

Parameters driving consumer demand in Brazil

by

Krishna Rajendran

Bachelor of Engineering, Telecommunication, PES Institute of Technology, 2011

SUBMITTED TO THE PROGRAM IN SUPPLY CHAIN MANAGEMENT  
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF ENGINEERING IN LOGISTICS

AT THE

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

JUNE 2016

© 2016 Krishna Rajendran. All rights reserved.

The authors hereby grant to MIT permission to reproduce and to distribute publicly paper and electronic copies of this thesis document in whole or in part in any medium now known or hereafter created.

**Signature redacted**

Signature of Author.....

Master of Engineering in Logistics Program

May 6, 2016

**Signature redacted**

Certified by.....

Dr. Matthias Winkenbach

Director, Megacity Logistics Lab

Research Associate, Center for Transportation and Logistics

Thesis Supervisor

**Signature redacted**

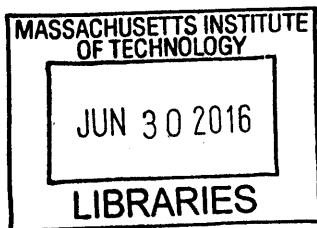
Accepted by.....

Dr. Yossi Sheffi

Director, Center for Transportation and Logistics

Elisha Gray II Professor of Engineering Systems

Professor, Civil and Environmental Engineering



**ARCHIVES**

Parameters driving consumer demand in Brazil

by

Krishna Rajendran

Submitted to the Program in Supply Chain Management  
on May 8, 2016 in Partial Fulfillment of the  
Requirements for the Degree of Master of Engineering in Logistics

ABSTRACT

What are the key store related parameters that drive sales for large retail chains? This question has become increasingly important to Lojas Americanas, the sponsor company. In the last few years, the company has expanded rapidly to cater to a larger group of consumers in a wide range of locations across Brazil. With this expansion, it wishes to determine the key parameters that drive sales for each department and modify its assortment policy accordingly for each store, so as to optimize total sales. This thesis investigates the sales impact of a wide range of store related parameters such as location, size, and socio-economic profile of the surrounding population. Stepwise regression analysis is used here. For this regression, AIC and the p-value threshold are used as the criteria to identify statistically significant store related parameters that influence sales. Furthermore, cross validation is performed to check the explanatory power of the model. The analysis performed yields useful results. A total of 36 different retail departments are analyzed and an adjusted R-squared value (for the validation set) of over 0.6 is obtained for a vast majority of them, indicating that the model performs well in determining the key parameters that drive sales. Furthermore, for each department, the statistically significant set of parameters is obtained and for the company's overall revenue a set of 11 key parameters is identified as highlighted in the Discussion section of the thesis. LA can use the results of this analysis to guide its product assortment policy.

Thesis Supervisor: Dr. Matthias Winkenbach

Title: Director, Megacity Logistics Lab

## **Acknowledgements**

I would like to thank Dr. Matthias Winkenback for his constant guidance and support. I would also like to make a special mention about my main point of contact at the thesis company - Dr. Thomas Meinel, for his timely responses, encouragement, and for the courtesy extended during the site visit to Brazil. Furthermore, I wish to express my gratitude to Dr. Bruce Arntzen (Executive Director, MIT's Supply Chain Management program) for giving me the opportunity to work on this thesis project. Lastly, I wish to thank my classmates, family, and friends for all their support.

## **Table of Contents**

<b>List of Tables .....</b>	<b>5</b>
<b>List of Figures.....</b>	<b>6</b>
<b>1. Introduction.....</b>	<b>7</b>
<b>2. Literature review .....</b>	<b>11</b>
<b>2.1 Selecting store parameters.....</b>	<b>11</b>
<b>2.2.1 External store parameters .....</b>	<b>11</b>
<b>2.2.2 Internal Store Parameters.....</b>	<b>12</b>
<b>2.2 Selecting a statistical analysis technique .....</b>	<b>14</b>
<b>3. Methodology .....</b>	<b>18</b>
<b>3.1 Project scope definition .....</b>	<b>18</b>
<b>3.2 Parameters to analyze .....</b>	<b>19</b>
<b>3.3 Data cleaning.....</b>	<b>25</b>
<b>3.4 Regression model .....</b>	<b>27</b>
<b>3.4.1 Removing highly correlated parameters .....</b>	<b>27</b>
<b>3.4.2 Cross-validation .....</b>	<b>28</b>
<b>3.4.3 Stepwise backward elimination regression .....</b>	<b>30</b>
<b>3.4.4 Running the regression model.....</b>	<b>31</b>
<b>4. Results .....</b>	<b>33</b>
<b>5. Discussion.....</b>	<b>48</b>
<b>5.1 Overall results analysis.....</b>	<b>48</b>
<b>5.2 Most significant parameters that affect sales.....</b>	<b>49</b>
<b>6. Conclusion .....</b>	<b>54</b>
<b>6.1 Limitations of the model .....</b>	<b>54</b>
<b>6.2 Fruitful areas of future research .....</b>	<b>55</b>
<b>References .....</b>	<b>57</b>
<b>Appendix A.....</b>	<b>59</b>
<b>Appendix B.....</b>	<b>94</b>

## **List of Tables**

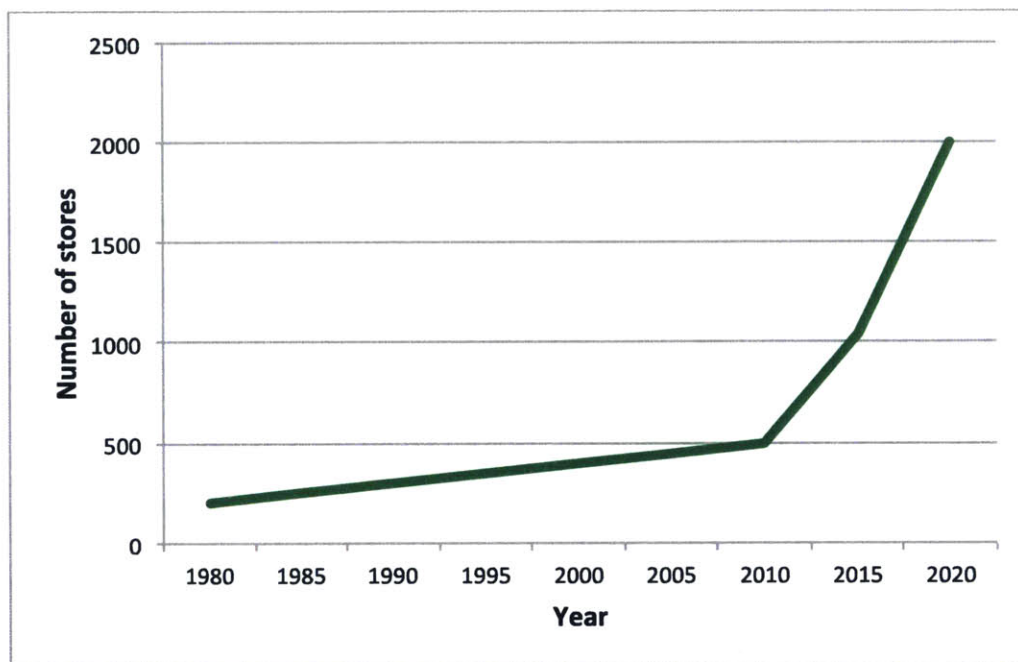
Table 1: List of internal store parameters.....	20
Table 2: List of establishments present near the store.....	21
Table 3: List of all city related parameters.....	23
Table 4: List of all socio-economic related parameters .....	25
Table 5: Percentage of stores in a given region .....	26
Table 6: Regression output table for Department 1 .....	35
Table 7: Summary of fit table .....	36
Table 8: Crossvalidation .....	37
Table 9: R-squared validation set frequency distribution.....	48
Table 10: Key statistically significant parameters across multiple departments .....	51
Table 11: Regression output for Department 3 (Children’s clothes) .....	59
Table 12: Regression output for Department 4 (Lingerie).....	60
Table 13: Regression output for Department 5 (Socks & Scarves) .....	61
Table 14: Regression output for Department 7 (Footwear) .....	62
Table 15: Regression output for Department 9 (Briefs and Underwear).....	63
Table 16: Regression output for Department 11 (Baby products).....	64
Table 17: Regression output for Department 12 (Women’s clothing) .....	65
Table 18: Regression output for Department 13 (Bed and Bath) .....	66
Table 19: Regression output for Department 15 (Clothes for men) .....	67
Table 20: Regression output for Department 16 (Stationary items) .....	68
Table 21: Regression output for Department 17 (Books and Magazines) .....	69
Table 22: Regression output for Department 19 (Cleaning products).....	70
Table 23: Regression output for Department 21 (Kitchen utensils) .....	71
Table 24: Regression output for Department 22 (Table and Bar).....	72
Table 25: Regression output for Department 23 (Toys) .....	73
Table 26: Regression output for Department 24 (Music) .....	74
Table 27: Regression output for Department 25 (Audio & Video Equipment) .....	75
Table 28: Regression output for Department 27 (Home Appliances).....	76
Table 29: Regression output for Department 28 (Perfumes and Hygiene) .....	77
Table 30: Regression output for Department 30 (Electronics/ IT Equipment) .....	78
Table 31: Regression output for Department 31 (Hair Coloring products).....	79
Table 32: Regression output for Department 32 (Cosmetics) .....	80
Table 33: Regression output for Department 33 (Sports and Beach Equipment) .....	81
Table 34: Regression output for Department 34 (Imported Toys) .....	82
Table 35: Regression output for Department 40 (Candies) .....	83
Table 36: Regression output for Department 43 (Easter).....	84
Table 37: Regression output for Department 44 (Plastic Utilities) .....	85
Table 38: Regression output for Department 45 (Videogames) .....	86
Table 39: Regression output for Department 46 (Chips and Cereal Bars) .....	87
Table 40: Regression output for Department 48 (Beverages) .....	88
Table 41: Regression output for Department 49 (Biscuits and Cookies) .....	89
Table 42: Regression output for Department 51 (Diapers) .....	90
Table 43: Regression output for Department 56 (Christmas Event).....	91
Table 44: Regression output for Department 63 (Telephones) .....	92
Table 45: Regression output for Department 72 (Movies) .....	93

## **List of Figures**

Figure 1: Total number of stores by year .....	7
Figure 2: Distribution of LA's stores across Brazil .....	8
Figure 3: Product Hierarchy .....	18
Figure 4: Breakdown of parameters by category .....	20
Figure 5: Distribution of stores by region .....	27
Figure 6: Splitting the entire dataset into a training and validation set .....	29
Figure 7: High-level overview of regression model .....	32
Figure 8: R-squared validation set frequency distribution .....	48
Figure 9: R-squared adjusted validation set cumulative frequency distribution .....	49

## 1. Introduction

Lojas Americanas (LA) is one of the largest retail chains in Brazil, founded in 1929. LA sells products a wide range of departments, from home appliances and clothes to perfumes and chocolates. It has grown rapidly in the last five years from 500 to more than 1000 stores. It wishes to continue this rapid rate of expansion and double the number of stores by 2020. In Figure 1 below, the year-over-year growth in the number of stores is shown. Currently, LA has 36 departments and over 1,000 product lines. LA's rapid expansion has brought with it many challenges, especially with respect to Operations and Supply Chain Management.



**Figure 1: Total number of stores by year**

Up until 2010, most of LA's stores were located in large cities. However, in the last 5-6 years, it has started to open up more number of smaller stores in smaller cities. With its' rapid growth, LA now operates stores all over Brazil and caters to a large diverse group of consumers. In Figure 2, the distribution of LA's stores in Brazil is shown.



**Figure 2: Distribution of LA's stores across Brazil<sup>1</sup>**

LA's executives believe that customers in various locations differ in the kinds of products they wish to purchase. Until 2010, LA had an assortment policy of stocking more or less the same products on the shelves of all of its stores. Since most stores were similar in size, this policy could be followed. But since LA started opening more number of smaller stores, this policy had to change. From an operational standpoint, the same set of products cannot be stocked at all stores since the new stores were significantly smaller than the older stores. For example, LA's stores vary in size from 100 m<sup>2</sup> to 5,000 m<sup>2</sup>, so the same volume of products cannot be stocked at all stores. Thus, due to shelf-space constraints, LA needs to have different product assortments at different stores. Furthermore, since LA was opening up new stores in smaller cities and towns, it had to cater to a more diverse group of customers as well. So LA wanted to devise a new assortment policy that could optimize

---

<sup>1</sup> Source: Google Maps.



sales by selecting the right kind of products to stock at stores with different sizes and locations.

LA wants to understand how to best serve the needs of its now diverse customer base by identifying the key store-related parameters that drive sales for each of its 36 departments. With this information, LA can understand its customers better and subsequently adopt an improved assortment policy to optimize sales by stocking the right set of products at each of the different stores. Up until now, LA has not adopted a rigorous quantitative approach to analyzing data to guide its product assortment decisions. This thesis attempts to fill this gap.

This thesis identifies the statistically significant store parameters that drive sales for each of the company's 36 departments and subsequently quantifies the impact of these parameters on the department's sales. The next step is to combine this quantitative analysis with practical on-the-ground insights from LA's executives to identify the key takeaways for each department. This thesis develops a statistical analysis model to provide an estimate of both the impact of each of the statistically significant parameters on sales and the overall impact of all of the statistically significant parameters combined on sales.

LA's executives requested that this thesis focus on data analysis at the department-level and not at the product line level or SKU level. They requested a model that could be seamlessly extended in the future for do product-line level and SKU level analysis. The SKU level analysis could help drive product assortment decisions i.e. deciding what products to stock at what types of stores. Furthermore, LA could do more research in the future to study the statistically significant parameters in greater detail. Future areas of research are discussed in the conclusion section of this thesis.

The literature review section presents a review of the most relevant research on the parameters that drive retail sales and of the statistical analysis tools used to study correlations between these parameters and sales. Next, in the methodology section the steps taken to build the regression model is discussed. Following this, in Sections 4 and 5, results and relevant points of discussion are presented. Lastly, in the conclusion section , areas of future research and the limitations of the model are discussed.

## **2. Literature review**

The literature review focuses on determining two main aspects of the subsequent analysis:

- A list of store parameters that would be most useful to analyze.
- An analysis tool to identify the statistically significant parameters and quantify the impact of these parameters on sales.

### **2.1 Selecting store parameters**

The first step in the literature analysis is to identify both the external and internal store parameters that greatly impact sales for retail stores.

#### **2.2.1 External store parameters**

Jayasankara Prasad and Ramachandra Aryasri (2011) find that the socio-economic traits of consumers such as age, profession, and salary level determine the type of stores that they prefer to visit. The researchers interviewed customers at various different stores and then used inferential analysis techniques to arrive at their findings. Kumar and Karande (2000) study the impact of the demographic profile of the population around the store on sales, using a linear regression model. They find that sales are greater when any of these parameters are higher – salary level, number of individuals in a house, and proportion of people who possess homes.

Another location related parameter that could affect sales is the presence of establishments around a store. Yoo and Chang (2005) find that good access to parking lots and to nearby public transport has a strong influence (p-value less than 0.05) on store loyalty.

Furthermore, Artz and Stone (2006) find that large supercenter stores negatively affect sales of nearby retail stores. Lastly, Konishi (2005) analyzes whether a store will perform better if it is far away from competing stores or if it is closer. He investigates the tradeoff between the positive impact of increased foot traffic as a result of stores clustering together and the negative impact of price wars that stores located close by may engage in. He finds that sales of stand alone stores are likely to be negatively impacted by the presence of a nearby shopping mall with similar stores.

Thus, two broad factors - socio-economic characteristics of consumers and the presence of establishments around the store - appear to have a strong correlation to retail store sales. Based on the data provided by LA and a subsequent analysis, identifying the specific aspects of these two parameters that have the most statistically significant impact will be the key for this thesis.

### **2.2.2 Internal Store Parameters**

There are three main internal store parameters that affect sales – stockouts, promotions, and assortment size.

Stockouts can play a key role in impacting retail sales. Anderson, Fitzsimons and Simester (2006) investigate the consequences of stockouts by analyzing customers' past purchases and observing how many units of products are backordered after customers were told that the product is unavailable. The researchers find that a stockout for one product sometimes negatively impacts sales for other products as well, as some products may be closely linked to each other. Furthermore, the researchers conclude that customers may start looking for other stores that have all products they want.

Promotions appear to be a key factor that impacts retail sales and brand image. Retail companies have promotions for various reasons including special holidays or to dispose unsold inventory. Palazón-Vidal and Delgado-Ballester (2005) investigate how promotions can influence a consumer's impression of a brand. The researchers study the impact of price-related and non-price related promotions on commonly used day-to-day products such as laundry detergents and on luxury products such as perfumes. They find that overall non-price related promotions have a greater impact in growing brand loyalty. Furthermore, they find that non-price related promotions have a similar impact on both day-to-day products and luxury products whereas price-related promotions have a greater impact on day-to-day products than luxury products. Promotions at LA are applicable for some (and not all) SKUs and product lines, inside specific departments. This thesis focuses on analysis at the department level and not at the product line level or SKU level. While promotions do have an effect on the department level data, attempting to factor it into the model without considering product line and SKU level data could lead to potentially unreliable results. Thus, promotions are not considered and this is a limitation of the model developed in this thesis.

Product assortment is another key factor in determining retail sales. However, whether or not a larger assortment leads to higher sales appears to be a topic of debate among researchers. Oppewal and Koelemeijer (2005) find that consumers have a predisposition towards a large assortment irrespective of how differentiated the products (in the assortment) are with respect to their features. However, Boatwright and Nunes (2001) find that in one experiment a reduction in the assortment size actually caused sales to increase.

Thus, even though the direction of the effect of assortment size is ambiguous, assortment size is an important factor that needs to be investigated.

## **2.2 Selecting a statistical analysis technique**

The next step of the literature review is to find an analysis tool to identify the statistically significant factors from the initial list of parameters that are considered for the analysis.

Hastie, Tibshirani, Friedman, and Franklin (2005) discuss a commonly used statistical analysis technique – Regression. Regression is used to analyze the impact that a set of predictor/independent variables (store parameters in this thesis) have on the dependent variable/variable of interest (sales in this thesis). Here, for each of the 36 departments, a separate regression model is developed.

In this thesis, a large number of predictor variables (105) is used for the analysis. One type of regression that could be used for this thesis is stepwise regression. According to Huang and Townshend (2003), stepwise regression is a good tool to use when there are a large number of independent variables because it helps identify and select a smaller subset of statistically significant independent variables, while still maintaining a high R-squared value. In this thesis, for each of the 36 departments, 105 predictor variables from 682 different stores are factored into the analysis. For such a large volume of data, stepwise regression is a good fit.

