Implementation of a predictive model for apartment prices in Porto Alegre using machine learning

Max Cohen

2019

TABLE OF CONTENTS

List of Figures

List of Tables

1. Introduction

The Brazilian economy has been in crisis for some years and continues to show no signs of improvement in the short term. In the period between 2012 and 2013 property prices reached the highest historical high. At that moment the market had a strong demand. Today the scenario is quite different.

After the boom, prices began to fall steadily and, so far, uninterrupted. Faced with the deterioration of the economy, many Brazilians began to sell their real estate, or to pay off debts or to leave the country. Day after day real estate prices were not only different but also smaller. I found myself in this situation when I began to offer the sale of my apartment and I had to reduce the price systematically. However, at that moment, a question remained: "how much is my apartment worth at the current market moment?"

In the academic literature, as well as in several technical papers, it is possible to find many articles, theses and reports on the application of machine learning to predict real prices (BERNÁDEZ, 2018; KOMAGOME-TOWNE, 2016; NGUYEN, 2018; PARK & BAE, 2018, PIERRE, 2018). It is also possible to find work using neural networks (deep learning) (YU, 2018). The greatest benefit in using machine learning to solve this type of problem is the possibility of using several algorithms and choosing the one with the best performance. In other words, it is possible to have an optimized algorithm for the proposed problem.

In view of the worsening of the Brazilian economy and the constant fall in prices of used apartments, it is difficult to identify, in a scientific and quick way, what is the fair price for an apartment. The target of this study was the city of Porto Alegre (RS). Thus, the main problem of this research was: What is the selling price of a used apartment in the city of Porto Alegre?

To address the work, the following secondary questions were listed:

1) What is the best algorithm for the sample studied?
2) What attributes have the most influence on the sale price?

The "fair sale price" stated here refers to the price practiced at a given moment and compatible with the reality of the market. The "timing" for this research was determined with a specific day. In turn, the price of an apartment can be estimated from the characteristics of the property, such as size, number of rooms, parking spaces, etc. It is understood, therefore, that this is a supervised problem, a regression type, whose "selling price" will be the dependent variable, with its value expressed in monetary units.

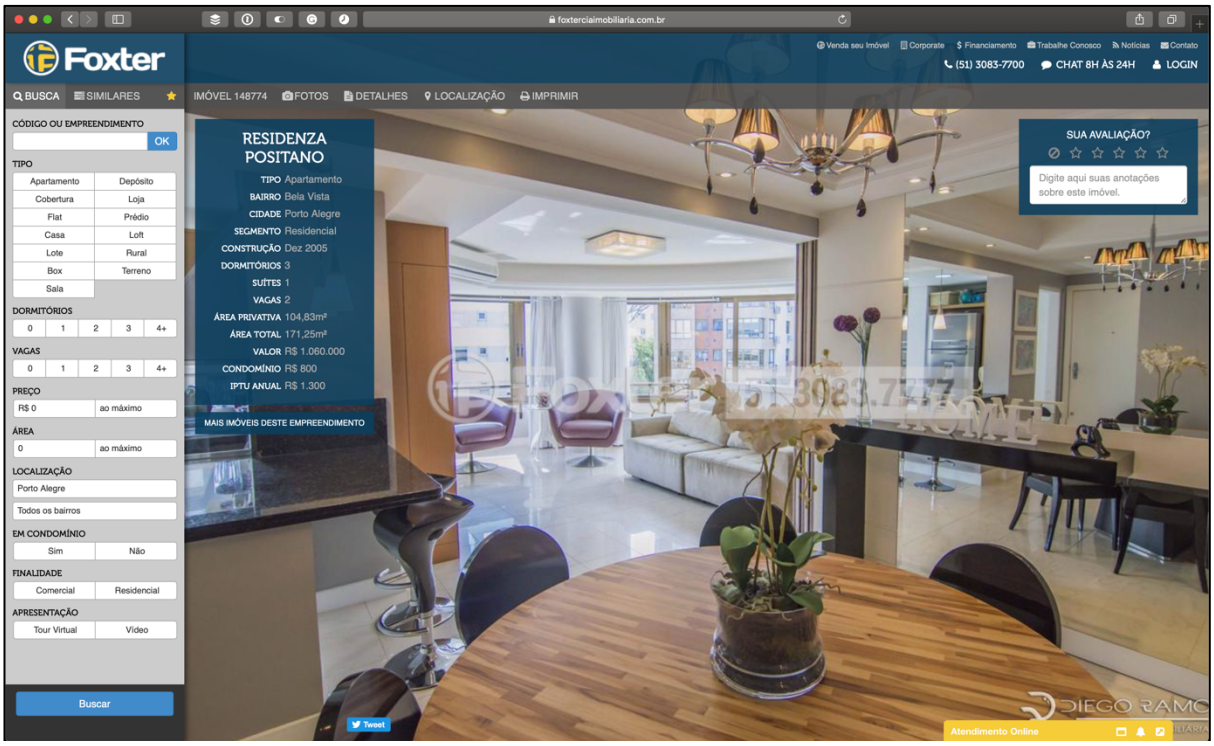
2. Methodology

The project was developed in three stages. In the first step was built a web scraper that captured the data available on Foxter's website – a real estate agency, and stored in a csv file. The second step consisted in reading the csv, cleaning and data formatting. The third step was the exploratory data analysis. And finally, the fourth step consisted in the creation of the model and evaluation of the algorithms.

2.1. The inputs and data sets

In recent years, small and large real estate agents in Porto Alegre have been active in the sale of new and used real estate, and their portfolios have been made publicly available on the Internet. Given this, the data set worked on this project was composed by the data available on the Foxter's website.

Figure 1 – Screen with an apartment description from Foxter[1]



A web scraper was created specifically with the aim of extracting the descriptive data of the apartments available and populated in a dataset. The apartment data was available in a blue background rectangle on the real estate agency query screen, as in the example above. Thus, the data collected were:

[1] https://www.foxterciaimobiliaria.com.br/imovel/148774/residencial-porto-alegre-bela-vista-apartamento-residenza-positano-3-dormitorios-zona-norte

1. Identification:
   1.1. real state code;
   1.2. url;

2. Target:
   2.1. price;

3. Attributes:
   3.1. area;
   3.2. neighborhood;
   3.3. city;
   3.4. type;
   3.5. segment;
   3.6. condominium’s price;
   3.7. tax;
   3.8. number of rooms;
   3.9. number of parking spaces.

The amount of real estate offered by the site should vary daily because of sales and the insertion of new offers. The data used in this study were collected by the web scraper on 01/20/2019. The study was limited to data from the Auxiliadora, Bela Vista and Mont'Serrat neighborhoods, as they are neighbors and maintain similarities. In this way, the web scraper application has been adjusted to include this delimitation. Data were collected from 595 properties.

2.2. Data Preprocessing

Data were preprocessed prior to exploratory analysis to verify this data. The first action was reading the csv file created by the web scraper and storing the data in a Pandas Data Frame. It totaled 595 properties. Then we checked whether or not there were duplicate properties from their identification code. No repeated properties found. The unnecessary columns 'Unnamed: 0' and 'id' have been deleted from the data frame.

When viewing the attributes by neighborhood, it was identified that some properties were priceless. This was an indication that there were missing values in the set. Other neighborhoods, besides the three delimited by the work, appeared and were removed from the data frame. It was also checked whether the properties were really from Porto Alegre (city), apartment (type), residential (segment). The “rooms” attribute that stored the number of rooms had a unique odd value that was dropped.

Then the columns that would not be part of the analysis were deleted from the data frame: 'city', 'type', 'segment', 'url', 'date'. As well as outliers that were detected. Finally, the data frame accounted for a total of 226 properties.

2.3. Techniques and Algorithms

In order to reach the objective of the work, in estimating the sale price of an apartment in Porto Alegre, we sought to apply a regression algorithm, whose set of attributes taken into consideration were:

1. Area of the apartment (‘area’)
2. Condominium value (‘condominium’)
3. IPTU value (‘iptu’)
4. Number of parking spaces (‘box’)
5. Number of rooms (‘rooms’)
6. Neighborhood (‘district’)

The analysis focused on three neighborhoods: Auxiliadora, Bela Vista and Mont'Serrat. In this way, the neighborhood identification variable was transformed into three dummies variables.

