Lighthouse Indicium Data Science Challenge

Statistical Analyses Report

Introduction

This report describes all the results, business insights and answers obtained by statistical analyses conducted using the dataset named 'teste_indicium_precificacao.csv'. The aim of these analyses is to provide a better understanding of data patterns related to short-term rental prices in New York City.

Results

The first step of the statistical analyses was a general inspection of the data provided, including checking for missing values as well as outliers and distribution of numerical data. Outliers' values identified in the variables' 'price' and 'minimo_noites' were excluded using a criterion based on percentile values.

The Figure 1 explores a possible linear relationship between characteristics of the listings such as minimum stay, total number of reviews, number of reviews per month, number of listings per host, and availability. It was not observed a clear linear relationship between any of those characteristic and short-term rental prices. The number of reviews show a tendency to an inverse relationship with price, this could indicate that listings with a higher number of reviews are those with lower prices.

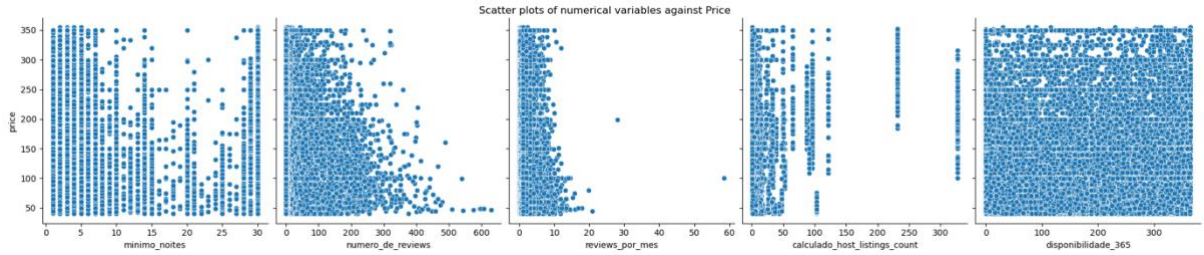

Figure 1. Scatter plots of numerical variables and price.

Similar results were showed in the correlation matrix. None of the variables presented a high correlation coefficient with price (Figure 2).

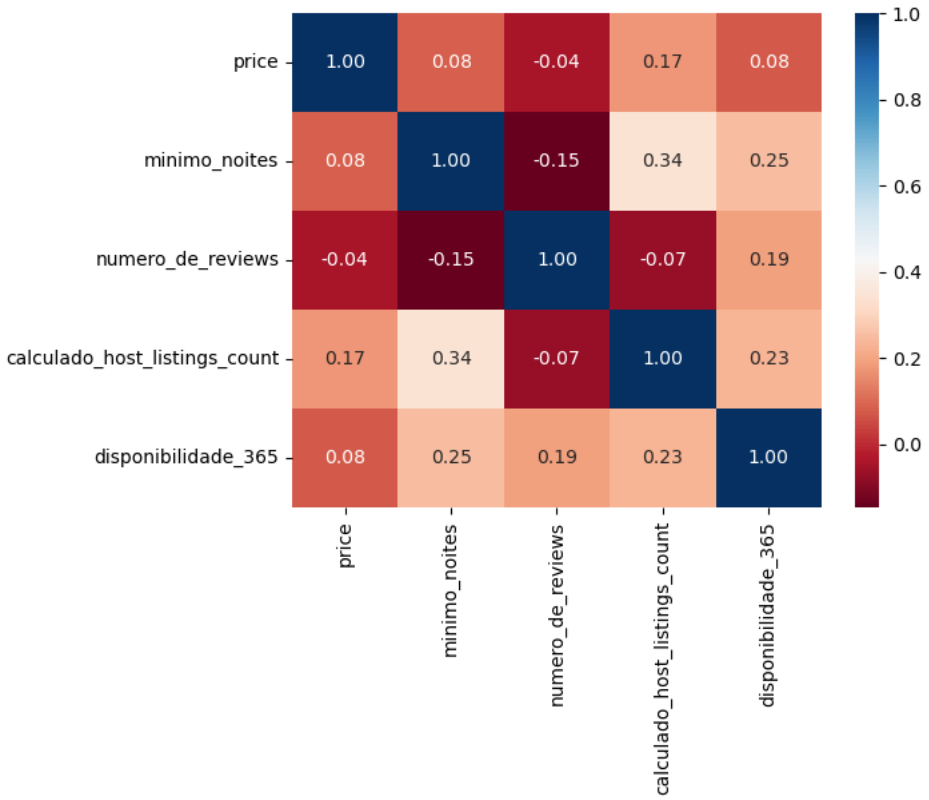

Figure 2. Correlation matrix of numerical variables.

