***PREDICTING THE PREICES OF AIRBNB IN NEWYORK – TRAN TUAN VU***

1. Introduction
   1. Background

If for some reason you don’t already know, Airbnb is a internet marketplace for short-term home and apartment rentals. It allows you to, for example, rent (list) out your home for a week while you’re away, or rent out your empty bedroom. One challenge that Airbnb hosts face is determining the optimal nightly rent price. In many areas, renters (hosts) are presented with a good selection of listings and can filter by criteria like price, number of bedrooms, room type, and more

* 1. Problem

Airbnb pricing is important to get right, particularly in big cities like London, New York, Tokyo,…. where there are lots of competition and even small differences in prices can make a big difference. It is also a difficult thing to do correctly — price too high and no one will book. Price too low and you’ll be missing out on a lot of potential income. This project aims to solve this problem, by using machine learning and deep learning to predict the base price for properties in New York.

* 1. Interest

Obviously, Airbnb teams would be very interested in accurate prediction of player improvement, for competitive advantage and business values. Others who are interested in rent price in Airbnb such as traveler, business man, student, supplier,…

1. Data acquisition and cleaning
   1. Data sources:

The dataset used for this project comes from Kaggle (<https://www.kaggle.com/dgomonov/new-york-city-airbnb-open-data>). Since 2008, guests and hosts have used Airbnb to expand on traveling possibilities and present more unique, personalized way of experiencing the world. This dataset describes the listing activity and metrics in NYC, NY for 2019. This public dataset is part of Airbnb, and the original source can be found on this website (<http://insideairbnb.com/>)

This dataset has around 49,000 observations in it with 16 columns and it is a mix between categorical and numeric values.

* 1. Data cleaning

Data downloaded from public dataset of Airbnb. The data is quite messy, and has some limitations. The major one is that it only includes the advertised price (sometimes called the ‘sticker’ price). The sticker price is the overall nightly price that is advertised to potential guests, rather than the actual average amount paid per night by previous guests. The advertised prices can be set to any amount by the host.

Each row in the data set is a listing available for rental in Airbnb’s site for the specific city (observations). The columns describe different characteristics of each listing (features).

