Airbnb Listing Revenue Analysis

Akshat Kumar
ak1648@rutgers.edu

May 5, 2020

1 Introduction

Most analysis on Airbnb revolves around finding the perfect listing for travellers or sentiment analysis or general exploratory data analysis. This project project takes a step further in this analysis and rather than thinking from a customer or host point of view, it would provide analysis from a business point of view for Airbnb. The project evaluates the and measures the financial performance of a listing by taking into account several factors such as location, price, reviews, availability besides other factors. This will be beneficial for Airbnb as it can reward, penalize or remove the listing as per it's financial performance and it will be great for the host to review their performance and take steps to improve it. The project will use the New York City Airbnb Open Data which describes the listing activity and metrics in New York City. This data file includes all needed information to find out more about listings, hosts, geographical availability and for our predictions.

You can find the code and data of the project from the Github repository mentioned below.

https://github.com/akshatism/data-wrangling-final-project

2 Dataset

The NYC Airbnb Open dataset contains 16 columns and 48895 unique values and the full data can be obtained from http://insideairbnb.com/. I wanted to take the Airbnb dataset from two different sources that is one the listing dataset and the other one is reviews dataset but I got the dataset which had a combination for both of them and required subsequent amount of cleaning. It contains really important features which are listing ID, name of the listing, host ID, name of the host,location, area, latitude coordinates, longitude coordinates, listing space type, price in dollars, amount of nights minimum, number of reviews, latest review, number of reviews per month, amount of listing per host, number of days when listing is available for booking

3 Data Preparation

First I have introduced three important variables into our problem which is really important for our case. These are

- Price Group - KPI for classifying based on price
- Usage group - Listings use based on the number of reviews
- Total Revenue - product of price, reviews and nights and the decision class for our problem

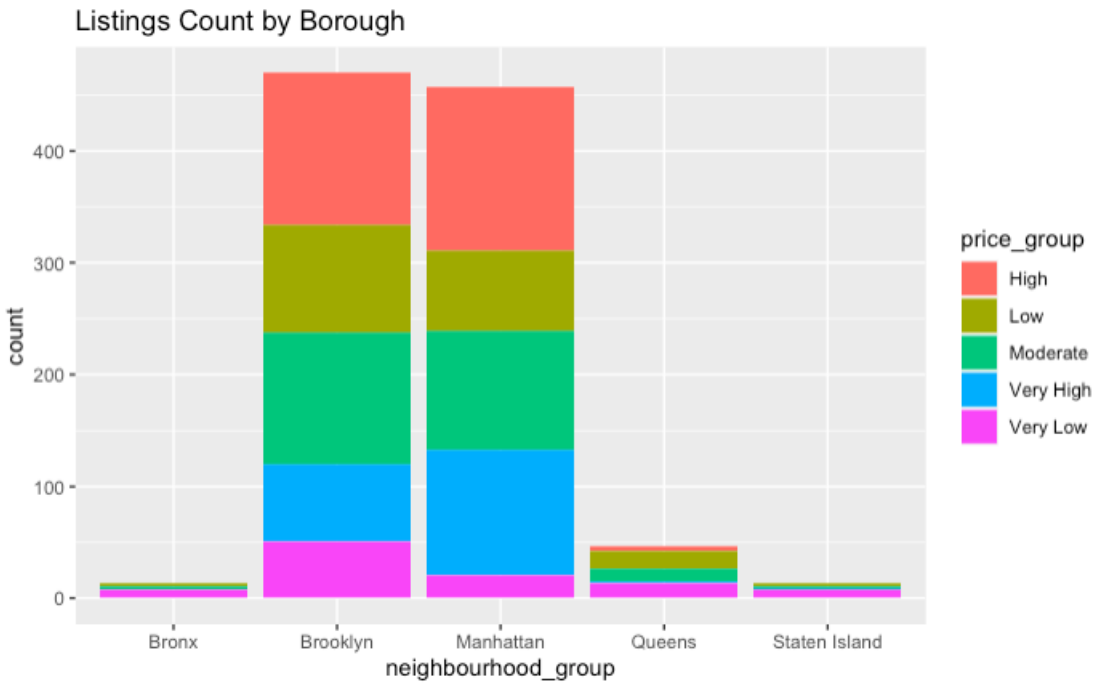
Then in order to get a clean corpus for further analysis we need to clean or mine the reviews. So we will do the following steps for that

After this we move on to the exploratory data analysis (EDA) or data visualization part which will extract meaningful information for our problem.