Dattalo (2013) discusses two ways of doing stepwise regression – the forward selection method and the backward elimination method. In the forward selection method, initially the model does not contain any variable. The variables that most improve the explanatory

power of the regression model are added one by one till the optimum value of the explanatory power of the regression is reached. In the backward elimination method, initially all independent variables are put into the model. Then, one by one, variables are eliminated, based on which elimination most improves the explanatory power of the regression model. Again, like in the case of forward selection, this iterative process runs till the optimum value of the explanatory power of the regression model is reached.

Stepwise regression appears to be a popularly used statistical analysis technique for research efforts that focus on consumer purchasing behavior, similar to the area of focus for this thesis. Carpenter and Moore (2006) use stepwise regression to analyze how retail store parameters impact customer preferences for various types of stores. The authors use a minimum p-value threshold of 0.05 as the criterion for adding parameters into the regression model. Thus, for stepwise regression, a selection criterion can be used to control the variables to be added to the model. Verbeke (2006) also adopts stepwise regression to analyze how consumers' socio-economic characteristics influence their preferences for certain food items.

Stepwise regression does have a drawback. Montgomery, Peck, and Vining (2012) discuss one drawback of regression that can produce unreliable results – multicollinearity. Multicollinearity occurs when two or more independent predictor variables have strong linear relationships with each other. This issue could be relevant for this thesis since there are some independent variables that may potentially have strong linear associations with each other. So for this thesis, multicollinearity must be effectively dealt with.

Dormann et al. (2013) discuss multiple ways to minimize the effect of multicollinearity. One such technique is to use Principal Component Analysis (PCA) clustering. However, the authors in their experiments find that this technique did not perform well even for low levels of multicollinearity. The authors state that the technique of eliminating variables with a correlation coefficient magnitude greater than 0.7 works well in minimizing multicollinearity. Götz, Liehr-Gobbers, and Krafft (2010) state that another technique to minimize multicollinearity is to use the Variance Inflation Factor (VIF). VIF is a numerical scale that provides an indication of the extent to which multicollinearity has affected each predictor variable used in the regression model. Hair, Black, Babin, Anderson, and Tatham (2006) recommend using a VIF threshold of 10 for the independent variables, above which multicollinearity is high. Based on this threshold, certain variables can be eliminated to reduce multicollinearity.

The final step of regression analysis is to assess how the model performs on new data (i.e. cross validation). Refaeilzadeh, Tang, and Liu (2009) discuss two possible ways to do cross validation – hold-out method and k-fold cross validation. In the hold-out method, one set of datapoints are used to build the model and another previously unseen set of datapoints are used to assess the predictive power of the model. In the k-fold cross validation method, the entire dataset is divided into k subsets. For each run, one new subset is used for validation and the remaining subsets are used to construct the model. This process is run k times. Koufakou, Georgiopoulos, Anagnostopoulos, and Kasparis (2001) state that when extensive data is to be analyzed, (as is the case in their experiment) the hold-out method is preferred since it is computationally less expensive. Hawkins, Basak, and Mills (2003) state that one drawback of the hold-out method is that the model developed may not have strong



explanatory power since some datapoints are reserved for validation and there are fewer data points available for building the model. This is an issue when a lot of datapoints are not available. However, for this thesis, a large number of data points is available, so the hold-out method can be used.

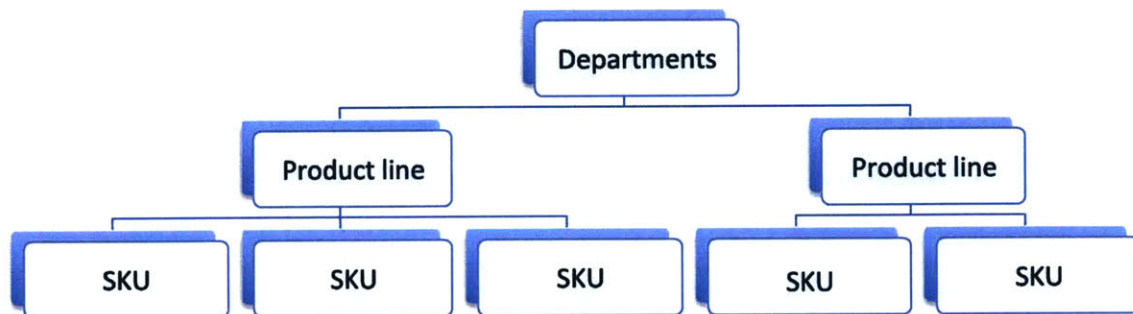
Most of the current literature on the analysis of parameters that drive retail sales focus on analyzing one or two key statistically significant parameters in detail. However, this thesis identifies the subset of statistically significant parameters from an initial larger list of potentially significant factors. The step by step procedure used to develop the model used for this analysis will be discussed in more detail in Section 3.

### **3. Methodology**

The methodology section is divided into four sub-sections. Section 3.1 discusses the process of breaking down LA’s requirements into smaller modules and subsequently defining the scope for this thesis. Section 3.2 presents the process of selecting the store parameters that would be most useful to analyze. Section 3.3 discusses the process of cleaning the data received from LA to ensure that that the data used to build the model is reliable. Finally, Section 3.4 discusses the step-by-step process of building the stepwise regression model.

#### **3.1 Project scope definition**

The first step of the methodology involves defining the scope of the project. LA has 1,041 stores and 36 departments. Each of these departments has many product lines (ranging from 10 to 50) and each product line in turn has individual SKUs. This hierarchy is pictorial represented in Figure 3 below.



**Figure 3: Product Hierarchy**

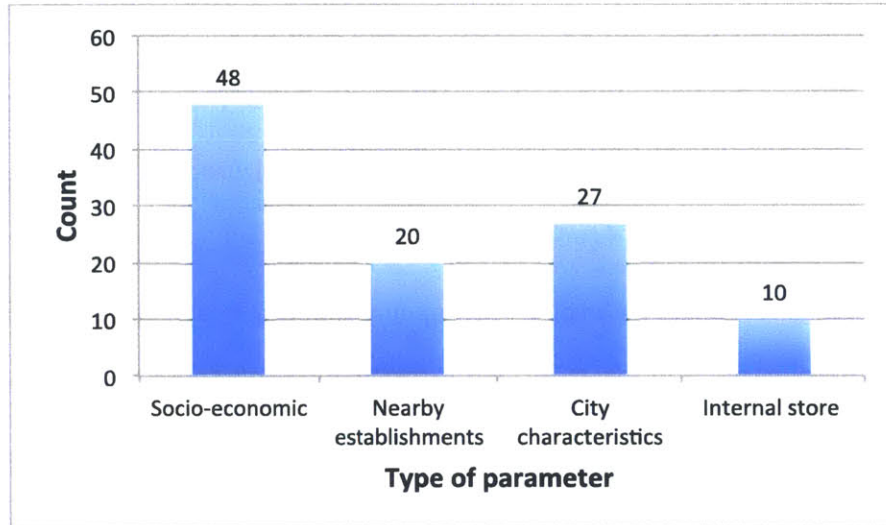
LA’s long-term goal is to improve its product assortment policy to optimize sales. This is a broad requirement that needs to be broken down into smaller and more specific targets. The focus of this thesis is to analyze sales data at the department level. This analysis will

yield useful insights about the key parameters that are strongly correlated to sales and also help identify some of the key parameters that drive sales across many departments. This is the first major step towards helping LA develop an optimal product assortment policy. Product assortment decisions, however, cannot be made just by analyzing products at the department level or product line level. LA has a policy that requires all departments and all product lines to be present at all stores. Only at the SKU level can products be selectively removed. Thus, to help LA develop an optimal assortment policy, it is necessary to analyze and identify the parameters that drive sales at all three levels – department, product line, and SKU. Based on this analysis, product assortment recommendations can be made for each store based on its parameters, to optimize sales. This thesis focuses on analyzing data only at the department level due to time constraints.

In the future, a similar analysis could be performed for each product line and for the SKUs in each product line. Based on such an analysis, LA could make better assortment decisions and do more advanced analysis such as shelf-space optimization.

### **3.2 Parameters to analyze**

LA provided data for 105 parameters for each of its 36 departments across all of the company's 1,041 stores. Thus, in total, there are about 3.9 million data points. There are four broad categories of store parameters – socio-economic characteristics of the population surrounding the store, presence of establishments nearby, characteristics of the city the store is located in, and internal store features. Figure 4 below shows the breakdown of the parameters by category.



**Figure 4: Breakdown of parameters by category**

Detailed information about each of the categories is presented in Tables 1 through 4 below.

1) Internal Store parameters

Internal store parameters	Explanation (if required)
Sales area of the store	-
Store Status (Premium = 1)	Set as a nominal regression variable. Premium stores are high-end stores.
Stock area (in meter square)	Total backroom area available to store inventory
Assortment Size of Department	Size of the assortment for a given department
Average winter temperature (C°)	-
Average summer temperature (C°)	-
SKUs	Total number of SKUs present in the store
Days since last refurbishment	-
Store Type (Conventional =1)	Set as a nominal regression variable. The company internally classifies stores as conventional and non-conventional.
Days since Inauguration of store	Number of days since the store was opened. The end date is assumed to be 1 <sup>st</sup> January, 2016.

**Table 1: List of internal store parameters**

2) Parameters related to number of each type of establishment present near the store

For all the establishments considered in the analysis, Portuguese names are provided below in addition to English names to make it easier for LA's executives to understand the content.

Portuguese name	English name
Faculdades	Colleges
Agências Bancárias	Banks
Escola Privada	Private schools
Farmacias	Pharmacies
Perfumaria e Cosméticos	Perfumes/Cosmetic Shops
Super e Hiper Mercados	Supermarkets
Gyms	Fitness centers/gyms
Bar - Diversos	Bars
Cursos	Universities
Escola Pública	Public schools
Hospitais	Hospitals
Hotel	Hotels
Postos Combustível	Fuel stations
Restaurantes	Restaurants
Shopping Center	Shopping centers
Geographic Region - Sul (South)	Nominal regression variables. Refers to the region where the store is located. Only one of these columns will have a value of 1. All others columns will be set to zero.
Geographic Region - Centro-Oeste (Central region)	
Geographic Region - Sudeste (South East)	
Geographic Region - Nordeste (North East)	
Location - SHC/Rua (Rua = 0)	Nominal regression variables. Refers to the location of the store. SHC (stands for shopping mall), RUA (stands for street)

**Table 2: List of establishments present near the store**

Data for the number of establishments within both a 0.5 km and 1 km radius was available. LA's executives mentioned that recently the company opened many new stores in smaller towns, where there aren't too many establishments present close to the stores. Thus, based on expert opinions from LA's executives, a larger (1 km) radius was selected as the standard for the model in this thesis.

In Table 2 above, four regions are listed: Central, North East, South East, and South. Each of the four regions listed in the table represent one parameter of the regression model. These

parameters can have a value of 0 or 1, depending on the location of the store. For instance, if a store is in the Central region, then the value of the central parameter for this store is 1 and the values for the other region parameters are set to 0.

Even though five regions parameters are present, for the regression model only four parameters are used. The Region parameter North is not included in the model. If a store is present in the North region, it will have 4 zeros in all 4 region parameter values. We could go a step further, by adding a fifth region parameter 'North' and setting this value to 1 for a store in the North region. By doing so, now the 4 existing region parameters taken together will be perfectly negatively correlated with the North column. This would lead to multicollinearity. To avoid this issue, we use only 4 region columns.

### 3) Parameters related to city data

These parameters refer to the characteristics of the city the store is located in. Portuguese names and English translations are provided to make it easier for LA's executives to understand the content.

<b>Parameter (in Portuguese)</b>	<b>English Translation and brief explanation</b>
<b>Apartamento</b>	Number of apartments
<b>Homem</b>	Total number of men
<b>Mulher</b>	Total number of women
<b>PIB per Capita Total (2012)</b>	Per capita GDP for the year 2012
<b>Casa</b>	Number of homes
<b>Casa em vila ou condomínio</b>	Number of residential complexes/condos
<b>Densidade demográfica</b>	Population density
<b>Renda Média 2014</b>	Average annual income of the people in 2014
<b>Área Km<sup>2</sup></b>	Total area of the city in square-kilometers
<b>População (2014)</b>	Total population in 2014
<b>Domicílios (2014)</b>	Total number of households in 2014
<b>Salário médio mensal (2014)</b>	Average monthly income of the people in 2014
<b>Renda Nominal (2014)</b>	Sum of the monthly income of all the households in 2014

<b>PEA Density (2007)</b>	Density of the economically active people in 2007
<b>Pop. Turística (2013)</b>	Total number of tourists that visited the city in 2013
<b>Segmento Turístico (A01 -6, A02 -5, B01-4, B02-3, B03-2, C-1)</b>	The English translation is tourist segment. Set as an ordinal regression variable. Company segregated the tourist attractiveness of a city on a 6 point scale. Level A01 is the highest and level C is the lowest.
<b>Universitário Segmento Level (A-5, B-4, C-3, D-2, E-1)</b>	The English translation is university segment. Set as an ordinal regression variable. Company segregated the the level of education of a city on a 6 point scale. A1 is the highest level and E is the lowest level.
<b>TGCA População (%) (2010-2014)</b>	Geometric average of the annual population growth for the city
<b>TGCA Renda Média (%) (2010-2014)</b>	Geometric average of the annual income growth for the city
<b>Superior completo</b>	Total number of people who have graduated from universities
<b>Médio completo e superior incompleto</b>	Total number of people who completed high school but have not graduated from university
<b>Fundamental completo e médio incompleto</b>	Total number of people who have completed elementary school but have not completed high school
<b>Sem instrução e fundamental incompleto</b>	Total number of people who are not educated and who have not completed elementary school
<b>IDH - Dimensão Educação (2010)</b>	Education Index for the city (2010)
<b>IDH - Dimensão Longevidade (2010)</b>	Longevity Index for the city (2010)
<b>IDH - Dimensão Renda (2010)</b>	Income Index for the city (2010)
<b>IDH - Índice de Desenv. Humano(2010)</b>	Human Development Index for the city (2010)

**Table 3: List of all city related parameters**

4) Socio-economic parameters:

These parameters refer to the characteristics of the population present in a fixed radius around the store. On the advice of LA's executives, for stores located in shopping malls a 9 km radius was used and for stores located on the street a 3km radius was used.

<b>Age group</b>	Number of people in each age group is taken as an input for the model
0 to 9 years old	
10 to 14 years old	
15 to 19 years old	
20 to 24 years old	
25 to 34 years old	
35 to 49 years old	

50 to 59 years	
60+ years old	

<b>Monthly Income (in Brazilian Reais) received by workers (Portugese - Trabalhadores)</b>	Number of workers in each income bracket is taken as an input for the model. The workers here refer to people who are working in the area around the store. They may not necessarily be people who live here. The income ranges were chosen by LA.
Less than 415	
Between 415 and 830	
Between 830 and 1,245	
Between 1,245 and 2,075	
Between 2,075 and 4,150	
Between 4,150 and 6,225	
Between 6,225 and 8,300	
Above 8,300	

<b>Monthly Income (in Brazilian Reais) received by Inactive residents (Portugese – Pop.Res. Inactive)</b>	Number of inactive residents in each income bracket is taken as an input for the model. Inactive residents are mostly people who are retired. The income ranges were chosen by LA.
Less than 415	
Between 415 and 830	
Between 830 and 1,245	
Between 1,245 and 2,075	
Between 2,075 and 4,150	
Between 4,150 and 6,225	
Between 6,225 and 8,300	
Above 8,300	

<b>Monthly Income (in Brazilian Reais) received by Economically Active People (Portugese – PEA)</b>	Number of economically active people in each income bracket is taken as an input for the model. Economically active people include people who are salaried and not salaried. The income ranges were chosen by LA.
Less than 415	
Between 415 and 830	
Between 830 and 1,245	
Between 1,245 and 2,075	
Between 2,075 and 4,150	
Between 4,150 and 6,225	
Between 6,225 and 8,300	
Above 8,300	



Income Level for Private Permanent Homes (Portugese - Domicílio Particulares Permanentes por Faixa de Renda )	
Level A1	Number of private permanent homes in each level of the scale was taken as an in input for the model. An ordinal scale was used here. A1 is the highest income level and E is the lowest income level. The scale was devised by LA.
Level A2	
Level B1	
Level B2	
Level C1	
Level C2	
Level D	
Level E	

Income level for entire population (Portugese - População por Faixa de Renda)	This column refers to the monthly income range of the population for each of the levels in the scale. Number of people in each scale was taken as an input for the model. LA wanted the ordinal scale to be used to better understand if the scale needed to be recalibrated.
A1 level	Above 19,016
A2 level	Between 10,953 and 19,016
B1 level	Between 6,490 and 10,953
B2 level	Between 3,328 and 6,490
C1 level	Between 1,896 and 3,328
C2 level	Between 1,270 and 1,896
D level	Between 724 and 1,270
E level	Below 724

**Table 4: List of all socio-economic related parameters**

### 3.3 Data cleaning

LA provided data for 1,041 stores. Some stores are not factored into the model:

- Annual sales data for 2015 is the dependent variable, so stores that opened in 2015 are not considered since sales data for the whole year is not available.
- Stores present at the company's distribution centers are not considered since these stores are not representative of the stores open to the public.
- Some stores, in the data received, had negative or null values for the number of

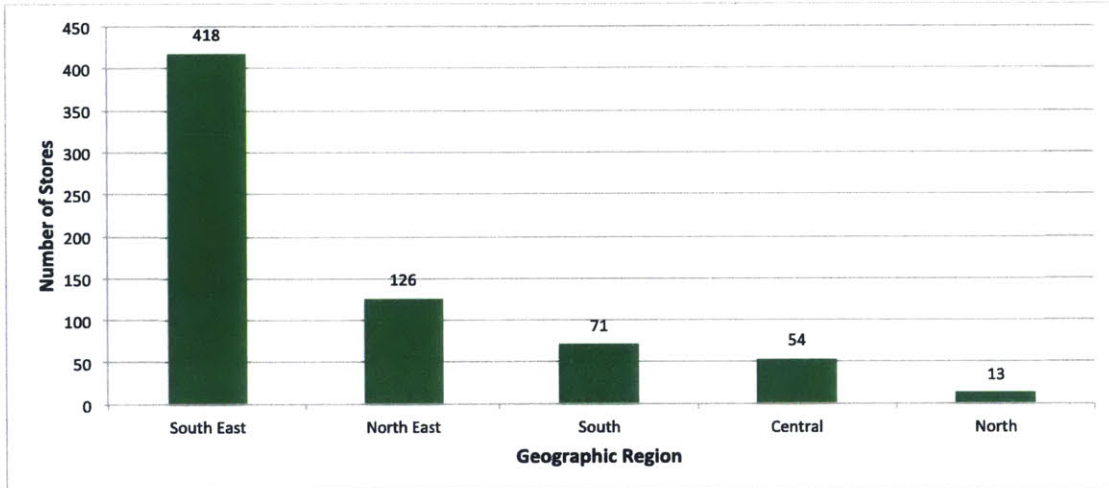
people (in the area around the store) such as the number of workers, inactive residents, and economically active people. LA’s executives mentioned that these null values are incorrect since there are actually a significant number of people present. They mentioned that the tool they used does not have the latest data for smaller cities. Thus, these stores are not considered for the model to avoid unreliable results.

