# Prediction of the Total Price of the Houses in Brasil

*Bennur Kaya*
*Middle East Technical University*
*Ankara, Turkey*

***Abstract*— This article presents the relationship between the total price and properties of houses in Brazil. Prices are estimated using different machine learning methods such as Support Vector Machine, Random Forest. In this article, I tried to find answers for the research questions in the direction of my interest. After the data was tidied and cleaned, the models are executed. The estimated performances of the models are evaluated by RMSE. R-Studio programming language is used to execute all the procedures.**

**Keywords—Multiple Linear Regression, SVM, Random Forest, Airbnb, House Price, Brazil.**

## INTRODUCTION

The aim of this study is to examine the relationship between the total price and the features of the houses in Brazil.

In the work, the total prices of the houses in Brasil, are predicted by using several regression model such as Multiple Linear Regression, Support Vector Machine, Random Forest. The min-max accuracy and RMSE of each model are calculated and compared to identify their performance.

## LITERATURE REVIEW

Since 2015, Property Price in Brazil has started to decline (Global Property Guide). Nowadays, Brazil's housing market remains weak, despite improving economic conditions. Along with the economic and sectoral effects, the price of the houses are affected by the features of the houses, their locations and many other factors.

A multiple linear regression model is suggested to predict the total prices of the Brazilian Houses. In this way, it was seen that Homeowners association tax, fire insurance and property tax had a significant effect on the total price.

## METHODOLOGY

### *A. Dataset*

The name of the dataset is Brazilian Houses to Rent. This dataset was renewed as version 2 in March, 2020. The data has 10692 observations for 13 different variables (features of the houses).

This dataset contains 10962 houses to rent with 13 different features. These features are City, Area, Rooms, Bathroom, Parking Spaces, Floor, Animal, Furniture, HOA (R$), Rent Amount (R$), Property Tax (R$), Fire Insurance (R$) and Total (R$) respectively. At the beginning of the analysis, only Floor variable had "NA". At the end of EDA, missing data imputations were made.

*Fields in the dataset:*

• City – Factor with levels "Belo Horizonte", "Campinas", "Porto Alegre", "Rio de Janeiro", "São Paulo".
• Rooms - Quantity of the rooms.
• Bathroom – Quantity of bathrooms
• Parking Spaces - Quantity of parking spaces
• Floor - The floor the property is on.
• Animal - Represents whether it is allowed to feed animals at home or not.
• Furniture - Represents whether the house is furnished or not.
• HOA (R$) - Homeowners Association Tax
• Rent Amount (R$) - Rent Price
• Property Tax represents the property
• Fire Insurance (R$) - Fire insurance fee
• Total (R$)

A. *Descriptive Statistics*

As it can be seen below, descriptive statistics table for numerical variables is shown. At the beginning of the analysis, it was obtained because it gives an idea of the dataset at the beginning of the discovery.

| city | area | rooms | bathroom |
|---|---|---|---|
| Length:10692 | Min. : 11.0 | Min. : 1.000 | Min. : 1.000 |
| Class :character | 1st Qu.: 56.0 | 1st Qu.: 2.000 | 1st Qu.: 1.000 |
| Mode :character | Median : 90.0 | Median : 2.000 | Median : 2.000 |
| | Mean : 149.2 | Mean : 2.506 | Mean : 2.237 |
| | 3rd Qu.: 182.0 | 3rd Qu.: 3.000 | 3rd Qu.: 3.000 |
| | Max. :46335.0 | Max. :13.000 | Max. :10.000 |

| parking_spaces | floor | animal |
|---|---|---|
| Min. : 0.000 | Min. : 1.000 | Length:10692 |
| 1st Qu.: 0.000 | 1st Qu.: 2.000 | Class :character |
| Median : 1.000 | Median : 5.000 | Mode :character |
| Mean : 1.609 | Mean : 6.583 | |
| 3rd Qu.: 2.000 | 3rd Qu.: 9.000 | |
| Max. :12.000 | Max. :301.000 | |
| | NA's :2461 | |