The analysis of short-term rental prices according to room type clearly showed that entire home/apartment listings have the higher prices, followed by private room and shared room (Figure 3).

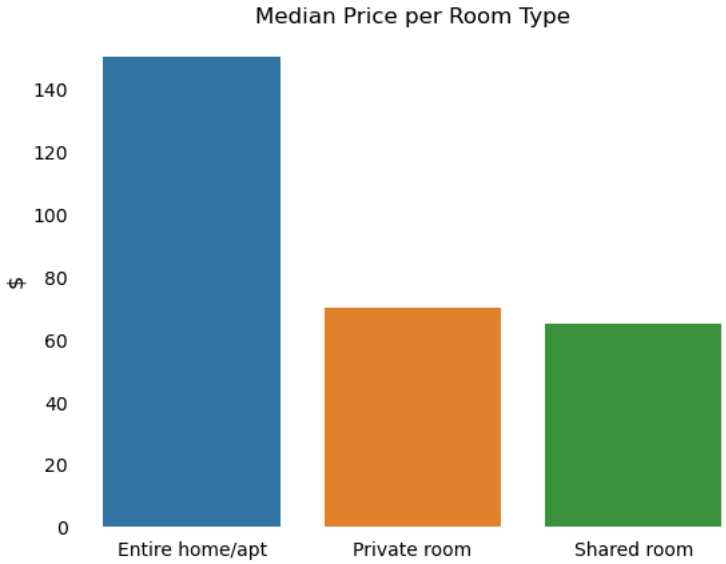

Figure 3. Median short-term rental price according to room type.

Similarly, the analysis of short-term rental prices according to neighbourhood groups showed that Manhattan and Brooklyn have the higher median prices, while Staten Island, Queens, and Bronx have the lower prices when compared to the first two neighbourhoods (Figure 4).

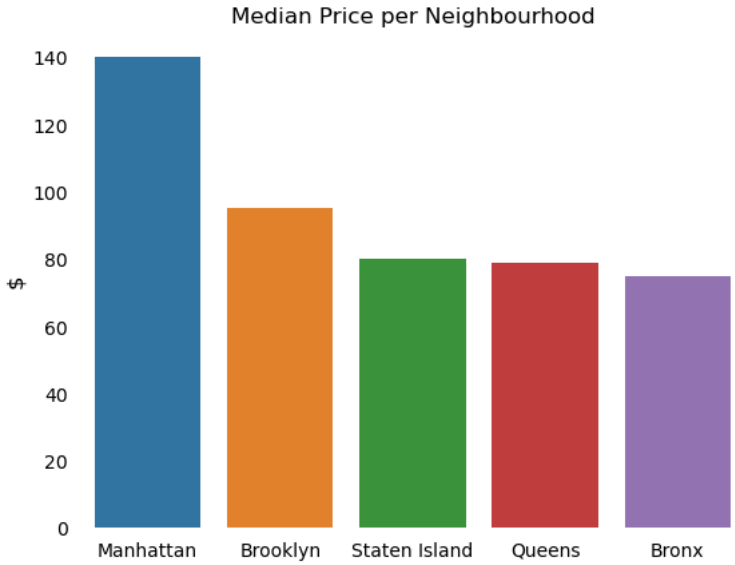

Figure 4. Median short-term rental price across neighbourhhods.

Manhattan and Brooklyn also have the higher number of listings (Figure 5) and predominantly offer listings of entire home/apartment and private room (Figure 6).