- Remove unwanted characters
- Convert all words to lowercase
- Filter unnecessary large sentences
- Remove stopwords that neither add positive or negative impact
- Lemmatization that is converting words to their root form

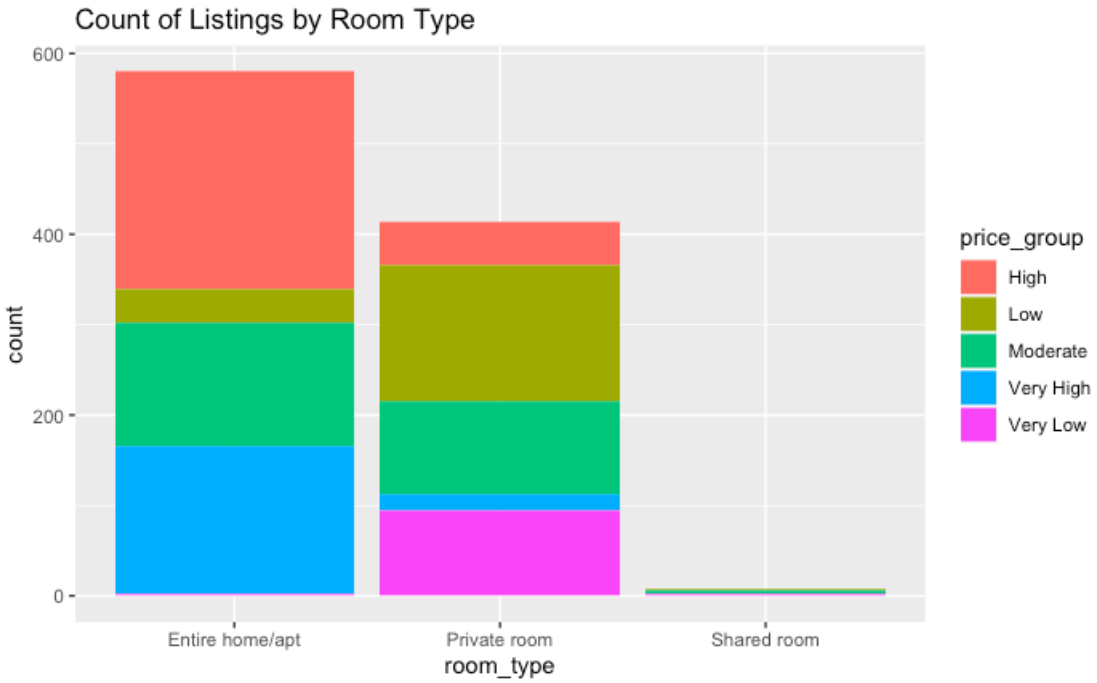
Then comes the final preparation step to convert into term document matrix for Exploratory Data Anaysis step