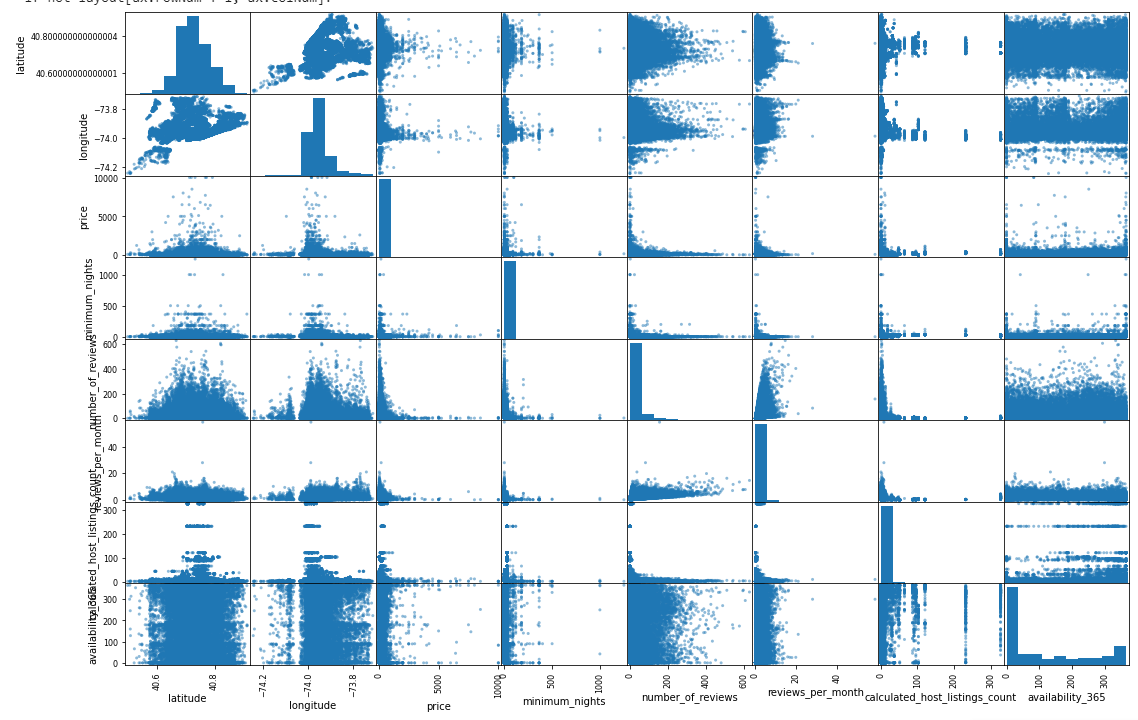
There was 4 features which had missing values. It includes : “name, host\_name, last\_review, review\_per\_month”. Because name and host\_name are account for ~ 0,04% in data, I removed the rows which has these columns. Otherwise, we can look into the nature of our dataset we can state further things: columns “name” and “host\_name” are irrelevant and insignificant to our data analysis, columns “last\_review” and “review\_per\_month” need very simple handling. To elaborate, “last\_review” is date; if there were no reviews for the listing – date simply will not exist. In our case, this column is irrelevant and insignificant therefore appending those values is not needed. For “review\_per\_month” column, we can simply append it with 0.0 for missing values; we can see that in “number\_of\_review” that column will have a 0, therefore following this logic with 0 total reviews there will be 0.0 rate of reviews per month. Therefore, let’s proceed with removing colunms that are not important and handling of missing data. Last\_review and review\_per\_month are account for ~ 20% in data.

* 1. Feature selection

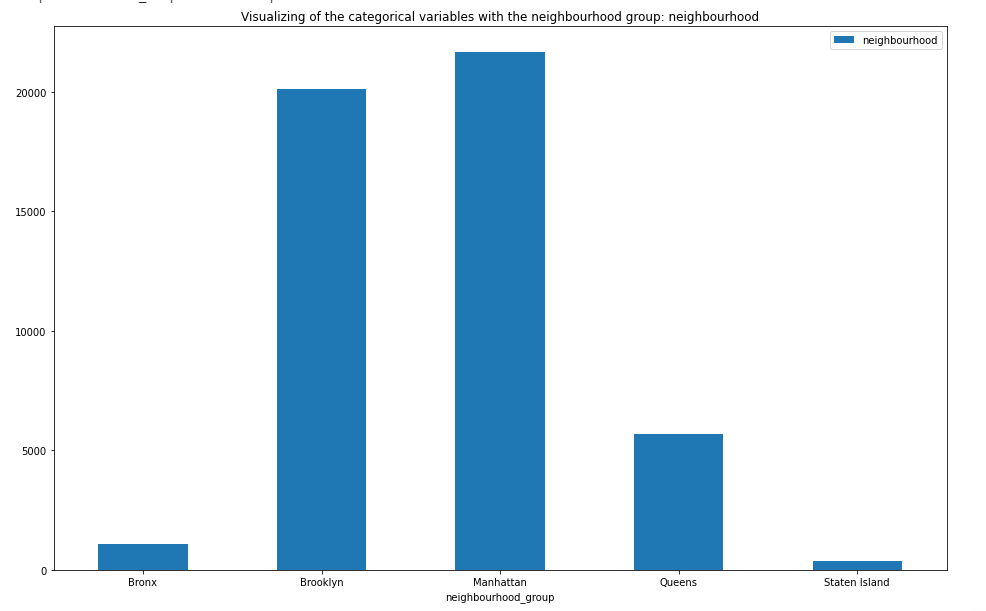
There were 48895 samples and 16 feature in the data. Upon examining the meaning of each feature, it was clear that there was some redundancy in the features. For example, there was a feature of the name, and another feature of the host name. These two features contained very similar information (determine for hotel/motel). It is insignificant and ethical reasons when we were in process building model. There should be no reasoning to continue data exploration and model training (which we will be doing later) towards specific individuals based on their names. Why is that? Those names are assigned to actual humans, also they present no security threat or military/governmental interest based on the nature of the dataset, therefore names are unimportant to us. The host\_id feature is the same with this reason.