To evaluate the models to be generated, three conditions were initially determined:

1. Data split strategy for cross-validation was determined to use 10 sets[2];
2. KFold[3] method was chosen as cross-validation strategy;
3. A random number has been set to 7 so that the work can be repeated and comparable.

The original data were divided into four sets, two for training (X_train, y_train) and two for testing (X_test, y_test). The price attribute of the data frame is now part of the "y" and the other attributes the "X". The split was determined so that 25% of the total data was for testing and the remainder for training.

The linear regression[4] algorithm was selected as a benchmark model for the work. Linear regression is a classic statistical tooling technique, widely used for determining continuous dependent variables.

A set of regression algorithms was created to identify which would perform best. *A priori* there was no information on which would be best for the case under study. Therefore, we started from a random selection[5] strategy, where eleven algorithms were listed and tested. The strategy had as its benefits[6] the speed obtained for dealing with different algorithms; objectivity in applying different algorithms to a single defined problem; and obtaining comparable results. The algorithms were:

1. Linear Regression[7]
2. Lasso[8]

[2] cv=10, https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.cross_val_score.html
[3] https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.KFold.html
[4] https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html
[5] Estratégia “spot-checking” (p.76, BROWNLEE, 2018).
[6] Da mesma forma como relatado por Brownlee (2014).
[7] https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html
[8] https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.Lasso.html

3. Elastic Net[9]
4. K-Neighbors Regressor[10]
5. Decision Tree Regressor[11]
6. Support Vector Regression[12]
7. AdaBoost Regressor[13]
8. Gradient Boosting Regressor[14]
9. RandomForestRegressor[15]
10. ExtraTreesRegressor[16]
11. XGBRegressor[17]

Regression is a well-known and widely used statistical technique. It describes the relationship of variables, where the model is built with coefficients to minimize the error by the difference of the set of observations and the values predicted by the linear approximation. The existence of highly correlated variables leads to the situation of multicollinearity.[18].

Least Absolute Shrinkage and Selection Operator (Lasso) regression, in turn, is a variation of linear regression, where the loss function is modified to minimize model complexity, reducing the number of variables[19].

Elastic Net regression combines two types of regressors: Ridge and Lasso. It seeks to minimize the complexity of the model from the coefficients L2 (the sum of the squares of the coefficient values) and L1 (the absolute sum of the coefficient values). Useful when there are several correlated attributes[20].

K-Neighbors acts nonlinearly by locating k similar instances in the training set. It is a classic algorithm, simple and easy to understand[21]. From k neighbors the mean or median of the variables is generated and becomes the predictor.

CART, or Decision Tree Regressor, is a nonparametric supervised learning algorithm used for both classification and regression. It uses the dataset to select the best points, dividing the data to minimize the cost of the performance metric (which is usually the mean square error - MAE). The advantage is: its use and interpretation are simple; created trees can be viewed; requires little data preparation; capable of handling numerical and categorical data; can handle multiple output issues[22].

SVR, or Support Vector Regression, aims to find a function that has the smallest possible error determined by a range. This function seeks to separate data by a hyperplane. The

[9] https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.ElasticNet.html
[10] https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsRegressor.html
[11] https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeRegressor.html
[12] https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVR.html
[13] https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.AdaBoostRegressor.html
[14] https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.GradientBoostingRegressor.html
[15] http://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html
[16] https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.ExtraTreesRegressor.html
[17] https://xgboost.readthedocs.io/en/latest/index.html
[18] https://scikit-learn.org/stable/modules/linear_model.html#ordinary-least-squares
[19] https://scikit-learn.org/stable/modules/linear_model.html#lasso
[20] https://scikit-learn.org/stable/modules/linear_model.html#elastic-net
[21] https://bit.ly/2Ur4aSc
[22] https://scikit-learn.org/stable/modules/tree.html#tree

optimization of the solution is to maximize the dividing margin, that is, to achieve the largest possible spacing of two distinct data sets separated by the hyperplane (BHATTACHARYYA, 2018). In the case of SVR, the error is accounted for by the sum of the classification errors and the margin errors.[23]. For data that is not linearly separable in the original dimensional space, it may be linearly separable in a higher dimensional space. This approach is called a "kernel trick"[24] (KANDAN, 2017).

AdaBoost (or Adaptive Boosting) is an algorithm that is part of a method called Boosting Ensemble that creates a "strong" classifier from a number of "weak" classifiers. Initially a model is trained, followed by a second model that corrects the errors of the first. Templates are added until you have a perfect set or limited by a maximum number of templates added. AdaBoost was initially created for classification and then to act as a regressor (BROWNLEE, 2016).

Gradient Boosting Regressor is an algorithm that builds an additive model progressively, that is, at each stage a regression tree is fitted[25]. It is a stimulus / impulse generalization for arbitrarily differentiable loss functions[26]. It is used for both classification and regression. Its advantages are the handling of mixed data (heterogeneous); predictive force; robustness to handle outliers

Because it is a regression, the performance evaluation of the generated model was compared with a previous benchmark model and computed the specific metrics usually employed (p.299, Müller & Guido, 2017) (Swalin, 2018), where: MAE (Mean Absolute Error)[27], MSE (Mean Squared Error)[28] and R2 (also called Coefficient of Determination)[29]. The metrics equations are[30,31,32]:

$$\mathrm{MAE}(y, \hat{y}) = \frac{1}{n_{\text{samples}}} \sum_{i=0}^{n_{\text{samples}}-1} |y_i - \hat{y}_i| .$$

$$\mathrm{MSE}(y, \hat{y}) = \frac{1}{n_{\text{samples}}} \sum_{i=0}^{n_{\text{samples}}-1} (y_i - \hat{y}_i)^2 .$$

[23] https://bit.ly/2UrmBWL
[24] https://bit.ly/2FW4jtH
[25] https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.GradientBoostingRegressor.html
[26] https://scikit-learn.org/stable/modules/ensemble.html#gradient-boosting
[27] https://scikit-learn.org/stable/modules/generated/sklearn.metrics.mean_absolute_error.html
[28] https://scikit-learn.org/stable/modules/generated/sklearn.metrics.mean_squared_error.html
[29] https://scikit-learn.org/stable/modules/generated/sklearn.metrics.r2_score.html
[30] https://scikit-learn.org/stable/modules/model_evaluation.html#mean-absolute-error
[31] https://scikit-learn.org/stable/modules/model_evaluation.html#mean-squared-error
[32] https://scikit-learn.org/stable/modules/model_evaluation.html#r2-score

$$R^2(y, \hat{y}) = 1 - \frac{\sum_{i=0}^{n_{\text{samples}}-1}(y_i - \hat{y}_i)^2}{\sum_{i=0}^{n_{\text{samples}}-1}(y_i - \bar{y})^2}$$

The algorithm with the best performance of metric R^2 was selected and had its parameters adjusted for refinement and even higher performance. To this end, it used GridSearchCV[33], which, from an initial configuration, performed an exhaustive search to find a combination of parameters that enhance the performance of the tested algorithm.

The motivation for choosing the R2 statistic was to provide "(...) a measure of how well future samples are likely to be predicted by the model"[34].

[33] https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html
[34] https://scikit-learn.org/stable/modules/model_evaluation.html#r2-score

3. Exploratory Data Analysis

The data used in the study were collected from the real estate website Foxter[35], one of the largest in Porto Alegre. In order to delimit the research, the neighborhoods Auxiliadora, Bela Vista and Mont'Serrat were chosen as targets, because they are neighbors (see Figure 2), sharing common streets and bus lines, and with infrastructure (supermarkets, pharmacies etc.) similar.

Figure 2 – Map[36]

[35] https://www.foxterciaimobiliaria.com.br
[36] Fonte: Google Maps.

The sample totaled 226 properties. Figure 3 shows the number of properties per neighborhood, of which 101 are from the Auxiliadora neighborhood, 77 from Bela Vista and 48 from Mont'Serrat.