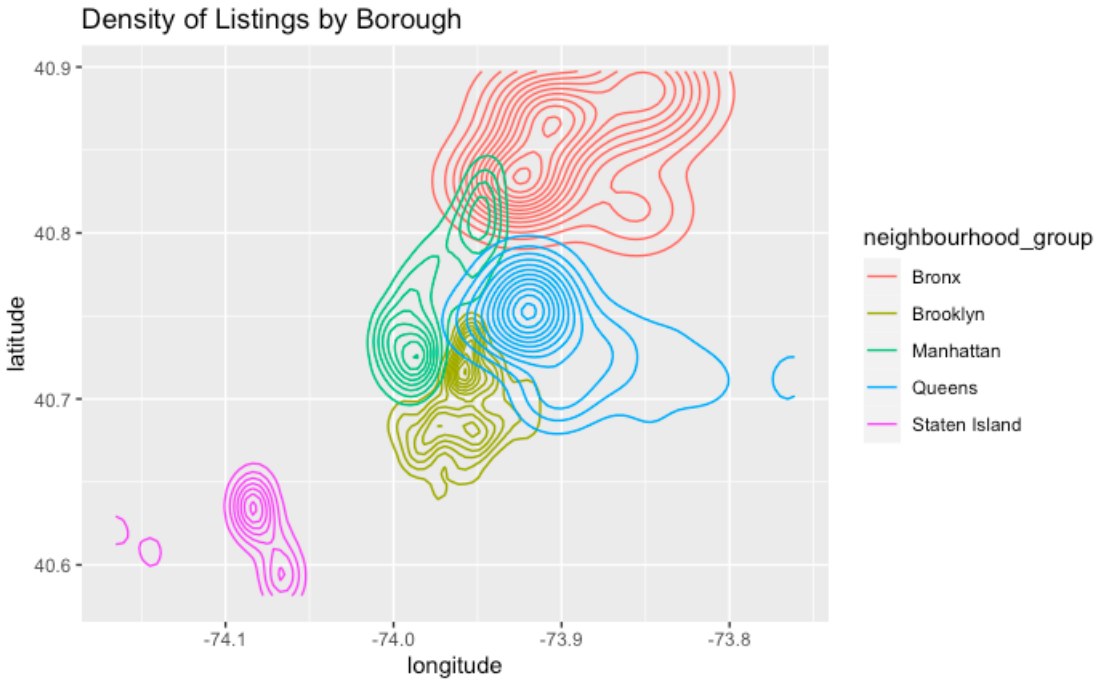
4 Analysis



Manhattan contains the most amount of listings; additionally, it looks like it is the most expensive. Next,Brooklyn appears to be next popular; however,with a more reasonable distribution of price listings. High - Very High seem to only take up 20% of the population. Queens has only 5000 listings, and appears very cheap. Lastly, the Bronx and State Island do not appear very popular for Airbnb.

We can see from the count of listings by room type graph that entire home/apt have high price range and shared have generally low range for all the price groups with private in the middle.

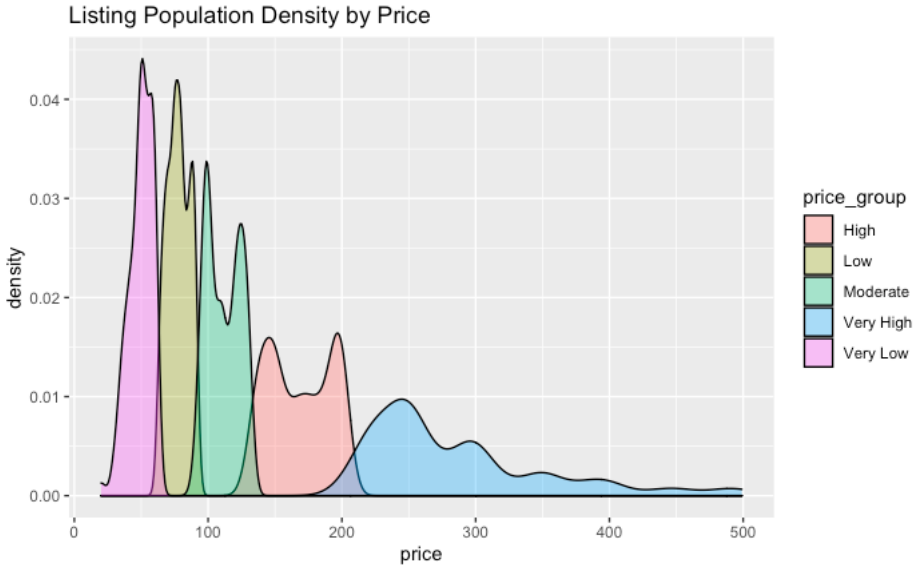

Count of Listings by Room Type
600
400
200
0
count
Entire home/apt
Private room
Shared room
room_type
price_group
High
Low
Moderate
Very High
Very Low


