000 | 1.00 | 0.058 | 10.44 | 3.88 | 9.57 | 3.95 |
| 700 | 6 | 0.30 | 0.000 | 1.00 | 0.058 | 10.44 | 3.88 | 9.57 | 3.95 |
| 900 | 6 | 0.30 | 0.000 | 1.00 | 0.058 | 10.44 | 3.88 | 9.57 | 3.95 |
| 1100 | 6 | 0.30 | 0.000 | 1.00 | 0.058 | 10.44 | 3.88 | 9.57 | 3.95 |
| 1300 | 6 | 0.30 | 0.000 | 1.00 | 0.058 | 10.44 | 3.88 | 9.57 | 3.95 |
| 1500 | 6 | 0.30 | 0.000 | 1.00 | 0.058 | 10.44 | 3.88 | 9.57 | 3.95 |

*Table 28: Best hyperparameters obtained for XGBoost Model 1, iterating nround.*

As it can be observed, results did not get better when incrementing **nround**. Therefore, we proceed to iterate model's hyperparameters in a neighborhood of baseline's hyperparameters. Table 29 shows best 10 combinations.

| nround | max_depth | eta | gamma | colsample_bytree | MSE | MAPE | MdAPE | sMAPE | sMdAPE |
|---|---|---|---|---|---|---|---|---|---|
| 100 | 6 | 0.30 | 0.000 | 1.00 | 0.0579 | 10.07 | 4.21 | 9.28 | 4.23 |
| 100 | 6 | 0.30 | 0.000 | 1.00 | 0.0582 | 10.33 | 4.29 | 9.43 | 4.26 |
| 100 | 6 | 0.30 | 0.000 | 0.98 | 0.0572 | 10.48 | 4.31 | 9.62 | 4.30 |
| 100 | 6 | 0.30 | 0.000 | 1.00 | 0.0583 | 10.29 | 4.39 | 9.43 | 4.32 |
| 100 | 6 | 0.30 | 0.000 | 1.00 | 0.0572 | 10.35 | 4.35 | 9.50 | 4.33 |
| 100 | 6 | 0.30 | 0.000 | 1.00 | 0.0567 | 10.18 | 4.38 | 9.42 | 4.35 |
| 100 | 6 | 0.30 | 0.001 | 0.97 | 0.0556 | 10.46 | 4.42 | 9.81 | 4.40 |
| 100 | 6 | 0.30 | 0.000 | 0.97 | 0.0588 | 10.53 | 4.45 | 9.69 | 4.43 |
| 100 | 6 | 0.30 | 0.000 | 1.00 | 0.0589 | 10.40 | 4.53 | 9.50 | 4.52 |
| 100 | 6 | 0.30 | 0.002 | 0.97 | 0.0593 | 10.60 | 4.74 | 9.59 | 4.71 |

*Table 29: Best hyperparameters obtained for XGBoost Model 1, iterating in a neighborhood of baseline's hyperparameters.*

Again, performance obtained with default hyperparameters could not be enhanced. Hence, we define *XGBoost Model 1* equal to the baseline model already developed (same hyperparameters).

#### 6.2.2.2. XGBoost Model 2: Wrap Around Cases

As well as we proceed for shelf ready and regular cases, we search the optimum hyperparameters for this model, by choosing 40 random combinations from a range of possible values. In particular, the window chosen is the same as the one chosen for shelf ready and regular cases, which can be observed in table 26.

In table 30 we present the best 10 combinations of hyperparameters obtained, sorted descending by sMdAPE.

| nround | max_depth | eta | gamma | colsample_bytree | MSE | MAPE | MdAPE | sMAPE | sMdAPE |
|---|---|---|---|---|---|---|---|---|---|
| 1795 | 7 | 0.37 | 0.012 | 0.87 | 0.009 | 8.77 | 3.77 | 8.35 | 3.76 |
| 492 | 5 | 0.34 | 0.019 | 0.87 | 0.008 | 10.57 | 5.40 | 9.86 | 5.33 |
| 736 | 5 | 0.25 | 0.011 | 0.84 | 0.008 | 9.04 | 5.21 | 8.60 | 5.35 |
| 390 | 5 | 0.34 | 0.011 | 0.97 | 0.009 | 9.29 | 5.46 | 8.84 | 5.46 |
| 1261 | 4 | 0.35 | 0.023 | 0.90 | 0.010 | 10.74 | 6.16 | 10.04 | 6.07 |
| 412 | 6 | 0.26 | 0.025 | 0.86 | 0.009 | 11.90 | 6.14 | 10.95 | 6.07 |
| 1798 | 6 | 0.29 | 0.018 | 0.87 | 0.009 | 10.41 | 6.32 | 9.72 | 6.23 |
| 953 | 8 | 0.30 | 0.024 | 0.88 | 0.010 | 11.51 | 6.91 | 10.74 | 6.68 |
| 927 | 6 | 0.24 | 0.024 | 0.98 | 0.010 | 11.88 | 6.73 | 10.97 | 6.81 |
| 1863 | 8 | 0.23 | 0.028 | 0.97 | 0.010 | 12.44 | 7.36 | 11.42 | 7.44 |

*Table 30: Best hyperparameters obtained for XGBoost Model 2, iterating over a range of possible values.*

As occurred with shelf ready and regular cases, none of the hyperparameters combinations used deliver better results than the one obtained for *XGBoost Baseline 2*, which uses hyperparameters set by default. So, in a second attempt to enhance the performance of the model, we set the values of **max_depth**, **eta**, **gamma** and **colsample_bytree** by default, and we iterate over different values of **nround**. The results are presented in table 31.

| nround | max_depth | eta | gamma | colsample_bytree | MSE | MAPE | MdAPE | sMAPE | sMdAPE |
|---|---|---|---|---|---|---|---|---|---|
| 300 | 6 | 0.30 | 0.000 | 1.00 | 0.058 | 10.44 | 3.88 | 9.57 | 3.95 |
| 500 | 6 | 0.30 | 0.000 | 1.00 | 0.058 | 10.44 | 3.88 | 9.57 | 3.95 |
| 700 | 6 | 0.30 | 0.000 | 1.00 | 0.058 | 10.44 | 3.88 | 9.57 | 3.95 |
| 900 | 6 | 0.30 | 0.000 | 1.00 | 0.058 | 10.44 | 3.88 | 9.57 | 3.95 |
| 1100 | 6 | 0.30 | 0.000 | 1.00 | 0.058 | 10.44 | 3.88 | 9.57 | 3.95 |
| 1300 | 6 | 0.30 | 0.000 | 1.00 | 0.058 | 10.44 | 3.88 | 9.57 | 3.95 |
| 1500 | 6 | 0.30 | 0.000 | 1.00 | 0.058 | 10.44 | 3.88 | 9.57 | 3.95 |