After discarding redundant features, I inspected the correlation of independent variables, and found several pairs that were highly correlated (Pearson correlation coefficient > 0.9). From these highly correlated features, only one was kept, others were dropped from the dataset. After all, the features were kept include: 'neighbourhood\_group', 'latitude', 'longitude', 'room\_type', 'price', 'minimum\_nights', 'number\_of\_reviews', 'reviews\_per\_month', 'calculated\_host\_listings\_count', 'availability\_365'.

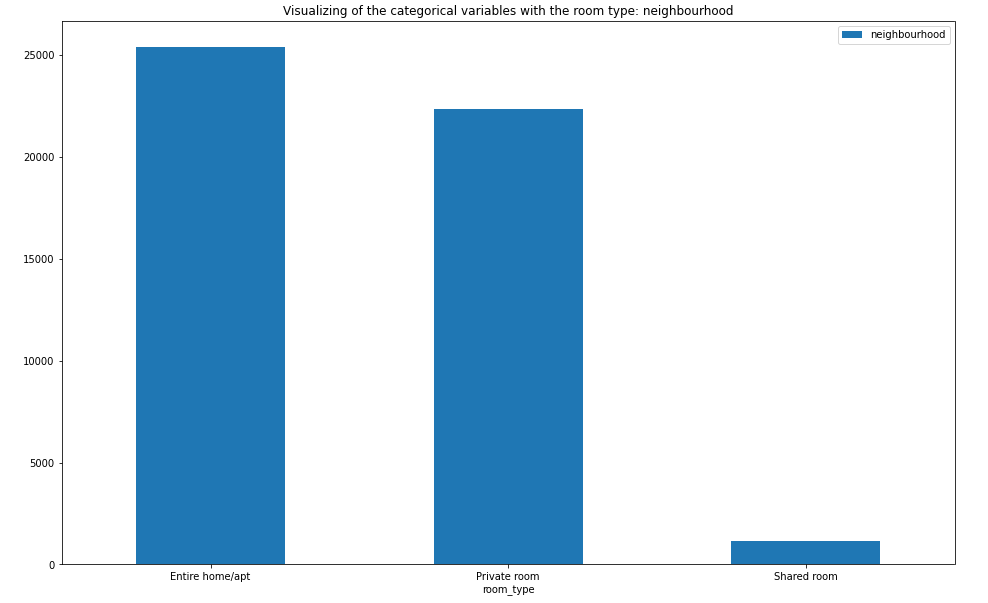
1. Exploratory Data Analysis
   1. Scatter Matrix with Continuous variables