Density of Listings by Borough
40.9
40.8
40.7
40.6
latitude
-74.1
-74.0
-73.9
-73.8
longitude
neighbourhood_group
Bronx
Brooklyn
Manhattan
Queens
Staten Island

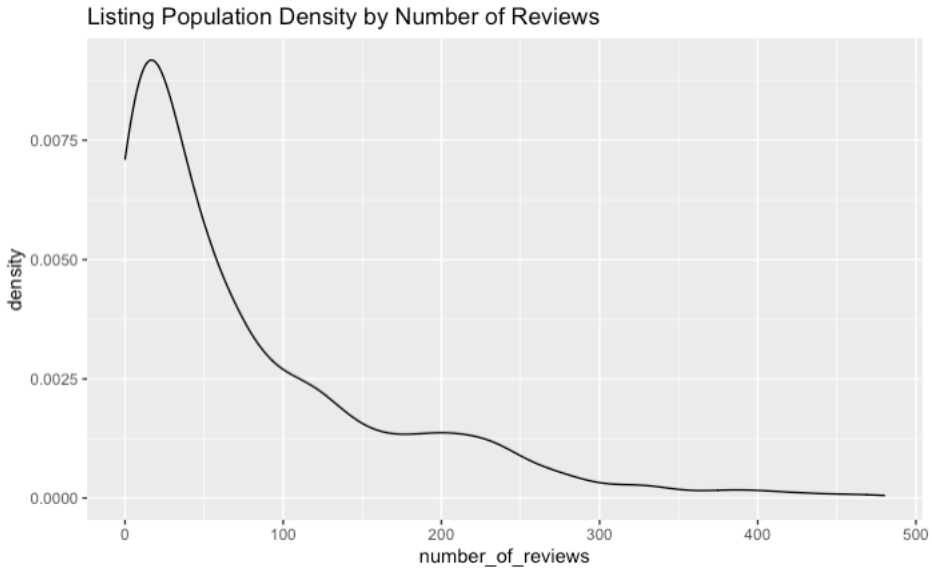
Now we can focus on the density of the listings with respect to latitude and longitude and see the distribution by borough and price group. We can see that Manhattan, Bronx and Brooklyn are highly populated in comparison to the other boroughs. Price group density shows the high price density listings with the distribution of low price.

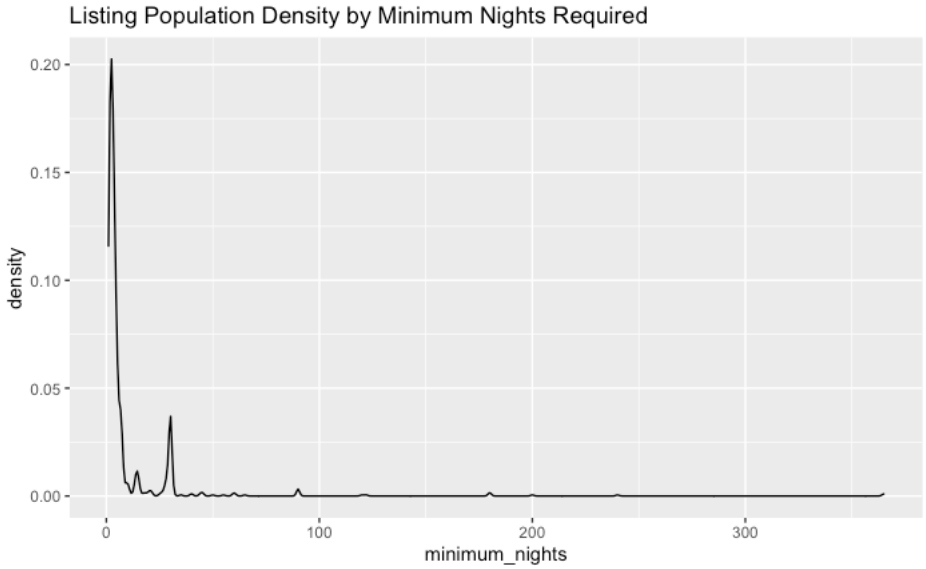
For customer centric companies like Airbnb, word clouds can help them find customers' pain points. If they collect feedback from your customers, they can