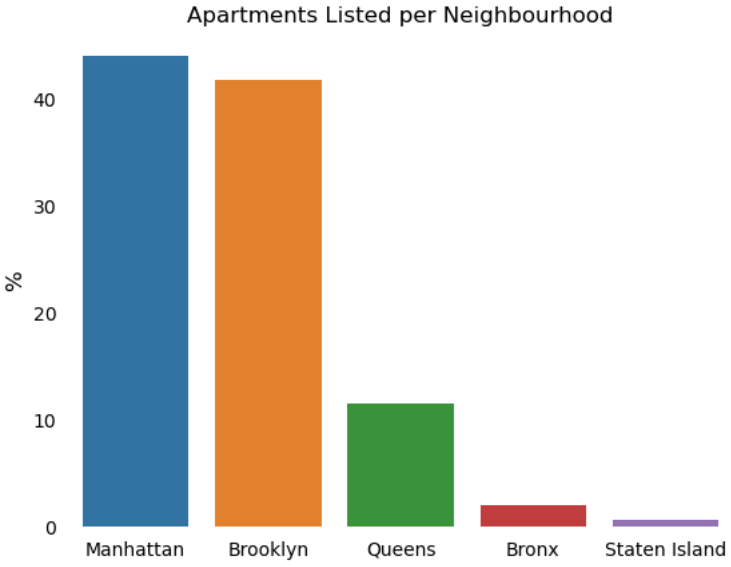

Figure 5. Percentage of listings across neighbourhoods.

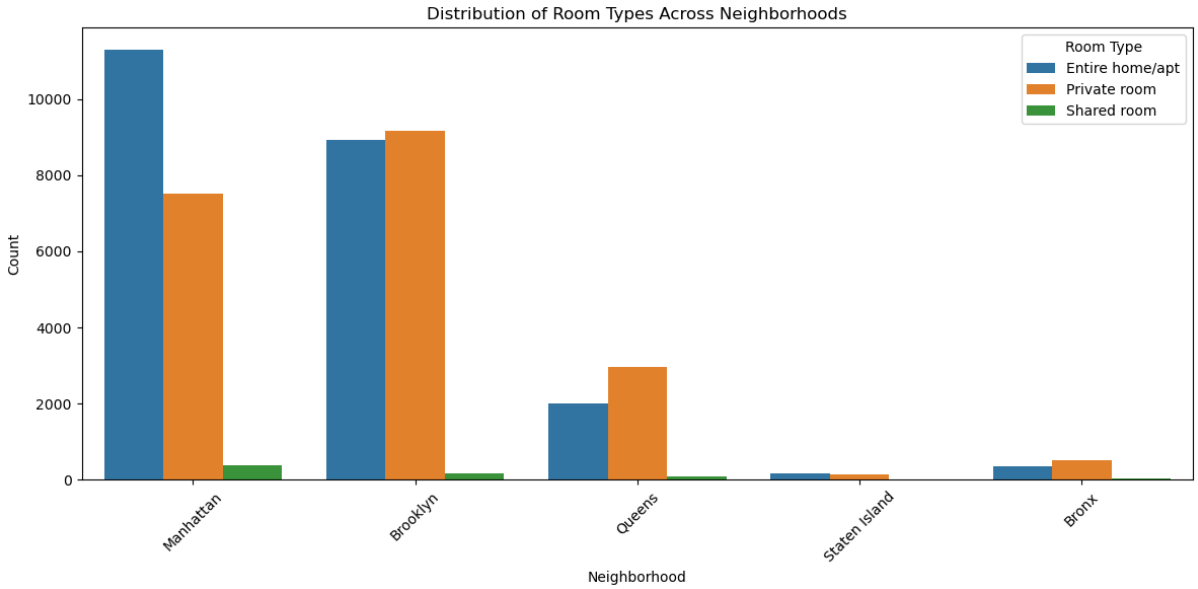

Figure 6. Distribution of room type across neighborhoods.

Analysing the number of reviews and availability according neighborhood, it was observed that Staten Island presented the higher number of reviews and the highest availability. These results arise some hypotheses. This could indicate older properties thar are listed for longer time in the platform, but do not have high demand currently. On the opposite, Manhattan showed the lowest number of reviews and availability. This could reflect that this area has properties listed more recently and a high demand for short-term rent. The same pattern is observed for Brooklyn.

Based on this data analysis, we can affirm that Manhattan and Brooklyn are the neighborhoods with higher demand despite higher prices. This indicate that these neighbourhoods are good options when considering to buy an apartment to list as sort-term rental property (Figures 7 and 8).