| furniture | hoa_tax | rent_amount |
|---|---|---|
| furnished :2606 | Min. : 0 | Min. : 450 |
| not furnished:8086 | 1st Qu.: 170 | 1st Qu.: 1530 |
| | Median : 560 | Median : 2661 |
| | Mean : 1174 | Mean : 3896 |
| | 3rd Qu.: 1238 | 3rd Qu.: 5000 |
| | Max. :1117000 | Max. :45000 |

| property_tax | fire_insurance | total |
|---|---|---|
| Min. : 0.0 | Min. : 3.0 | Min. : 499 |
| 1st Qu.: 38.0 | 1st Qu.: 21.0 | 1st Qu.: 2062 |
| Median : 125.0 | Median : 36.0 | Median : 3582 |
| Mean : 366.7 | Mean : 53.3 | Mean : 5490 |
| 3rd Qu.: 375.0 | 3rd Qu.: 68.0 | 3rd Qu.: 6768 |
| Max. :313700.0 | Max. :677.0 | Max. :1120000 |

*Table 1. Summary Statistics*

City is a categorical variable with 5 different levels as Belo Horizonte, Campinas, Porto Alegre, Rio de Janeiro, São Paulo. Area is a numerical variable which indicates the Area of the property. The average of Area is 149.2 m2. Rooms and Bathroom are both numerical variables as well. Lastly, Total is a numerical variable with a minimum value as 499(R$) and the maximum value as 1120000(R$).

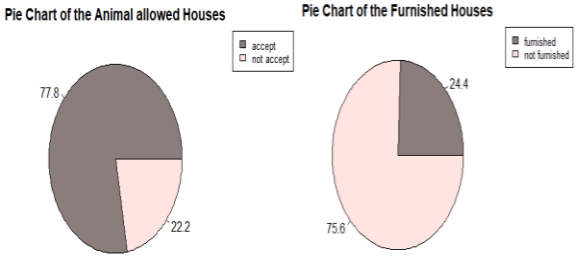
***Correlation Plot***


*Graph 1. Correlation Plot*

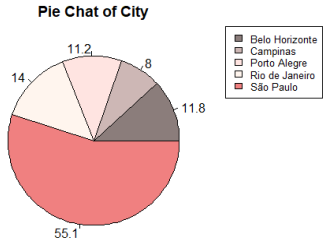
As it can be seen from the correlation table, Rooms, Bathroom, Parking Spaces and Floor have a positive correlation with Total. However, they are not strong positive correlations.

Since these 4 variables are physically related with the houses, it can expected that they are directly related to Rent amount. By looking at the correlation table, it can be said that Room, Bathroom and Parking Spaces have higher correlation rate with Rent Amount than Total.

After Data Cleaning & Tidying part, some visualizations are made to have a better understanding on categorical data.




*Graph 2-3. Pie Chart of the Animal allowed Houses and Furnished Houses*




*Graph 4. Pie Chart of City*

As in the pie charts, animals are accepted in 77.8 percent of the houses. Also, 75.6% percent of the houses are not furnished. Lastly, Alegre, Rio de Janeiro, São Paulo. Also, 55.1% of the houses in the data were located in São Paulo and it is followed by Rio de Janeiro with 14% and Belo Horizot 11.2% respectively.

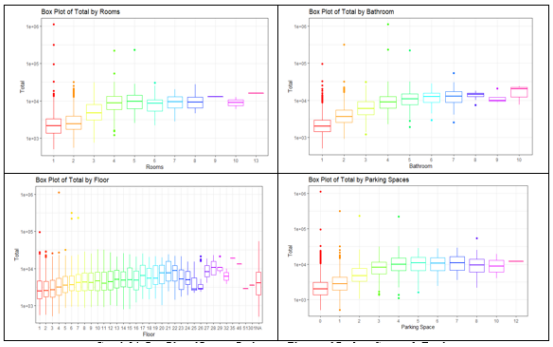
Also, all continuous variables are examined by visualizing graphs to have a better understanding.