In the end, 682 stores are considered for the final analysis. Table 5 indicates the percentage of the stores in a given geographic region both in the original sample of 1,041 stores and the revised sample of 682 stores. The number of stores removed from each region is more or less proportional to the total number of stores in these regions. The stores considered for the model, thus, have a distribution that resembles the original sample of 1,041 stores.

<b>Percentage of stores in a given region</b>		
	<b>Original Sample</b>	<b>Sample used for model</b>
<b>Central</b>	8	8
<b>North East</b>	21	18
<b>North</b>	3	2
<b>South East</b>	57	61
<b>South</b>	11	10

**Table 5: Percentage of stores in a given region**

The final distribution of stores across all the regions is as shown in below in Figure 5.



**Figure 5: Distribution of stores by region**

### **3.4 Regression model**

In regression analysis, there is a set of input predictor (independent) parameters and an output dependent variable. For the model in this thesis, the 105 external and internal store parameters are the predictor parameters and the 2015 annual department sales is the output dependent variable. For each of the 36 departments, the steps in the following sections were followed to build the model.

#### **3.4.1 Removing highly correlated parameters**

As discussed in the literature review section, multicollinearity occurs when independent predictor variables have strong linear relationships with each other. Strong linear relationships indicate that the variables have a high degree of correlation. Correlation coefficient values close to -1 and +1 indicate that two variables are highly negatively or positively correlated respectively. A correlation coefficient close to 0 indicates that two

variables have very little correlation with each other. The first stage of the regression analysis is to remove highly correlated predictor variables. The steps in this first stage are as follows:

- A correlation matrix is constructed in which the correlation coefficients between each pair of predictor variables is analyzed. In total there are 105 predictor variables, so there are 10,920 relationships that are analyzed here for each of the 36 departments.
- As discussed in the literature review section, a correlation coefficient of 0.7 is used as an upper threshold above which the effects of multicollinearity hamper the overall predicative ability of the model. When a pair of predictor variables has a correlation coefficient greater than 0.7, one of variables is removed.
- The predictor variable that is less correlated to the dependent variable (2015 department sales) is removed.

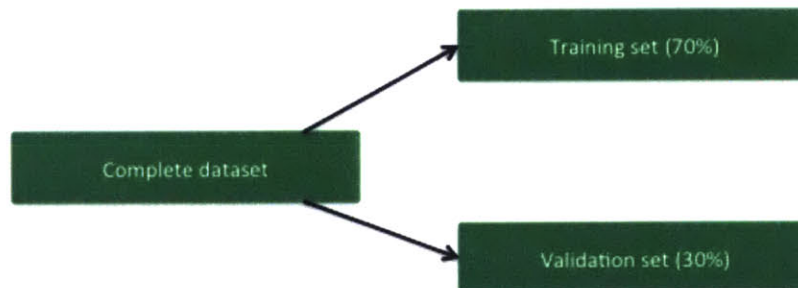
### **3.4.2 Cross-validation**

To perform cross-validation, first, the entire dataset for 682 stores is divided into a training set and a validation set. The training set is used to build the regression model. The model that is developed is then run for the validation set. The R-squared value is an indicator of the explicatory power of the regression model i.e. it helps indicate how well the regression model can explain the variation in the output variable for the given dataset. For cross-validation, the R-squared value of the validation set is the key indicator. Cross-validation helps check the predictive capability of the regression model on new and previously unseen

data. Without cross-validation, the stepwise regression model could potentially over fit the data and produce a model that may not act as a good predicative model for new data.

For cross-validation in this thesis, the hold-out method is used for reasons specified in literature review section. To determine what ratio to split the 682 data points into the training and validation set, a methodology similar to the one followed by Crowther and Cox (2005) is adopted. The researchers use three sets of data – training, test, and validation. They experimented with different training to test ratios and determined the ratio that gave them the best output performance. The model developed in this thesis contains only a training and validation set. Different training to validation set ratios are experimented with to arrive at the ratio that maximizes the value of the predictive ability of the model (i.e. the R-squared of the validation set). For the analysis of each of the 36 departments, four different training to validation set ratios are used – 75:25, 70:30, 65:35, 60:40. The 75:25 ratio is first used and all the steps mentioned in Section 3.4.4 are carried out. Similarly, for each of the other three ratios, the steps mentioned in Section 3.4.4. are carried out. The best-fit model is the model that has the highest R-square adjusted validation set value.

It is observed that for the vast majority of departments, a 70:30 ratio resulted in the highest R-squared value.



**Figure 6: Splitting the entire dataset into a training and validation set**

### 3.4.3 Stepwise backward elimination regression

According to Dattalo (2013), backward elimination is more useful than forward selection when a combination of different predictors are able to explain the variation in the dependent variables, but the predictor variables by themselves are not able to do so. This phenomenon is observed in the vast majority of cases in this thesis when the explanatory power of different combinations of predictor variables is tested. Thus, for the model in this thesis, backward elimination stepwise regression is chosen over forward selection stepwise regression.

For stepwise regression, independent variables are entered (or deleted) in a stepwise sequence based on a specified selection criterion. For the model in this analysis, two selection criteria were used:

- P-value threshold: The p-value threshold determines the number of iterations for the stepwise model. For instance, for backward elimination, one could set the selection rule as 'Minimum p-value threshold = 0.1'. With each iteration, the predictor variable that least improves the fit of the model (i.e. the one with the highest p-value) will be removed. This cycle will keep running till the p-value of the least significant variable goes below 0.1.
- AIC: AIC stands for Akaike information criterion. According to Burnham and Anderson (2002), the AIC criterion helps compare a group of models and identify the relatively most optimal model. The best fit model is the one with the lowest AIC value. With each iteration, the variable that helps to most reduce the AIC value is removed. This process continues till the lowest AIC value is obtained.

#### **3.4.4 Running the regression model**

For each of the 36 departments, the stepwise procedure adopted to run the regression model is as follows:

1. First, the highly correlated predictor variables are removed based on the process explained in section 3.4.1. Then, the entire dataset is divided into a training and validation set in the ratio 75:25.
2. Three separate stepwise backward elimination regression models are constructed with the stopping criteria 'Minimum p-value threshold' of 0.05, 0.1, and 0.15, respectively. For each model, cross-validation is performed and the model with the highest R-squared value of the validation set is chosen.
3. A new stepwise backward elimination regression model is constructed with the selection criteria set to minimize the AIC value. This model is constructed and cross-validation is performed. The R-squared of the validation set is noted.
4. The R-squared value from step 2 and step 3 are compared, and the model with the highest R-squared validation value is chosen.
5. The predictor variable with the highest VIF is eliminated as long as the highest VIF is greater than 10. If the highest VIF is less than 10, go to step 8.
6. The model is run again to check for changes in the R-squared of the validation set.
7. Step 5 and 6 are repeated till the highest VIF value falls below 10.
8. Note down the R-squared adjusted validation set value for this model.
9. Similarly perform steps 1- 8 for the training to validation ratios – 70:30, 65:35, 60:40.

10. The final best fit model is the one that has the highest R-squared adjusted validation set value amongst the four models constructed using different training to validation ratios.

The regression model helped identify the list of statistically significant factors that affected sales for each of LA's 36 departments. In Figure 7 below, the step-by-step process of the regression analysis is shown. In the next section, the results of the regression analysis are presented.

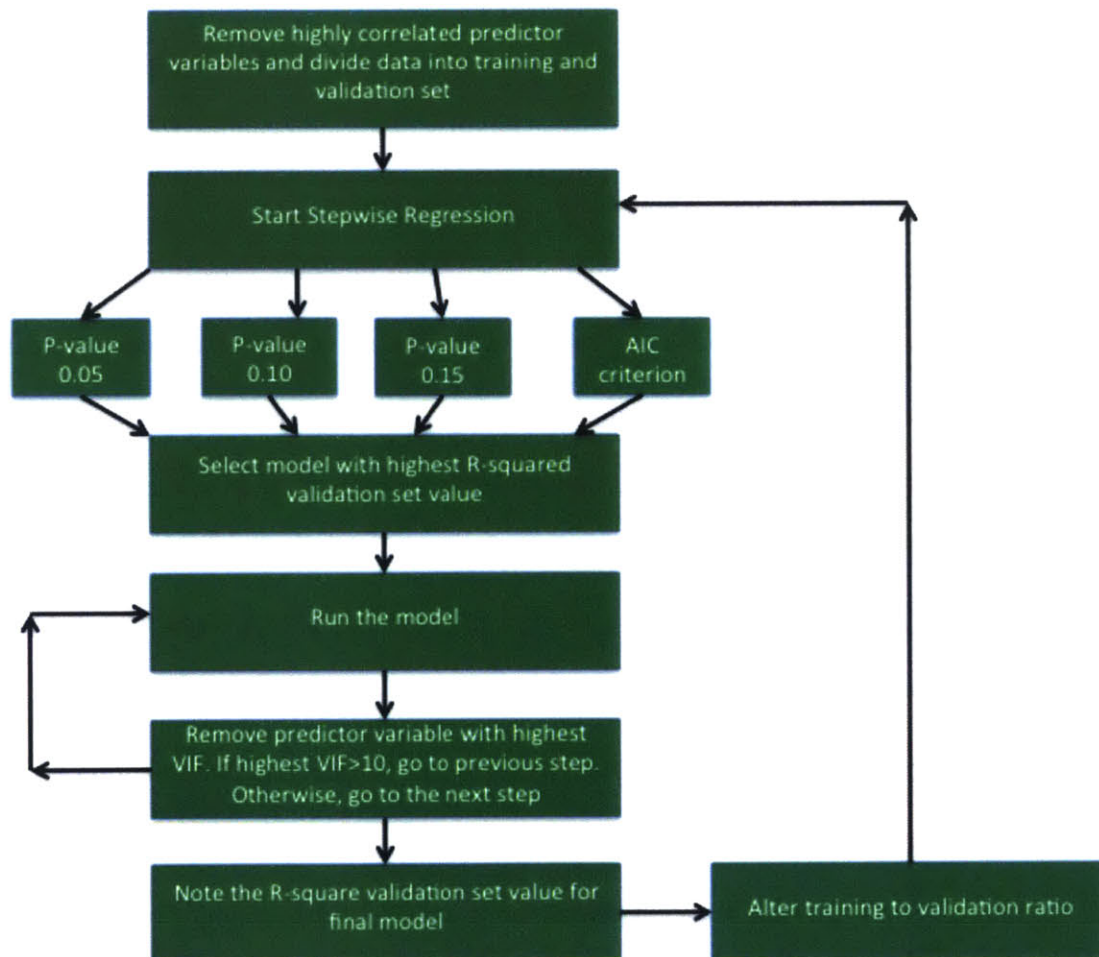


Figure 7: High-level overview of regression model



## 4. Results

This section has two parts:

1. First, the regression output tables for the first department (chocolates) are shown, the terms used in the tables are explained, and the key points of analysis for the chocolate department are presented.
2. Second, the key points of analysis for departments that yielded noteworthy and interesting insights are presented.

The key points of analysis for the rest of the departments are presented in Appendix B.

Furthermore, the regression output tables for all departments are presented in Appendix A.

The equation of the linear regression model is as follows:

$$Y = p1 * a1 + p2 * a2 + p3 * a3 + \dots$$

where

Y is the output dependent variable

p1, p2, p3,... are the independent predictor variables

a1,a2,a3,... indicate the weights for the predictor variables p1,p2, p3,.. respectively.

The weights (also called parameter estimates) indicate the amount the output variable will increase for a unit increase in the corresponding predictor parameter variable.

As illustrated in the methodology section, four separate stepwise regression models are created for each department. Three models use the minimum p-value threshold as the selection criterion (the models have value thresholds of 0.05, 0.1, and 0.15 respectively).

The fourth model uses the AIC criterion. Out of these four models, the best fit is the one that has the highest adjusted R-squared validation set value.

Details about the regression output table:

- Table 6 below presents the list of statistically significant parameters that affect sales for each department.
- For this table and the regression output tables in Appendix A, the cells highlighted in green indicate the parameters that LA's executives found most noteworthy/interesting and these are the parameters that are discussed in the key points of analysis for each department.
- The rows in these tables are arranged in ascending order based on the p-value i.e., the most significant parameter is present in the first row.

**Department 1 - Chocolates**

The parameter estimate value tells us how much the sales for the department will increase for each unit increase in the value of the corresponding predictor parameter.

Parameter	Parameter estimate	P-value
University Segment: Level A	136,946.1623	<. 0001
Hospitals (1km)	-23,494.6512	<. 0001
Public schools (1km)	23,143.1645	<. 0001
Sales Area (in meter square)	879.9767	<. 0001
Days since Inauguration of store	43.7731	<. 0001
Store Type: Conventional	-289,896.9525	<. 0001
Geographic Region - NOT South East	-293,664.4856	<. 0001
PIB per Capita Total (2012)	-3.9253	0.0153
Store Status: NOT Premium	-63,477.4435	0.0115
Geographic Region - NOT Central	-129,384.8987	0.0121
Shopping Center (1km)	53,401.3897	0.0129
Stock area (in meter square)	208.7668	0.0215

Store Location - Street	-61,159.8487	0.0324
Private schools (1km)	4,145.2265	0.0532
Tourist Segment: Level C	-77,980.0254	0.0541
Days since last refurbishment	-13.8445	0.0689
PEA Density (2007)	18.9676	0.0754
Perfume and Cosmetic shops (1km)	-17,589.1867	0.0767
Intercept	573,766.3657	0.0812
Average summer temperature (C°)	-11,664.5267	0.0842
Geographic Region - NOT North East	-54,772.2298	0.0963
University Segment: Level C	-74,425.1432	0.0985

**Table 6: Regression output table for Department 1**

Key points of analysis for Department:

- Stores located in areas surrounded by a higher number of schools have greater sales. The public school parameter has a greater statistical significance and a larger parameter estimate value than the private school parameter. Thus, public schools have a greater positive impact on sales. Furthermore, stores located in cities where the level of education is the highest (University Segment: Level A) tend to have higher sales.
- Stores present in warmer regions tend to have lower sales. This could be because stores in warmer regions have less variety of chocolates. LA's executives mentioned that some suppliers do not deliver to warmer regions as they cannot guarantee the quality of their chocolate since the risk of the chocolate melting is much higher.
- Stores located close to shopping malls/centers have significantly higher sales. According to LA's executives, the company does not have a lot of competitors for chocolate in shopping malls since large supermarkets chains are usually not present in these shopping malls and most of the chocolate stores in shopping malls are high-end stores. So customers who want to buy regular chocolates (which is what LA

predominantly sells) in a shopping mall do not have many options.

Table 7 shows the overall regression analysis results.

### Summary of Fit

RSquare	0.75243
RSquare Adj	0.738583
Root Mean Square Error	439688.2
Mean of Response	949379.9
Observations (or Sum Wgts)	473

Table 7: Summary of fit table

#### Interpretation of Table 7:

- R-squared & R-squared adjusted: R-squared indicates to what degree the model can explain the variation in sales. For instance, an R-squared of 0.7524 indicates that about 75% of the variation in sales can be explained by the model. One drawback of using the R-squared (unadjusted) is that it will always improve or stay the same even if predictor variables that have no influence whatsoever on sales are added. R-squared adjusted value also helps explain the variance as unadjusted R-squared does, but the adjusted R-squared imposes a penalty if a parameter is added that does not explain any variation in sales. Thus, R-squared adjusted value is a stronger indicator of predictive ability and is preferred over the unadjusted R-squared value. According to Wooldridge (2015), the formula to transform R-squared to R-squared adjusted is as follows:

$$R^2 \text{ adjusted} = 1 - (1 - R^2 \text{ unadjusted})(n - 1)/(n - k - 1)$$

Where n is the sample size and k is the total number of explanatory parameters

- Observations: Indicates the number of data points used to construct the training set.

### **Crossvalidation**

<b>Source</b>	<b>RSquare</b>	<b>RASE</b>	<b>Freq</b>
Training Set	0.7524	427433	473
Validation Set	0.7224	411867	209

**Table 8: Crossvalidation**

**Interpretation of Table 8:**

- **R-squared of validation set:** This is the adjusted R-squared value of the validation set. The main objective here is to maximize the adjusted R-squared value of the validation set.
- **RASE (Root Average Squared Error or Root Mean Squared Error):** The square root of the mean value of all the errors. This gives an indication of the magnitude of the error in the model.
- **Frequency:** This indicates the number of data points used to build the training and validation set. There are a total of 682 data points from 682 stores. Here the 682 data points are divided into the training and validation set in the ratio 70:30.

For this department, the analysis has yielded a high adjusted R-squared validation set value of 0.7224. So the model performs well in identifying the parameters that drive sales.

For the rest of this section, the key points of analysis for departments that yielded noteworthy and interesting insights are presented.

### **Department 5 - Socks and Scarves**

- Sales are highly positively impacted by sales area and assortment size (even more so than for other departments) implying that customers like to have more variety of socks and scarves when they are making their purchasing decisions.
- Sales are negatively influenced by higher average summer temperature. This could be because when the weather is warmer, people may not need to wear scarves.
- Sales are negatively influenced by a higher number of economically active people in the lower income range in the area around the store (PEA with monthly income between 415 and 830 Reais). This could be because people in this range have less disposable income and consequently a lower willingness to spend on accessories such as scarves.
- Older and conventional stores have higher sales indicating that customers prefer to buy socks and scarves from well established stores.

### **Department 7 - Footwear**

- Sales are positively correlated to higher average temperatures in both summer and winter. One reason for this could be that LA mostly sells open footwear called havaianas and not closed footwear. So when the temperature is higher, people tend to buy more open footwear.
- Sales increase significantly when the area around the store has more number of higher income people. (PEA with monthly income above 8,300 Reais and A1 Income Level for Private Permanent Homes).

- Sales are strongly negatively impacted by the presence of competitors. This could be explained by the fact that in Brazil most of the super and hypermarkets also sell open footwear as LA does.
- The number of people in the age group 15 to 19 have a statistically significant impact on sales. This could be a key target market for LA for this department.