*Table 31: Best hyperparameters obtained for XGBoost Model 2, iterating nround.*

Again, as occurred with shelf ready and regular cases, sMdAPE metric was not enhanced by a greater value of **nround**. Therefore, we proceed to iterate model's hyperparameters in a neighborhood of baseline's hyperparameters. Table 32 shows best 10 combinations.

| nround | max_depth | eta | gamma | colsample_bytree | MSE | MAPE | MdAPE | sMAPE | sMdAPE |
|---|---|---|---|---|---|---|---|---|---|
| 100 | 6 | 0.30 | 0.0002 | 0.96 | 0.0079 | 6.36 | 1.54 | 6.22 | 1.54 |
| 100 | 6 | 0.30 | 0.0003 | 0.90 | 0.0072 | 6.40 | 2.31 | 6.27 | 2.31 |
| 100 | 6 | 0.30 | 0.0007 | 0.98 | 0.0087 | 6.49 | 2.39 | 6.36 | 2.36 |
| 100 | 6 | 0.30 | 0.0007 | 0.92 | 0.0085 | 6.59 | 2.41 | 6.41 | 2.41 |
| 100 | 6 | 0.30 | 0.0009 | 1.00 | 0.0084 | 6.66 | 2.56 | 6.53 | 2.53 |
| 100 | 6 | 0.31 | 0.0009 | 0.95 | 0.0077 | 6.57 | 2.67 | 6.41 | 2.67 |
| 100 | 6 | 0.29 | 0.0009 | 0.93 | 0.0071 | 6.48 | 2.77 | 6.28 | 2.80 |
| 100 | 6 | 0.31 | 0.0022 | 0.96 | 0.0085 | 6.87 | 2.84 | 6.71 | 2.89 |
| 100 | 6 | 0.31 | 0.0040 | 0.92 | 0.0075 | 7.15 | 2.95 | 6.89 | 2.97 |
| 100 | 6 | 0.30 | 0.0010 | 0.96 | 0.0085 | 6.60 | 3.08 | 6.49 | 3.03 |

*Table 32: Best hyperparameters obtained for XGBoost Model 2, iterating in a neighborhood of baseline's hyperparameters.*

As it can be observed, 3 combinations with a better performance that *XGBoost Baseline Model 2* were found. In particular, the following values delivered the best performance (sMdAPE=1.54), hence they are defined as definitive hyperparameters for *XGBoost Model 2*:

- **nround** = 100
- **max_depth** = 6
- **eta** = 0.3
- **gamma** = 0.0002
- **colsample_bytree** = 0.96

### 6.2.3. XGBoost Features Importance

A benefit of using boosting is that, after boosted trees are constructed, it is relatively straightforward to retrieve importance scores for each attribute. Generally, importance provides a score that indicates how useful or valuable each feature was in the construction of the boosted decision trees within the model. The more an attribute is used to make key decisions with decision trees, the higher its relative importance. This importance is calculated explicitly for each attribute in the dataset, allowing attributes to be ranked and compared to each other [32].

#### *6.2.3.1. XGBoost Model 1: Shelf Ready and Regular Cases*

Figure 58 shows the importance of the predictor variables for this model.

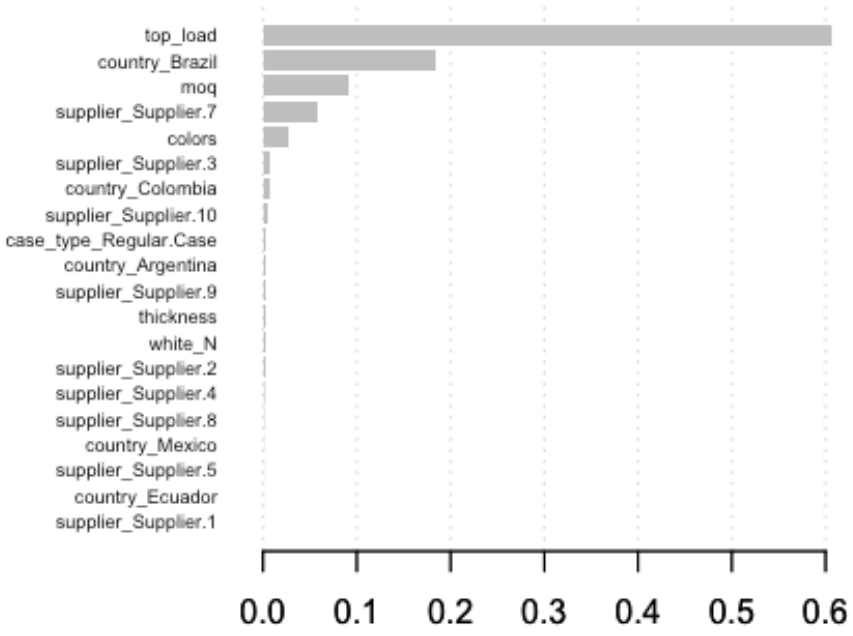


*Figure 58: Feature importance for XGBoost Model 1 (Shelf Ready and Regular Cases).*

It can be observed that variables that were significant in linear and log-log regressions such as *top_load* and *country* still hold a relevant importance in boosting. Surprisingly, variables as *moq* and *colors* which were not relevant in regressions, play an important role in this method as well.

Features that do not have any importance on the decision rules established by the algorithm are not displayed in the plot. Finally, it is worth mentioning a key difference between XGBoost and Random Forest methods to measure categorical feature importance: while the first considers importance of one hot encoded regressors (e.g. *country_Brazil*, s*upplier_Supplier.7*), the second considers categorical features importance as a whole (e.g. *country, supplier*).

#### 6.2.3.2. XGBoost Model 2: Wrap Around Cases

In the same way feature importance was measured for *XGBoost Model 1*, feature importance for *XGBoost Model 2* was measured and is displayed in figure 59.

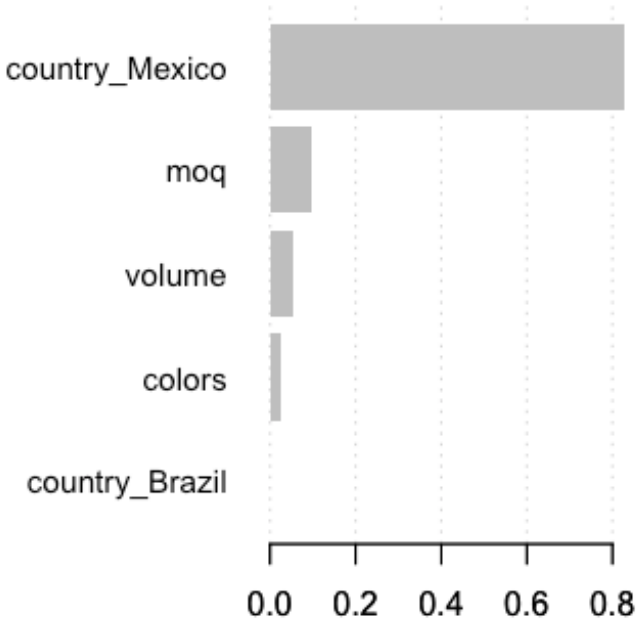


*Figure 59: Feature importance for XGBoost Model 2 (Wrap Around).*

As it was expected, the most important variable is *country_Mexico*; it was observed in the exploratory analysis that wrap around cases from this country are significantly more expensive that the same type of cases in other countries. Less important but relevant features for the algorithm rules are *moq, volume* and *colors*.