*B. Explanatory Data Analysis*

In this part of the study, there exists six research questions. The solutions has been given by using appropriate visualization graphs and suitable statistical methods. This section gives general information about researcher's interest. Also, It provides an important basis for the further studies.

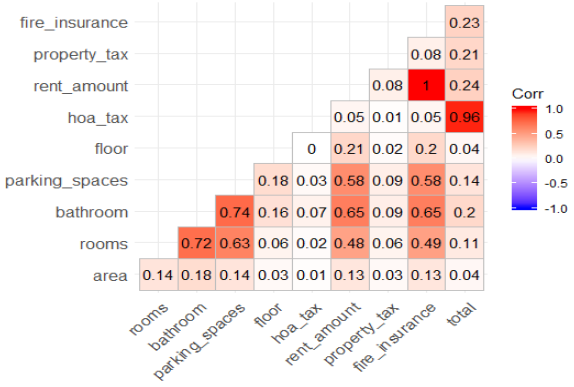
*RESEARCH QUESTIONS*

1) Is there any relationship between rooms, bathroom, floor, parking places & total?




*Graph 5-6-7-8. Plot of Rooms, Bathroom, Floor and Parking Spaces & Total respectively*

Generally, as the quantity of the variables increases, Total also shows increasing trend. For Rooms, Bathroom and Parking spaces, untill the quantity reaches 5, there exists positive relationship obviously. After 5, the trend is stable. For Floor, after the 22nd floor, it seems Total has fluctuated trend.

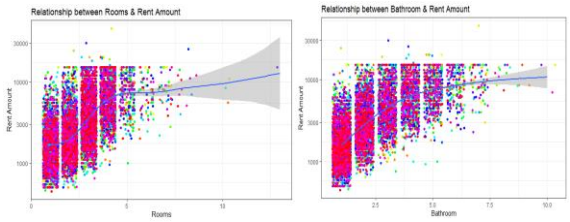



*Graph 9. Correlations of Variables*

When we look at the correlation table, we see that Rooms, Bathroom, Parking Spaces and Floor have a positive correlation with Total. However, they are not strong positive correlations.

Note: The value of 1 (correlation btw. Rent Amount & Fire Ins.) was 0.99 before NAs were omitted.

2) Is there any significant relationship between rooms, bathroom, parking places & rent amount?

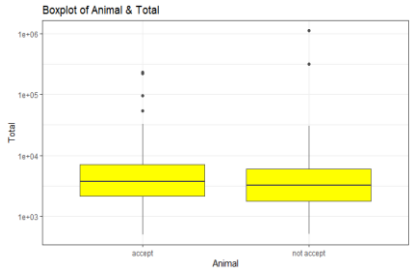



*Graph 9-10. Scatter Plots of Rooms and Bathroom & Rent Amount*


*Graph 11. Scatter Plot of Parking Space & Rent Amount*

As it can be seen in scatter plots, Rooms, Bathroom and Parking Spaces have a positive relationship with Rent Amount. Kendall's Rank Correlation Test was used to see wheter the relationships were significant or not. Since, p values are less than 0.05, it can be said that all three variables have **significant** positive relationship between Rent Amount.

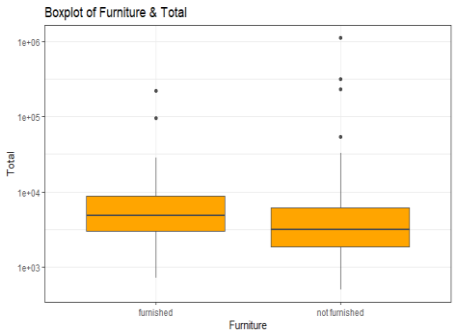
3) Is there a significant difference between the total amount of houses which accept animals and do not accept animals?




*Graph 12. Box Plot of Total by Animal*

By looking at the median of total in boxplots, it can be said that the prices of houses that accept animals are higher. To be sure, t test was conducted. All the assumptions of t test is provided. As a result, since p value is higher than alpha value, we fail to reject null hypothesis in favor of the alternative hypothesis. We conclude that the two means are not significantly different.