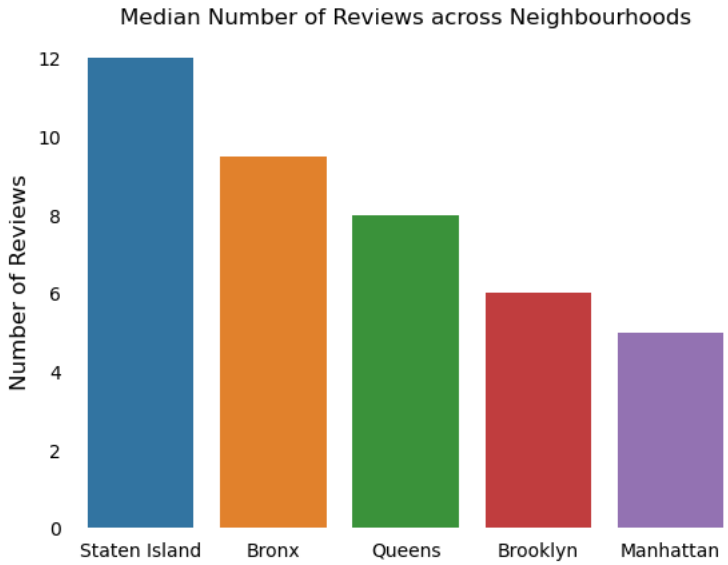

Figure 7. Median number of reviews across neighbourhoods.

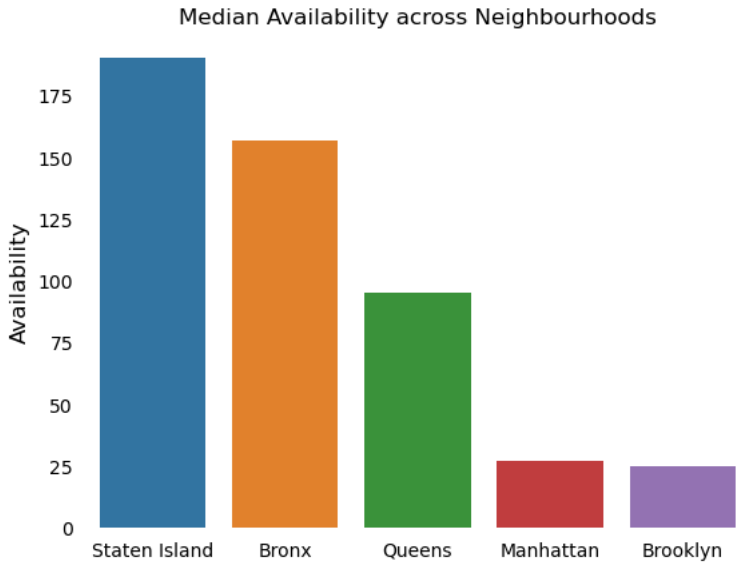

Figure 8. Median availability across neighbourhoods.

Considering the influence of minimum stay and availability of the listing on price, it was not observed a clear relationship between these characteristics and price.

Possible patterns on the listing names in higher price listings were also investigated. The five most common words found were bedroom, apartment, village, luxury, and spacious, reflecting not only listings offering the entire apartment but also the quality of the property. The word cloud presented below shows the 20 most common words found on listing names of properties with prices above the 95th percentile.

Prices Prediction Model – Linear Regression

Prices prediction is a regression problem and a linear regression model was applied. To find the best model fit, few steps were conducted to transform the variables. After the exploratory data analysis, the following features (independent variables) were selected to build the model: 1) 'bairro_group', 2) 'room_type', 3) 'minimo_noites', 4) numero_de_reviews. The neighbourhood and room type available in the listing presented a clear relationship with short-

term rental prices, while minimum stay nights and number of reviews did not present a clear relationship with price, they were selected considering the characteristics of short-term rental business.

Variables´ transformation

The numerical variables were standardized to the same scale using the function StandardScaler. This procedure was conducted to ensure certain features won't dominate the model due to their magnitude. The categorical variables were transformed into numerical variables using the function LabelEncoder, considering that linear regression model only takes numerical inputs into consideration. Regarding the transformation of the dependent variable, two models were tested. The first one with the original price values, and the second with the log transformed price values.

To evaluate the models, it was applied the R^2, and the Root Mean Squared Error (RMSE). The first model with the original price values showed a R^2 of 0.35 and a RMSE of 57.8 dollars. The second model used the log transformed price values and showed a R^2 of 0.42 and a RMSE of 58.7 dollars. The second model was chosen as the best one, considering the improvement in the R^2 value and the assumptions of linear regression with a dependent variable with normal distribution.

Considering the final model and the listing features presented in the challenge, the price prediction was 138 dollars.

Conclusion

Even though the final model presented a satisfactory evaluation, additional data could be used to improve the price explanation. Size of the property, seasonality, and reviews score could be useful information to improve the prediction of short-term rental prices in NY city.