## 6.3. Models Performance Comparison

Figure 60 shows the performance of all developed models, both for Shelf Ready - Regular Cases and Wrap Around sub-datasets, according to the sMdAPE metric.

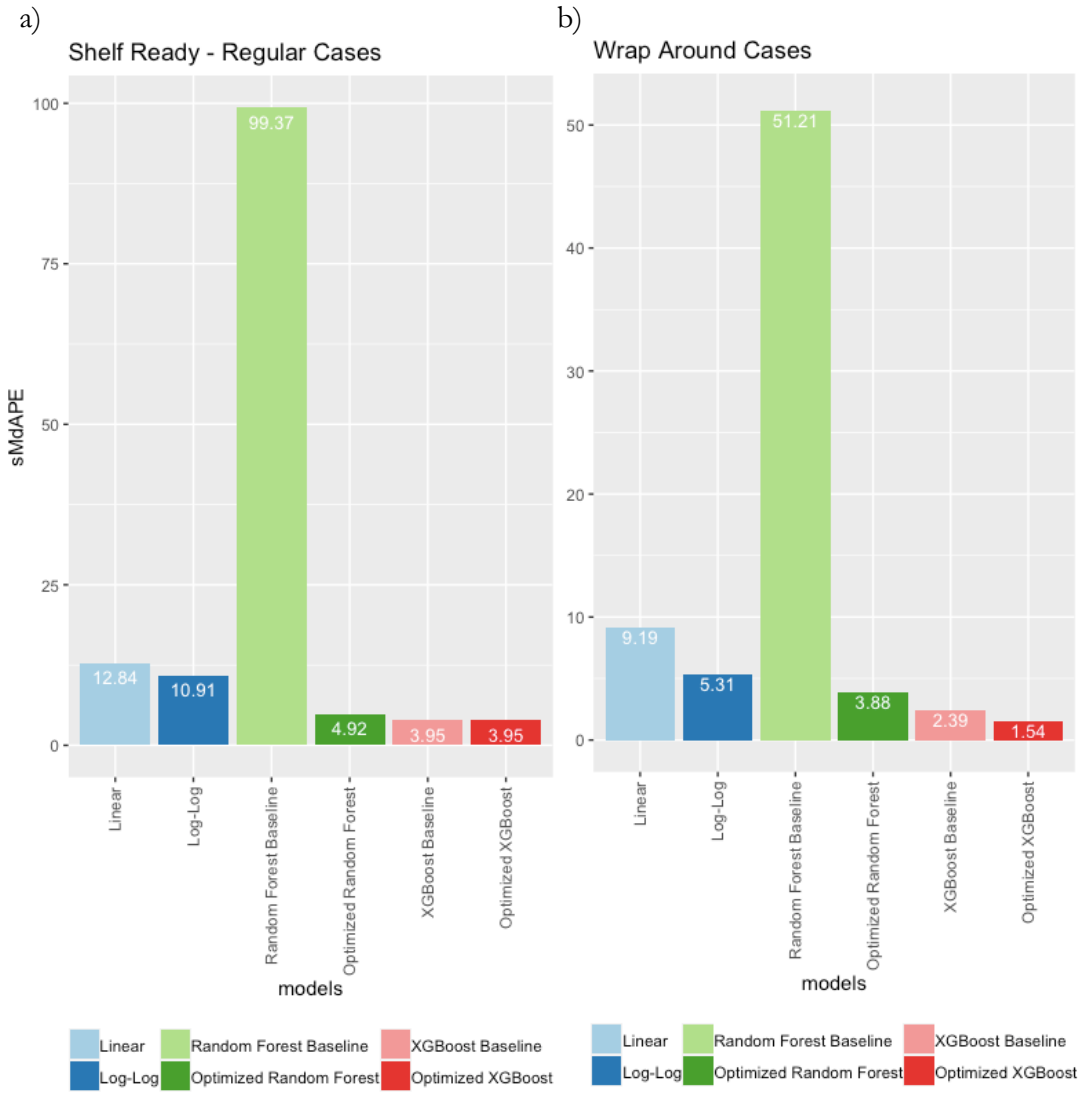


*Figure 60: Performance of all developed models.*
*a) Shelf Ready – Regular Cases; b) Wrap Around.*

It can be observed that the best performances are provided by optimized predictive models. XGBoost achieved the best sMdAPE, not being sensible to hyperparameters optimization in the case of Shelf Ready – Regular Cases, but slightly enhanced by this technique for Wrap Around sub-dataset. On the contrary, Random Forest showed a great influence of chosen hyperparameters for both sub-datasets; however, the performances achieved are worse than the ones obtained with XGBoost. As it was expected, log-log and linear regression models showed inferior performances than optimized predictive models.

## 6.4. Error Distribution for XGBoost

As was shown in the previous section, XGBoost presented the best performances according to the sMdAPE metric, both for Shelf Ready - Regular Cases and Wrap Around sub-datasets. Therefore, for both XGBoost models developed, we proceed to analyze the distribution of the error function. In order to be aligned with the sMdAPE metric, we define the error function as follows:

$$e_i = 200 * \frac{y_i - \hat{f}(x_i)}{y_i + \hat{f}(x_i)} \quad (34)$$

### 6.4.1. XGBoost Model 1: Shelf Ready and Regular Cases

Figure 61 shows the distribution of the error function for *XGBoost Model 1* predictions over the Shelf Ready – Regular Cases sub-dataset.