generate a word cloud using customers' language to help identify what is most important to them. Imagine if "long wait time" cropped up as major emphasis words in customer feedback. That should ring a warning bell. Luxury, Loft, and Village pop out for high priced listings.

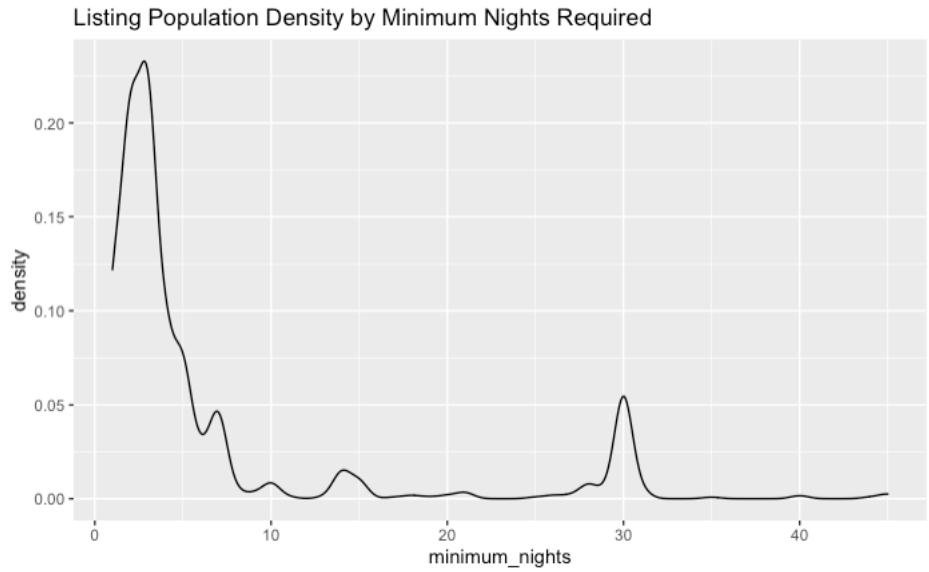


We can see from the population density by price graph that listing with lower price ranges have the most dense area and that density is inversely proportional to the price group of the listings.