4) Is there a significant difference between the Total amount of houses which are furnished and not furnished?

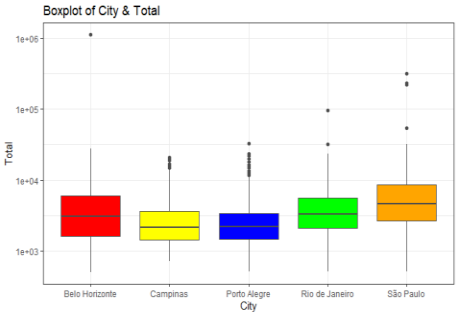



*Graph13. Box Plot of Total by Furniture*

By looking at the median of Total in boxplots, it can be said that the total prices of furnished houses higher than that are not furnished.

To be sure, t test was conducted. As a result, since p value is less than 0.05, Ho is rejected. That means, two means are significantly different. By serial correlation test, it can be said that there is positive correlation between Furniture & Total with 0.038.

5) Does the city have a significant effect on Total?

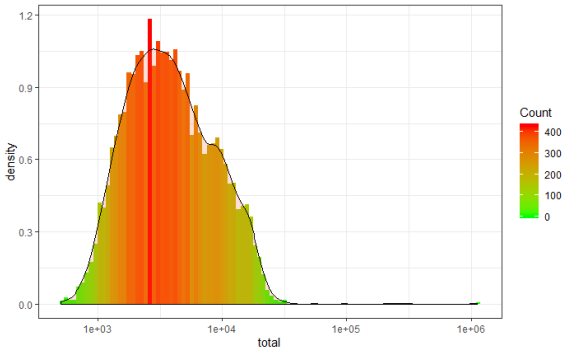



*Graph 14. Box Plot of Total by City*

By looking at the box plots, it can be said that the highest Total is seen in São Paulo and it is followed by Rio de Janerio and Belo Horizonte respectively.

To be sure if there is significant difference between cities, kruskal test was conducted. Since the p-value is less than 0.05, it can be concluded that there exist significant differences for Total Price of the houses according to cities.

6) Does "total" have a normal distribution? If not, what is the distribution of "total"?




*Graph 15. Histogram of Total with Density Line*

By looking at the histogram, it can be said that Total is not normally distributed. To be sure, Anderson Darling test is conducted and since p value is less than 0.05 we reject Ho where Ho: Total follows the normal distribution. That means, Total is not normally distributed.

The skewness here is 58.9. This value implies that the distribution of the data is skewed to the right or positively skewed. It is skewed to the right because the total value has some outliers.

For the kurtosis, we have 3927.183 and it seem really high. High kurtosis in a data set is an indicator that data has heavy tails or outliers. If there is a high kurtosis, then, we need to investigate why we have so many outliers.

## *C. Missingness*

At the beginning of the analysis, only Floor variable had "NA" and there were no relationship between the missingness of the data and any values, observed or missing. So, there was nothing systematic going on that makes some data more likely to be missing than others.

In this dataset, the Missing Data Mechanism is Missing Completely at Random (MCAR). At the end of EDA, missing data imputations were made. The mean imputation is a chosen method.

### *D. Modelling*

The data set is suitable to conduct a model to estimate the total price of the houses. However, before we create a model, we divide the data into two parts, train data and test data.

Therefore, we use 8021 observations which is the 75% of the data set as a train data to construct a model. In addition, we accept 2671 observations as a test data that are 25% of the data set.

#### *1. Linear Regression*

In this study, firstly, multiple regression analysis was established. Although the model was significant with p-value<0.05, many explanatory variables seemed as non-significant. For this reason, stepwise regression was established and the proposed model was examined. The stepwise model is also significant since the p-value is less than 0.05. As a result, both multiple regression model and stepwise regression model show the same predictor variables as significant ones. The output shows the suggested variables by stepwise regression model.