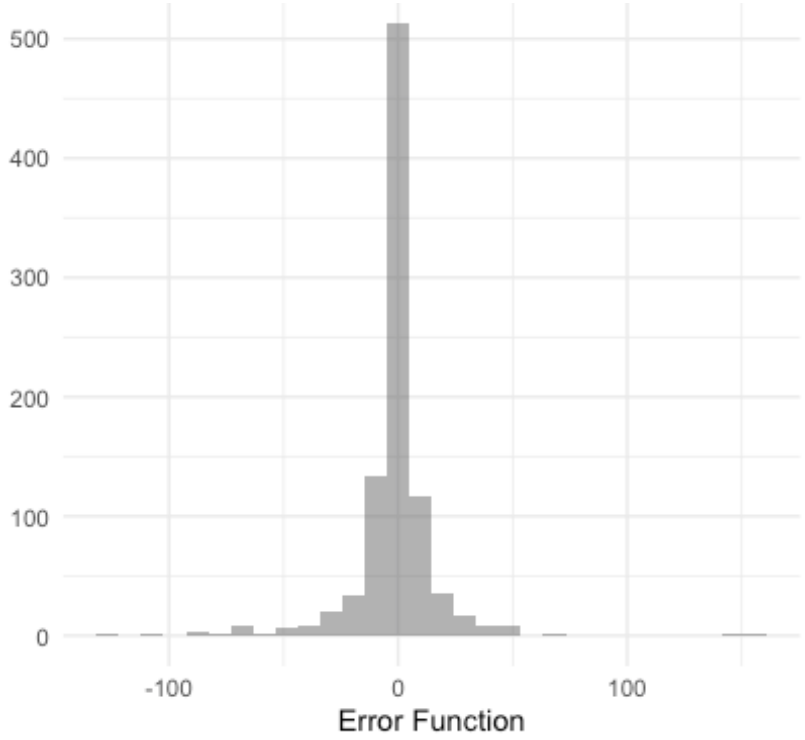


*Figure 61: Error function distribution for XGBoost Model 1 predictions (Shelf Ready and Regular Cases).*

It can be observed that the error function is normally distributed. The mean is -0.92 (close to 0) and the standard deviation is 18.73. With the objective of understanding whether the mean of the error distribution might have a bias towards a value different than 0, we run a two-sided t-test with the following structure:

$$H_0: \mu_{e_i} = 0 \quad (35)$$

$$H_1: \mu_{e_i} \neq 0 \quad (36)$$

The result is that we cannot reject the $H_0$ hypothesis with a confidence level of 95%, as the p-value of this t-test is 0.2211. Therefore, we cannot state that the mean of the distribution is different than 0 with this confidence level.

### 6.4.2. XGBoost Model 2: Wrap Around Cases

Figure 62 shows the distribution of the error function for *XGBoost Model 2* predictions over the Wrap Around sub-dataset.

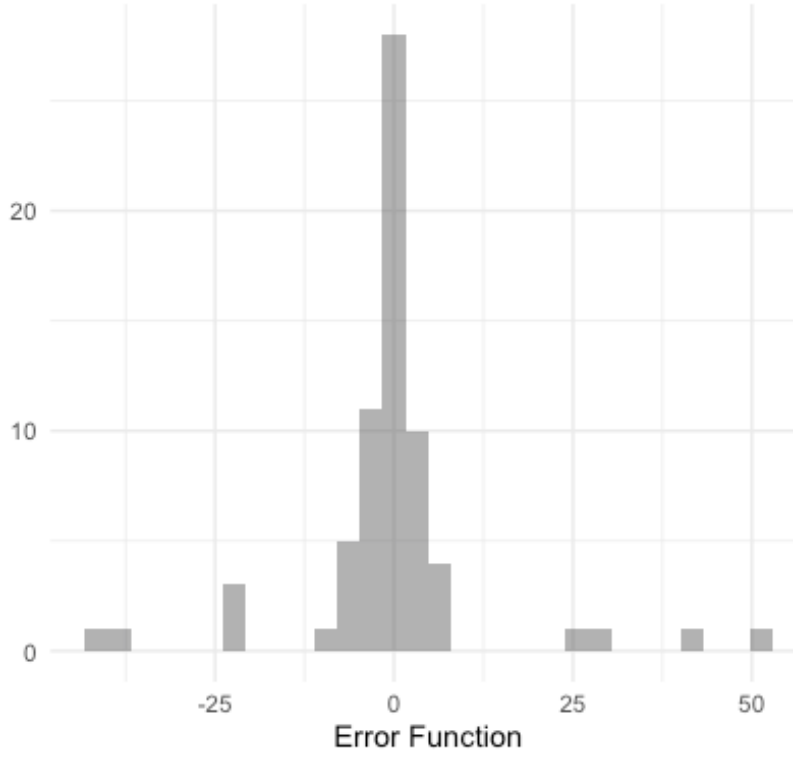


*Figure 62: Error function distribution for XGBoost Model 2 predictions (Wrap Around).*

As occurred with *XGBoost Model 1*, *XGBoost Model 2* error function is also normally distributed. The mean is -0.31 and the standard deviation is 12.77.

We ran an analogous t-test analysis, according to equations (35) and (36), to understand the possibility for the mean to have a bias towards a value different than 0. The result is that we cannot reject the $H_0$ hypothesis with a confidence level of 95%, as the p-value of this t-test is 0.8424. Therefore, we cannot state that the mean of the distribution is different than 0 with this confidence level.

# 7. Business Results

In this section we will use the most accurate models developed in section 6, *XGBoost Model 1* and *XGBoost Model 2*, to predict the *price_board_area* of the current SKUs. For each observation, we will calculate the error value $e_i$, given by equation (34). When this value exceeds the 95% confidence interval given by the error function normal distribution, we will define a negotiation opportunity. In particular, we will be interested in detecting negotiation opportunities to reduce prices; that is, when the predicted values of *price_board_area* are lower than the real values of *price_board_area*[2] ($e_i > 0$).

A 95% confidence interval for a normal distribution can be defined as follows:

$$95\%\ CONF\ INT\ (X) = \bar{X} \pm 1.96 * sd(X) \quad (37)$$

Thus, given values calculated in section 6, we define the two following 95% confidence intervals:

- XGBoost Model 1: **( -37.63 , 35.79 ).**
- XGBoost Model 2: **( -25.34 , 24.72 ).**

## 7.1. Methodology

### 7.1.1. Shelf Ready and Regular Cases

1. Training of *XGBoost Model 1* with all shelf ready and regular cases observations, except the 1st observation.
2. Prediction of the 1st observation, with the model already trained.
3. Iteration of steps 1 and 2 for all shelf ready and regular cases observations.
4. Calculation of the error value $e_i$ given by equation (34), for each observation in the sub-dataset.

A positive value of $e_i$ means that the predicted price of the given case is lower than the real price; hence, indicating a negotiation opportunity for price reduction.

5. Segmentation of observations for which $e_i$ > 35.79, being 35.79 the upper limit of the 95% confidence interval for $e_i$ distribution in *XGBoost Model 1*.
6. For segmented observations, calculation of negotiation opportunities (NO), in euros per year. This is given by equation (38):

$$NO_i = (price\_board\_area_i - predicted\_price\_board\_area_i) * board\_area_i * volume_i \quad (38)$$

[2] It is interesting to notice that in this case we are focused on negotiation opportunities derived from low predicted prices, because we have the perspective of the buyer firm. However, suppliers could make the same analysis and focus on negotiation opportunities derived from high predicted prices, looking for opportunities to increment current prices.

