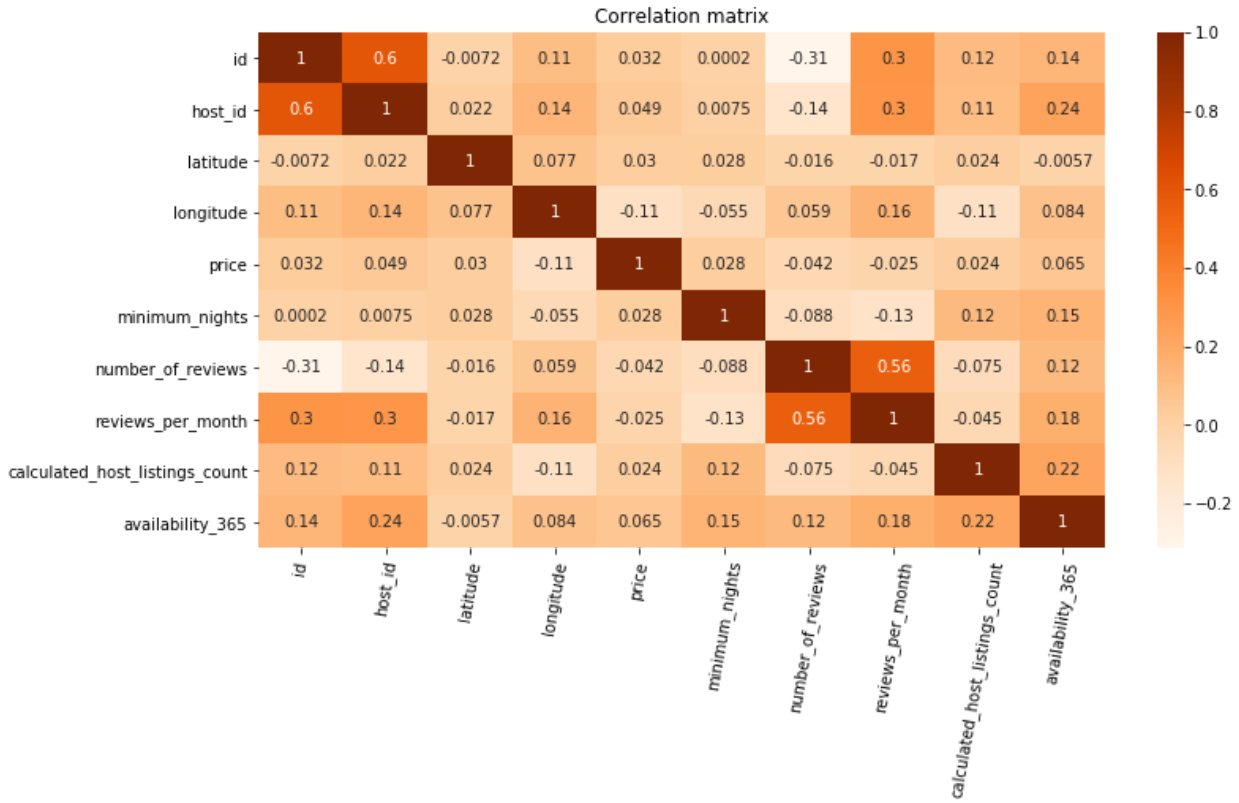
Data Wrangling: Estimating rental price for lodging, homestay or tourist houses in New York City

Capstone Project -1

By Muluemebet Ayalew
March, 2020

Data collection and Feature engineering

The main data about rental listing is downloaded from the web, Inside Airbnb[1], and contains information about host, location, room type, price, review and availability. The objective is to predict the price of rental from relevant predictor variables. However, the features available from this dataset are not highly correlated with price(see the heat map below). Therefore, other data sets have been looked at and thorough feature engineering was done.



For example, subway entrance location data, geographic location of some point of interest such as times square, Penn Station and similar data was collected from the web. Then, the distance

[1] Dec, 2019, "*Get The Data"*, Inside Airbnb, Accessed Jan 30, 2020. http://insideairbnb.com/get-the-data.html

from each listing to selected point of interest and to the nearest subway were computed and added as new predictor features.

Other feature engineering tasks such as feature encoding for categorical variables and new labeling were done. The data contains information about room type, Whether it is the entire home/apartment, private room, shared room or hotel room. However, there are hotel listings that are not specified as hotel rooms. In some cases, they are listed as private rooms and in other cases entire homes/apartments. A new variable is created to distinguish a listing as hotel or non hotel. To do this, hosting names containing the word hotel and listings listed as hotel rooms are labeled as hotel (encoded as 1) and the rest as none hotel(encoded as 0). So that we can have deeper granularity.
But this data has to be verified from Airbnb if they have detailed information about the host.

Data cleaning

The listing data contains some missing values(see Table 1).There are four variables with missing values. Among these, host name and list name were not considered relevant for price prediction and were ignored. Missing values for the variable reviews per month was filled by zero. Missingness for this variable was due to the fact that the number of reviews for those listings was in the first place zero.

Table 1: Number of missing values by variable name

Variable Name	**Number of missing values**
name	17
host_name	563
last_review	10220
reviews_per_month	10220

The price and other numerical variables such as “minimum nights” are highly skewed. The price also has zero value. One reason could be the listings were inactive and intentionally left as zero price. There could also be other reasons. In any case, including the zero price in the model might be misleading, therefore only the non- zero priced listings were considered for the analysis . Outliers are also handed by taking only data points within three standard deviations from the mean.

After all the collection ,data cleaning and feature engineering tasks were done, the cleaned data was saved as a csv file for the next step of the analysis. The cleaned data was also randomly splited for training(80%) and test(20%) data and saved as separate files.