Figure 3 – Number of properties per neighborhood

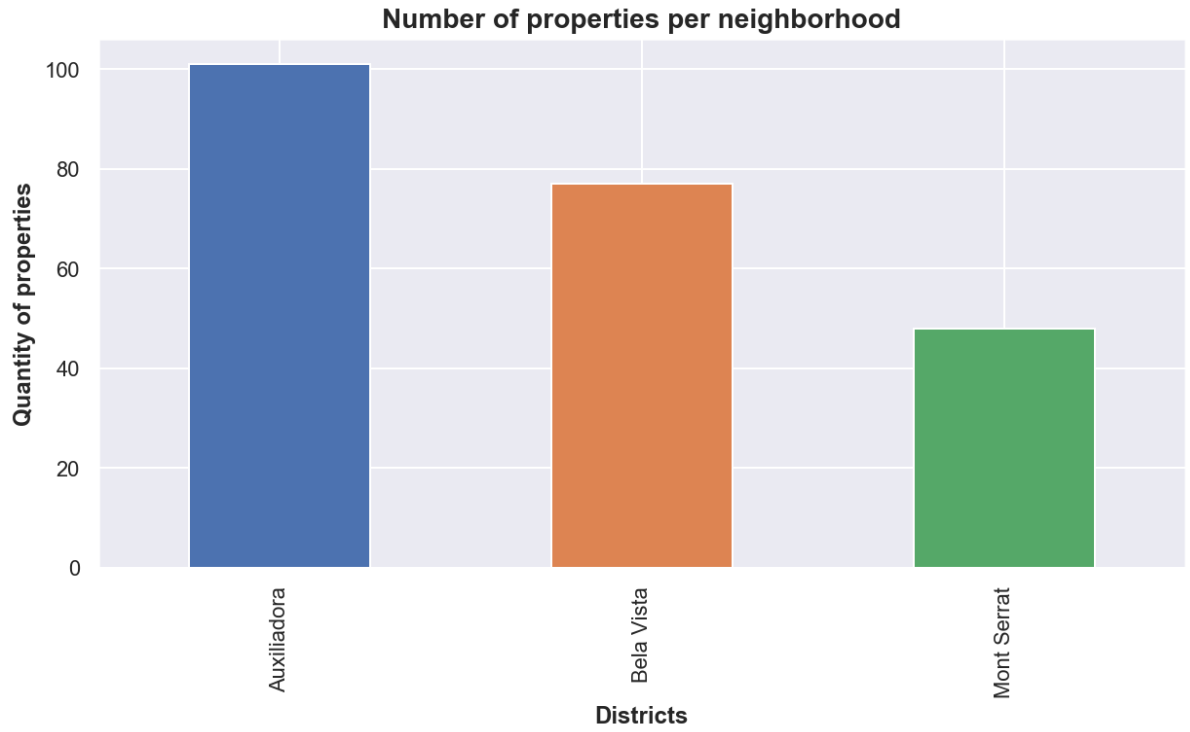
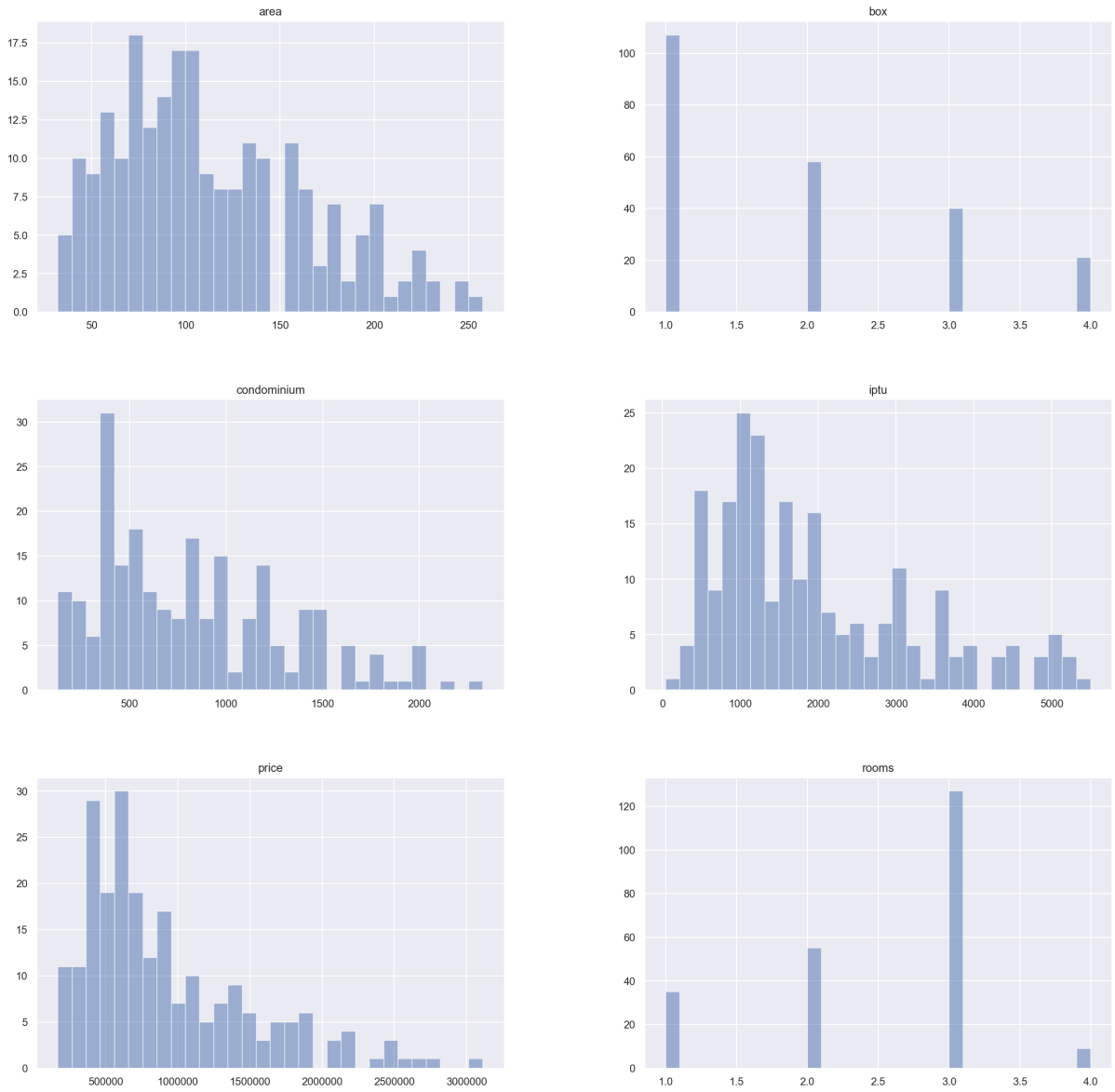


Figure 4 shows the histograms of the six quantitative variables. None of them behaved at or near the normal curve. Variables indicating monetary values, such as condominium, property tax and property price, as well as the variable that recorded the area in square meters, showed deviations to the right. This would be expected since the largest apartments are those with higher selling prices, property tax and condominium prices, and at the same time are in smaller quantities.

It is also possible to observe that in terms of parking spaces, the variable “box” registered that the number of apartments with one parking space is practically equal to those of two parking spaces. Apartment with three vacancies are few. And as for the amount of rooms, three-bedroom apartments make up the most advertisements.

Figure 4 – Histograms of area, box, condominium value, property tax, number of rooms and price



In the following table are the descriptive statistics of the quantitative variables. There is a large variation in prices, with a minimum of R$170 thousand and a maximum of R$3 million. The value of the condominium and property tax also vary widely, with values between R$130 and R$2,330, and from R$43 to R$5,500, respectively.