### 7.1.2. Wrap Around Cases

1. Training of *XGBoost Model* 2 with all wrap around cases observations, except the 1st observation.
2. Prediction of the 1st observation, with the model already trained.
3. Iteration of steps 1 and 2 for all wrap around cases observations.
4. Calculation of the error value $e_i$ given by equation (34), for each observation in the dataset.
5. Segmentation of observations for which $e_i > 24.72$, being 24.72 the upper limit of the 95% confidence interval for $e_i$ distribution in *XGBoost Model 2*.
6. For segmented observations, calculation of negotiation opportunities (NO), in euros per year, according to equation (38).

## 7.2. Saving Results Analysis

With this exercise we found 23 opportunities negotiation opportunities within the dataset. Figure 63.a shows the distribution of opportunities by country. Given that the dataset is not balanced regarding the quantity of observations per country, in figure 63.b we show the quantity of opportunities per 100 observations for each country.

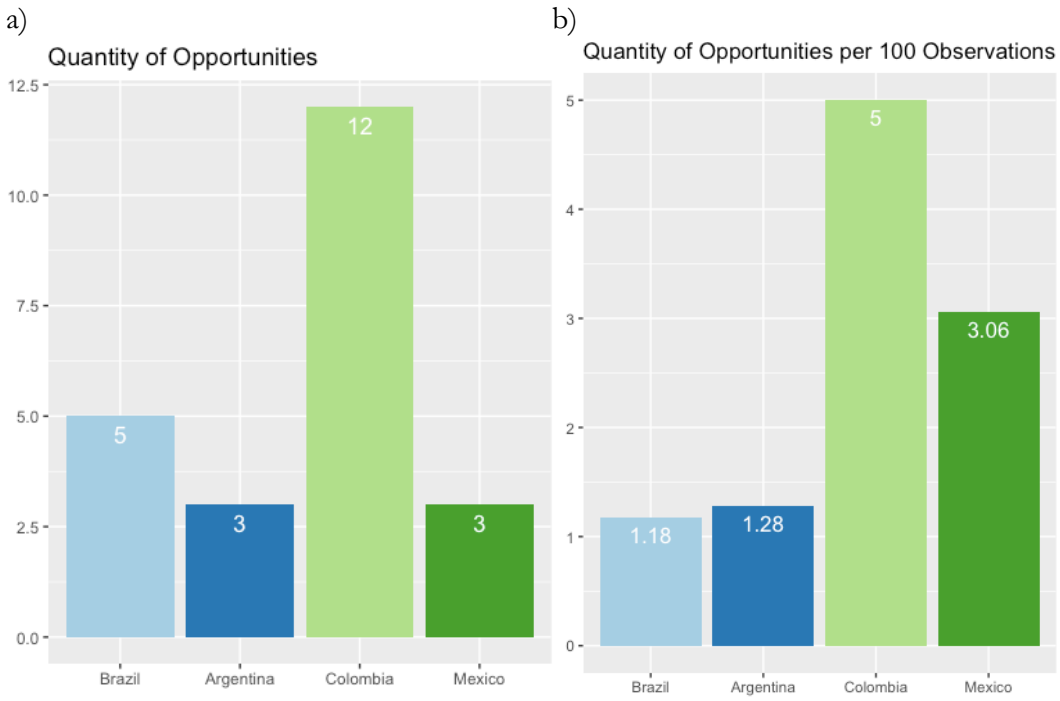


*Figure 63: a) Quantity of opportunities per country;*
*b) Quantity of opportunities per 100 observations for each country.*

It can be observed that most potential opportunities come from Colombia (12), followed by Brazil (5), and finally Argentina (3) and Mexico (3). In terms of opportunities per 100 observations, Colombia still leads the ranking (5), followed by Mexico (3.06), Argentina (1.28) and Brazil (1.18) respectively. Ecuador does not present saving opportunities.

In terms of potential savings, we found opportunities for € 906,000. Figure 64.a shows the potential savings splitted by country. It can be observed that most savings come from

Colombia (€ 785,000), followed by Mexico (€ 69,000), Brazil (€ 46,000) and Argentina (€ 6,000) respectively. In addition, figure 64.b shows the percentual weight of savings over the respective spends. Colombia has the biggest opportunities compared with its spend (23.84%), followed by Mexico (0.76%), Brazil (0.12%) and finally Argentina (0.03%).

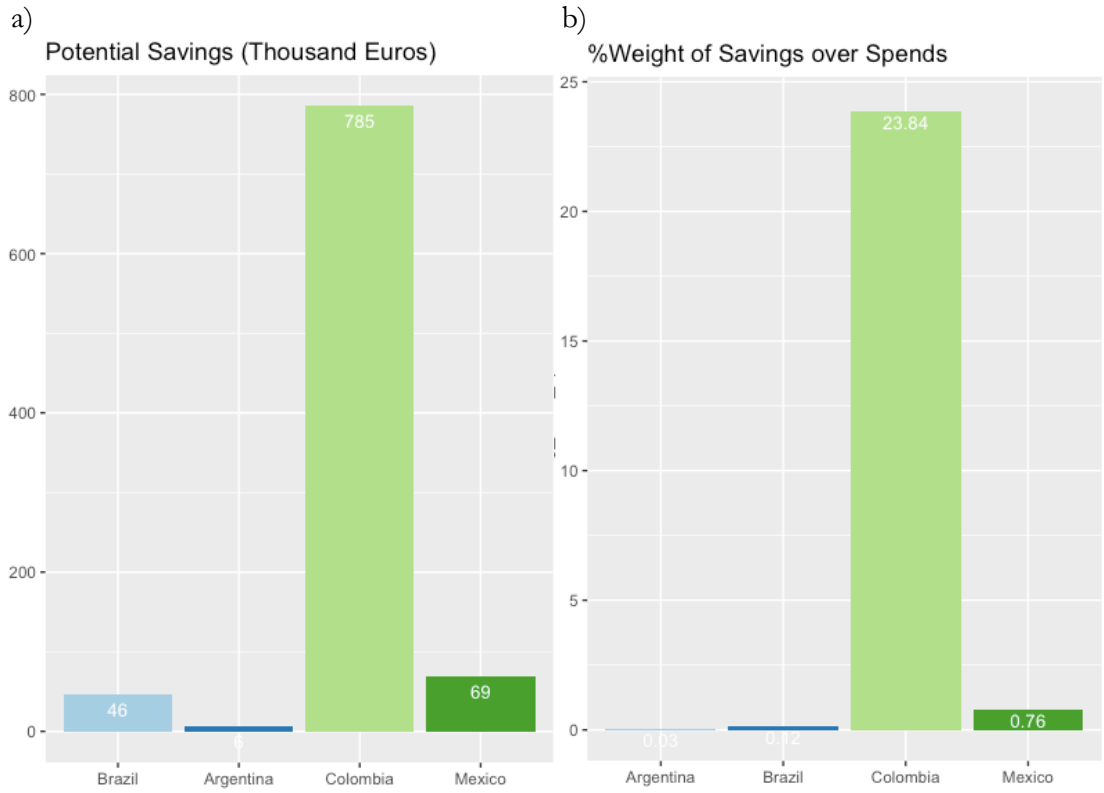


*Figure 64: a) Potential savings by country; b) %Weight of savings over countries spends.*

Saving opportunities are presented, by *country* and *case_type*, in figure 65. For each SKU, the letter *R* is shown in case of a regular case, the letter *W* in case of a wrap around case, and the letter *S* in case of a shelf ready case. There are 19 opportunities corresponding to the first group, 3 opportunities corresponding to the second group, and 1 opportunity corresponding to the last group.