#### **Department 9 - Briefs and underwear**

- Sales are highly sensitive to (even more so than for other departments) and positively impacted by sales area and stock storage area. One reason for this could be that stores that have higher inventory levels (safety stock) are able to handle volatilities in demand much better than stores with a smaller stock storage area.
- Sales are higher when the number of people in the lower income group increases (B2 Income Level for Private Permanent Homes) and sales are lower when the number of people in the higher income group increases (People with A1 income level and Inactive Residents with monthly income above 8,300 Reais). It appears that the brands present at LA's store appeal more to lower income customers than higher income customers.

#### **Department 11 - Baby products (refers to baby formula and special toys)**

- Sales for baby products are positively correlated to the number of people in the age group 25 to 34. One explanation for this could be that people in this age group are most likely to have babies and subsequently spend money on baby products. Yet, it is surprising that there is no statistical significance between sales and the number of individuals in the age group 0 to 9.

- Sales are positively correlated to stores located in cities where the level of education is the highest (University Segment Level A). This could indicate that education raises awareness among people about the importance of buying specific baby products.
- Sales are lower for non-premium stores. This indicates that people prefer to buy baby products from more upscale and high-end premium stores.

### **Department 12 – Women’s clothing**

- Sales area has a high statistical significance and positive correlation (even more so than for other departments) to this department’s sales. One explanation for this could be that women prefer to have more options available to them when it comes to buying clothes. So sales are greater when the sales area is larger.
- Sales are higher for older stores. This could be the case because women like to shop at the same store for longer periods of time. Once they pick a store, they like to stick to that store.
- Sales are higher when there are more people in the lower income category next to the store (E Income Level for Private Permanent homes). This indicates that the brands available at the store are more appealing to people with lower incomes than higher incomes.
- Sales decrease when there are more shopping centers in the area around the store. This could be because women would have more choices available.

### **Department 15 – Men’s clothing**

- Sales are higher when there are more shopping centers close to the store. This is in sharp contrast to sales for women's clothes, which are negatively correlated to the



presence of shopping centers. One reason for this could be that men prefer to purchase clothes quickly at LA's stores, while they visit shopping malls for other reasons.

- Sales for men's clothes do not increase as much as sales for women's clothes when there is more sales area, as indicated by the lower sales area parameter estimate value for men's clothes. This could be because men do not require as many clothing options as women want.
- Sales are higher when there are more people in the in the age group 50 to 59 years and when there are more economically inactive people with higher incomes, present near the store. Thus, older men and men who have possibly retired but are nevertheless wealthy could be key target demographics for this department.

#### **Department 17 – Books and Magazines**

- Sales are higher when there are more private schools close to the store. Furthermore, sales are higher when there are more young children – age group 0 to 9 and age group 10 to 14 – living close to the store.
- Compared to other departments, for this department, sales do not increase much when the sales area is larger, as evidenced by the low parameter estimate value for the sales area.
- The parameter estimate values for the lowest income category (PEA with monthly income below 415 Reais) and the highest income category (PEA with monthly income above 8,300 Reais) are positive and similar. LA's executives mentioned that in Brazil, books are quite expensive. So it is surprising that the parameter estimate

value for the lowest income category is positive and similar to the estimate value for the highest income category.

#### **Department 24 – Music**

- Sales are higher when the assortment size is greater indicating that customers prefer to have more options available when purchasing music.
- Sales are higher in stores located in cities where the level of education is the highest (University Segment : Level A). Furthermore, sales are higher in stores located close to more private schools. These points indicate that education plays a strong role in increasing music sales.
- Sales appear to be strongly positively correlated to households in the highest income range (A1 Income level for private permanent homes) indicating that a large portion of customers for music come from the higher income range.

#### **Department 25 – Audio and Video equipment**

- Sales are much higher in stores located in areas that have a larger number of economically inactive people with high incomes (Inactive residents with monthly income above 8,300 Reais). So this demographic is very important for sales here. Furthermore, sales are negatively affected when there are more people in the income range 4,150 to 6,225 Reais. Thus the positive effect is only for the highest income group.
- Similar to the results for department 24 (music), sales are much higher in cities where the level of education is highest (University Segment: Level A) and sales are

much lower in cities where the level of education is the lowest (University Segment Level C and D).

- Another interesting point to note here is that sales are higher for stores present near super and hypermarkets. One may have expected that LA's stores would sell less if competitors are present in close proximity, but that does not appear to be the case for this department. One explanation for this could be that the presence of supermarkets drives more foot traffic towards LA's stores.

#### **Department 31 – Hair coloring products**

- Sales are higher when the store has more backroom stock area. This could indicate that these stores are able to better handle demand volatilities by having more safety stock.
- Sales are higher when there are more inactive people in the lower income category near the store(Inactive residents with monthly incomes between 415 and 830 Reais).
- Surprisingly, sales here increase when there are more cosmetic/perfume shops near the store. This is surprising because one may have expected that sales would decrease with the presence of competitors nearby. Furthermore, sales decrease when super and hypermarkets are present nearby. This could indicate that the real competitors LA has for this department are super and hypermarkets, and not other cosmetic shops and pharmacies.

#### **Department 33 – Sports and Beach equipment**

- Sales are lower when there are more gyms nearby. This could indicate that people prefer to go to gyms instead of playing sports. So they buy less sports equipment.

- Sales are higher when there are more private schools near the store. This indicates that an important target audience for sales of sports equipment is school going children.
- Sales increase when there are more higher income households (A1 Income Level for Private Permanent Homes) and higher income individuals (PEA with monthly income above 8,300 Reais) in the region surrounding the store. This could be the case because sports items are considered as luxury goods, which only the higher income individuals may be willing to spend more money on.
- Sales increase when the average winter temperature increases. One explanation for this could be that when the winters are warmer, customers spend more time outdoors. So they spend more money on beach outfits and sports equipment.

#### **Department 40 – Candies**

- Sales increase when the number of higher income individuals increases in the area surrounding the store (PEA with monthly income between 6,225 and 8,300 Reais).
- Sales are higher for stores that have been recently refurbished as indicated by the negative value for the parameter 'Days since last refurbishment'.
- Sales are higher for stores located around more residential complexes and areas where the population density is higher.
- Sales of candies are positively correlated to the number of hotels nearby. But the presence of hotels does not have a significant impact on related items in other departments such as chocolates and biscuits. LA's executives said this is an interesting phenomenon that requires more detailed investigation.

### **Department 43 – Easter**

- Sales are higher for stores located close to shopping centers. LA's executives mentioned that this could be the case because people like to go to shopping malls to do all their Easter shopping. Thus the increased foot traffic could cause more people to buy Easter items from LA's stores located close by.
- Sales are higher for stores located in cities that have more tourists.
- Sales are higher when there are more people in the lower income category next to the store (E income level people and Inactive residents with monthly incomes between 830 and 1245 Reais). This is quite surprising since Easter products are very expensive. LA's executives mentioned that more investigation should be done to see why sales are positively correlated to the number of people in the lower income category.
- Sales decrease as the average summer temperature increases. LA's executives mentioned that there could be one explanation for this phenomenon. The main products for this department are chocolate Easter eggs, which tend to melt or break easily when the weather is warmer. So suppliers do not supply as many chocolate eggs when the weather is warmer since they cannot guarantee the quality. This may be responsible for the lower sales.

### **Department 46 – Chips and Cereal Bars**

- Sales are higher when there are more gyms present near the store. One explanation for this could be that people who go to gyms tend to buy more cereal/energy bars.
- Sales increase when there are more residential complexes nearby.

- Sales do not vary much based on the income of the households. For A1 (highest income), B1 (mid level), and E (lowest) income level for private permanent homes, the parameter estimate values are more or less the same indicating that sales is more or less equally impacted by the different income ranges. LA's executives mentioned that a reason for this could be that even though chips and cereal bars are quite expensive in Brazil, LA sells some basic cheaper versions as well.

#### **Department 48 – Beverages**

- Sales for beverages increase when there are more hotels present near the store.
- Stores located in cities that have a greater tourist population have greater sales indicating that this is a very important target audience for beverages.
- Sales for beverages decrease when the average summer temperature goes up. LA's executives mentioned that this phenomenon was interesting since they expected that warmer temperatures would result in greater sales. More research needs to be done to investigate why sales decrease when the temperature goes up.

#### **Department 51 – Diapers**

- Sales decrease significantly when there are more perfume/hygiene related shops and super markets near the store, indicating the negative effects of competition.
- Stores that have not been refurbished recently tend to have lower sales. This could be because parents want to buy diapers from stores that are newer.
- Stockouts appear to have a significant impact on sales. Stores that have more backroom storage area (to hold safety stock) may have greater sales since they are able to better handle the volatility in demand.

### **Department 63 – Mobile phones**

- Sales are lower for non-premium stores since phones are high value items. LA's executives mentioned that in non-premium stores, the latest phone models are not very likely to be stocked. Currently, LA's suppliers are unable to meet the high demand for mobile phones. So the phones that are received go first to the high-end premium stores, where the probability of selling the phones is the highest.
- Sales increase when there are more lower income people and lower income households in the area near the store (E Income level for private permanent homes and PEA with monthly income below 415 Reais). This is interesting since it seems contrary to what one would expect. Phones are high value items so one would not expect lower income people to be a target demographic. More investigation should be done to understand the reason for this phenomenon.

### **Department 72 – Movies**

- Sales are higher when there are more people in the age group 50-59 in the area next to the store. LA's executives mentioned that one possible explanation for this could be that people in this age group prefer to purchase the movies at the store in the traditional manner, as opposed to buying movies online through streaming services on the internet.
- Sales are lower in non-premium stores and stores located in the street (as opposed to a shopping mall).
- Sales are higher when there are more workers in the higher income category (Workers with monthly income between 6,225 and 8,300 Reais).

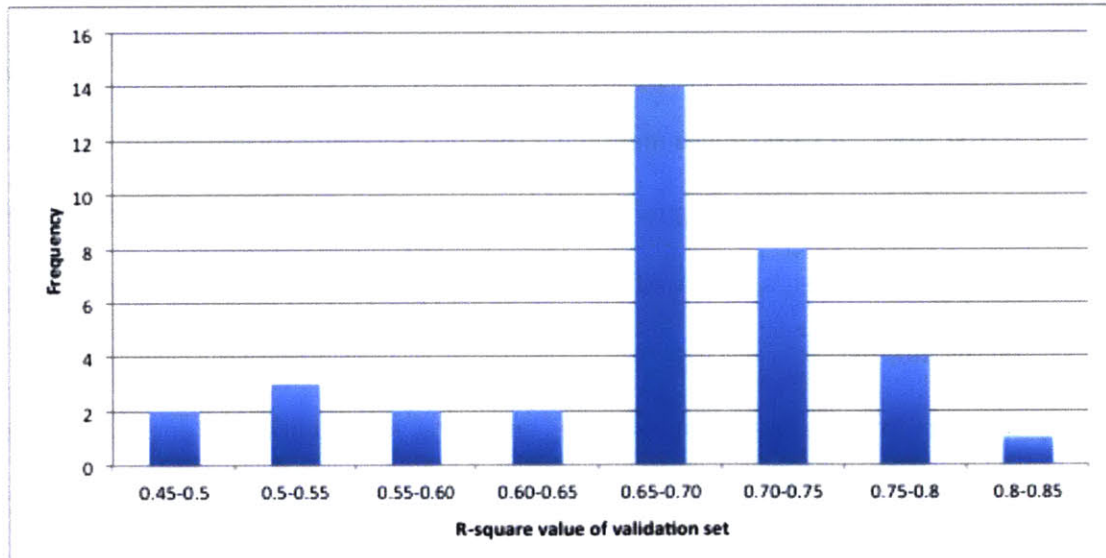
## 5. Discussion

### 5.1 Overall results analysis

The key performance metric for the model used in this thesis is the R-squared adjusted value of the validation set. The distribution of the R-squared value for each of the 36 departments is as shown below in Table 9. The frequency and the cumulative frequency of this distribution are also shown below in Figure 8 and Figure 9.

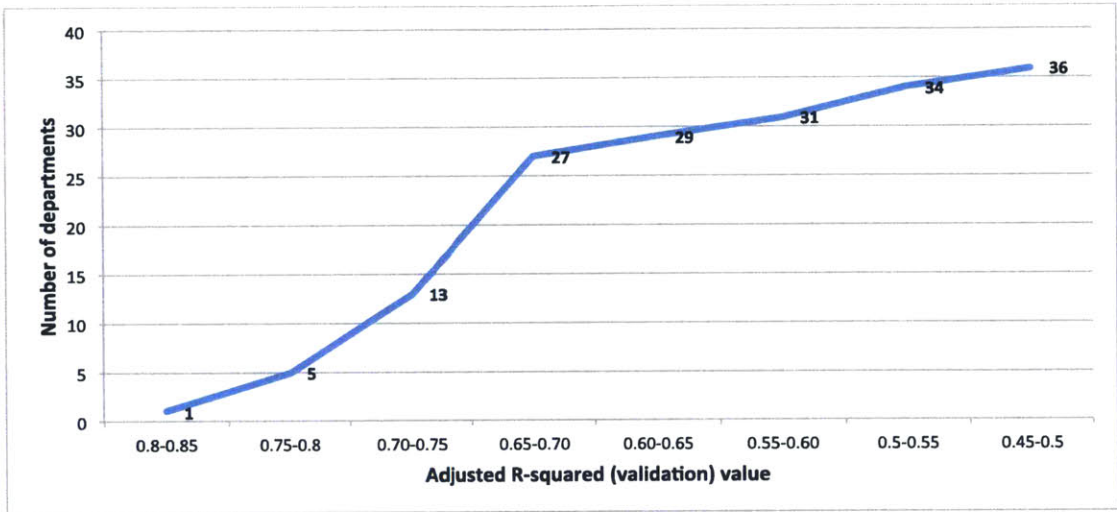
R-squared of validation set	Frequency
0.45-0.5	2
0.5-0.55	3
0.55-0.60	2
0.60-0.65	2
0.65-0.70	14
0.70-0.75	8
0.75-0.8	4
0.8-0.85	1

**Table 9: R-squared validation set frequency distribution**



**Figure 8: R-squared validation set frequency distribution**





**Figure 9: R-squared adjusted validation set cumulative frequency distribution**

The models developed above have yielded useful insights according to LA’s executives. For the vast majority of departments, an R-squared value greater than 0.6 was achieved. However, for 7 departments, the R-squared value is below 0.6. Most of the departments whose R-squared valued is below 0.6 are electronics related including audio/video equipment, mobile phones, Electronic/IT Equipment, and videogames. A more detailed investigation needs to done to understand why the R-squared values for this department family in particular are low.

**5.2 Most significant parameters that affect sales**

Based on the results of the previous section, it is observed that certain key store parameters have a statistically significant and noteworthy (as indicated by the green highlights in the regression output tables) impact across multiple departments. Table 10 below shows the significant parameters that were found to impact sales for more than 5 departments (the

number of departments impacted by the parameter is shown by the frequency row highlighted in yellow at the bottom).

Legend for Table 10:

Red color– Indicates that parameter had a negative estimate value for this department

Green color– Indicates that parameter had a positive estimate value for this department

No color – Indicates that parameter did not have a statistically significant impact for this department.

	High income	Low income	Super market	Sales Area	Shopping centers	Temp.	Inventory Area	Age of store	Education level	Schools	50 -59 age group
Department 1 (Chocolates)			✓		✓	✓		✓	✓	✓	
Department 3 (Children clothes)	✓		✓					✓		✓	
Department 4 (Lingerie)	✓				✓						
Department 5 (Socks&Scarves )		✓	✓	✓		✓		✓			
Department 7 (Footwear)	✓		✓			✓					
Department 9 (Innerwear)	✓	✓		✓			✓				
Department 11 (Baby Products)									✓		
Department 12 (Women's clothing)		✓		✓	✓			✓			
Department 13 (Bed and Bath)						✓					
Department 15 (Men's clothes)	✓			✓	✓						✓
Department 16 (Stationary)	✓										
Department 17 (Books)	✓	✓		✓						✓	
Department 19 (Cleaning)		✓	✓			✓					
Department 21 (Kitchen utensils)			✓								
Department 22 (Table and Bar)	✓				✓						
Department 23 (Toys)	✓			✓							
Department 24 (Music)	✓								✓	✓	
Department 25 (Audio & Video Equipment)	✓		✓						✓		

Department 27 (Home appliances)	✓	✓	✓		✓		✓				✓
Department 28 (Perfumes)	✓	✓	✓	✓	✓		✓	✓			
Department 30 (Electronics & IT)		✓		✓	✓		✓				
Department 31 (Hair Coloring)		✓	✓	✓			✓				
Department 32 (Cosmetics)			✓	✓		✓	✓				
Department 33 (Sports & Beach)	✓					✓				✓	
Department 34 (Imported Toys)	✓										✓
Department 40 (Candies)	✓							✓			
Department 43 (Easter)		✓			✓	✓					
Department 44 ( Plastic utilities)		✓	✓								
Department 45 ( Video games)		✓		✓							
Department 46 (Chips& Cereal)	✓	✓									
Department 48 (Beverages)						✓					
Department 49 (Biscuits, Cookie s)	✓	✓	✓								
Department 51 (Diapers)					✓		✓	✓	✓		
Department 56 (Christmas)	✓										✓
Department 63 (Telephones)		✓									
Department 72 (Movies)	✓										✓

Frequency	20	15	13	11	10	10	7	7	5	5	5
-----------	----	----	----	----	----	----	---	---	---	---	---

**Table 10: Key statistically significant parameters across multiple departments**

Table 10 provides useful insights. Eleven key parameters have a statistically significant impact across 5 or more departments. Thus, these 11 key parameters have the greatest impact on LA's overall revenues. Furthermore, nearly half of these 11 parameters have different impacts (positive/negative) for different departments. The reasons for these effects are discussed in more detail in the results section for the corresponding

departments. Furthermore, from Table 10, we can also generalize the overall results:

- Income level of the people surrounding the store has a significant impact on the sales of many departments. For many departments, sales increase significantly when there is a higher number of people with high incomes living in the area around the store. However, for certain departments such as innerwear, female clothes, hair coloring products, sales increase significantly when there is a higher number of people with low incomes living in the area around the store.
- The presence of shopping centers (within a 1 km radius around LA's stores) has a significant impact on many departments. For some departments, the presence of shopping centers helps increase sales for LA stores by increasing overall foot traffic. However, for other departments the presence of supermarkets decreases sales for LA stores. For instance, for men's clothes, LA stores located close to shopping centers have higher sales whereas for women's clothes, LA stores located close to shopping centers have lower sales.
- The proximity of supermarkets to LA's stores has a significant influence on many departments. Supermarkets present close to LA's stores have a negative influence on sales for some departments (such as perfumes, hair coloring, and cosmetics) and positive influence on sales for some other departments (such as audio/video equipment, and kitchen utensils).
- Inventory storage area has a statistically significant positive impact on sales for certain departments such as Innerwear and Cosmetics. When more inventory storage area is available, more safety stock could be available which could help minimize stockouts (or lost sales). However, while a larger inventory storage area could help increase sales, it may not necessarily result in higher profits since costs

are also likely to increase.