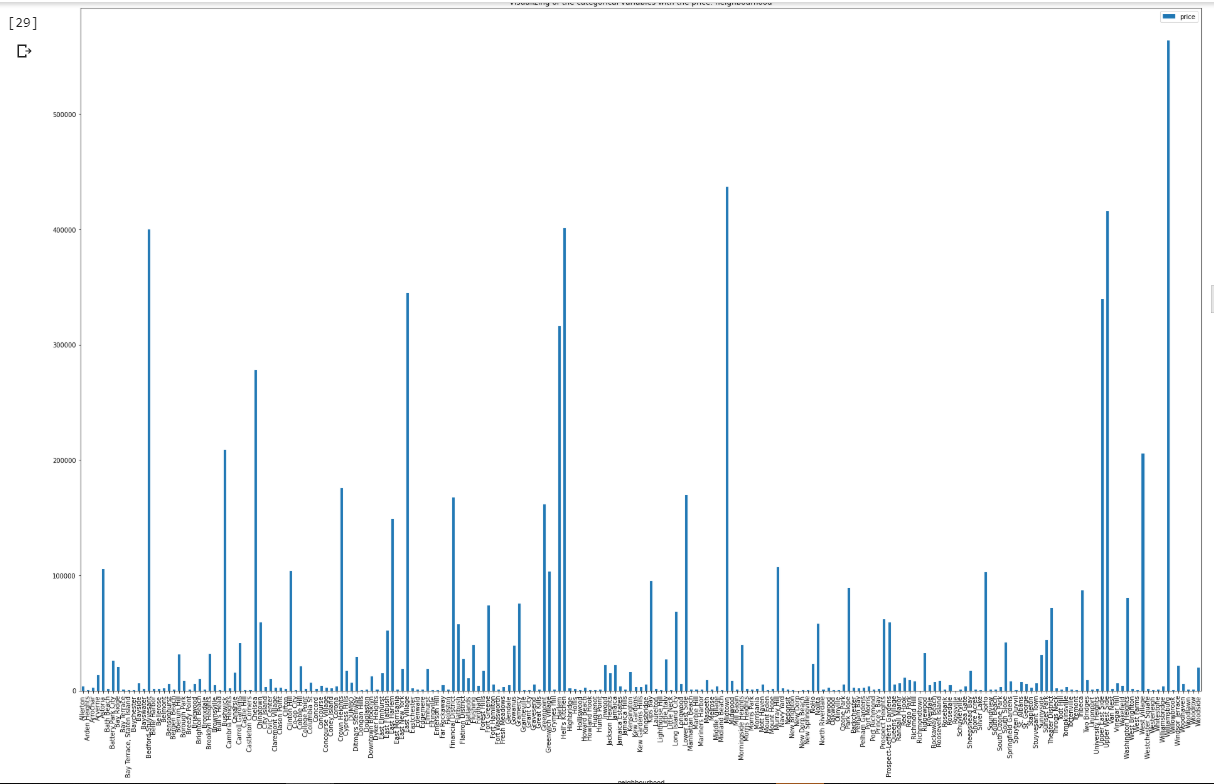
* 1. Visualization total price with categorical variables
     1. Price and Neighbourhood group