Table 1 – Descriptive statistics of price, area, condominium, property tax, rooms and garage

	Price	Area	Condominium	IPTU	Rooms	Garage
Quantity	226.00	226.00	226.00	226.00	226.00	226.00
Mean	923,572.89	112.76	825.88	1,946.77	2.49	1.89
Std	585,765.75	50.89	484.64	1,266.43	0.80	1.01
Min	170,000.00	32.13	130.00	43.00	1.00	1.00
25%	499,250.00	73.31	418.50	1,002.50	2.00	1.00
50%	728,151.50	103.11	740.00	1,500.00	3.00	2.00
75%	1,222,500.00	144.29	1,190.00	2,784.50	3.00	3.00
Max	3,114,542.00	257.60	2,330.00	5,500.00	4.00	4.00

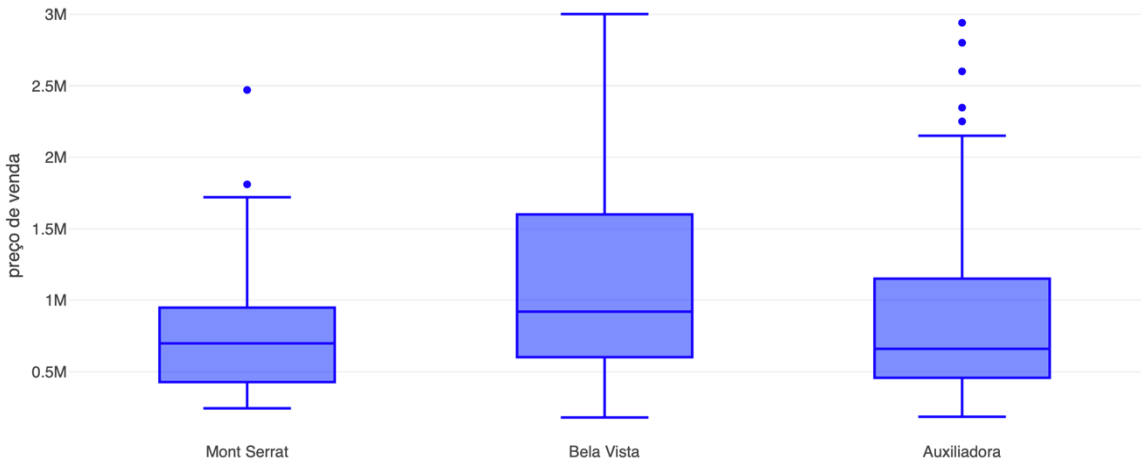
In the following table are the means of the quantitative variables. Note that the Bela Vista neighborhood has the highest averages. It is therefore the neighborhood with the largest and most expensive apartments among the three analyzed.

Table 2 – Averages of price, area, condominium, property tax, rooms and garage identified by neighborhood

Districts	Area	Garage	Condominium	IPTU	Price	Rooms
Auxiliadora	105.74	1.69	733.61	1,705.47	818,909.97	2.51
Bela Vista	127.32	2.27	975.48	2,386.32	1,140,049.92	2.53
Mont Serrat	104.19	1.69	780.04	1,749.42	796,535.88	2.35

Regarding prices, Bairro da Bela Vista has the highest average price of the studied set. In the following boxplot, you can see that both Mont'Serrat and Auxiliadora present properties with high values similar to Bela Vista. However, the same boxplot draws attention when it identifies such values as discrepant with its original neighborhood set.

Figure 5 – Sale price per neighborhood boxplot



Districts	Quant	Mean	Std	Min	25%	50%	75%	Max
Auxiliadora	101.00	818,909.97	516,647.81	185,000.00	460,000.00	650,000.00	1,100,000.00	2,800,000.00
Bela Vista	77.00	1,140,049.92	658,071.36	170,000.00	660,000.00	920,000.00	1,600,000.00	3,114,542.00
Mont Serrat	48.00	796,535.88	507,560.21	180,000.00	415,000.00	644,000.00	961,250.00	2,470,000.00

Figure 6, Figure 7, Table 3 and Table 4 describe the number of rooms of the properties analyzed. It is possible to observe in the graph that the three-bedroom apartments are the largest in the three neighborhoods analyzed. Four-bedroom apartments are the smallest. They are more common in Bela Vista than in the other two neighborhoods.

Figure 6 – Number of Rooms by Neighborhood

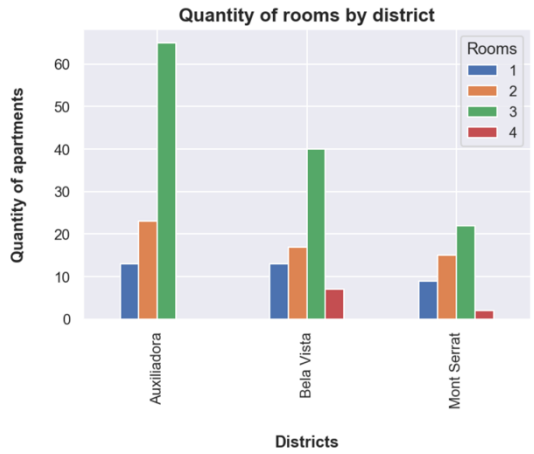


Figure 7 – Number of Apartments by Number of Rooms and Neighborhood

Table 3 – Number of apartments by number of rooms and neighborhood

Districts	Auxiliadora	Bela Vista	Mont Serrat	Total
Rooms				
1	13.00	13.00	9.00	**35.00**
2	23.00	17.00	15.00	**55.00**
3	65.00	40.00	22.00	**127.00**
4	-	7.00	2.00	**9.00**
Total	101.00	77.00	48.00	**226.00**

Table 4 – Descriptive statistics for number of rooms per neighborhood

Districts	Quant	Mean	Std	Min	25%	50%	75%	Max
Auxiliadora	101.00	2.51	0.72	1.00	2.00	3.00	3.00	3.00
Bela Vista	77.00	2.53	0.88	1.00	2.00	3.00	3.00	4.00
Mont Serrat	48.00	2.35	0.84	1.00	2.00	2.50	3.00	4.00

All offered apartments have at least one parking space. Altogether, 107 apartments have one, 58 have two, 40 have three and 21 have four.

Figure 8 – Garagens por Bairro

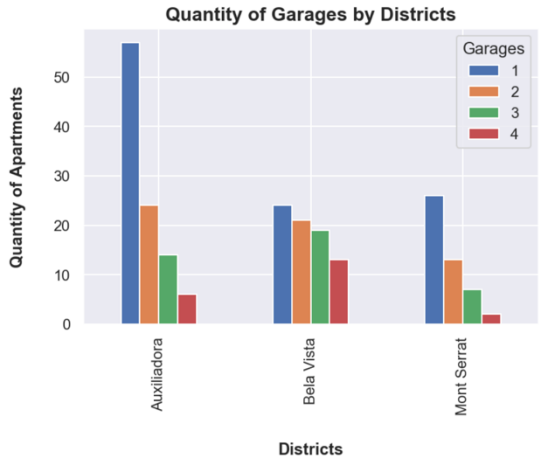


Figure 9 – Bairros por Garagens

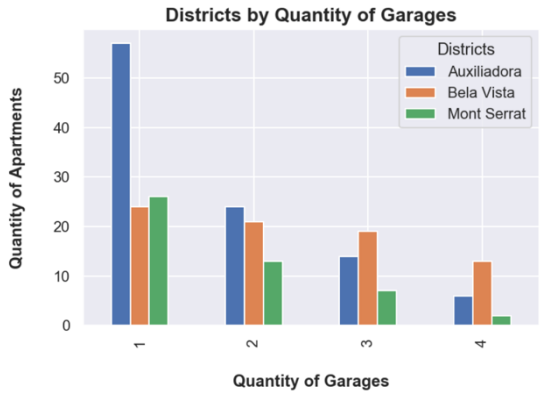


Table 5 – Number of apartments by number of garages and neighborhood

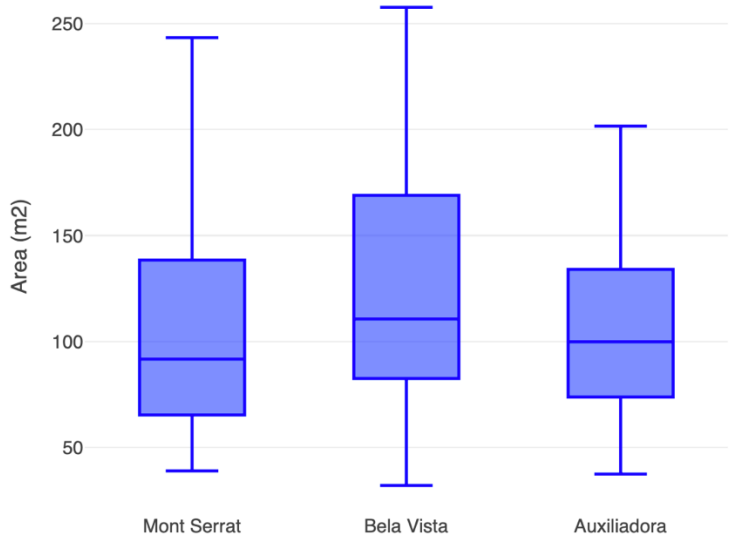
Districts	Auxiliadora	Bela Vista	Mont Serrat	Total
Garages				
1	57.00	24.00	26.00	107.00
2	24.00	21.00	13.00	58.00
3	14.00	19.00	7.00	40.00
4	6.00	13.00	2.00	21.00
All	101.00	77.00	48.00	226.00

Table 6 – Descriptive statistics for number of garages per neighborhood

Districts	Quant	Mean	Std	Min	25%	50%	75%	Max
Auxiliadora	101.00	1.69	0.92	1.00	1.00	1.00	2.00	4.00
Bela Vista	77.00	2.27	1.08	1.00	1.00	2.00	3.00	4.00
Mont Serrat	48.00	1.69	0.88	1.00	1.00	1.00	2.00	4.00

By viewing the following boxplot, you can see that the areas of the apartments are similar. Bela Vista has the highest average for this variable, with average m^2 equal to 127.32. The neighborhood also has small apartments (minimum 32.13), smaller than Mont'Serrat. This fact leads to the highest standard deviation among the three neighborhoods.

