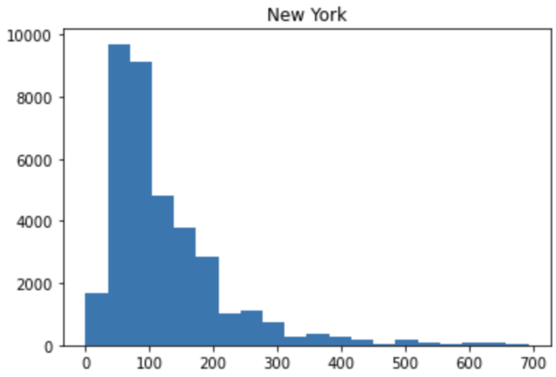
ORIE 5741 Midterm Report

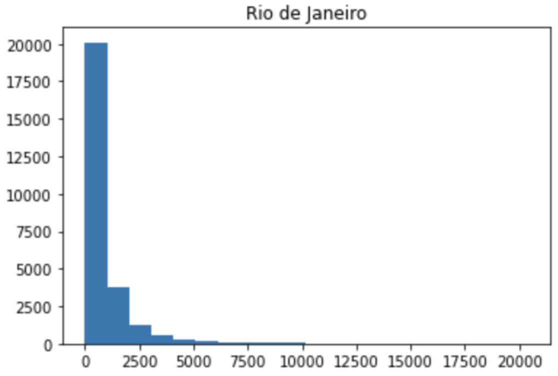
Jiongjiang Duan(jd2253), Sirawich Tippawanich(st895), Jiaqi Zhang (jz2232)

October 28, 2021

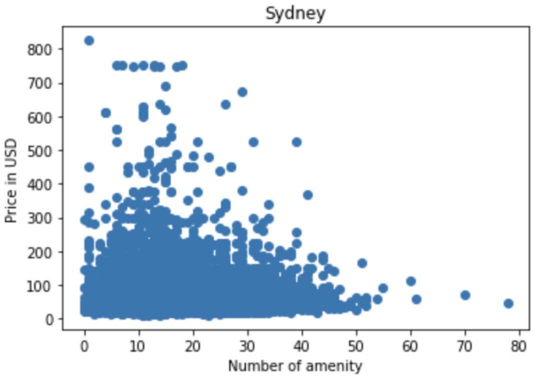
1 Data Exploration and Visualization

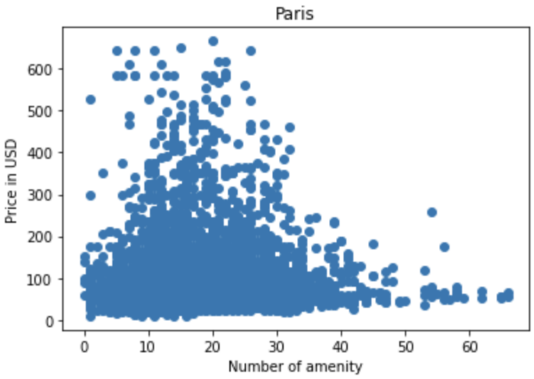
Histograms are a powerful tool in representing the distribution of a variable within a data set. Plots of Airbnb listing prices against their frequencies show that prices follow right-skewed distributions for all cities within the data set. Some cities feature price distributions where the modes are at the lowest price bin (Rio De Janeiro, Cape Town, Mexico City and Istanbul) . Paris, New York, Bangkok , Sydney, Rome and Hong Kong feature distributions where the lowest price bin was not the mode.





The relative differences in pricing is also worth noting. The average listing prices for all cities are between $20 to $100, suggesting that they are within the same order of magnitude. As expected, higher averages are associated to cities in countries with higher costs of living. As a product that provides visitors with lodging spaces, available amenities at the properties can be an influential factor in determining how prices are set. Number of amenities for each property was extracted from the lists of amenities within each listing and plotted against prices. There is a clear observation across all cities that there are a few low-price listings who provide high number of amenities. The most expensive listings in each city provide between 10 - 30 amenities. The trend suggests that some lower-priced hosts try to create a draw towards value by listing longer lists amenities.