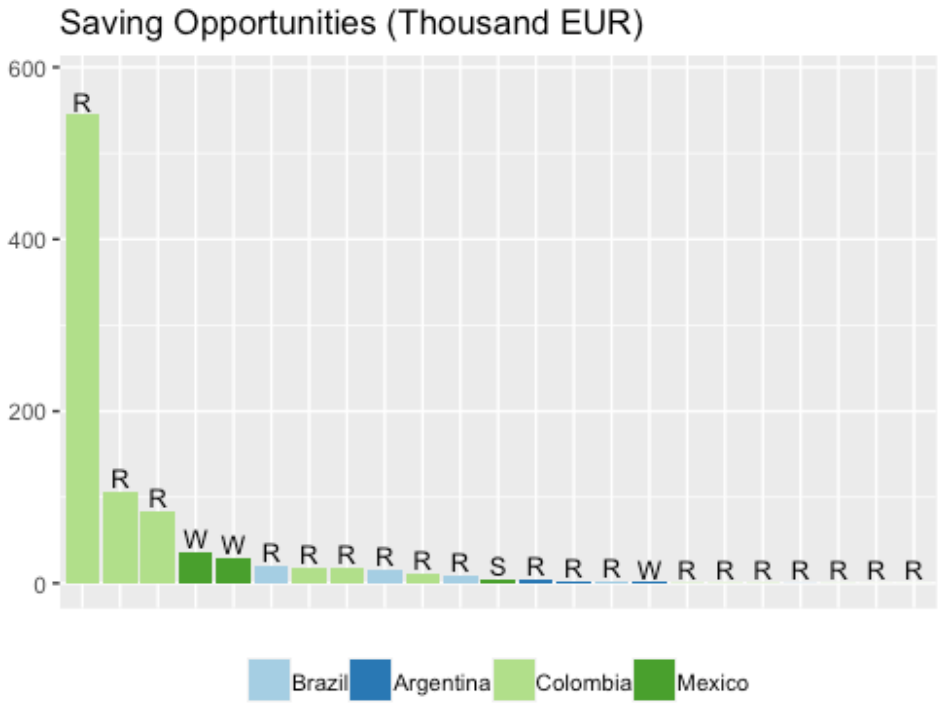


*Figure 65: Saving opportunities, by country and case_type.*

Is interesting to notice that 7 of the 8 *price_board_area* outliers detected in the exploratory analysis with Rosner's test (figure 21) were detected as negotiation opportunities by XGBoost algorithm: 3 observations from Colombia, 3 from Mexico and 1 for Argentina. Only 1 outlier from Mexico was not detected by the model.

Finally, figure 66 shows the distribution of savings by category.

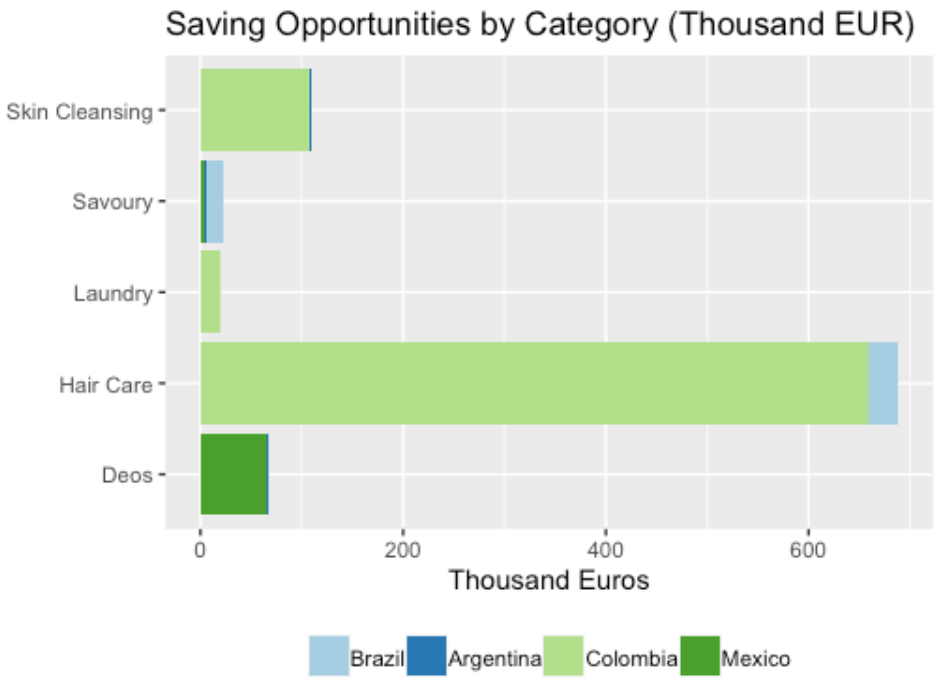


*Figure 66: Saving opportunities, by category and country.*

It can be observed that most potential savings come from Hair Care Colombia's observations. This is of no surprise, as 3 big outliers from the dataset correspond to this country and this category. There are also opportunities in Colombia for Skin Cleansing category. In addition, relevant potential savings come from Deos' category in Mexico. Again, this was expected, as 4 *price_board_area* outliers were detected in the exploratory analysis, from which 3 became saving opportunities.

# 8. Conclusions

Using a dataset with technical and commercial features of a portfolio of corrugated cases of an international FMCG firm in LATAM, this project focused on two main objectives:

- Measuring the influence of the different features on the prices of the cases.
- Predicting what prices should be charged for each SKU, to unveil negotiation opportunities.

For both activities, the target variable was *price_board_area,* which represents the price of a given case in euros, divided by the board area needed for manufacturing this case. In this way, our analysis and predictions were not biased by the size of the respective cases.

## 8.1. First Objective: Influence of Features on Prices

### 8.1.1. Quantitative Variables

In the first place, it was observed that shelf ready cases are more expensive than regular cases, being this effect bigger in Brazil that in Argentina (there are no shelf ready cases in other countries), and that, generally, wrap around cases have the most competitive prices in the portfolio. However, we discovered that this rule does not apply to Mexico, where wrap around cases have the highest prices.