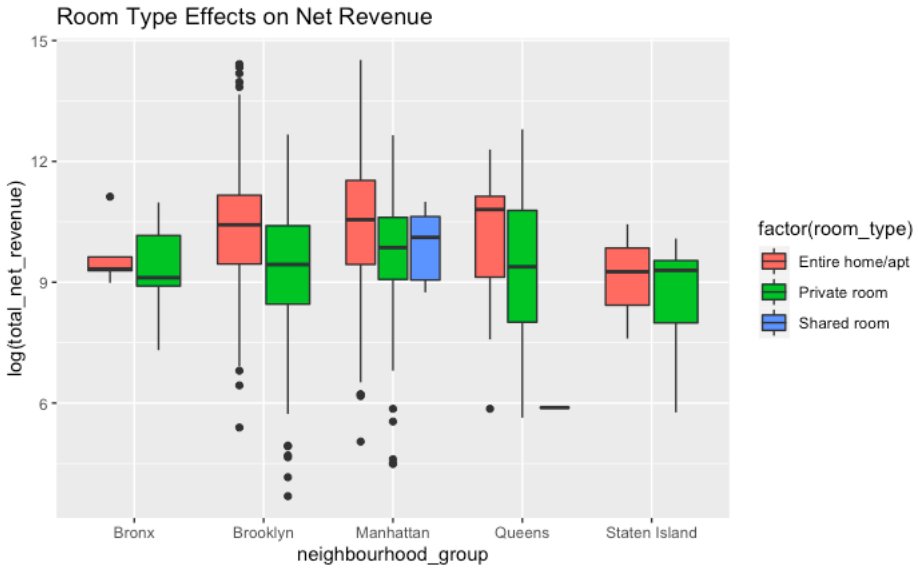




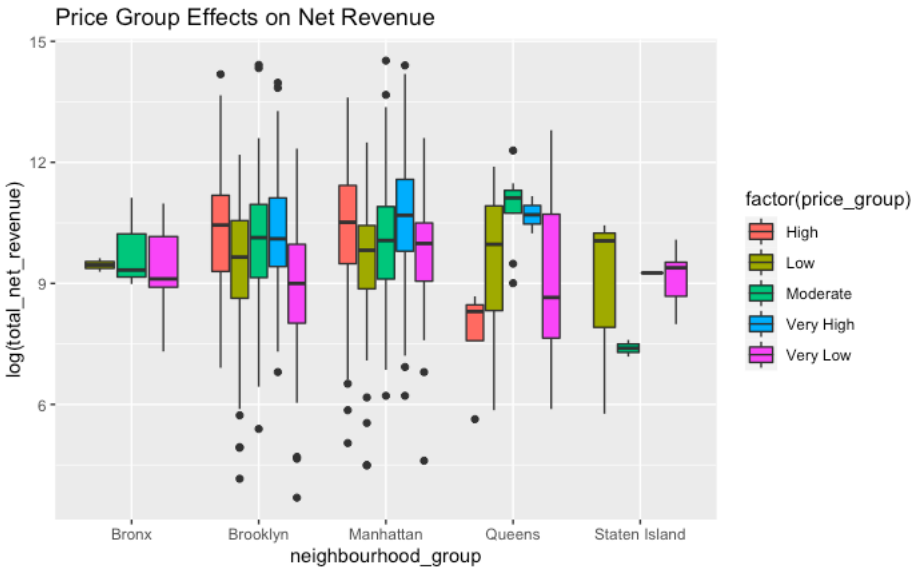
It also looks like 7% of our population has 0 reviews. Also we can observe that less dense populated areas have large reviews. We can observe the same behaviour by looking a the plot of population density with minimum nights required and both these graphs show the inverse relationship with population density. The first population density vs minimum nights graph is really skewed so we will adjust it putting a filter of less than 50 nights required to get a more reasonable graph as below.



Now we will observe the effects of of room type on the net revenue through the boxplot. As we can see the revenue generally comes from Entire home/apt in all the cities because travellers usually prefer that for their stays but we can see one exception in the form of Bronx in which private takes the major chunk of the revenue.