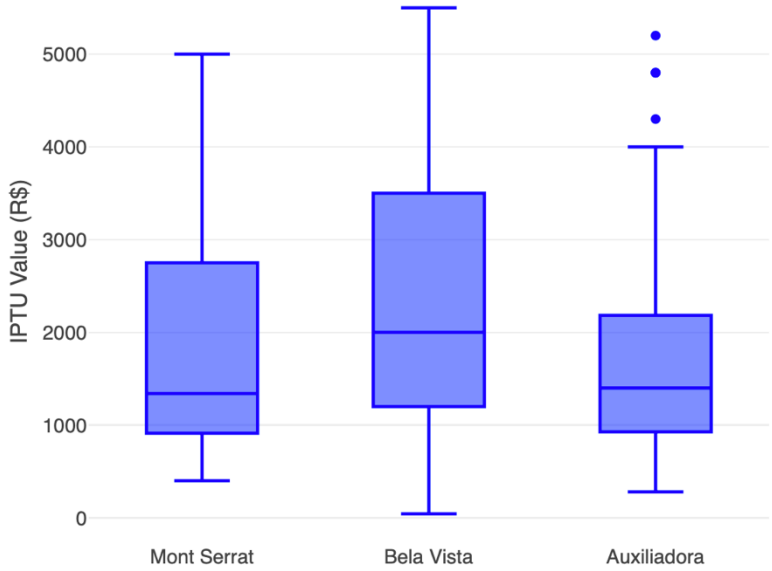
Figure 10 – Area Boxplot (m^2)



Districts	Quant	Mean	Std	Min	25%	50%	75%	Max
Auxiliadora	101.00	105.74	41.83	37.50	74.00	99.86	134.00	201.58
Bela Vista	77.00	127.32	57.10	32.13	82.93	110.63	168.91	257.60
Mont Serrat	48.00	104.19	53.59	39.00	65.40	91.69	137.64	243.28

The minimum amount paid as a condominium was identified in the neighborhood of Mont'Serrat, being R$130.00, followed by Bela Vista with R$140.00. The largest value came from the Auxiliadora neighborhood, with R$2,330.00. Bela Vista has the highest average value, R$975.48.

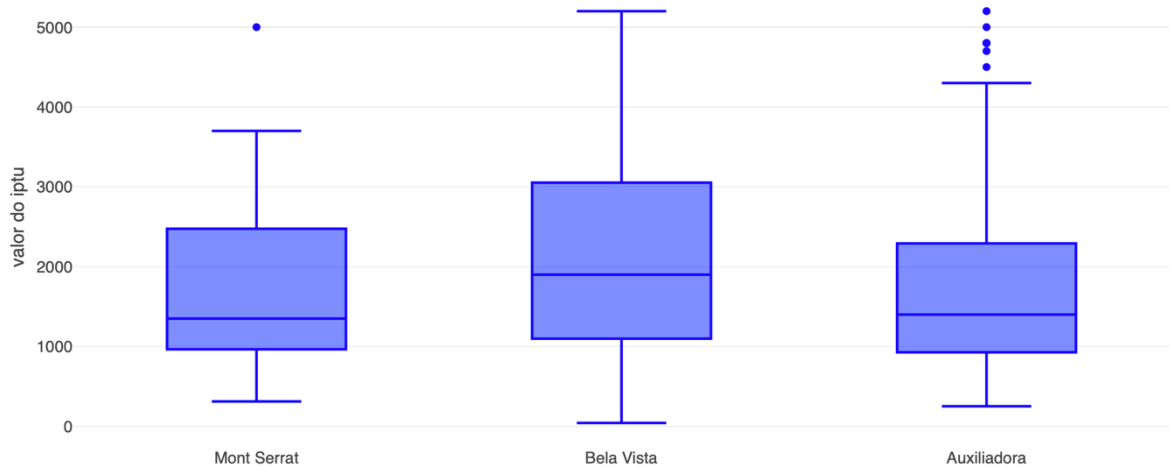
Figure 11 – Condominium Value Boxplot



Districts	Quant	Mean	Std	Min	25%	50%	75%	Max
Auxiliadora	101.00	1,705.47	1,098.58	280.00	930.00	1,400.00	2,178.00	5,200.00
Bela Vista	77.00	2,386.32	1,425.97	43.00	1,200.00	2,000.00	3,500.00	5,500.00
Mont Serrat	48.00	1,749.42	1,157.89	400.00	921.00	1,340.00	2,725.00	5,000.00

The following boxplot shows the property tax for the three neighborhoods. The three are similar. One point that drew attention was the minimum value for Bela Vista, of R$43.00. It is a very low value for a neighborhood that has shown to be the most expensive neighborhood of the three analyzed. The maximum values are very close. Bela Vista returns to present the highest average value with R$2,149.37.