2 Feature Engineering

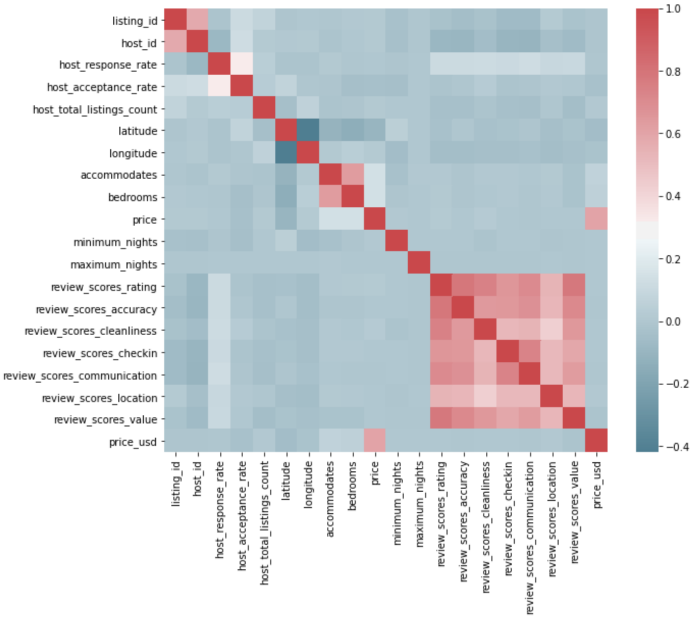
2.1 Dealing with Missing Values

There are 279,712 observations in the Airbnb listings where 128,617 observations has at least one null value in the variables. The goal of the project is to investigate whether the review has an impact on the price of listing, so we only kept observations whose review score rating is not null. After that we used methods below to deal with other null values:

1. Delete columns which has more than 80 percent missing values: district
2. Replace the missing values with the mean: bedrooms
3. For the informative missing values such as response rate and acceptance rate, we first replace the null values with 'unknown', treating them as a categorical value and then transform into dummy variables.
4. Group the values into bracket and treat the null value as an categorical value. For example, both host response rate and acceptance rate are skewed to the right. If the rate is smaller than 0.3 in the host response rate column, it's a low response rate. If it's between 0.3-0.7, it's a medium response rate. If it's over 0.7, then it's a high response rate. The null value in host response rate is still null but is in a new bracket called unknown response rate.

The total number observations after dealing with the missing values is 188,199, which is large enough for regression modeling.

2.2 Correlation Analysis



Correlation Map Among Features

To run a regression analysis, the features need to be independent otherwise it will run into the multicollinearity issue. From the correlation map above, the review score rating is closely related to review score accuracy, review score cleanliness, review score check-in, review score communication, review score location and review score value. Therefore, we deleted all review score related features except review score rating. The longitude and latitude have high correlation as well but since city can also represent geographical information, we deleted the longitude and latitude.

3 Preliminary Linear Regression and Analysis

A preliminary linear regression was performed with all the processed features as inputs and price (in USD as output). The entire data set was used to fit the model to gauge the performance. As an evaluation metric, mean squared error was computed between predicted and real listing price. The preliminary model with all features provided a training error of 37,750. The high mean squared error suggests that simple linear regression is not capable of capturing the relationship between the features and predicted variable.

The p-value of each variables in the regression model was examined. Notably, no statistical significance was observed in variables such as maximum nights and the binary variables representing the city locations of New York and Sydney. Other features show p-values lower than a universally accepted threshold of 0.05, suggesting that they have significant relationship with the listing price.

Examination of coefficients for categorical variables provide some hints about their impact on pricing. There is no meaningful difference in the coefficients for the variable indicating whether a host is a ”super host” or not.

The coefficients of the regression model for numerical variables are also in line with expectation. The number of bedrooms and people that listings accommodate have positive coefficients in magnitudes that agree with how much price scales with capacity of accommodations. Host experience, measured as days since initial host registration, also has a positive coefficient. As mentioned earlier, counts of amenities in the listings are spread out among the range of pricing. The coefficient for count of amenities is positive and small in magnitude, suggesting that although the number does influence pricing positively, the effect is small to negligible. A positive coefficient for review scores rating (the overall score given in reviews) hints that higher prices are associated to higher scores given by visitors.

Fitting a linear regression to a data set with numerous one-hot encoded features is far from ideal. However, it enables primitive analyses of coefficients and statistical significance for some (mostly numerical) variables. A benchmark error performance metric was also achieved from the linear model. A natural next step is to implement non-linear models in order to further explore the relationship between individual features and listing pricing.

4 Forward Plan

So far we have completed exploratory analysis, feature engineering and a linear regression model. In the next, we will first add a L1 regularization to prevent over-fitting and selectively assign less weight to unimportant features. To deal with the multi-collinearity of dummy variables, we will use one-hot encoding technique. We can also experiment with decision tree and XGBoost models to see which feature has the most drop in metrics such as Gini index (impurity). After running all the models, we will do post-model exploratory analysis to see how effective the review system will change the price. In addition to the review scores, we will also capture other factors that impact he price system. For example, we can break down the model by region and property type to analyze whether the location and property will impact the price of Airbnb listings.