- The presence of schools around the store and the education level of the city have a statistically significant impact in increasing sales for some departments. For instance, sales for departments such as chocolates, children's clothes, books, and music increase when more schools are present near the LA store. Furthermore, for stores present in cities where the level of education is high, sales for departments such as baby products, diapers, and music are higher. These points illustrate that education plays an important role in raising awareness for certain products.
- Temperature has a significant impact on the sales of many departments. For certain departments such as cosmetics, sales are higher when the temperature is higher since consumers are likely to buy more personal care products such as skin lotions. However, for certain departments such as chocolates and easter, sales are lower when the temperature is higher since these products have a higher probability of getting spoiled when the temperature is higher.
- The age of the store also has a statistically significant impact on the sales for certain departments. For instance, for perfumes and diapers, sales are higher for newer stores. This could be because these stores appear more hygiene in the eyes of the consumer (irrespective of how hygienic the store actually is) and for departments such as diapers, parents may place more importance on hygiene. For certain other departments such as female clothes, older stores have higher sales since women may have greater store loyalty.

In the next section, the concluding points for this thesis are presented including areas for future scope and some of the limitations of the model developed in this thesis.

## **6. Conclusion**

Most of the earlier research efforts on the factors that drive retail sales have focused on analyzing a few statistically significant parameters in detail. However, LA did not have a clear-cut idea about the key parameters that should be studied in detailed. In this thesis, a large number of internal and external store parameters are analyzed and the statistically significant parameters that influence sales are presented in the results section. In the future, LA can build on these results and go into more detailed analysis about these parameters.

### **6.1 Limitations of the model**

- For the analysis in this thesis, a correlation coefficient of 0.7 was used as a threshold for eliminating highly correlated predictor variables. To eliminate multicollinearity further, the VIF criterion was used. But despite performing this step, multicollinearity still hasn't been fully eliminated. The parameter estimates may still be skewed. Thus, more importance should be given to the list of parameters identified as statistically significant and the direction in which these parameters work (negative/positive), rather than the actual parameter estimate values. To find out more accurate parameter estimate value, more complex regression analysis tools must be used as as Ridge and Lasso Regression.
- The research presented in this work was more exploratory in nature. The goal of this thesis was to identify the key statistically significant parameters that drive sales from a large initial list of parameters. The analysis performed here attempts to find the statistically significant factors based purely on numbers. In some cases, this yielded results that could not be practically explained. In other words, a parameter showed up

as statistically significant but it was hard to make a practical connection between the parameter and the department that was being analyzed. For example, the presence of banks was shown to have a negative correlation with sales for toys. Another example was that the presence of fuel stations was shown to have a negative impact for men's clothes. More detailed analysis should be done in the future to investigate why certain parameters are statistically significant, even though there appears to be no clear practical explanation. In the next sub-section, some more fruitful areas of future research are presented.

## **6.2 Fruitful areas of future research**

Possible areas of future research include:

- For each department, the analysis done has yielded a list of the statistically significant parameters that impact department sales. In the future more research could be done by studying these parameters in more detail.
- LA could follow the methodology present in this thesis to analyze data at the product line and SKU level. With this analysis, LA can make optimal product assortment decisions.
- The model developed in this thesis hasn't factored in the effect of some key internal store parameters – promotions and pricing. Future research effects could focus on how to incorporate the effects of these parameters into the model developed in this thesis. This would make the model more comprehensive and holistic. So future research efforts could focus on collecting suitable data for these parameters and then incorporating the effect of these two parameters, to arrive at a more comprehensive model.

For some departments, the analysis yielded surprising results which go contrary to the established beliefs of LA's executives and which need to be investigated further. Some of these findings include:

- Sales for mobile phones and easter products are higher for stores surrounded by more lower income people. This is surprising, because both mobile phones and easter products are expensive.
- Sales of candies are positively correlated to the number of hotels nearby. But the presence of hotels does not have a significant impact on related items such as chocolates and biscuits. This is an interesting phenomenon, which requires more detailed investigation.
- Sales for videogames are positively correlated with the number of people in the age group 25 to 34. LA's executives mentioned that this result was quite surprising since they expected that the major target demographic for videogame sales would be students in schools and colleges.
- Sales for hair coloring products increase when there are more cosmetic/perfume shops and pharmacies nearby. Furthermore, sales for audio/video equipment are higher for stores near super and hypermarkets. These points are surprising because one may have expected that sales would decrease with the presence of competitors nearby.
- Sales for beverages decrease when the summer temperature increases.



## References

Anderson, E. T., Fitzsimons, G. J., & Simester, D. (2006). Measuring and mitigating the costs of stockouts. *Management Science*, 52(11), 1751-1763.

Artz, G. M., & Stone, K. E. (2006). Analyzing the impact of Wal-Mart supercenters on local food store sales. *American journal of agricultural economics*, 88(5), 1296-1303.

Boatwright, P., & Nunes, J. C. (2001). Reducing assortment: An attribute-based approach. *Journal of marketing*, 65(3), 50-63.

*Burnham, K. P.; Anderson, D. R. (2002), Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach (2nd ed.), Springer-Verlag, ISBN 0-387-95364-7.*

Carpenter, J. M., & Moore, M. (2006). Consumer demographics, store attributes, and retail format choice in the US grocery market. *International Journal of Retail & Distribution Management*, 34(6), 434-452.

Crowther, P. S., & Cox, R. J. (2005, September). A method for optimal division of data sets for use in neural networks. In *Knowledge-based intelligent information and engineering systems* (pp. 1-7). Springer Berlin Heidelberg.

Dattalo, P. (2013). *Analysis of multiple dependent variables*. Oxford University Press.

Dormann, C. F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., ... & Münkemüller, T. (2013). Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. *Ecography*, 36(1), 27-46.

Götz, O., Liehr-Gobbers, K., & Krafft, M. (2010). Evaluation of structural equation models using the partial least squares (PLS) approach. In *Handbook of partial least squares* (pp. 691-711). Springer Berlin Heidelberg.

Hair, J. F., Black, W. C., Babin, B. J., Anderson, R. E., & Tatham, R. L. (2006). *Multivariate data analysis* (Vol. 6). Upper Saddle River, NJ: Pearson Prentice Hall.

Hastie, T., Tibshirani, R., Friedman, J., & Franklin, J. (2005). The elements of statistical learning: data mining, inference and prediction. *The Mathematical Intelligencer*, 27(2), 83-85.

Hawkins, D. M., Basak, S. C., & Mills, D. (2003). Assessing model fit by cross-validation. *Journal of chemical information and computer sciences*, 43(2), 579-586.

Huang, C., & Townshend, J. R. G. (2003). A stepwise regression tree for nonlinear approximation: applications to estimating subpixel land cover. *International Journal of Remote Sensing*, 24(1), 75-90.

Jayasankara Prasad, C., & Ramachandra Aryasri, A. (2011). Effect of shopper attributes on retail format choice behaviour for food and grocery retailing in India. *International Journal of Retail & Distribution Management*, 39(1), 68-86.

Konishi, H. (2005). Concentration of competing retail stores. *Journal of Urban Economics*, 58(3), 488-512.

Koufakou, A., Georgiopoulos, M., Anagnostopoulos, G., & Kasparis, T. (2001). Cross-validation in Fuzzy ARTMAP for large databases. *Neural Networks*, 14(9), 1279-1291.

Kumar, V., & Karande, K. (2000). The effect of retail store environment on retailer performance. *Journal of Business Research*, 49(2), 167-181.

Montgomery, D. C., Peck, E. A., & Vining, G. G. (2012). *Introduction to linear regression analysis* (Vol. 821). John Wiley & Sons.

Oppewal, H., & Koelemeijer, K. (2005). More choice is better: Effects of assortment size and composition on assortment evaluation. *International Journal of Research in Marketing*, 22(1), 45-60.

Palazón-Vidal, M., & Delgado-Ballester, E. (2005). Sales promotions effects on consumer-based brand equity. *International Journal of Market Research*, 47(2), 179-204.

Refaeilzadeh, P., Tang, L., & Liu, H. (2009). Cross-validation. In *Encyclopedia of database systems* (pp. 532-538). Springer US.

Verbeke, W. (2006). Functional foods: Consumer willingness to compromise on taste for health?. *Food Quality and Preference*, 17(1), 126-131.

Wooldridge, J. (2015). *Introductory econometrics: A modern approach*. Nelson Education.

Yoo, S., & Chang, Y. (2005). An exploratory research on the store image attributes affecting its store loyalty.

## Appendix A

Table 11: Regression output for Department 3 (Children's clothes)

Parameter	Parameter Estimate	P-value
Private schools (1km)	383.1243	<. 0001
Sales Area (in meter square)	110.6334	<. 0001
0 to 9 years old	10.8543	<. 0001
A2 Income Level for Private Permanent Homes	6.9342	<. 0001
Residential complexes	1.4743	<. 0001
Banks (1km)	-719.7153	<. 0001
Hospitals (1km)	-3,193.2435	<. 0001
Intercept	-56,790.693	<. 0001
Hotel (1km)	1,242.3134	0.0235
Days since Inauguration of store (Assume end date - 1st Jan, 2016)	-254.4534	0.0267
Tourist Population(2013)	0.0132	0.0567
Pharmacies (1km)	485.0835	0.0832
A1 Income Level for Private Permanent Homes	0.5267	0.087
PEA with monthly income above 8,300 Reais	0.0643	0.0885
A2 Income Level people	0.6912	0.0935

### Summary of Fit

RSquare	0.716658
RSquare Adj	0.706716
Root Mean Square Error	52875.6
Mean of Response	63295.46
Observations (or Sum Wgts)	473

### Crossvalidation

Source	RSquare	RASE	Freq
Training Set	0.7167	51917	473
Validation Set	0.6897	50005	209

**Table 12: Regression output for Department 4 (Lingerie)**

Parameter	Parameter Estimate	P-value
Perfume and Cosmetic shops (1km)	55,032.7012	<. 0001
Sales Area (in meter square)	583.0806	<. 0001
A1 Income Level for Private Permanent Homes	188.3184	<. 0001
Residential complexes	-7.028	<. 0001
Banks (1km)	-5,993.6829	<. 0001
Super and Hyper Markets (1km)	-7,838.8993	<. 0001
Fuel station(1km)	-22,773.6469	<. 0001
Store Type (Conventional)	-71,698.3075	<. 0001
Geographic Region - Not Nordeste	-89,700.8846	<. 0001
Private schools (1km)	8,433.6143	0.0003
Days since Inauguration of store (Assume end date - 1st Jan, 2016)	17.799	0.0006
PEA with monthly income between 6,225 and 8,300 Reais	251.4662	0.0021
PEA with monthly income above 8,300 Reais	5.86	0.0033
Pharmacies (1km)	12,242.5512	0.0080
Intercept	-11,1526.793	0.0198
Shopping Center (1km)	-49,349.6591	0.0415
Restaurants(1km)	1,274.4691	0.0497
Store Location - Street	-34,804.3391	0.0799

### Summary of Fit

RSquare	0.693529
RSquare Adj	0.682776
Root Mean Square Error	324296.8
Mean of Response	425741.8
Observations (or Sum Wgts)	473

### Crossvalidation

Source	RSquare	RASE	Freq
Training Set	0.6935	318416	473
Validation Set	0.5496	349952	209

**Table 13: Regression output for Department 5 (Socks & Scarves)**

Parameter	Parameter Estimate	P-value
Sales Area (in meter square)	4,678.6886	<. 0001
Pharmacies (1km)	119.7007	<. 0001
PEA with monthly incomes between 415 and 830 Reais	-15,289.8216	<. 0001
Days since Inauguration of store (Assume end date - 1st Jan, 2016)	2,162.0417	0.0021
Hotel (1km)	-12,976.9581	0.0024
Average summer temperature (C°)	-3,609.8188	0.0026
Store Location - Street	-838.4209	0.0050
Perfume and Cosmetic shops (1km)	58.2986	0.0070
Geographic Region - NOT South East	0.9019	0.0074
Bars (1km)	-4,203.7833	0.0134
A1 Income Level for Private Permanent Homes	23.1621	0.0154
Hospitals (1km)	-11,030.5043	0.0156
Intercept	70,942.2329	0.017
Store Type - Conventional	4,131.8108	0.0206
Stock area (in meter square)	-0.6841	0.0305
Assortment Size	55.03206	0.0434
Banks (1km)	921.9607	0.0457

### Summary of Fit

RSquare	0.708975
RSquare Adj	0.697308
Root Mean Square Error	76402.97
Mean of Response	128984.2
Observations (or Sum Wgts)	468

### Crossvalidation

Source	RSquare	RASE	Freq
Training Set	0.7090	74836	468
Validation Set	0.6914	84929	214

**Table 14: Regression output for Department 7 (Footwear)**

Parameter	Parameter Estimate	P-value
Sales Area (in meter square)	100.8363	<. 0001
PEA with monthly income above 8,300 Reais	1.768	<. 0001
Workers with monthly income between 6,225 and 8,300 Reais	-1.0284	<. 0001
Banks (1km)	-914.2996	<. 0001
Pharmacies (1km)	2,863.7747	0.0001
A1 Income Level for Private Permanent Homes	17.5167	0.0001
Geographic Region - NOT South East	-34,600.5504	0.0008
Restaurants(1km)	328.2452	0.0011
Hospitals (1km)	-2,708.0603	0.0020
Days since Inauguration of store (Assume end date - 1st Jan, 2016)	2.3808	0.0022
Residential complexes	-3.3415	0.0031
Super and Hyper Markets (1km)	-4,030.7789	0.0031
Private schools (1km)	296.6289	0.0034
Geographic Region - NOT South	-32,685.414	0.0036
Average winter temperature (C°)	2,812.5128	0.0078
Intercept	-11,627.2287	0.0123
Average summer temperature (C°)	119.3532	0.0324
15 to 19 years old	2.2977	0.0625

### Summary of Fit

RSquare	0.68937
RSquare Adj	0.676759
Root Mean Square Error	50686.96
Mean of Response	78153.77
Observations (or Sum Wgts)	488

### Crossvalidation

Source	RSquare	RASE	Freq
Training Set	0.6894	49637	488
Validation Set	0.7164	54994	194

**Table 15: Regression output for Department 9 (Briefs and Underwear)**

Parameter	Parameter Estimate	P-value
A1 income level people	-7,164.383	<. 0001
Inactive residents with monthly income above 8,300 Reais	-37,345.2078	<. 0001
B2 Income Level for Private Permanent Homes	149.5761	<. 0001
25 to 34 years old	-2.6732	<. 0001
50 to 59 years	-5.855	<. 0001
Average income2014	-6,0993.204	<. 0001
Stock area (in meter square)	3.7442	0.0003
Casa	3.9714	0.0006
Store Type: Conventional	6.4479	0.0075
Fuel station (1km)	-1.0326	0.0123
Sales Area (in meter square)	1,875.4718	0.0138
Pharmacies (1km)	5.3074	0.0149
C1 income level people	-3,905.5475	0.0234
D income level people	-38.0326	0.0424
Perfume and Cosmetic shops (1km)	-0.8496	0.0485
Gyms (1km)	1.7256	0.0510
Intercept	2,119.5867	0.0697
10 to 14 years old	-1.8098	0.0969

### Summary of Fit

RSquare	0.665963
RSquare Adj	0.653932
Root Mean Square Error	63698.74
Mean of Response	87105.54
Observations (or Sum Wgts)	490

### Crossvalidation

Source	RSquare	RASE	Freq
Training Set	0.6660	62518	490
Validation Set	0.6715	66108	192

**Table 16: Regression output for Department 11 (Baby products)**

Parameter	Parameter Estimate	P-value
Geographic Region - NOT North East	-19,325.0784	<.0001
University Segment: Level A	21,808.9395	<.0001
25 to 34 years old	2.3111	<.0001
Store Status- NOT premium	-3,681.9781	<.0001
Women	5.2149	0.0006
Location - Street	-8,813.7416	0.0008
Days since Inauguration of store (Assume end date - 1st Jan, 2016)	2.3529	0.0028
Perfume and Cosmetic shops (1km)	2,174.399	0.0072
Geographic Region - NOT South East	-9,237.115	0.0080
Banks (1km)	-327.2094	0.0088
Residential complexes	1.2716	0.0152
Geographic Region - NOT Central	-11,116.9694	0.0253
University Segment: Level D	-8,692.3166	0.0290
University Segment: Level C	-16,337.4665	0.0488
Intercept	20,709.7577	0.0527
University Segment: Level E	-6,895.9362	0.0530
Fuel station(1km)	-2,139.9476	0.0678

### Summary of Fit

RSquare	0.795308
RSquare Adj	0.788384
Root Mean Square Error	47369.03
Mean of Response	104328.3
Observations (or Sum Wgts)	490

### Crossvalidation

Source	RSquare	RASE	Freq
Training Set	0.7953	46540	490
Validation Set	0.8300	44403	192



**Table 17: Regression output for Department 12 (Women's clothing)**

Parameter	Parameter Estimate	P-value
Sales Area (in meter square)	11,680.5153	<. 0001
Days since Inauguration of store (Assume end date - 1st Jan, 2016)	11,263.0393	<. 0001
E Income Level for Private Permanent Homes	115.7723	<. 0001
Geographic Region - NOT South	6.0786	<. 0001
Geographic Region - NOT South East	4.4596	<. 0001
Fuel station(1km)	-4,246.7635	<. 0001
Intercept	18,157.72	0.0014
Average summer temperature (C°)	1.5992	0.0029
Shopping Center (1km)	-6,452.1036	0.0050
Gyms (1km)	-1,882.7633	0.0193
Inactive residents with monthly income above 8,300 Reais	-16,333.6987	0.0487

**Summary of Fit**

RSquare	0.676145
RSquare Adj	0.669384
Root Mean Square Error	66046.63
Mean of Response	75867.23
Observations (or Sum Wgts)	490

**Crossvalidation**

Source	RSquare	RASE	Freq
Training Set	0.6761	65301	490
Validation Set	0.6371	80405	192