In the second place, it was observed that, for shelf ready and regular cases, Brazil is the most competitive country, followed by Colombia, Ecuador, Argentina and Mexico respectively. When focusing on wrap around cases, Brazil has again the lowest prices, followed by Argentina and finally Mexico. Colombia and Ecuador do not have wrap around cases in their respective portfolios.

Thirdly, it was observed that there is not a significant difference between prices of both suppliers of Argentina, but that suppliers from Brazil, Colombia and Mexico show different competitivity. In particular, when comparing to respective biggest suppliers in each country (baselines), Brazil has two more competitive suppliers, Colombia has one more competitive supplier, while Mexico has a more expensive alternative supplier.

Finally, we did not find a significant correlation between whiteboard manufactured cases and prices. This was a surprise, as a positive correlation between these variables was expected. The absence of this positive correlation might be explained by the low quantity of *white* observations in the dataset, which derives in an underrepresentation of this feature.

### 8.1.2. Qualitative Variables

When analyzing qualitative variables, we found that the most important one is *top_load*. In particular, we measured the percentual impact on *price_board_area* of incrementing this variable by a 1%. We found that the greatest positive impact is registered in Ecuador, followed by Brazil, Argentina, Mexico and Colombia respectively.

Influence of other qualitative variables on prices is not so unequivocal. Thickness has a strong positive impact in Colombia; however, this effect was not observed in other countries. In addition, *moq* has a very weak negative impact in Brazil, being not possible to extend this

result to the rest of the countries. Surprisingly, other variables as *volume* and *colors* do not have a significant influence on the prices of the cases.

## 8.2. Second Objective: Prediction of Prices and Saving Opportunities

For predicting prices of cases, we splitted the dataset in two, one containing shelf ready and regular cases and the other containing wrap around cases. In this way, it was possible to include all relevant variables for both type of observations.

For each sub-dataset, we developed a linear regression model, a log-log regression model, a Random Forest model, and a XGBoost model. The first two models were used as benchmarks, as we acknowledge that their predictive power is not high, while the last two models were optimized for maximizing prediction accuracy. We used LOOCV for calculating accuracy measures, and we used the sMdAPE measure as an objective criterion to decide whether a model is better than other.

When analyzing accuracy, we found that the best model was the XGBoost, followed by Random Forest, log-log regression, and linear regression respectively. For XGBoost, the search of optimum hyperparameters did not enhance the performance for Shelf Ready – Regular Cases sub-dataset but slightly did for Wrap Around sub-dataset. In the case of Random Forest, this technique was imperative to get a good performance for both sub-datasets.

XGBoost model was used to predict which prices should be charged for each SKU of the portfolio. When the error of a particular prediction exceeds the upper limit of the 95% confidence interval of the error function normal distribution, we interpreted that this difference was not due to the variance of the model, but rather to a negotiation imperfection; hence, detecting a negotiation saving opportunity.

In this way, we detected 23 negotiation opportunities with an associated potential saving of € 906,000 per year. When analyzing these opportunities, we discovered that most of them are concentrated in Colombia, representing 23.84% of its spend. In addition, we found that most opportunities in Colombia come from Hair Care's portfolio, followed by Skin Cleansing's portfolio. Finally, some relevant opportunities come from Deos' portfolio in Mexico, representing 0.76% of its spend.

## 8.3. Final Remarks and Next Steps

In the current project we posed the question of whether applying machine learning techiques to a corrugated cases portfolio dataset could propel the detection of prices outliers, in order to unveil negotiation opportunities. Not only this objective was accomplished, finding potential savings opportunities of € 906,000 per year, but also a full analysis of the current portfolio was developed. In this way, the influence of each technical and commercial parameter was measured, giving visibility on which could be influenced in a competitive direction.

Next steps naturally include the renegotiation of mapped opportunities, so that most of the potential savings are captured, but also:

- Review suppliers strategy in Brazil and Colombia, to allocate more volume in peripheric suppliers, which demonstrated to be more competitive.
- Understand if cases could be supplied from Brazil and Colombia to other countries, being these countries the most competitive ones.
- Understand if thickness of cases can be reduced in SKUs from Colombia, being this parameter relevant for this country.
- Understand if flute B wrap around cases in Brazil could be manufactured with flute E, hence reducing their respective prices.
- Challenge wrap around cases prices in Mexico, as these are not aligned with what is observed in other countries of the region.
- Challenge R&D team to reduce top load of current cases, being this parameter key in all countries of the region, especially in Ecuador and Brazil.

## 8.4. Contribution to Other Industries

The present work focused on predicting what prices should be charged for each corrugated case, to unveil negotiation opportunities. For this task, flexible machine learning algorithms as XGBoost and Random Forest were trained with current technical and commercial features. This project represents an important contribution to the use of machine learning techniques for packaging price predictions, as no similar bibliography was found for estimation of packaging costs from the buyer's perspective. However, not only the packaging industry can benefit from this kind of analysis, as the use of machine learning is not widespread in other similar industries as well.

For example, Dewhurst and Boothroyd (1988) developed cost models for products derived from machining and injection molding processes [33]. These models are only based on physics and mechanical engineering principles, and do not use any data analysis to assess whether the outputs are accurate or not. Furthermore, the influence of the different parameters on the final costs is assumed linear, being in this way easier to interpret. A machine learning technique could be applied over a dataset of these products, to develop a more accurate and robust model. In addition, a more realistic assessment over the influence of the technical parameters could take place. The same applies to the cost model developed by Boothroyd and Reynolds (1989) for turned parts [34]. A machine learning approach would be useful to develop a more accurate model, and to develop a better understanding of the parameters' effect on the final cost of the pieces.

Other industry that could take advantage of machine learning techniques to assess the correct price of a product is the steel pipe industry. Shtub and Versano did so by developing a neural network algorithm to predict prices in this industry with objective (technical) information and subjective information (e.g. an estimate of the quantities that will be manufactured each period during the product's life cycle) [35]. They came with a fast, inexpensive, yet accurate and objective way of estimating product costs. However, this model was developed from the manufacturer's perspective; a similar model to the one developed in the present project could be implemented, taking in consideration the buyer's perspective. In this way, in addition to the technical and subjective features, other variables such as *country* and *supplier* could be incorporated to support decisions regarding product sourcing from buying firms. The same applies to the work of Verlinden et al. (2008), who used regression techniques and artificial neural networks to estimate the price that should be charged for sheet metal parts [36].

To sum up, the present project exhibits a systematic machine learning approach to the price prediction problem, from the buyer's perspective, that could be replicated in plenty industrial environments. Being these techniques not common practices within many non-tech industries, they could represent a relevant differential advantage, becoming decisive in the competitiveness and profitability of many firms in the near future.