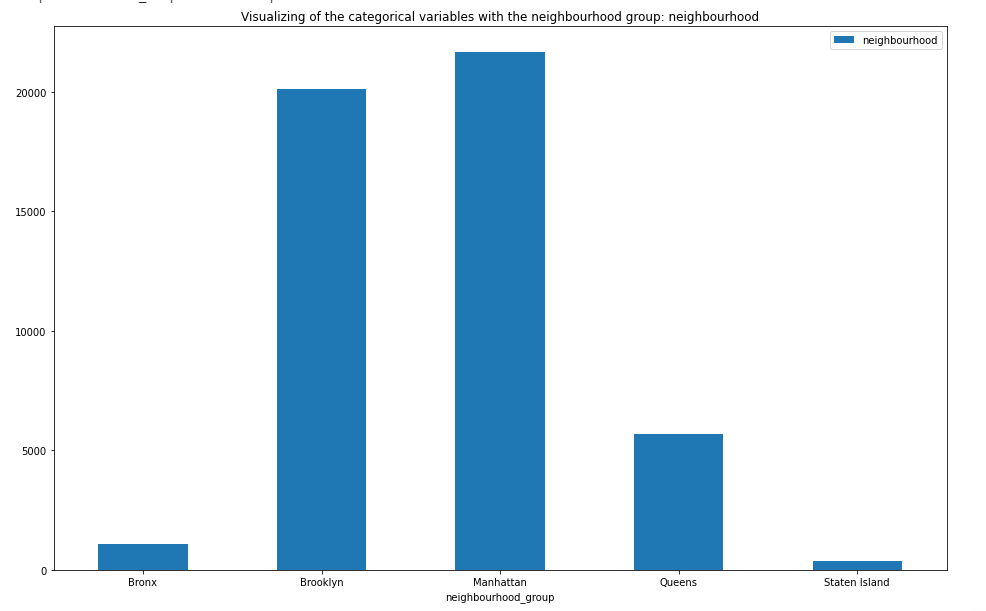


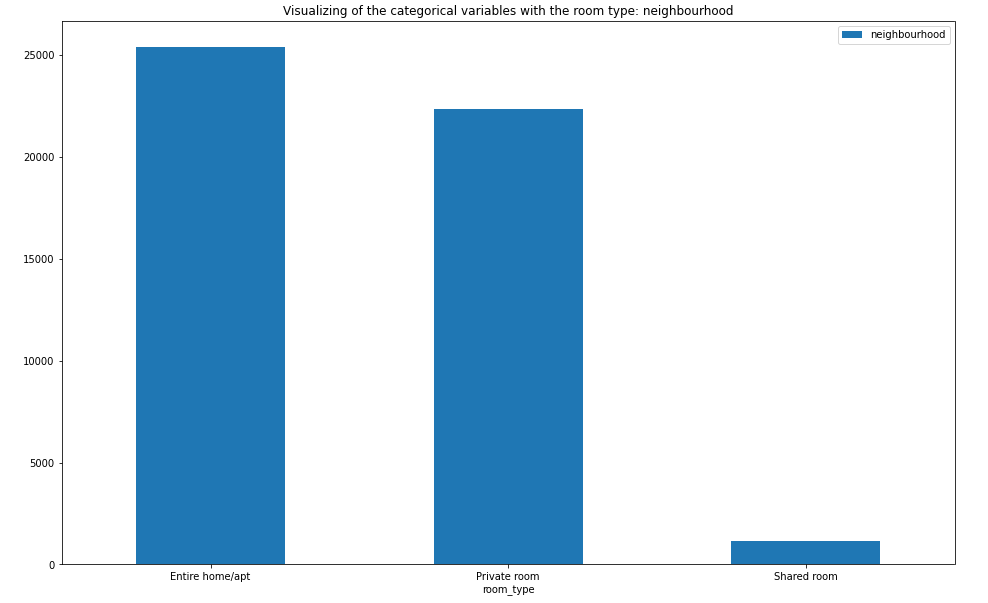
* + 1. Price and Room type



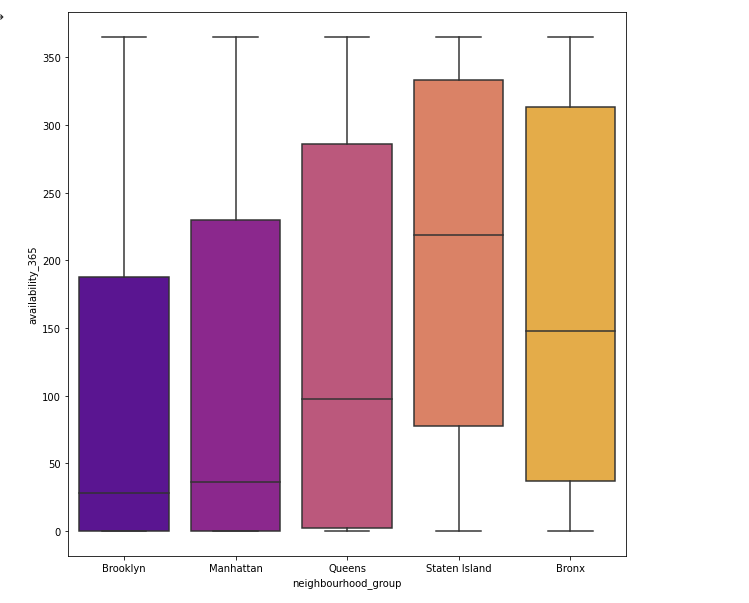