**Table 18: Regression output for Department 13 (Bed and Bath)**

Parameter	Parameter Estimate	P-value
Geographic Region - Not South	84,274.2367	<. 0001
Geographic Region - Not South East	50,226.8428	<. 0001
Pharmacies (1km)	13,948.1387	<. 0001
Sales Area (in meter square)	334.7247	<. 0001
D Income Level for Private Permanent Homes	30.1381	<. 0001
Banks (1km)	-3,129.9076	<. 0001
Fuel station(1km)	-12,135.2325	<. 0001
Average summer temperature (C°)	2,939.2294	<. 0001
PEA with monthly income above 8,300 Reais	7.2693	0.0011
15 to 19 years old	-10.6930	0.0017
Store Type: Conventional	-87,688.1317	0.0049
Hotel (1km)	-3,611.4076	0.0647
B1 income level people	1.8070	0.0670
PEA with monthly income between 6,225 and 8,300 Reais	-2.3073	0.0675
Gyms (1km)	-11,264.7309	0.0684
Store Location - Street	-15,441.1696	0.0693
PIB per Capita Total (2012)	-0.8137	0.0710
Residential complexes	-4.9973	0.0725
Intercept	-14,4415.5133	0.0755
High school graduates	0.0154	0.0810
Geographic Region - Not Central	18,654.0504	0.0825
Private schools (1km)	749.7187	0.0845
PEA Density (2007)	-2.4724	0.0856

**Summary of Fit**

<b>RSquare</b>	0.594932
<b>RSquare Adj</b>	0.57494
<b>Root Mean Square Error</b>	199662.3
<b>Mean of Response</b>	251073.4
<b>Observations (or Sum Wgts)</b>	490

**Crossvalidation**

Source	RSquare	RASE	Freq
Training Set	0.5949	194711	490
Validation Set	0.6593	172170	192

**Table 19: Regression output for Department 15 (Clothes for men)**

Parameter	Parameter Estimate	P-value
Sales Area (in meter square)	118.0797	<. 0001
Inactive residents with monthly income above 8,300 Reais	7.7628	<. 0001
Days since Inauguration of store (Assume end date - 1st Jan, 2016)	4.1594	<. 0001
Fuel station(1km)	-4,073.0186	<. 0001
50 to 59 years	5.6912	<. 0001
Shopping Center (1km)	10,187.162	<. 0001
Store Type - Conventional	-31,070.0216	0.0014
Gyms (1km)	-7,136.1724	0.0016
Hospitals (1km)	-2,653.9995	0.0089
Super and Hyper Markets (1km)	-2,652.91	0.0109
PEA Density (2007)	6.2448	0.0111
Densidade Demográfica (2014)	-7.7253	0.0138
Workers with monthly income between 6,225 and 8,300 Reais	34.1652	0.0278
Pharmacies (1km)	1,490.9357	0.036
E Income Level for Private Permanent Homes	1.3866	0.0464
PEA with monthly income above 8,300 Reais	0.2445	0.0524
Average summer temperature (C°)	-855.1017	0.0586
D Income Level for Private Permanent Homes	1.413	0.0596
Intercept	-3,126.6272	0.0948

<b>Summary of Fit</b>			
RSquare		0.666325	
RSquare Adj		0.652836	
Root Mean Square Error		63553.8	
Mean of Response		75631.96	
Observations (or Sum Wgts)		490	
<b>Crossvalidation</b>			
Source	RSquare	RASE	Freq
Training Set	0.6663	62243	490
Validation Set	0.6777	73576	192

**Table 20: Regression output for Department 16 (Stationary items)**

Parameter	Parameter estimate	p-value
Sales Area (in meter square)	325.1172	<. 0001
Store Location - Street	-35,225.8745	<. 0001
Geographic Region - Not North East	-57,600.118	<. 0001
PEA with monthly income above 8,300 Reais	2.5803	<. 0001
Geographic Region - Not Central	-61,332.8026	0.0001
15 to 19 years old	18.1396	0.0003
Intercept	98,404.1762	0.0004
Store Type - Conventional	-78,773.114	0.0006
Banks (1km)	-1,666.4058	0.0010
Geographic Region - Not Sudeste	-35,696.7458	0.0014
Perfume and Cosmetic shops (1km)	7,861.4127	0.0044
Stock area (in meter square)	-46.0646	0.0121
Store Status - premium	-11,948.2842	0.0121
TGCA Renda Média (%) (2010-2014)	6,078.2824	0.0216
Residential complexes	-2.5479	0.039
C1 income level people	-0.3499	0.0416
Days since Inauguration of store (Assume end date - 1st Jan, 2016)	4.5599	0.0636
Restaurants(1km)	-517.8507	0.0656

### Summary of Fit

RSquare	0.686854
RSquare Adj	0.674195
Root Mean Square Error	146173.6
Mean of Response	281085.6
Observations (or Sum Wgts)	490

### Crossvalidation

Source	RSquare	RASE	Freq
Training Set	0.6869	143159	490
Validation Set	0.7390	112914	192

**Table 21: Regression output for Department 17 (Books and Magazines)**

Parameter	Parameter Estimate	P-value
Intercept	40,856.868	<. 0001
Sales Area (in meter square)	67.8818	<. 0001
Days since Inauguration of store (Assume end date - 1st Jan, 2016)	3.1275	<. 0001
Store Location - Street	-13,188.2997	<. 0001
Área of city (Km <sup>2</sup> )	1.5845	0.0011
PEA with monthly income below 415 Reais	0.3616	0.0029
MonthlyAverage income(2014)	-4.7812	0.0034
0 to 9 years old	12.1955	0.0074
10 to 14 years old	12.4267	0.008
PIB per Capita Total (2012)	-0.2115	0.0123
PEA Density (2007)	1.5722	0.0132
Private schools (1km)	536.6925	0.0171
Hospitals (1km)	-1,325.9977	0.0247
PEA with monthly income above 8,300 Reais	0.3033	0.0309
Store Status: NOT Premium	-1,732.9215	0.0369
Geographic Region - NOT Central	-7,895.2321	0.047
PEA with monthly income between 830 and 1,245 Reais	-0.2231	0.0641
Geographic Region - NOT South East	-4,436.3056	0.0673
E income level people	-0.0563	0.0788
Inactive residents with monthly income between 415 and 830 Reais	-0.0268	0.0963

### Summary of Fit

RSquare	0.739999
RSquare Adj	0.729488
Root Mean Square Error	36496.43
Mean of Response	88049.26
Observations (or Sum Wgts)	490

### Crossvalidation

Source	RSquare	RASE	Freq
Training Set	0.7400	35744	490
Validation Set	0.7156	40097	192

**Table 22: Regression output for Department 19 (Cleaning products)**

Parameter	Parameter Estimate	P-value
Sales Area (in meter square)	157.9531	<. 0001
D Income Level for Private Permanent Homes	14.2269	<. 0001
PEA with monthly income below 415 Reais	1.0917	<. 0001
C1 income level people	-2.2669	<. 0001
Inactive residents with monthly income below 415 Reais	0.9501	<. 0001
Fuel station(1km)	-4,688.6975	0.0002
Store Location - Street	-15,783.4839	0.002
Private schools (1km)	836.2194	0.0179
Store Type: Conventional	-31,534	0.0254
Average winter temperature (C°)	3,092.867	0.0337
Gyms (1km)	-2,205.2849	0.0408
Intercept	-20,725.0448	0.0471
Densidade Demográfica (2014)	-2.6947	0.0537
Super and Hyper Markets (1km)	1,123.3385	0.0614
PEA Density (2007)	2.428848	0.0664
Perfume and Cosmetic shops (1km)	540.9696	0.0744
B2 income level people	0.5749	0.0851

### Summary of Fit

RSquare	0.631315
RSquare Adj	0.617225
Root Mean Square Error	91059.45
Mean of Response	147462.2
Observations (or Sum Wgts)	490

### Crossvalidation

Source	RSquare	RASE	Freq
Training Set	0.6313	89277	490
Validation Set	0.6516	73119	192

**Table 23: Regression output for Department 21 (Kitchen utensils)**

Parameter	Parameter Estimate	p-value
Super and Hyper Markets (1km)	1,028.8054	<.0001
Sales area (in meter square)	138.2672	<.0001
25 to 34 years old	3.1058	<.0001
Geographic Region - Not South East	-24,497.0549	<.0001
Geographic Region - Not North East	-49,803.4671	<.0001
Geographic Region - Not Central	-28,619.3736	0.0005
Women	3.2861	0.001
Store Location - Street	-10,135.2798	0.0131
PEA with monthly income between 830 and 1,245 Reais	-0.5015	0.0262
PEA with monthly income below 415 Reais	0.5514	0.0425
Intercept	101,068.3227	0.0444
Residential complexes	-2.828	0.0457
Perfume and Cosmetic shops (1km)	2,296.9377	0.0512
Inactive residents with monthly income below 415 Reais	-1.0512	0.0524
Days since Inauguration of store (Assume end date - 1st Jan, 2016)	1.9872	0.0544
Private schools (1km)	774.5372	0.0568
Average summer temperature (C°)	-2,376.7228	0.0589
Gyms (1km)	-1,812.9573	0.061
Workers with monthly income between 6,225 and 8,300 Reais	0.2913	0.0625
PEA Density (2007)	0.4156	0.071
PIB per Capita Total (2012)	0.03244	0.0725
Fuel station(1km)	-4,183.9697	0.085

<b>Summary of Fit</b>			
RSquare		0.701148	
RSquare Adj		0.688404	
Root Mean Square Error		73758.28	
Mean of Response		133020.4	
Observations (or Sum Wgts)		490	
<b>Crossvalidation</b>			
Source	RSquare	RASE	Freq
Training Set	0.7011	72160	490
Validation Set	0.7272	63692	192

**Table 24: Regression output for Department 22 (Table and Bar)**

Parameter	Parameter Estimate	P-value
Sales Area (in meter square)	213.4704	<. 0001
Geographic Region – Not South East	-44,445.0433	<. 0001
Geographic Region – Not Central	-63,614.4586	<. 0001
Geographic Region – Not North East	-47,467.2733	0.0006
A1 Income Level for Private Permanent Homes	82.8855	0.0023
PEA with monthly income between 6,225 and 8,300 Reais	60.9499	0.0045
Intercept	269,281.6834	0.0052
Average summer temperature (C°)	-6,656.8407	0.035
Residential complexes	-5.4458	0.0451
Private schools (1km)	1,737.6495	0.0476
Inactive residents with monthly income between 830 and 1245	-5.6350	0.0489
Universities (1km)	-2,822.3676	0.0519
Pharmacies (1km)	2,337.7380	0.0525
Shopping Center (1km)	10,679.8993	0.0568
PIB per Capita Total (2012)	-0.4456	0.0625
Gyms (1km)	-3,215.0014	0.0678
B2 income level people	0.1740	0.0788
Inactive residents with monthly income below 415 Reais	2.6619	0.0821

**Summary of Fit**

<b>RSquare</b>	<b>0.583081</b>
<b>RSquare Adj</b>	<b>0.569887</b>
<b>Root Mean Square Error</b>	<b>131736.2</b>
<b>Mean of Response</b>	<b>208246</b>
<b>Observations (or Sum Wgts)</b>	<b>490</b>

**Crossvalidation**

<b>Source</b>	<b>RSquare</b>	<b>RASE</b>	<b>Freq</b>
Training Set	0.5831	129568	490
Validation Set	0.6663	121070	192



**Table 25: Regression output for Department 23 (Toys)**

Parameter	Parameter Estimate	p-value
Sales Area (in meter square)	663.0054	<. 0001
A1 Income Level for Private Permanent Homes	161.7783	<. 0001
Days since Inauguration of store (Assume end date - 1st Jan, 2016)	27.4	<. 0001
Banks (1km)	-5,050.808	<. 0001
PEA with monthly income above 8,300 Reais	11.4521	0.0002
Perfume and Cosmetic shops (1km)	32,378.8789	0.0018
Intercept	-27,8756.5087	0.0037
Store Location - Street	-57,350.9212	0.0057
Store Status - NOT premium	-45,350.9645	0.0160
Assortment Size	327.0778	0.0172
Store Type - Conventional	-172,699.5574	0.0181
Geographic Region - NOT North East	-35,569.0822	0.0292
PEA with monthly income between 830 and 1,245 Reais	-4.0006	0.0732
Geographic Region - NOT South East	-43,086.1942	0.0748
PEA with monthly income below 415 Reais	2.7766	0.0881

**Summary of Fit**

RSquare	0.735906
RSquare Adj	0.727821
Root Mean Square Error	369968.2
Mean of Response	683960.8
Observations (or Sum Wgts)	506

**Crossvalidation**

Source	RSquare	RASE	Freq
Training Set	0.7359	364072	506
validation Set	0.6908	344890	176

**Table 26: Regression output for Department 24 (Music)**

Parameter	Parameter Estimate	P-value
Perfume and Cosmetic shops (1km)	7,649.4324	<. 0001
Sales Area (in meter square)	109.0017	<. 0001
A1 Income Level for Private Permanent Homes	26.2692	<. 0001
Days since Inauguration of store (Assume end date - 1st Jan, 2016)	5.1203	<. 0001
Banks (1km)	-639.6036	<. 0001
Store Location - Street	-22,688.3519	<. 0001
Store Type: Conventional	-51,897.5383	<. 0001
University Segment: Level A	26,934.0653	<. 0001
Assortment Size	90.0746	0.0002
Geographic Region - NOT South East	-13,004.7928	0.001
PIB per Capita Total (2012)	-0.4667	0.0159
University Segment: Level D	-36,313.0903	0.024
University Segment: Level C	-11,943.1341	0.0306
Intercept	-20,289.6683	0.0311
PEA Density (2007)	-0.919	0.0587
Private schools (1km)	186.2178	0.0601
Store Status: NOT premium	-5,730.2259	0.0806
PEA with monthly income below 415 Reais	0.5452	0.0983

### Summary of Fit

RSquare	0.751619
RSquare Adj	0.742967
Root Mean Square Error	64692.19
Mean of Response	148934.5
Observations (or Sum Wgts)	506

### Crossvalidation

Source	RSquare	RASE	Freq
Training Set	0.7516	63531	506
Validation Set	0.6709	65304	176

**Table 27: Regression output for Department 25 (Audio & Video Equipment)**

Parameter	Parameter Estimate	P-value
Sales Area (in meter square)	1,168.7809	<. 0001
PEA with monthly incomes between 415 and 830 Reais	-4.2592	<. 0001
TGCA População (%) (2010-2014)	-192,114.3847	0.0004
Store Type: Conventional	-374,832.6255	0.0015
University Segment: Level D	-591,070.5875	0.0017
Average winter temperature (C°)	34,767.629	0.0042
University Segment: Level C	-364,313.4092	0.0121
Inactive residents with monthly income above 8,300 Reais	50.7181	0.0167
Inactive residents with monthly income between 4,150 and 6,225 Reais	-37.5056	0.0280
Geographic Region - NOT South	149,800.8019	0.0315
Pharmacies (1km)	-8,511.4227	0.0322
Store Status: NOT premium	-38,089.7209	0.0354
University Segment Level B	-121,549.4964	0.0403
University Segment: Level A	97,946.6523	0.0448
Intercept	155,474.449	0.0535
Super and Hyper Markets (1km)	32,749.6708	0.0758

<b>Summary of Fit</b>			
<b>RSquare</b>		<b>0.556501</b>	
<b>RSquare Adj</b>		<b>0.541142</b>	
<b>Root Mean Square Error</b>		<b>789291.5</b>	
<b>Mean of Response</b>		<b>765326.6</b>	
<b>Observations (or Sum Wgts)</b>		<b>479</b>	
<b>Crossvalidation</b>			
<b>Source</b>	<b>RSquare</b>	<b>RASE</b>	<b>Freq</b>
Training Set	0.5565	775159	479
Validation Set	0.4580	652862	203

**Table 28: Regression output for Department 27 (Home Appliances)**

Parameter	Parameter Estimate	P-value												
Average winter temperature (C°)	21,184.2794	<.0001												
Sales Area (in meter square)	522.6772	<.0001												
50 to 59 years	40.9894	<.0001												
Geographic Region - NOT South East	-81,864.0431	<.0001												
A1 Income Level for Private Permanent Homes	7.9308	<.0001												
Intercept	-306,404.0881	0.0002												
Geographic Region - NOT Central	-93,518.5913	0.0016												
35 to 49 years old	19.7891	0.0035												
Hospitals (1km)	-14,143.5426	0.0042												
Store Type - Conventional Stores	-133,600.0215	0.0044												
Days since Inauguration of store (Assume end date - 1st Jan, 2016)	13.6367	0.0067												
Super and Hyper Markets (1km)	-12,746.6109	0.0123												
Private schools (1km)	2,795.02608	0.0164												
Shopping Center (1km)	28,603.6269	0.0168												
B2 income level people	-1.1828	0.0368												
Store Status: NOT premium	-13,682.148	0.0390												
B1 income level people	-1.3153	0.0443												
Stock area (in meter square)	-109.908	0.0649												
City Population (2014)	-0.0027	0.0652												
C2 Income Level for Private Permanent Homes	2.3275	0.0652												
Gyms (1km)	-3,650.9161	0.0699												
Inactive residents with monthly income below 415 Reais	4.8948	0.0759												
PEA with monthly income below 415 Reais	0.2676	0.0787												
People who completed elementary school but not high school	-0.0083	0.0831												
Perfume and Cosmetic shops (1km)	21,412.241	0.0943												
<p><b>Summary of Fit</b></p> <p>RSquare 0.596048</p> <p>RSquare Adj 0.577018</p> <p>Root Mean Square Error 299786.1</p> <p>Mean of Response 472860.2</p> <p>Observations (or Sum Wgts) 490</p> <p><b>Crossvalidation</b></p> <table border="1"> <thead> <tr> <th>Source</th> <th>RSquare</th> <th>RASE</th> <th>Freq</th> </tr> </thead> <tbody> <tr> <td>Training Set</td> <td>0.5960</td> <td>292666</td> <td>490</td> </tr> <tr> <td>Validation Set</td> <td>0.6580</td> <td>279029</td> <td>192</td> </tr> </tbody> </table>			Source	RSquare	RASE	Freq	Training Set	0.5960	292666	490	Validation Set	0.6580	279029	192
Source	RSquare	RASE	Freq											
Training Set	0.5960	292666	490											
Validation Set	0.6580	279029	192											

Table 29: Regression output for Department 28 (Perfumes and Hygiene)