| | **Estimate** | **Std. Error** | **Pr(>\|z\|)** |
|---|---|---|---|
| *(Intercept)* | 1.078e-01 | 1.172e-01 | 0.358 |
| *hoa_tax* | 1.000e+00 | 4.295e-06 | <2e-16 *** |
| *rent_amount* | 9.999e-01 | 1.443e-04 | <2e-16 *** |
| *property_tax* | 1.000e+00 | 1.225e-04 | <2e-16 *** |
| *fire_insurance* | 1.009e+00 | 1.029e-02 | <2e-16 *** |

*Table 2. The Results of Stepwise Regression Model*

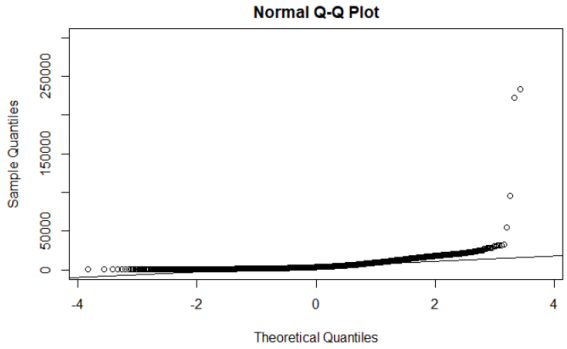
Residual standard error: 6.91 on 8016 dof
Adjusted R-squared: 0.9761
p-value: < 2.2e-16

Then, multicollinearity check was performed. Later, multicollinearity between rent_amount and fire_insurance is observed (VIF>10).

Furthermore, ridge regression which is a technique for analyzing multiple regression data that suffer from multicollinearity was applied. Accuracy values were examined between Stepwise regression and Ridge Regression to see if ridge regression have a positive effect on accuracy or not. While the accuracy value of the Stepwise model was 97.10%, the accuracy value increased to 99.10% with ridge regression. Clearly, in this case, ridge regression is successful in improving the accuracy by a minor but significant fraction. However, there still exist multicollinearity problem.

Hence, in logical way, if rent_amount is high, total price of the house will also high. So, the study was continued by removing rent_amount from multiple regression analysis to get rid of multicollinearity problem.

After multicollinearity is examined, assumptions checks are done.




*Graph 16. QQ-Plot for total*

Normal Q-Q plot shows the multiple regression model doesn't satisfy normality assumptions. It will not be used for prediction.

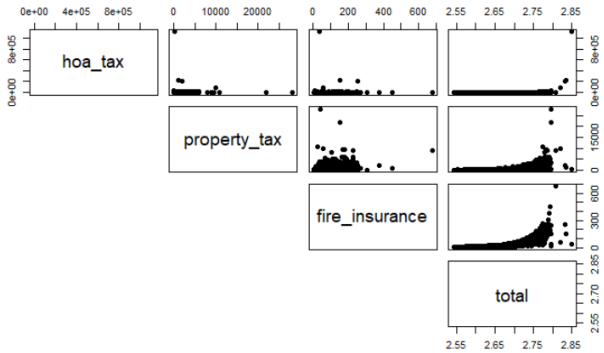
Transformation was applied to ensure the normality assumption. BoxCox transformation was made to the total variable. When the new model was created, all variables and intercept were significant and there was no multicollinearity problem. However, normality assumption was still not provided.

```
Shapiro-Wilk normality test

data:  sample(model3$residuals,
5000)
W = 0.9257, p-value < 2.2e-16
```

*Table 3. The Results of Shapiro Wilk Test*

For this reason, the relationship between explanatory variables and response variable was checked. There seems no linear relationship between explanatory variables and "total". Only "fire_insurance" seems to have smooth positive relationship.




*Graph 17. The relationship between explanatory variables and response variable*

So, transformation is needed for Xs. After boxcox transformation for independent variables, normality is still not achieved.

Lastly, Generalized additive models for non-parametric was conducted. Although, model is significant with significant variables, The plots showed that the model is not suitable for this data.

Several models are used to explain data. None of them were completely suitable for data. By comparing the RMSE values of the models, it can be said that Linear Regresssion Model after Box Cox transformations for both response and explanatory variables without non-significant variables and rent_amount has the smallest RMSE value. Also, since the observations seem not so far away from the line and the constant variance assumption is not violated, it will be used for the further work.