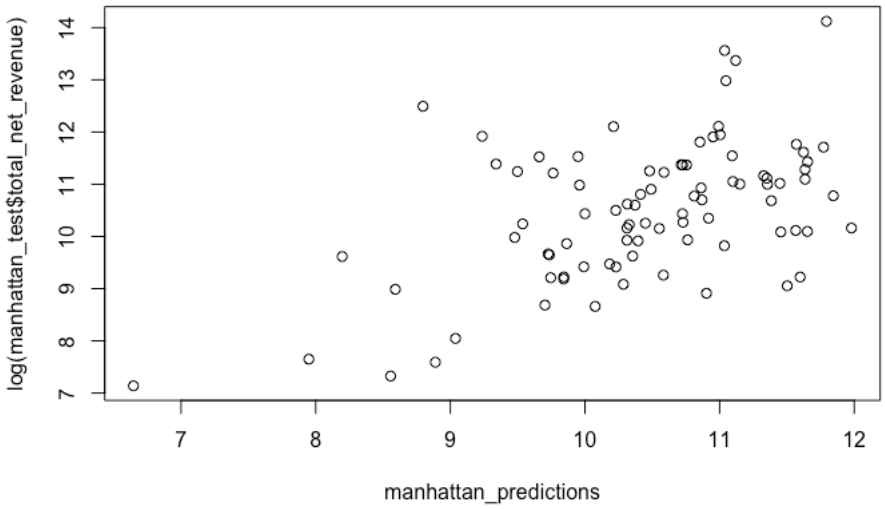
Now boxplot will tell us the relationship between price group and the total revenue generated by the listings. So there are a few things that we can analyze and these can't be generalized in one category. We see in Brooklyn and Manhattan high priced group listings bring out the major revenue for these cities. Staten Island's listings income come from very cheap priced listing maybe because they don't have that many high priced listings or people usually can't afford. In Bronc less than average listing account for increase in revenue. So we see that different cities have different modes of revenue generation.

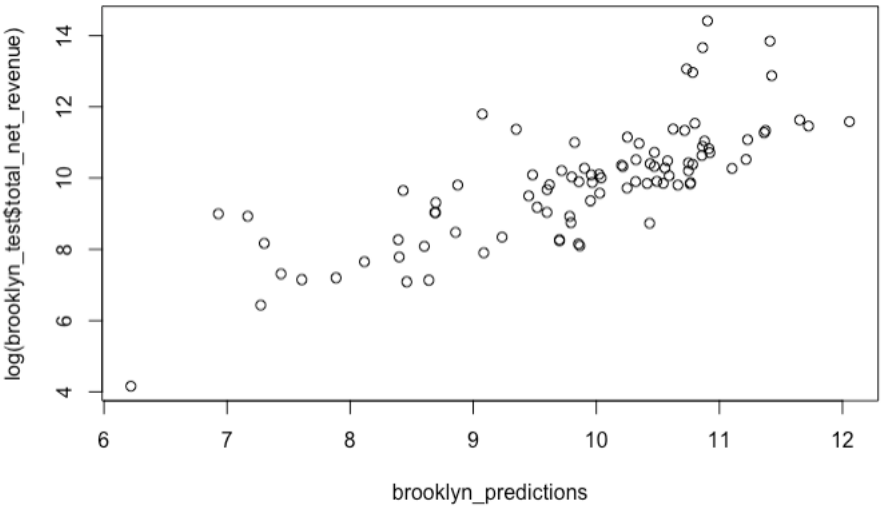


5 Modeling and Conclusion

The model evaluates the host success and measures the financial performance of a listing by taking into account several factors such as location, price, reviews, availability besides other factors.

We want to maximize our net revenue as a lister. Therefore, we want to model net revenue. We can control a lot of things such as the listing title, where our property is located, what kind of property we're leasing, the price range, and the availability. We will create a regression based model to appropriately approach this problem.
We will predict for the two most popular destinations in New York that is Brooklyn and Manhattan.





After running the regression model on the Manhattan data, we have an Adjusted R-Squared of 0.5478., and an RMSE of 17k. So the model is off by about 17k USD / prediction on average.
After running the regression model on the Brooklyn data, we have an Adjusted R-Squared of 0.6263., and an RMSE of 23k. So the model is off by about 23k USD / prediction on average.

It means that the model is performing better for Manhattan in comparison to Brooklyn These are simple regression models and can be further extended by including far more complex features and advanced machine learning and deep learning algorithms

6 References

- http://insideairbnb.com/get-the-data.html
- https://www.epi.org/publication/the-economic-costs-and-benefits-of-airbnb-no-reason-for-local-policymakers-to-let-airbnb-bypass-tax-or-regulatory-obligations/
- https://github.com/NikhilKumarMutyala/NYC-Airbnb-Data-Visualization-for-Classification-
- http://www.columbia.edu/ sg3637/airbnb_final_analysis.html