Parameter	Parameter Estimate	p-value
Geographic Region - NOT South	203,578.3595	<.0001
Geographic Region - NOT South East	178,838.3207	<.0001
Sales Area (in meter square)	738.3241	<.0001
Stock area (in meter square)	703.3334	<.0001
Residential complexes	-34.0996	0.0008
Store Type - Conventional	-218,588.3537	0.0025
Inactive residents with monthly income between 1,245 and 2,075 Reais	-30.3091	0.0081
Days since Inauguration of store (Assume end date - 1st Jan, 2016)	-55.005	0.0081
Super and Hyper Markets (1km)	-31,659.6591	0.0098
Days since last refurbishment	-12.7294	0.0108
Workers with monthly income between 4,150 and 6,225 Reais	6.2445	0.0139
Average summer temperature (C°)	-16,486.0058	0.0158
Private schools (1km)	7,373.5039	0.0167
Restaurants(1km)	-1,203.7593	0.0195
PEA Density (2007)	27.9514	0.0213
Shopping Center (1km)	61,409.7603	0.0335
Inactive residents with monthly income above 8,300 Reais	20.0362	0.0379
Intercept	262,528.72	0.0457
TGCA Average Income (%) (2010-2014)	8,300.6806	0.0598
C2 Income Level for Private Permanent Homes	4.5227	0.0638
Inactive residents with monthly income between 415 and 830 Reais	17.0603	0.0709
Houses	-0.9789	0.0715
PIB per Capita Total (2012)	-0.1089	0.0942

### Summary of Fit

RSquare	0.71605
RSquare Adj	0.702035
Root Mean Square Error	455537.9
Mean of Response	730867
Observations (or Sum Wgts)	490

### Crossvalidation

Source	RSquare	RASE	Freq
Training Set	0.7160	444242	490
Validation Set	0.5539	499610	192

**Table 30: Regression output for Department 30 (Electronics/ IT Equipment)**

Parameter	Parameter Estimate	P-value
Average winter temperature (C°)	31,553.1744	<. 0001
Sales Area (in meter square)	524.0384	<. 0001
PEA with monthly income below 415 Reais	1.9849	<. 0001
Intercept	-525,225.8993	<. 0001
60+ years old	-8.8421	0.0009
Days since Inauguration of store (Assume end date - 1st Jan, 2016)	19.7897	0.0022
Perfume and Cosmetic shops (1km)	-23,841.4034	0.0195
TGCA Average Income (%) (2010-2014)	-29,886.809	0.0248
Store Type: Conventional	-127,230.2538	0.0369
Stock area (in meter square)	157.1185	0.0431
A2 Income Level people	5.29	0.0441
E Income Level for Private Permanent Homes	15.2361	0.0446
Private schools (1km)	2,999.4452	0.0841
Shopping Center (1km)	33,418.1297	0.0866
PEA Density (2007)	-9.0972	0.0872
B2 income level people	1.5178	0.0883
Houses	1.3288	0.0899
Tourist Population(2013)	-0.0034	0.0905
Restaurants (1km)	-330.1201	0.0934
PEA with monthly income between 6,225 and 8,300 Reais	0.2845	0.0978
Residential complexes	-1.2857	0.0991

### Summary of Fit

RSquare	0.610301
RSquare Adj	0.592815
Root Mean Square Error	388695.6
Mean of Response	542648
Observations (or Sum Wgts)	490

### Crossvalidation

Source	RSquare	RASE	Freq
Training Set	0.6103	379870	490
Validation Set	0.5583	434498	192

**Table 31: Regression output for Department 31 (Hair Coloring products)**

Parameter	Parameter Estimate	p-value
Sales Area (in meter square)	208.5394	<.0001
Stock area (in meter square)	150.8407	<.0001
Inactive residents with monthly income between 415 and 830 Reais	2.3743	<.0001
PEA Density (2007)	11.6044	0.0004
Geographic Region - Not North East	-30,722.0105	0.0007
Days since Inauguration of store (Assume end date - 1st Jan, 2016)	6.0514	0.0017
Geographic Region - Not Central	-30,267.8506	0.0054
Perfume and Cosmetic shops (1km)	11,243.2343	0.0071
Área of the city (in Km2)	3.9519	0.0081
Private schools (1km)	1,667.9529	0.0304
Super and Hyper Markets (1km)	-6,601.4243	0.0317
Residential complexes	-5.7538	0.0319
PIB per Capita Total (2012)	-0.8271	0.0399
Store Type: Conventional	-36,818.8798	0.0411
Houses	-0.9008	0.0435
Workers with monthly income between 4,150 and 6,225 Reais	-1.1676	0.0456
B2 Income Level for Private Permanent Homes	-1.5095	0.0461
Intercept	111,967.2921	0.0598
Restaurants(1km)	-280.2088	0.0611
Average summer temperature (C°)	-3,255.4828	0.0621
A1 Income Level for Private Permanent Homes	1.1698	0.0632
E Income Level for Private Permanent Homes	1.8515	0.0911
Shopping Center (1km)	-1,742.5475	0.0923

### Summary of Fit

<b>RSquare</b>	0.748609
<b>RSquare Adj</b>	0.735634
<b>Root Mean Square Error</b>	114970.2
<b>Mean of Response</b>	192590.8
<b>Observations (or Sum Wgts)</b>	490

### Crossvalidation

Source	RSquare	RASE	Freq
Training Set	0.7486	111999	490
Validation Set	0.6903	129286	192

**Table 32: Regression output for Department 32 (Cosmetics)**

Parameter	Parameter Estimate	P-value
Sales Area (in meter square)	210.1076	<. 0001
Stock area (in meter square)	146.7229	<. 0001
PEA Density (2007)	11.7951	<. 0001
Geographic Region – Not North East	-30,669.7446	0.0003
Days since Inauguration of store (Assume end date - 1st Jan, 2016)	6.2048	0.0013
Perfume and Cosmetic shops (1km)	8,064.8461	0.0055
Geographic Region – Not Central	-29,024.9656	0.0075
Área of city (Km <sup>2</sup> )	3.9597	0.0077
Store Type: Conventional	-40,718.8965	0.0217
Super and Hyper Markets (1km)	-6,848.2271	0.0243
Private schools (1km)	1,852.8827	0.0264
60+ years old	-1.0519	0.0447
Residential complexes	-4.6862	0.0777
PIB per Capita Total (2012)	-0.6836	0.0812
Pharmacies (1km)	1,808.9362	0.0852
Banks (1km)	-463.7231	0.0865
E Income Level for Private Permanent Homes	1.7569	0.0867
High school graduates	-0.0129	0.0873
Average summer temperature (C°)	327.9533	0.09023
Restaurants(1km)	-398.9195	0.0911
A1 Income Level for Private Permanent Homes	2.4846	0.0914
Shopping Center (1km)	-604.0142	0.09401
Intercept	-2,716.7052	0.09735

### Summary of Fit

RSquare	0.749175
RSquare Adj	0.737359
Root Mean Square Error	114594.4
Mean of Response	192590.8
Observations (or Sum Wgts)	490

### Crossvalidation

Source	RSquare	RASE	Freq
Training Set	0.7492	111873	490
Validation Set	0.7004	127154	192



**Table 233: Regression output for Department 33 (Sports and Beach Equipment)**

Parameter	Parameter Estimate	P-value
Average winter temperature (C°)	10,977.5081	<. 0001
Sales Area (in meter square)	170.3580	<. 0001
A1 Income Level for Private Permanent Homes	44.9970	<. 0001
Intercept	-13,9277.3333	<. 0001
PEA with monthly income above 8,300 Reais	10.7304	0.0002
Geographic Region – Not South East	-25,229.7102	0.0003
Days since Inauguration of store (Assume end date - 1st Jan, 2016)	6.1436	0.0004
TGCA Population (%) (2010-2014)	-25,732.9323	0.0008
Gyms (1km)	-7,249.7766	0.0215
A2 Income Level people	-0.9251	0.0361
Store Status: NOT Premium	-9,251.9261	0.0382
TGCA Average Income (%) (2010-2014)	3,223.0799	0.0399
PEA Density (2007)	-2.2126	0.0438
Private schools (1km)	418.8781	0.0523
Shopping Center (1km)	2,818.2611	0.0665
Households in the city (2014)	0.0021	0.0746
PEA with monthly income below 415 Reais	0.0892	0.0750
Residential complexes	-0.6307	0.0789
Geographic Region - NOT Central	-784.5882	0.0940
0 to 9 years old	0.0046	0.0994

### Summary of Fit

RSquare	0.652379
RSquare Adj	0.639859
Root Mean Square Error	107303.5
Mean of Response	164076.6
Observations (or Sum Wgts)	490

### Crossvalidation

Source	RSquare	RASE	Freq
Training Set	0.6524	105314	490
Validation Set	0.6831	107674	192

**Table 34: Regression output for Department 34 (Imported Toys)**

Parameter	Parameter Estimate	P-value
Sales Area (in meter square)	224.1524	<. 0001
A1 Income Level for Private Permanent Homes	34.1903	<. 0001
Women	11.8943	<. 0001
Store Location - Street	-31,481.8873	<. 0001
Days since Inauguration of store (Assume end date - 1st Jan, 2016)	6.6829	0.0002
Intercept	95,906.6721	0.0004
50 to 59 years	10.6383	0.0004
Perfume and Cosmetic shops (1km)	7,943.5882	0.0006
25 to 34 years old	2.1088	0.0010
Banks (1km)	-893.5599	0.0014
Houses	-1.0119	0.0063
Store Type: Conventional	-44,326.4689	0.0068
10 to 14 years old	8.5943	0.0082
Stock area (in meter square)	-32.2539	0.0123
PEA Density (2007)	4.4364	0.0138
MonthlyAverage income(2014)	-6.3222	0.0142
Store Status - Not Premium	-7,493.0781	0.0174
Geographic Region - Not Central	-22,933.3488	0.0176
Inactive residents with monthly income between 1,245 and 2,075 Reais	-8.4912	0.0450
Inactive residents with monthly income below 415 Reais	1.2854	0.0597
Inactive residents with monthly income between 415 and 830 Reais	8.0461	0.0857

**Summary of Fit**

<b>RSquare</b>	0.727777
<b>RSquare Adj</b>	0.717972
<b>Root Mean Square Error</b>	104259.9
<b>Mean of Response</b>	255619.7
<b>Observations (or Sum Wgts)</b>	490

**Crossvalidation**

Source	RSquare	RASE	Freq
Training Set	0.7278	102327	490
Validation Set	0.7529	96540	192

**Table 35: Regression output for Department 40 (Candies)**

Parameter	Parameter Estimate	p-value
Sales Area (in meter square)	281.1672	<. 0001
A2 Income Level people	21.9102	<. 0001
Average income2014	8.2796	<. 0001
20 to 24 years old	-9.5245	<. 0001
Stock area (in meter square)	-131.8704	<. 0001
Store Location - Street	-80,252.9085	<. 0001
Store Type: Conventional	-75,429.6822	0.0008
PEA with monthly income between 6,225 and 8,300 Reais	54.4179	0.0009
Store Status: Not Premium	-23,932.8257	0.0013
Geographic Region - NOT South	33,691.7165	0.0029
Population Density of city (2014)	9.2132	0.0031
Average summer temperature (C°)	-8,974.2249	0.0046
Residential complexes	7.7571	0.0134
Intercept	211,867.7505	0.018
Hotel (1km)	2,357.6013	0.0491
Perfume and Cosmetic shops (1km)	4,544.1103	0.0799
Geographic Region - Not Central	-21,072.4916	0.0821
Days since last refurbishment	-3.7356	0.0841
Inactive residents with monthly income between 415 and 830 Reais	8.9850	0.0854
Days since Inauguration of store (Assume end date - 1st Jan, 2016)	4.1106	0.0873
E income level people	1.4012	0.0931
Universities (1km)	-630.4229	0.0942
City Population (2014)	0.0035	0.0957
TGCA Average Income (%) (2010-2014)	2,335.1285	0.0987

### Summary of Fit

RSquare	0.679532
RSquare Adj	0.664435
Root Mean Square Error	140058.4
Mean of Response	253864.7
Observations (or Sum Wgts)	490

### Crossvalidation

Source	RSquare	RASE	Freq
Training Set	0.6795	136732	490
Validation Set	0.6359	137015	192

**Table 36: Regression output for Department 43 (Easter)**

Parameter	Parameter Estimate	P-value
Geographic Region - NOT South	74,024.7167	<. 0001
Sales Area (in meter square)	523.8467	<. 0001
Days since Inauguration of store (Assume end date - 1st Jan, 2016)	21.9875	<. 0001
Stock area (in meter square)	-204.1683	<. 0001
E income level people	28.5056	<. 0001
Inactive residents with monthly income between 830 and 1245	17.7363	<. 0001
Tourist Population(2013)	54.0121	0.0003
Demographic Density (2014)	16.9080	0.0004
Average summer temperature (C°)	-16,256.7137	0.0010
Store Status: Not premium	-31,739.4479	0.0042
High school graduates	-0.0499	0.0048
Perfume and Cosmetic shops (1km)	12,128.3528	0.0178
Geographic Region - Not Central	-41,230.4349	0.0347
A1 Income Level for Private Permanent Homes	-4.8093	0.0675
Average income2014	7.0055	0.0689
B1 income level people	-3.1877	0.0780
Intercept	246,203.2881	0.0850
Shopping Center (1km)	20,477.3656	0.0860
C2 Income Level for Private Permanent Homes	-5.5479	0.0865
0 to 9 years old	2.1151	0.0911
TGCA Average Income (%) (2010- 2014)	-5,476.7793	0.0921
Gyms (1km)	2,332.8585	0.0982

**Summary of Fit**

<b>RSquare</b>	0.776024
<b>RSquare Adj</b>	0.765974
<b>Root Mean Square Error</b>	209184
<b>Mean of Response</b>	483965.8
<b>Observations (or Sum Wgts)</b>	490

**Crossvalidation**

Source	RSquare	RASE	Freq
Training Set	0.7760	204434	490
Validation Set	0.7823	234131	192

**Table 37: Regression output for Department 44 (Plastic Utilities)**

Parameter	Parameter Estimate	P-value
Sales Area (in meter square)	125.1968	<. 0001
D Income Level for Private Permanent Homes	9.7902	<. 0001
0 to 9 years old	-2.4503	0.0098
Banks (1km)	-860.1503	0.0128
Store Location - Street	-9,163.8334	0.0261
Average winter temperature (C°)	3,043.9285	0.0352
PEA with monthly incomes between 415 and 830 Reais	0.1668	0.0430
Days since Inauguration of store (Assume end date - 1st Jan, 2016)	2.1809	0.0460
Pharmacies (1km)	1,261.7757	0.0551
Geographic Region – NOT Nordeste	-7,858.4741	0.0556
Store Status – Not premium	-5,680.5279	0.0561
Average summer temperature (C°)	-3,000.4850	0.0566
High school graduates	0.0052	0.0622
Private schools (1km)	476.3683	0.0672
Super and Hyper Markets (1km)	-873.8389	0.0833
City Population (2014)	0.0011	0.0844
B2 income level people	-0.2021	0.0874
Intercept	29,424.4429	0.0899
PEA with monthly income above 8,300 Reais	-0.7359	0.0912
Universities (1km)	-68.0301	0.0951

<b>Summary of Fit</b>			
RSquare		0.648732	
RSquare Adj		0.632184	
Root Mean Square Error		75719.5	
Mean of Response		113194.9	
Observations (or Sum Wgts)		490	
<b>Crossvalidation</b>			
Source	RSquare	RASE	Freq
Training Set	0.6487	73921	490
Validation Set	0.7784	52300	192

**Table 38: Regression output for Department 45 (Videogames)**

Parameter	Parameter Estimate	P-value
Sales Area (in meter square)	575.1845	<. 0001
Stock area (in meter square)	-266.8757	<. 0001
Location - Not Rua	-10,0783.5329	<. 0001
PEA with monthly income between 830 and 1,245 Reais	-9.2252	<. 0001
E income level people	-60.0512	<. 0001
Pharmacies (1km)	-13,045.3273	0.0007
Store Type: Conventional	-162,977.7147	0.0018
Tourist Population of city (2013)	0.0034	0.0023
35 to 49 years old	-2.7007	0.0044
Intercept	-117,661.9840	0.0117
Average income2014	17.3052	0.0174
Perfume and Cosmetic shops (1km)	13,568.7881	0.0242
A1 Income Level for Private Permanent Homes	72.3333	0.0321
Área of city (in Km <sup>2</sup> )	2.9717	0.0399
60+ years old	0.1950	0.0447
A2 Income Level people	37.9456	0.0456
Geographic Region - Not North East	-17,196.5903	0.0482
Women	29.5804	0.0485
Days since Inauguration of store (Assume end date - 1st Jan, 2016)	11.1409	0.0485
Store Status - Not Premium	-33,865.6598	0.0517
25 to 34 years old	5.2245	0.0911
Geographic Region - Not Central	-48,706.8272	0.0924

### Summary of Fit

RSquare	0.581627
RSquare Adj	0.564714
Root Mean Square Error	330133.8
Mean of Response	430938
Observations (or Sum Wgts)	490

### Crossvalidation

Source	RSquare	RASE	Freq
Training Set	0.5816	323326	490
Validation Set	0.5311	353824	192

**Table 39: Regression output for Department 46 (Chips and Cereal Bars)**

Parameter	Parameter Estimate	P-value
Geographic Region - NOT South	21,849.8419	<. 0001
Sales Area (in meter square)	102.6593	<. 0001
A1 Income Level for Private Permanent Homes	30.7518	<. 0001
Days since Inauguration of store (Assume end date - 1st Jan, 2016)	6.7237	<. 0001
Average income2014	5.1969	<. 0001
Fuel station (1km)	-4,611.2497	<. 0001
B1 Income Level for Private Permanent Homes	29.5391	<. 0001
E Income Level for Private Permanent Homes	30.5189	<. 0001
Stock area (in meter square)	-43.1393	0.0006
Geographic Region - NOT North East	17,610.2707	0.0007
Private schools (1km)	1,215.7295	0.0066
Store Type : Conventional	-24,138.4697	0.0130
Days since last refurbishment	-2.6559	0.0147
0 to 9 years old	-1.7251	0.0207
Inactive residents with monthly income between 4,150 and 6,225 Reais	3.4028	0.0287
Perfume and Cosmetic shops (1km)	2,384.1460	0.0422
Tourist Population(2013)	0.0019	0.0623
High school graduates	-0.0094	0.0766
PEA with monthly income above 8,300 Reais	-0.6327	0.0787
PEA with monthly income below 415 Reais	0.2490	0.0796
Super and Hyper Markets (1km)	-1,802.1695	0.0812
Average summer temperature (C°)	-1,732.0311	0.0824
Gyms (1km)	1,909.6670	0.0865
A1 Income Level for Private Permanent Homes	-0.5893	0.0890
Intercept	14,278.3691	0.0895
Residential complexes	45.3651	0.0904
B2 income level people	0.0464	0.0905