* + 1. Price and neighbourhood



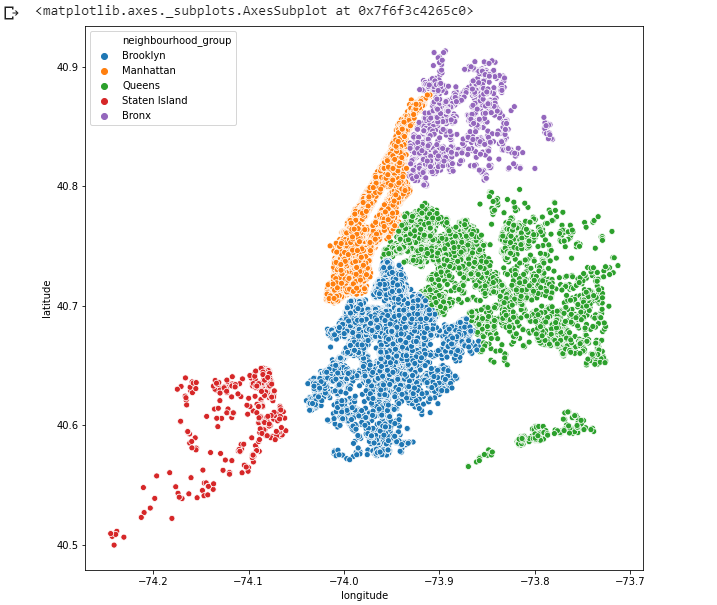
* 1. Plotting Neighbourhood group with Neighbourhood 
  2. Plotting Room Type with Neighbourhood