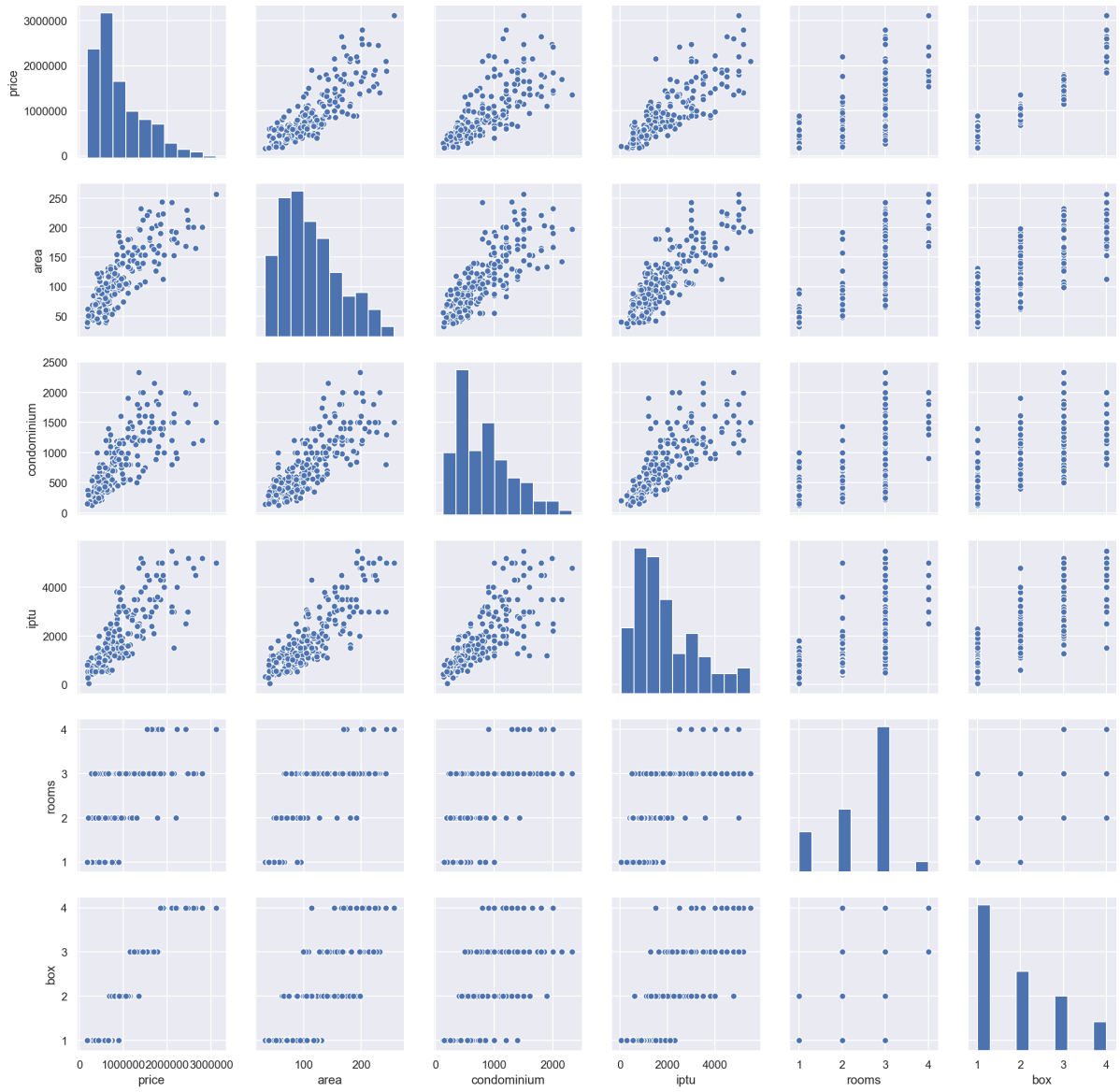
Figure 12 – IPTU Value Boxplot



Districts	Quant	Mean	Std	Min	25%	50%	75%	Max
Auxiliadora	125.00	1,778.84	1,194.30	252.00	930.00	1,400.00	2,288.00	5,200.00
Bela Vista	95.00	2,149.37	1,328.14	43.00	1,099.00	1,900.00	3,035.00	5,200.00
Mont Serrat	67.00	1,667.09	1,022.28	312.00	970.00	1,350.00	2,450.00	5,000.00

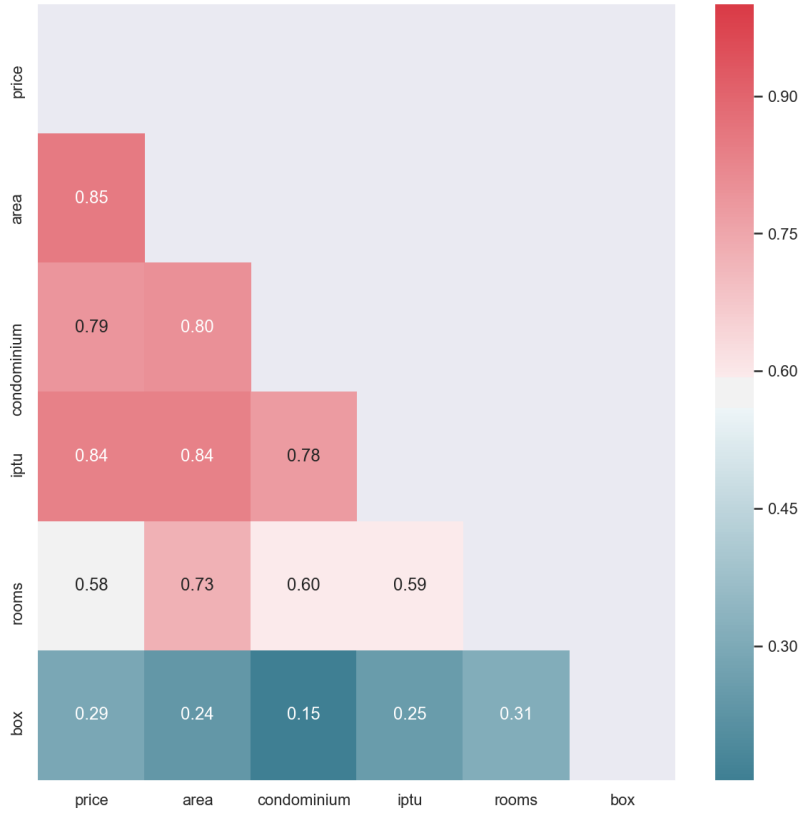
The *pairplot* below shows the relationships between variables from the graphical dispersion of points. You can see that price maintains a positive relationship with all other variables. Continuous variables (area, condominium value and property tax) also show positive relationships with each other.

Figure 13 – Pairplot: relationship between variables



The result of the correlations allows quantifying the degree of relationship between the variables and the meaning (positively or negatively). The following graph shows the correlation values. Most resulted in between 0.50 and 0.75, and strong above 0.750 values. We highlight the "box" variable, which brings together the number of parking spaces, which had a low correlation (from 0.15 to 0.31) with all other.

Figure 14 – Correlations



With a value of 0.83, area and price have a high correlation. In Figure 15 it is possible to observe such relationship, whose line generated by linear regression brings together the plotted points. The line shows an inclination close to 45 degrees, indicating that prices will rise in proportion to the area. Therefore, as expected, the increase in the area results in an increase in the price of the property.

Figure 16 shows the relationship of the area with the price of m2. In this graph you can see that apartments with the same area have different price values of m2. This can occur when comparing new and used properties of the same size, where it is common for new properties to have higher valuation m^2.

Figure 15 – Linear regression between area and price variables

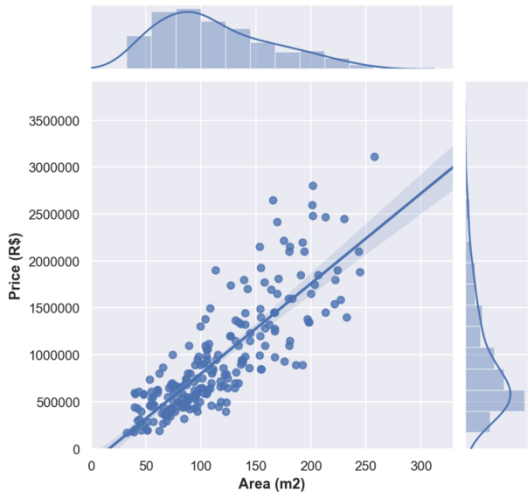
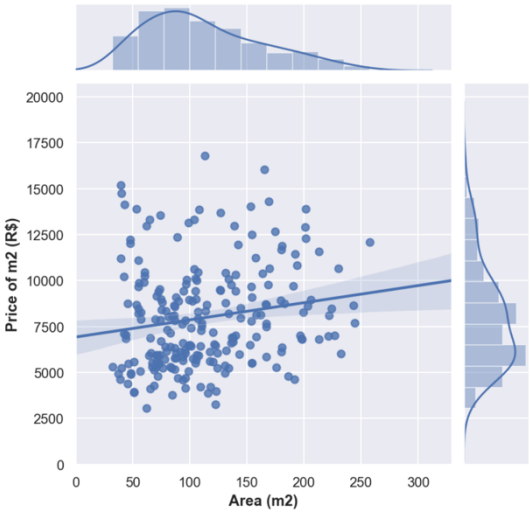
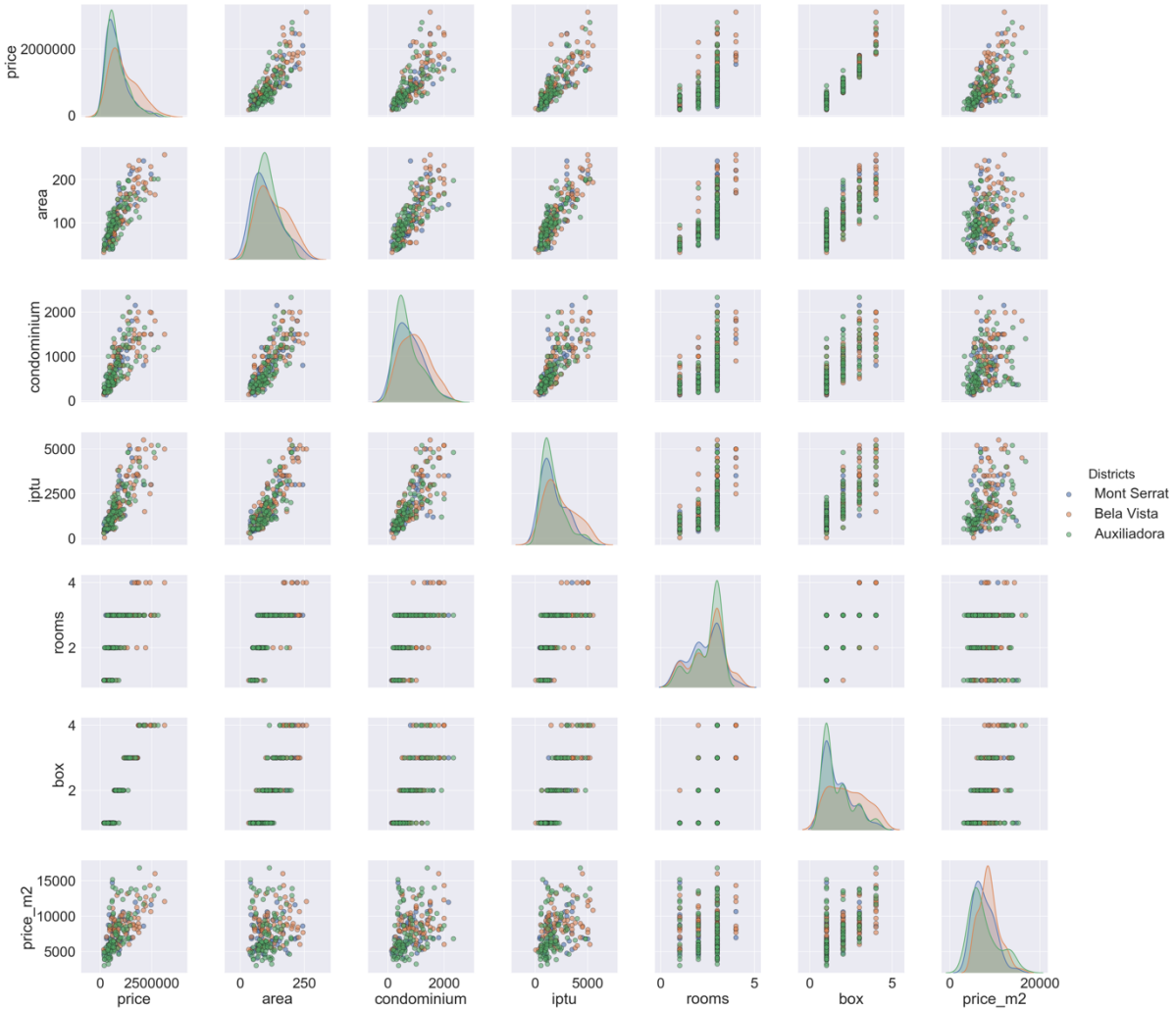