**Summary of Fit**

RSquare	0.632961
RSquare Adj	0.614017
Root Mean Square Error	62870.99
Mean of Response	127074.5
Observations (or Sum Wgts)	490

**Crossvalidation**

Source	RSquare	RASE	Freq
Training Set	0.6330	61246	490
Validation Set	0.7744	51599	192

**Table 40: Regression output for Department 48 (Beverages)**

Parameter	Parameter Estimate	P-value
Hotel (1km)	2,720.3100	<. 0001
Tourist Population(2013)	250.3453	<. 0001
Sales Area (in meter square)	150.7268	<. 0001
Average income2014	6.9883	<. 0001
Store Location - Street	-36,200.5190	<. 0001
Store Status: Not Premium	-17,767.7294	0.0002
Stock area (in meter square)	-65.7339	0.0003
Geographic Region – Not South East	-18,466.6937	0.0009
A2 Income Level people	-1.8590	0.0028
Average summer temperature (C°)	-6,176.9498	0.0046
Store Type: Conventional	-37,114.3177	0.0089
Intercept	153,498.9604	0.0122
Universities (1km)	-2,095.9129	0.0288
TGCA Average Income (%) (2010-2014)	-6,720.1738	0.0293
People who completed elementary school but not high school	-0.0241	0.0296
Inactive residents with monthly income between 830 and 1245	4.1841	0.0424
Perfume and Cosmetic shops (1km)	2,785.6053	0.0855
Geographic Region – Not Central	-14,787.4316	0.0860
35 to 49 years old	-0.6730	0.0881
B2 Income Level for Private Permanent Homes	1.9552	0.0891
B1 income level people	0.0693	0.0906
Population Density of city (2014)	-0.7045	0.0911

**Summary of Fit**

<b>RSquare</b>	0.620018
<b>RSquare Adj</b>	0.602968
<b>Root Mean Square Error</b>	89787.22
<b>Mean of Response</b>	141481
<b>Observations (or Sum Wgts)</b>	490

**Crossvalidation**

Source	RSquare	RASE	Freq
Training Set	0.6200	87748	490
Validation Set	0.5365	77544	192



**Table 41: Regression output for Department 49 (Biscuits and Cookies)**

Parameter	Parameter Estimate	P-value
Sales Area (in meter square)	487.4456	<. 0001
Average income (2014)	20.0547	<. 0001
Store Location - Street	-84,905.2142	<. 0001
A1 Income Level for Private Permanent Homes	65.8218	<. 0001
PEA with monthly income between 6,225 and 8,300 Reais	98.4284	<. 0001
Super and Hyper Markets (1km)	-9,937.9829	<. 0001
Store Type: Conventional	-144,500.6574	0.0001
Store Status: NOT premium	-38,391.2757	0.0019
Stock area (in meter square)	-125.2938	0.0091
Pharmacies (1km)	6,591.6759	0.0093
Days since last refurbishment	-9.1981	0.0219
Days since Inauguration of store (Assume end date - 1st Jan, 2016)	10.0822	0.0243
Geographic Region - Not North East	51,148.3960	0.0268
B1 income level people	2.9764	0.0362
E income level people	4.8323	0.0376
PEA with monthly income between 830 and 1,245 Reais	-3.1144	0.0396
A2 Income Level people	-3.5156	0.0407
Geographic Region - Not South	49,364.9552	0.0545
People who completed elementary school but not high school	-0.0533	0.0732
TGCA Renda Média (%) (2010-2014)	9,573.6851	0.0811
Intercept	-163,575.1047	0.0821
25 to 34 years old	-1.1546	0.0851
Nominal Income of city (2014)	0.0000	0.0911
Geographic Region – Not South East	-5,284.9356	0.0923
Average summer temperature (C°)	-1,287.7148	0.0945

### Summary of Fit

RSquare	0.667699
RSquare Adj	0.652788
Root Mean Square Error	233164.4
Mean of Response	445421.7
Observations (or Sum Wgts)	490

### Crossvalidation

Source	RSquare	RASE	Freq
Training Set	0.6677	227870	490
Validation Set	0.7425	206465	192

**Table 42: Regression output for Department 51 (Diapers)**

Parameter	Parameter Estimate	P-value
Geographic Region – Not South East	109,984.1500	<. 0001
Sales Area (in meter square)	402.1016	<. 0001
Stock area (in meter square)	361.0904	<. 0001
Store Type: Conventional	-22,0547.0051	<. 0001
Days since last refurbishment	-250.9422	<. 0001
Perfume and Cosmetic shops (1km)	-21,239.8208	<. 0001
Hospitals (1km)	-16,356.3779	0.0003
Intercept	-354,046.4841	0.0003
Geographic Region – Not South	132,876.3334	0.0004
Average winter temperature (C°)	16,422.6588	0.0018
Private schools (1km)	5,637.7417	0.0040
C2 Income Level for Private Permanent Homes	9.2643	0.0118
Days since Inauguration of store (Assume end date - 1st Jan, 2016)	12.4629	0.0178
A1 Income Level for Private Permanent Homes	7.5186	0.0197
Geographic Region – Not Central	82,269.4470	0.0216
Super and Hyper Markets (1km)	-16,852.6107	0.0255
Bars (1km)	-3,235.1571	0.0552
Superior completo	350.8472	0.0734
Residential complexes	-9.3869	0.0745
B2 income level people	-1.6171	0.0756
PEA with monthly incomes between 415 and 830 Reais	0.4224	0.0811
Shopping Center (1km)	24,797.5444	0.0845
MonthlyAverage income(2014)	-12.6525	0.0867
PEA Density (2007)	7.8668	0.0898
B1 income level people	-1.2380	0.0911
Workers with monthly income between 4,150 and 6,225 Reais	1.5339	0.0923
TGCA Average Income (%) (2010-2014)	-2,514.0226	0.0934

**Summary of Fit**

RSquare	0.683069
RSquare Adj	0.664548
Root Mean Square Error	276329.3
Mean of Response	352141
Observations (or Sum Wgts)	490

**Crossvalidation**

Source	RSquare	RASE	Freq
Training Set	0.6831	268318	490
Validation Set	0.7104	215113	192

**Table 43: Regression output for Department 56 (Christmas Event)**

Parameter	Parameter Estimate	P-value
Perfume and Cosmetic shops (1km)	7,243.4298	<. 0001
Sales Area (in meter square)	164.3023	<. 0001
50 to 59 years	9.8292	<. 0001
Average income2014	6.8826	<. 0001
PEA with monthly income above 8,300 Reais	5.8106	<. 0001
Stock area (in meter square)	-67.6175	<. 0001
Fuel station(1km)	-6,035.8960	<. 0001
Intercept	-73,470.1461	<. 0001
Days since Inauguration of store (Assume end date - 1st Jan, 2016)	5.2757	0.0010
Banks (1km)	-884.9755	0.0024
Geographic Region - Not South	22,581.0127	0.0082
Days since last refurbishment	-3.7979	0.0082
Store Location - Street	-11,986.6926	0.0134
Store Status: Not Premium	-7,686.5505	0.0803
Tourist Population(2013)	0.0017	0.0812
Gyms (1km)	-3,972.1811	0.0824
Inactive residents with monthly income between 2,075 and 4,150 Reais	1.9296	0.0832
B2 income level people	-0.5468	0.0848
25 to 34 years old	0.8155	0.0854
People who completed elementary school but not high school	-0.0097	0.0867
Geographic Region - Not North East	4,579.1046	0.0870
E income level people	-0.2915	0.0890
Geographic Region - Not South East	-1,495.8473	0.0921

**Summary of Fit**

<b>RSquare</b>	0.683278
<b>RSquare Adj</b>	0.669066
<b>Root Mean Square Error</b>	83262.66
<b>Mean of Response</b>	130796.8
<b>Observations (or Sum Wgts)</b>	490

**Crossvalidation**

Source	RSquare	RASE	Freq
Training Set	0.6833	81372	490
Validation Set	0.6685	88430	192

**Table 244: Regression output for Department 63 (Telephones)**

Parameter	Parameter Estimate	P-value
Sales Area (in meter square)	1,506.7605	<. 0001
E Income Level for Private Permanent Homes	47.1779	<. 0001
PEA with monthly income below 415 Reais	14.0000	<. 0001
MonthlyAverage income(2014)	-231.7489	<. 0001
Store Status: NOT premium	-120,331.4422	<. 0001
TGCA Average Income (%) (2010-2014)	-176,593.6173	<. 0001
Days since Inauguration of store (Assume end date - 1st Jan, 2016)	66.7772	0.0002
Perfume and Cosmetic shops (1km)	-89,728.0168	0.0009
Geographic Region - NOT South	304,935.7328	0.0017
Intercept	862,746.8902	0.0048
60+ years old	-17.5886	0.0237
Store Type: Conventional	-417,482.3758	0.0319
Households in the city (2014)	0.1578	0.0423
PEA Density (2007)	71.5747	0.0542
Store Location - Street	-135,913.4573	0.0543
Bars (1km)	-12,206.5972	0.0639
Stock area (in meter square)	320.5269	0.0734
Houses	-7.9537	0.0867
B2 income level people	3.6641	0.0898
A1 Income Level for Private Permanent Homes	3.9346	0.0911
	0.0107	0.0923

**Summary of Fit**

RSquare	0.608824
RSquare Adj	0.591309
Root Mean Square Error	1242886
Mean of Response	1655872
Observations (or Sum Wgts)	491

**Crossvalidation**

Source	RSquare	RASE	Freq
Training Set	0.6088	1214722.2	491
Validation Set	0.4695	1181078.8	191

Table 45: Regression output for Department 72 (Movies)

Parameter	Parameter Estimate	P-value
Sales area ( in meter square)	133.7668	<. 0001
Workers with monthly income between 6,225 and 8,300 Reais	122.5099	<. 0001
Women	106.8124	<. 0001
50 to 59 years	100.2773	<. 0001
Days since Inauguration of store (Assume end date - 1st Jan, 2016)	4.3682	<. 0001
Store Status: NOT Premium	-9,532.3566	<. 0001
Store Location - Street	-35,597.5360	<. 0001
Stock area (in meter square)	-51.7200	0.0002
PEA with monthly income below 415 Reais	0.7194	0.0016
Banks (1km)	-689.6567	0.0018
Store Type: Conventional	-30,654.9200	0.0081
25 to 34 years old	124.1632	0.0403
Geographic Region - NOT South	11,737.2218	0.0503
Intercept	27,118.2042	0.0518
60+ years old	0.5094	0.0632
High school graduates	0.0083	0.0665
Geographic Region -NOT Central	6,104.7463	0.0687
TGCA Average Income (%) (2010-2014)	-1,691.5334	0.0695
PEA with monthly income above 8,300 Reais	0.3253	0.0712
PEA de 1245,00 a 2075,00	0.1357	0.075
Tourist Population(2013)	-0.0002	0.0812
Geographic Region - NOT North East	812.8542	0.0825
Houses	-0.0358	0.0845

### Summary of Fit

RSquare	0.700907
RSquare Adj	0.68818
Root Mean Square Error	73847.91
Mean of Response	163738.5
Observations (or Sum Wgts)	491

### Crossvalidation

Source	RSquare	RASE	Freq
Training Set	0.7009	72251	491
Validation Set	0.6576	65765	191

## **Appendix B**

In this section, key points of analysis are presented for departments that did not yield results as noteworthy as the departments discussed in the results section (Section 4) of this thesis.

### **Department 3 - Clothes for children**

- For this department, sales are higher when there are more number of young children (0 to 9 years) in the area around the store.
- Sales are higher for stores located in cities that have more higher income people in the area near the store (A1 and A2 Income Level for Private Permanent Homes, PEA with monthly income above 8,300 Reais). Customers with higher incomes may be willing to spend more money on clothes for young children.
- We see that the 'Days since inauguration' parameter is negative. This could imply that parents prefer to buy clothes for their children at newer stores.
- Stores surrounded by more private schools tend to have greater sales.

### **Department 4 - Lingerie**

- Sales are higher when there are more perfume shops and pharmacies located close to the store. One reason for this could be that women prefer to buy lingerie in places where perfumes shops are located close by so that they can shop for both items in one trip.
- Sales are higher when there are more higher income people in the area near the store (PEA with monthly income above 8,300 Reais, A1 Income Level for Private Permanent Homes, PEA with monthly income between 6,225 Reais and 8,300 Reais). People with higher incomes may have a greater willingness to spend on lingerie.

- The presence of nearby super/hypermarkets and rival stores in nearby shopping centers has a statistically significant negative influence on sales.

#### **Department 13 – Bed and Bath**

- Sales are higher when the average summer temperature is higher. The reason for this phenomenon is not clear but LA's executives mentioned that this is an interesting phenomenon to investigate.
- Sales are higher when pharmacies (where one could buy items such as soaps, toothpaste etc.) are present nearby. One explanation for this could be that when people go to pharmacies, they also prefer to do their shopping for bed and bath products at LA's stores.

#### **Department 16 – Stationary items**

- Sales are higher when there are more people in the age group 15-19 near the store. This indicates that students in high school and college drive up sales of stationary items.
- Sales are higher when there are more high-income individuals in the area around the store (PEA with monthly income above 8,300 Reais).

#### **Department 19 - Cleaning Products**

- Sales are higher when there are more super markets near the store.
- Sales increase when the average winter temperature is higher.
- Sales are higher when there are more people in the lower income categories around the store (D Income level private permanent homes and and PEA with monthly income below 415 Reais).

### **Department 21 - Kitchen Utensils**

- Sales are higher when there are more people in the age group 25 to 34 in the area surrounding the store.
- Sales tend to be higher when there are more super and hypermarkets nearby. LA's executives mentioned that this may be due to a complementary effect - when people buy food items in supermarkets, they may also tend to buy kitchen utensils at nearby LA stores. Nevertheless, more research must be done to see if the complementary effect does in fact exist.

### **Department 22 – Table and bar**

- Sales are higher when there are more higher income people (A1 Income level for private permanent and PEA with monthly income between 6,225 and 8,300 Reais) in the area surrounding the store.
- Sales are higher for stores near shopping malls. This could indicate that shopping malls generate a lot of foot traffic that drives people to LA's stores as well.

### **Department 23 – Toys**

- Sales area and assortment size have a strong positive influence on sales. This could indicate that sales tend to increase when more sales space is dedicated (i.e. there are more toy options available to consumers).
- Sales are higher when there are more high income individual present near the store (PEA with monthly income above 8,300 Reais and A1 Income level for Private Permanent Homes).
- Sales are lower for stores not present in the North East and South East region.



### **Department 27 – Home appliances**

- Sales are higher when there are more people in the age group 35 to 49 and 50 to 59 indicating that this could be the major demographic that buys home appliances.
- Sales are higher for stores located close to shopping centers indicating that shopping centers may generate a lot of overall foot traffic that drives customers to LA's stores as well. Sales are lower when more competitors - super and hypermarkets - are present close to the store.
- Sales are higher in areas surrounded by a large number of higher income households (A1 Income Level for Private Permanent Homes). Sales do not increase much when there are more number of people in the lower income group (PEA with monthly income below 415 Reais).

### **Department 28 – Perfumes and Hygiene**

- Sales are lower when there are more super or hypermarkets nearby indicating that LA is losing customers when competitors are present nearby. Stores that are newer and refurbished more recently tend to have higher sales.
- Sales are higher when the store has more backroom stock area. This could indicate that these stores are able to better handle demand volatilities since they have more safety stock.

### **Department 30 – Electronic/ IT Equipment (Computers, Tablets, Mouse, USB Sticks)**

- Sales are higher for stores located close to more shopping malls. One reason for this could be that shopping centers increase overall foot traffic, which subsequently increases sales for LA's stores.

- Sales for this department are positively influenced by a larger stock area, indicating that demand may be volatile and stores that have more safety stock are able to better minimize the negative effects of stockouts.
- Sales are higher when there are more people in the lower income category present near the store (PEA with monthly income below 415 Reais and E Income Level for Private Permanent Homes).

### **Department 32 – Cosmetics**

- Surprisingly, sales here increase when there are more cosmetic/perfume shops and pharmacies nearby. It is surprising because one may have expected that sales would decrease with the presence of competitors nearby. Furthermore, sales decrease when super and hypermarkets are nearby. This indicates that the real competitors are super and hypermarkets, and not other cosmetic shops and pharmacies.
- Sales increase here as the average summer temperature increases. This could be because as it gets warmer, people may buy more products such as skin lotion.
- Sales are higher when the store has more backroom stock area. This could indicate that these stores are able to better handle demand volatilities since they have more safety stock.

### **Department 34 – Imported Toys**

- Sales are much higher when there are more high income households (A1 Income Level for Private Permanent Homes) in the area near the store. This could be because imported toys are more expensive.

- Sales are higher when there are more children in the age group 10 to 14 years in the area around the store.
- Sales are higher when there are more people in the age group 50 to 59 in the area near the store. According to LA's executives, this is an interesting phenomenon that needs to be investigated further.

#### **Department 44 – Plastic Utilities**

- Sales are lower when more super and hypermarkets are present around the store.
- Sales are higher when there are more lower income people in the area near the store (D Income level for Private Permanent Homes and PEA with monthly incomes between 415 and 830 Reais).
- Sales are positively correlated to higher average winter temperatures and negatively correlated to higher average summer temperatures.

#### **Department 45 – Videogames**

- Sales are lower when there are more people in the lower income category (PEA with monthly income between 830 and 1,245 Reais and E income level people) present near the store. This could be because people with lower incomes do not have a high willingness to spend on luxuries such as videogames.
- Sales are lower for stores not located in the Central or North East region.
- Sales are higher when there are more people in the age group 25 to 34 present in the area near the store. LA's executives mentioned that this is quite surprising since they expected the target demographic for videogame sales to be students in schools and colleges.

#### **Department 49 – Biscuits and Cookies**

- Sales are lower when there are more super and hypermarkets near the store.
- Sales increase significantly when the TGCA Average Income (geometric average of the annual income growth) of the city the store is located in increases.
- Sales are higher when there are more high-income people and households in the area around the store (A1 income level for private permanent homes and PEA with monthly income between 6,225 and 8,300 Reais).

#### **Department 56 – Christmas Event**

- Sales are higher when there are more people in the age group 50 to 59 years in the area near the store. Based on practical insights, LA's executives mentioned that people from this age group generally spend more money on Christmas decorations.
- Sales are higher for stores located in cities with higher average incomes. Furthermore, sales are higher when there are more high-income individuals (PEA with monthly income above 8,300 Reais) present in the area near the store.