# 9. Bibliography


[1] The Center for Paper Business and Industry Studies. (n. d.). Retrieved March 18, 2021, from http://www.paperstudies.org/pricetool/references/Factors%20Affecting%20Containerboard%20Pricing.pdf

[2] CFI. (n.d.). *Economies of Scale.* Retrieved March 18, 2021, from https://corporatefinanceinstitute.com/resources/knowledge/economics/economies-of-scale/

[3] Nilsson, N. J. (1998). *Introduction to Machine Learning.* Retrieved March 18, 2021, from https://ai.stanford.edu/~nilsson/MLBOOK.pdf

[4] Gareth, J., Witten, D., Hastie, T., Tibshirani, R. (2013). *An Introduction to Statistical Learning with Applications in R.* New York: Springer. pp. 1.

[5] Bontempi, G., Taieb, S., Le Borgne, Y. (2013). Machine Learning Strategies for Time Series Forecasting. *Lectures Notes in Business Information Processing, 138,* 62-77.

[6] One Tech. (2020). *Using Machine Learning Algorithms to Predict Pricing Trends.* Retrieved March 18, 2021, from https://www.onetech.ai/en/blog/using-machine-learning-algorithms-to-predict-pricing-trends

[7] Tarallo, E., Akabane, G. K., Shimabukuro, C. I., Mello, J., Amancio, D. (2019). Machine Learning in Predicting Demand for Fast-Moving Consumer Goods: An Exploratory Research. *IFAC-PapersOnLine, 52,* 737-742.

[8] Gareth, J., Witten, D., Hastie, T., Tibshirani, R. (2013). *An Introduction to Statistical Learning with Applications in R.* New York: Springer. pp. 16-29.

[9] Hyndman, R., Athanasopoulos, G. (2020). *Forecasting: Principle and Practice.* Retrieved March 18, 2021, from https://otexts.com/fpp3/accuracy.html

[10] Hyndman, R., Koehler, A. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting, 22,* 679–688.

[11] Makridakis, S. (1993). Accuracy measures: theoretical and practical concerns. *International Journal of Forecasting, 9,* 527-529.

[12] Gareth, J., Witten, D., Hastie, T., Tibshirani, R. (2013). *An Introduction to Statistical Learning with Applications in R.* New York: Springer. pp. 34-35.

[13] Gareth, J., Witten, D., Hastie, T., Tibshirani, R. (2013). *An Introduction to Statistical Learning with Applications in R.* New York: Springer. pp. 180-182.

[14] Gareth, J., Witten, D., Hastie, T., Tibshirani, R. (2013). *An Introduction to Statistical Learning with Applications in R.* New York: Springer. pp. 60-73.

[15] Zheng, A., Casari, A. (2018). *Feature Engineering for Machine Learning.* Sebastopol (California): O'Reilly Media. pp. 68-81.

[16] Stock, J., Watson, M. (2014). *Introduction to Econometrics: Third Edition Update*. New York: Pearson. pp. 269.

[17] Gareth, J., Witten, D., Hastie, T., Tibshirani, R. (2013). *An Introduction to Statistical Learning with Applications in R*. New York: Springer. pp. 130.

[18] R documentation. (n. d.). *rpart.control: Control for Rparts Fits*. Retrieved March 18, 2021, from https://www.rdocumentation.org/packages/rpart/versions/4.1-15/topics/rpart.control

[19] Gareth, J., Witten, D., Hastie, T., Tibshirani, R. (2013). *An Introduction to Statistical Learning with Applications in R*. New York: Springer. pp. 320-327.

[20] Gareth, J., Witten, D., Hastie, T., Tibshirani, R. (2013). *An Introduction to Statistical Learning with Applications in R*. New York: Springer. pp. 191-192.

[21] Hastie, T., Tibshirani, R., Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction (Second Edition)*. New York: Springer. pp. 587-604.

[22] R documentation. (n. d.). *randomForest: Classification and Regression with Random Forest*. Retrieved March 18, 2021, from https://www.rdocumentation.org/packages/randomForest/versions/4.6-14/topics/randomForest

[23] XGBoost developers. (n. d.). *XGBoost Parameters*. Retrieved March 18, 2021, from https://xgboost.readthedocs.io/en/latest/parameter.html

[24] Bergstra, J., Bengio, Y. (2021). Random Search for Hyper-Parameters Optimization. *Journal of Machine Learning Research, 13,* 281-305.

[25] Regattieri, A., Santarelli, G., Accorsi, R., Mora, C., Pareschi, A. (n. d.). *A Mathematical Model for Packaging Cost Evaluation along the Supply Chain*. Retrieved May 29, 2021, from http://www.summerschool-aidi.it/edition-2015/images/paper2012/4.5.pdf

[26] Zang, Y., Fuh, J. (1998). A Neural Network Approach for Early Cost Estimation of Packaging Products. *Computers and Industrial Engineering, 34,* 433-450.

[27] Stats and R. (2020). *Outliers Detection in R*. Retrieved May 29, 2021, from https://statsandr.com/blog/outliers-detection-in-r/#:~:text=The%20Grubbs%20test%20allows%20to,highest%20value%20is%20an%20outlier

[28] CFI. (n. d.). *Just in Time (JIT) Method*. Retrieved March 18, 2021, from https://corporatefinanceinstitute.com/resources/knowledge/accounting/just-in-time-jit-method/

[29] TED Packaging Bags & Pouches. (n. d.). *Stand Up Pouches: The Ultimate Guide*. Retrieved March 18, 2021, from https://www.tedpc.com/stand-up-pouches/

[30] STHDA. (n. d.). *Correlation Test Between Two Variables in R*. Retrieved March 18, 2021, from http://www.sthda.com/english/wiki/correlation-test-between-two-variables-in-r

[31] R documentation. (n. d.). *Importance: Extract Variable Importance Measure.* Retrieved May 29, 2021, from https://www.rdocumentation.org/packages/randomForest/versions/4.6-14/topics/importance

[32] Brownlee, J. (2016). *Feature Importance and Feature Selection with XGBoost in Python.* Retrieved May 29, 2021, from https://machinelearningmastery.com/feature-importance-and-feature-selection-with-xgboost-in-python/

[33] Dewhurst, P., Boothroyd, G. (1988). Early Cost Estimating in Product Design. *Journal of Manufacturing Systems, 7,* 183-191.

[34] Boothroyd, G., Reynolds, C. (1989). Approximate Cost Estimates for Typical Turned Parts. *Journal of Manufacturing Systems, 8,* 185-193.

[35] Shtub, A., Versano, R. (1999). Estimating the Cost of Steel Pipe Bending, a Comparison Between Neural Networks and Regression Analysis. *International Journal of Production Economics, 62,* 201-207.

[36] Verlinden, B., Duflou, J., Collin, P., Cattrysse, D. (2008). Cost Estimation for Sheet Metal Parts Using Multiple Regression and Artificial Neural Networks: A Case Study. *International Journal of Production Economics, 111,* 484-492.