Figure 16 – Linear regression between the variables area and price of m^2



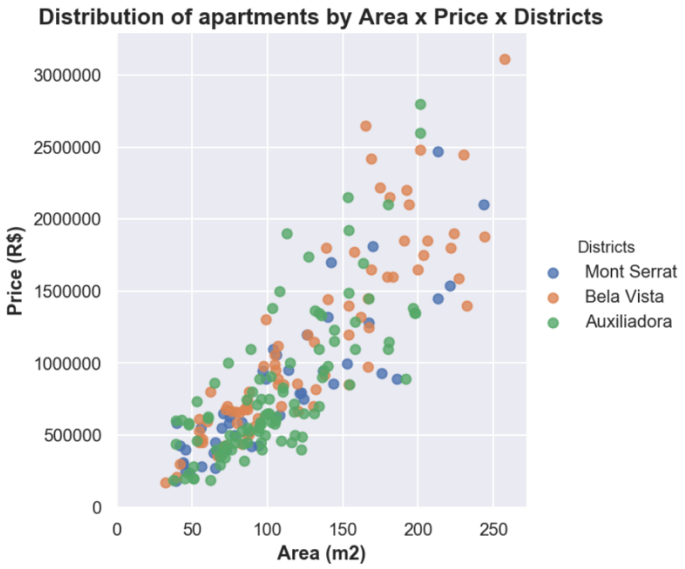
Using multivariate analysis, the following graph shows the relationships of data between variables with emphasis on neighborhoods. In it you can see that the data of the neighborhood Bela Vista (pink) stand out from the other two, always appearing to the right. This is caused by your properties having higher prices, larger areas, higher condominium values and higher property taxes.

Figure 17 – Pairplot: relationship between all variables highlighting the neighborhoods



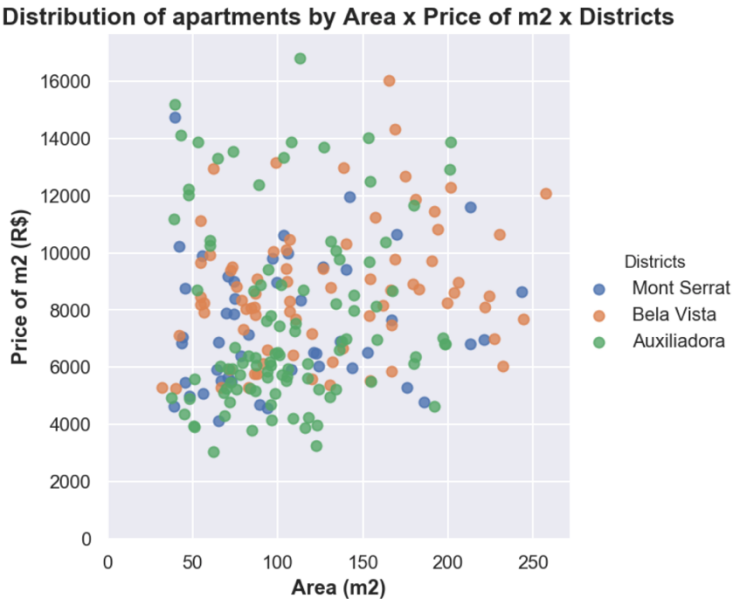
In the following graph, we seek to highlight the relationship between area, price and neighborhood. In a larger view than the previous pairplot, it is observed that Bela Vista's orange points stand out in the upper right quadrant.

Figure 18 – Distribution of real estate by area x price x neighborhood



Analyzing graphically (next figure) the same distribution of points, but using the price of m^2 on the y axis instead of the price of the apartment, important information emerges. You can see that there are apartments of the same size and in the same neighborhood with different m^2 prices. One hypothesis for this question would be that new apartments generally have higher m^2 prices than used ones.

Figure 19 – Distribution of real estate by area x price of m^2 x neighborhood



And in the next graph, there is a set of three plots comparing the area and price ratio separately for each neighborhood.

Figure 20 – Linear regression between area and price with grouping by neighborhood

4. Results

After the initial data processing and exploratory analysis, the work was to develop, the main objective, a predictive model for the price of an apartment. Having five attributes plus the variable "price" (model target), a benchmark model was first determined using the Linear Regression algorithm. The attributes used were:

1. Apartment area ('area')
2. Condominium value ('condominium')
3. IPTU tax ('iptu')
4. Quantity of parking spaces ('box')
5. Quantity of rooms ('rooms')

The Linear Regression Benchmark model performed well with R^2, whose performance measurement statistics using the test data set were:

- MSE of benchmark: 68,944,067,819.49
- MAE of benchmark: 191,058.33
- EVS of benchmark: 0.87
- R2 of benchmark: 0.87

Eleven regression algorithms were then executed in their default settings using the training data. With the exception of SVR, all the others showed close results. AdaBoost and ExtraTree got the best scores, with R^2 equal to 0.894. The scores were:

LR : Linear Regression =	0.870
LASSO : Lasso =	0.870
EM : Elastic Net =	0.795
KNN : KNeighbors Regressor =	0.618
CART : Decision Tree Regressor =	0,69
SVR : Support Vector Regression =	-0.179
AB : AdaBoost Regressor =	0.894
GBR : Gradient Boosting Regressor =	0.883
RF : Random Forest Regressor =	0.881
ET : Extra Trees =	0.894
XGB : XGBoost =	0.871

The next step was to refine the AdaBoost and ExtraTree parameters in order to obtain an even higher model. We started to use GridSearchCV, whose initial parameters were:

For AdaBoost

```
param_grid = {
        'n_estimators':range(20,151,10),
        'learning_rate':[1, 0.5, 0.1],
       }
```

And for ExtraTrees

```
param_grid = {
        'n_estimators':range(20,151,10),
        'max_depth':[1, 3, 5, 7, 9],
        'min_samples_split':[2, 4, 8],
        'min_samples_leaf':[1, 3, 5, 7, 9]
       }
```

After adjustments, the final setting for GridSearchCV search was:

For AdaBoost

```
param_grid = {'n_estimators':range(20,251,10),
        'learning_rate':[0.05, 0.01],
       }
```

And for ExtraTrees

```
param_grid = {'n_estimators':range(20, 51,10),
        'max_depth':[9],
        'min_samples_split':[4, 6],
        'min_samples_leaf':[1]
       }
```

With these search settings GridSearchCV found the best optimized setting for ExtraTrees in the training set, resulting in R^2 equal to 0.91. The optimized parameters were:

{'max_depth': 9, 'min_samples_leaf': 1, 'min_samples_split': 4, 'n_estimators': 40}

With this result, we proceeded to the final model configuration using ExtraTrees. First, the predictive model was generated with the training data set. Then the ExtraTrees model created with the test data was executed. The model evaluation statistics were:

- MSE: 49,298,699,950.05
- MAE do benchmark: 154,588.43
- EVS do benchmark: 0.91
- R2: 0.91

The R^2, 0.91, obtained with the test set, was close to the 0.87 achieved with the training data. With this, it is understood that there was no overfitting. In comparing ExtraTrees with the benchmark model, ExtraTrees was superior in the four valuation statistics, MSE, MAE, EVS, and R^2.