According to this model, it can be said that hoa_tax, property_tax and fire_insurance have significant effect on the Total price of the Houses.

After it is seen that which variables affect the total variable, the machine learning algorithms are applied to make predictions.

*2. Support Vector Machine*

Support Vector Machine is a common used machine learning algorithm with an aim to classify data into different classes.

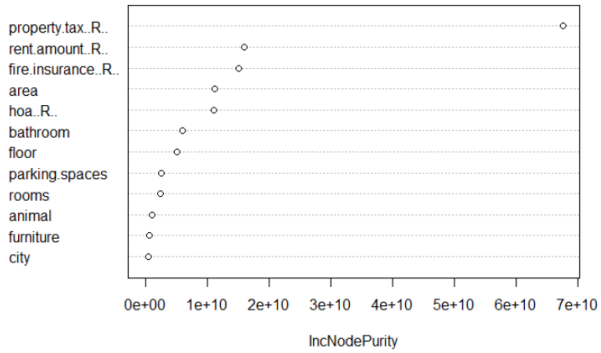
In SVM method, prediction of the total price of Brazilian Houses considering the City, Area, Rooms, Bathroom, Parking Spaces, Floor, Animal, Furniture, HOA (R$), Rent Amount (R$), Property Tax (R$), Fire Insurance (R$) and Total (R$) variables.

As a result, Support Vector Machine method gives the SVM-Type as eps-regression type and as SVM-kernel function as radial. Number of Support Vectors equals to 504. Lastly, Cost parameter is tuned as 10 and gamma parameter is tuned as 0.1.

*3. Random Forests*

Random Forest machine learning method is used to construct multiple decision trees to make a prediction about the given data. The final output is chosen by the random forest and this is based on the majority of those trees.

In this method, the number of trees are chosen as 500. Also, the same features, which are used in Support Vector Machine method, are used to make a prediction. The variability of the model equals to 48.65%.




*Graph 18. Varible Importance*

As a result of variable importance plot, property tax has the highest importance for total price of the house. After that, rent amount and fire

insurance has the highest importance, respectively. This is logical because, as a result of all models, these three variables were determined as significant explanatory.

### *E. Perfomance Comparison on Train and Test Data*

In this project, RMSE value of the models are used to identify the performance of the methods. Sensitivity, Specificity and Accuracy didn't preferred since the response is not binary.

*RMSE Value*

It means root mean square error. It is a standard way to measure the error of a model in predicting quantitative data.

## IV. RESULTS

In this part, the results are expressed for following models;

1. Multiple Linear Regression
2. Support Vector Machine
3. Random Forests

| | ***RMSE*** | |
|---|---|---|
| | *Train* | *Test* |
| ***MLR*** | *0.412* | *0.4113* |
| ***SVM*** | *0.404* | *0.4038* |
| ***RF*** | *0.453* | *0.4527* |

*Table 4. Performance Comparison*

Multiple Linear Regression model gives the best root means square error value according to Support Vector Machine and Random Forest.

## V. CONCLUSION

In this study, after data cleaning & tidying process, both categorical and numerical variables were examined one by one. Later, after applying mean imputation method, statistical solutions with both visual and formal tests are made for six different research questions. Then, the most suitable model was searched for the response variable which is Total Price of Brazilian Houses. Several models have been developed. Assumption violation such as multicollinearity and normality was tried to be removed from data. For this, Box Cox transformation was made for both the response and explanatory variables. Then, the most suitable one was selected by looking at the RMSE value among the models. Finally, predictions were made using the machine learning methods. Multiple Linear Model has the lowest RMSE value among Support Vector Machine and Random Forest methods. So, according to the data, hoa_tax (Homeowners association Tax), property_tax (Property Tax) and fire_insurance (Fire Insurance Fee) have significant effects on the Total price of the Houses.

## VI. REFERENCES

Global Property Guide. (2018, June 10). Brazilian Home Prices, Graphed as Time-Series. Retrieved from https://www.globalpropertyguide.com/Latin-America/Brazil/Home-Price-Trends