* 1. Visualizing bar plot between neighbourhood\_group and availablity of room



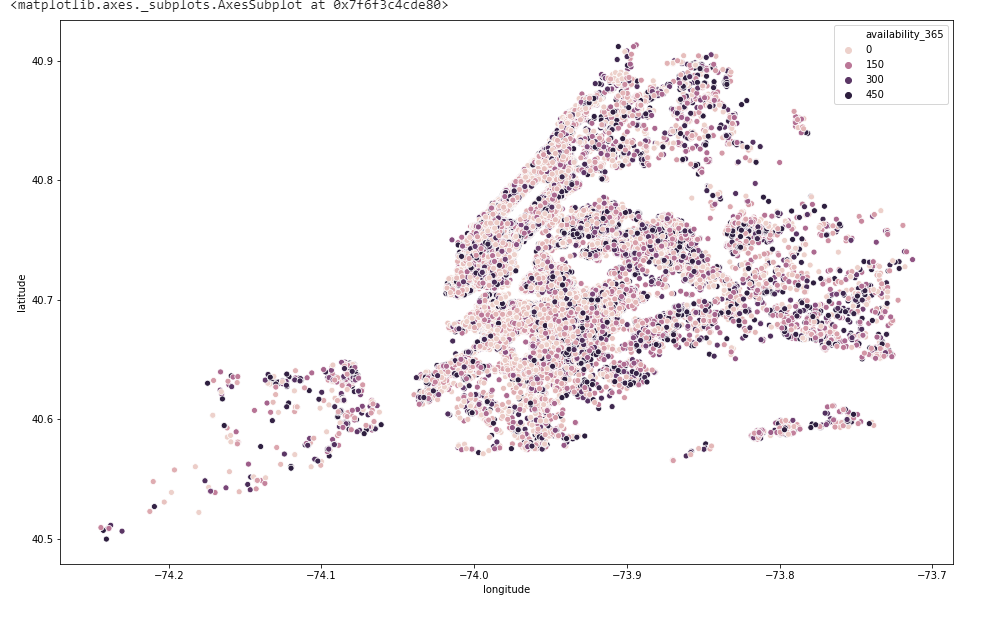
* 1. Map of Neighbourhood Group



* 1. Map of Neighbourhood



* 1. Availability of Room



1. Predictive Modeling

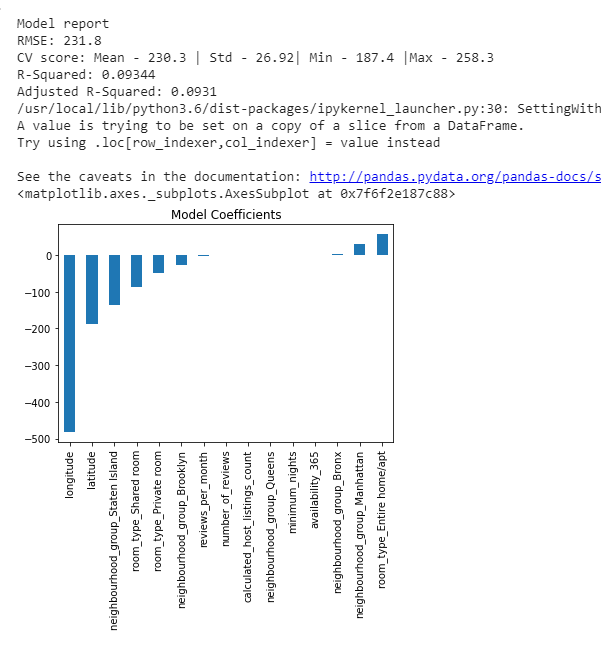
There are two types of models, regression and classification, that can be used to predict player improvement. Regression models can provide additional information on the amount of improvement, while classification models focus on the probabilities a player might improve. The underlying algorithms are similar between regression and classification models, but different audience might prefer one over the other. In this case, I used regession models.

I applied linear models (linear regression, Ridge regression, and Lasso regression), random forest to the dataset, using root mean squared error (RMSE) as the tuning and evaluation metric. The results all had the same problems.

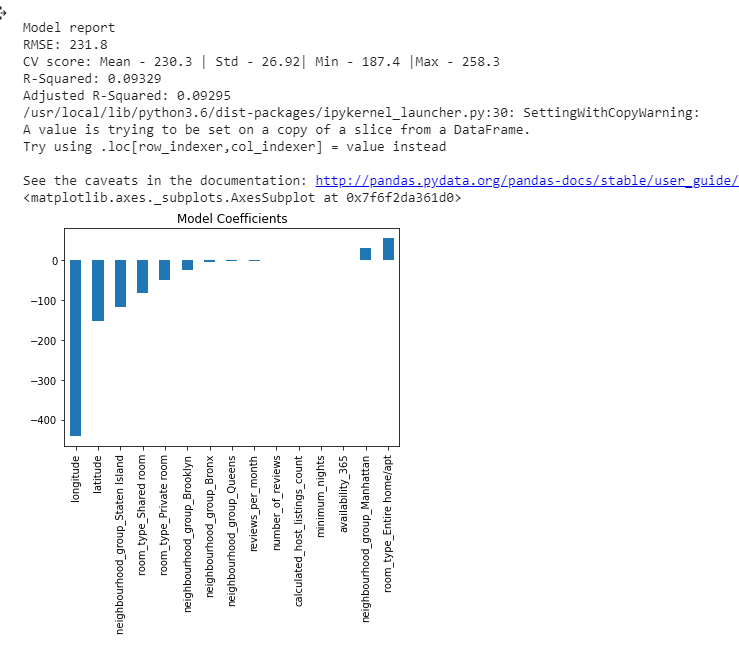
Performances of different models showed in the table below:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Linear Regression | Rigde regression | Decision tree | Random forest |
| Weighted RMSE | 231.8 | 231.8 | 223.6 | 226.6 |