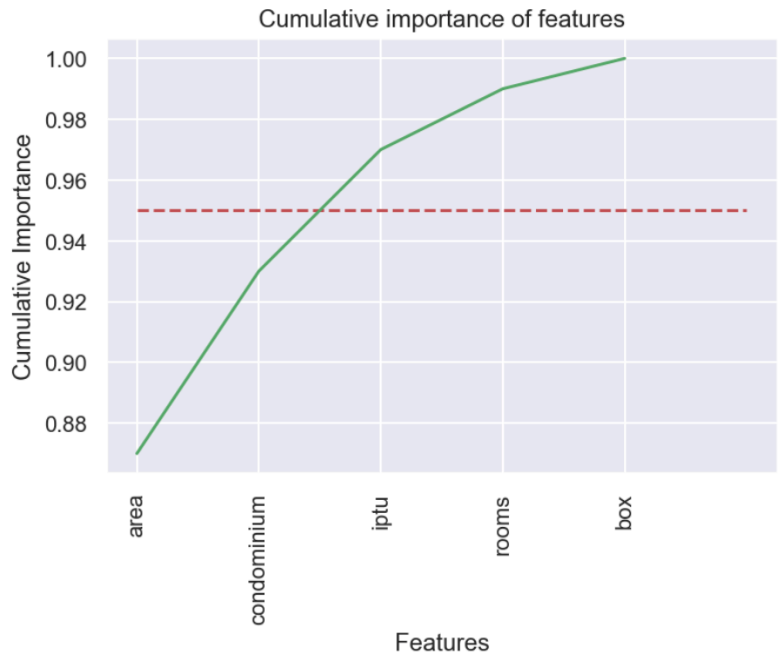
The degrees of importance of each attribute used in the ExtraTrees model were identified. Apartment area was the main feature, accounting for 87% of the total importance of the features. The value of the condominium comes next with 6%. And property tax, the number of rooms and garages were the last, with 4%, 2% and 1%, respectively. The following graphic shows the attributes and their importance in bar form. The values of the contributions of each attribute to the model were:

- `Feature: area` `Importance: 0.87`
- `Feature: condominium` `Importance: 0.06`
- `Feature: iptu` `Importance: 0.04`
- `Feature: rooms` `Importance: 0.02`
- `Feature: box` `Importance: 0.01`

Figure 21 – Features Importance

The next graph, in turn, shows the cumulative importance of the attributes of the model. It is noted that the red dotted line indicates the 95% level so that it can be viewed when the curve crosses that level.

Figure 22 – Cumulative importance of features



A specific objective listed at the beginning of the work was to predict the price of an apartment. Thus, having concluded the final model, it was estimated that an apartment of 119m^2, with 2 parking spaces, with R$750.00 of condominium value, with R$2,100.00 of property tax, with 3 bedrooms, has its estimated value at R$798,789.00.

5. Conclusions

5.1. Results obtained

This work aimed to predict the sale price of an apartment in the city of Porto Alegre. Two secondary objectives were listed: identifying the best prediction algorithm and identifying the most important attributes. The first stage of the work consisted of collecting data on the Foxter real estate website through a web scaper built for this purpose. In the second step the data was read from csv, cleaned and formatted. The exploratory analysis was done in the third stage. And finally, the fourth step was modeling.

Eleven regression algorithms were identified and used. One, the linear regression, was selected to be the benchmark of the work. At the end of the tests, the ExtraTrees Regressor had the best performance of all, and had an R^2 with 0.91 with the final model. Its result was even better than the benchmark.

The attributes related to the size of the apartments and the number of parking spaces were identified as the two main attributes for the model. Their amounts to the final model were 87% and 6%, respectively, totaling 93%. In turn, the value of the property tax, the number of rooms and garages were the least.

5.2. Model Improvement

Two aspects were identified during the work that can be explored in the future in search of model improvement. During the exploratory analysis, it was found that there are apartments of the same size, in the same neighborhood, but with different price values of m^2. The identification of possible causes and their categorization will allow to separate the properties and, thus, have a better explanation for price difference. The second aspect would be the treatment of the text included in the field “property description” on the real estate website. All the words (usually positive about the property) may contribute to a better explanation of the price formation.

6. References

ALFIYATIN, Adyan Nur, FEBRITA, Ruth Ema, TAUFIQ, Hilman, MAHMUDY, Wayan Firdaus. (2017). **Modeling House Price Prediction using Regression Analysis and Particle Swarm Optimization Case Study**: Malang, East Java, Indonesia International Journal of Advanced Computer Science and Applications (IJACSA), 8(10). https://bit.ly/2FzdIHf

BALDOMINOS, Alejandro, BLANCO, Iván, MORENO, Antonio José, ITURRARTE, Rubén, BERNÁRDEZ, Óscar, AFONSO, Carlos. (2018). **Identifying Real Estate Opportunities Using Machine Learning**. *Appl. Sci.* 2018, *8*, 2321. https://bit.ly/2QVRTTA

BHATTACHARYYA, Indresh. (2018). **Support Vector Regression Or SVR**. https://bit.ly/2Sk04Ow

BROWNLEE, Jason. (2014). **Why you should be Spot-Checking Algorithms on yout Machine Learning Problems**. https://bit.ly/2CNwruW

BROWNLEE, Jason. (2016). **Boosting and AdaBoost for Machine Learning**. https://bit.ly/2MFBQsG

BROWNLEE, Jason. (2018). **Machine Learning Mastery with Python**: understand your data, create accurate models and work projects ent-to-end. E-Book. https://bit.ly/2MDN6WA

KANDAN, Harish. (2017). **Understanding the kernel trick.** Towards Data Science. 30/08/2017. https://bit.ly/2B9MvqS

KOMAGOME-TOWNE, Anh. (2016). **Models and visualizations for housing price prediction**. California State Polytechnic University. Thesis. https://bit.ly/2swsg1S

MÜLLER, Andrea C. & GUIDO, Sarah. (2017). **Introduction to Machine Learning with Python**: a guide for data scientists. O'Reilly.

NGUYEN, An. (2018). **Housing Price Prediction**. Union College. Final Project. https://bit.ly/2RO6lSf

PARK, Byeonghwa, BAE, Jae Kwon. (2015). **Using machine learning algorithms for housing price prediction**: The case of Fairfax County, Virginia housing data. Expert Systems with Applications, Volume 42, Issue 19, 1 November 2015, Pages 6806. https://bit.ly/2Cqxlxo

PIERRE, Rafael. (2018). **Going Dutch: How I Used Data Science and Machine Learning to Find an Apartment in Amsterdam**—Part I. Towards Data Science. https://bit.ly/2HI3ReZ

SWALIN, Alvira. (2018). **Choosing the Right Metric for Evaluating Machine Learning Models(Part 1)**. https://bit.ly/2HwlUtT

YU, Li, JIAO, Chenlu, XIN, Hongrun, WANG, Yan, WANG, Kaiyang. (2018). **Prediction on Housing Price Based on Deep Learning**. International Journal of Computer and Information Engineering Vol:12, No:2, 2018. https://bit.ly/2MhHNfg