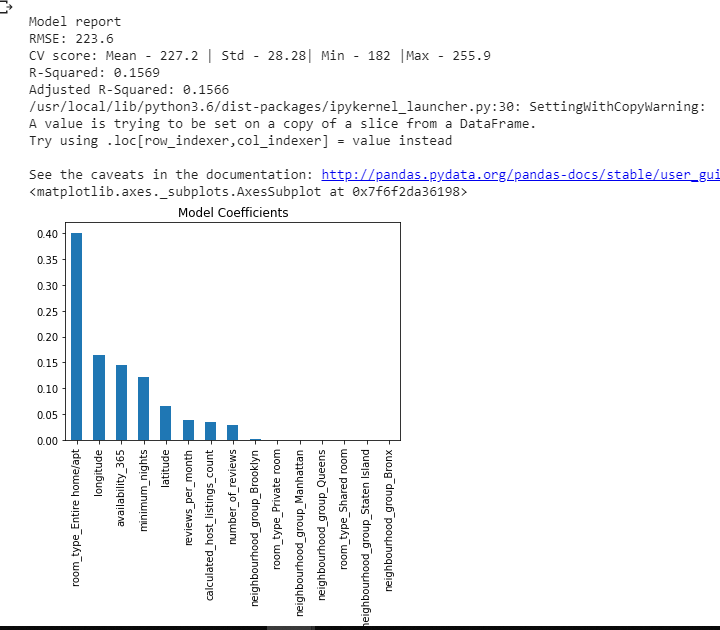
* 1. Linear Regression



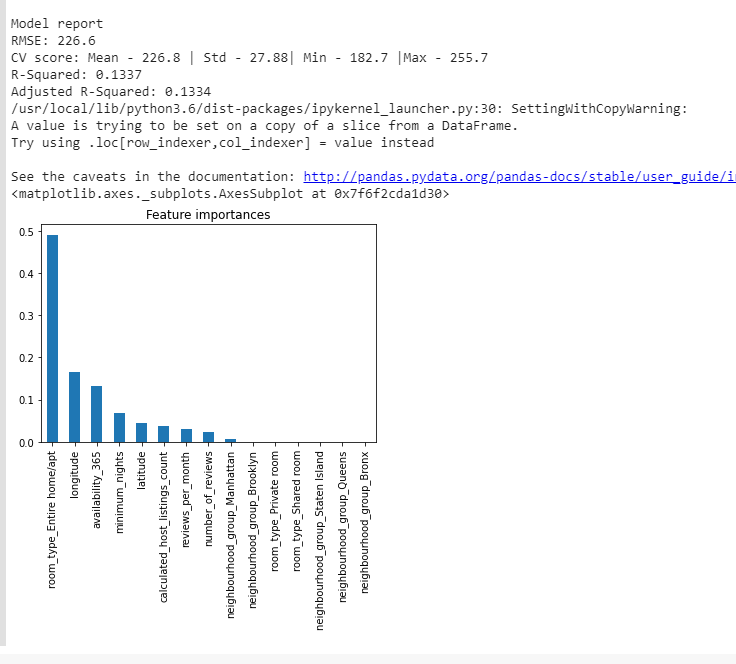
* 1. Ridge Regression



* 1. Decision Tree



* 1. Random Forest



1. Conclusion

This Airbnb ('AB\_NYC\_2019') dataset for the 2019 year appeared to be a very rich dataset with a variety of columns that allowed us to do deep data exploration on each significant column presented. First, we have found hosts that take good advantage of the Airbnb platform and provide the most listings. After that, we proceeded with analyzing boroughs and neighborhood listing densities and what areas were more popular than another. Next, we put good use of our latitude and longitude columns and used to create a geographical heatmap color-coded by the price of listings. Further, we came back to the first column with name strings and had to do a bit more coding to parse each title and analyze existing trends on how listings are named as well as what was the count for the most used words by hosts. Lastly, we found the most reviewed listings and analyzed some additional attributes.

This is one of those situations where machine learning simply is necessary for prediction, and a machine learning model performs just as well.

However, even in the best performing model, the models was only able to explain around 0.09% to 0.15% with R-score.