ISYE 4031 Regression/Forecasting

Regression Analysis on New York City Airbnb Open Data

Milandi Bezuidenhout, Erica Cho, Savannah Chunn
Ethan Kreager, Karen Loscocco, Samantha White

December 3, 2019

Contents

1 Executive Summary

In this paper, we examine a dataset related to Airbnb rental properties in New York City and perform regression analysis to determine the best predictors of price. Regression is a powerful approach used in a wide variety of applications for predictive analysis in both industry and academia. After exploring multiple regression models using R, we have come to the conclusion that the type of room, availability, location, number of reviews, and the amount of nights in a visit are the best predictors of price. Below, we include our data exploration, mathematical regression models, and implementations in R, as well as our solutions and comments on the applications of our findings.

2 Introduction

Since its inception in 2008, Airbnb has become one of the world's largest marketplaces to connect people who want to rent out their homes with travelers. It began as a small operation in downtown San Francisco, but quickly grew into the multinational corporation that it is today. Airbnb is very unique because the corporation itself does not actually own any of the properties that it rents out. Instead they are owned and maintained by common people, who make money by renting out their space to temporary tenants. The price of one night's rent at an Airbnb property can vary wildly based on factors such as location, reviews, and amenities. Our project attempts to determine which factors have the most impact on the final listed price.

3 Problem Description

Our analysis proposes to determine which variables have the most effect on the observed differences in price. We tested various multiple linear regression models with price as the dependent variable in order to find the model which contains the most statistically significant predictor variables. We also made observations on the data itself and included visuals to give a deeper understanding of the data. We also identified outliers and included potential justifications for why these outliers could exist.

4 Data Exploration

4.1 Data Description

The dataset includes 48,895 listings for rental properties in New York City from the year 2019. There are 16 columns that are a mix of categorical and numerical values, each of which has an impact of some degree on the price to rent the property. The following is a breakdown of the columns and their respective meanings:

Column	Meaning	Datatype
id	listing ID	int
name	name of the listing	str
host_id	host ID	int
host_name	name of the host	str
neighbourhood_group	location	str
neighbourhood	area	str
latitude	latitude coordinates	float
longitude	longitude coordinates	float
room_type	listing space type	str
price	price in dollars	int
minimum_nights	amount of nights minimum	int
number_of_reviews	number of reviews	int
last_review	latest review	date
reviews_per_month	number of reviews per month	float
calculated_host_listings_count	amount of listing per host	int
availability_365	number of days when listing is available	int

After initial inspection of the dataset, we immediately noticed there were missing values. We edited the data to remove all rows with null values. We also decided to cut down our data to be more aligned with the scope of this project. We randomly selected a subset of 5000 data points in order to ensure the distribution of datapoints and prices were roughly the same for each location.

A distribution of price by neighbourhood before and after we cut down the data can be seen in figures (4.1) and (4.2) below. A display of the datapoints from each neighbourhood group for before and after sampling is also shown in figures (4.3) and (4.4). From this analysis, we can assume that the random sampling preserved the integrity of the data.

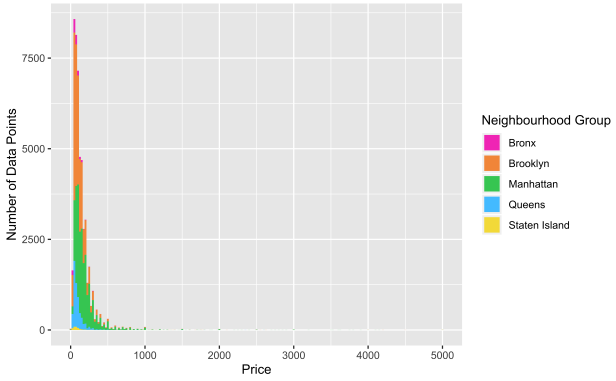

Figure 4.1: Before Random Sampling

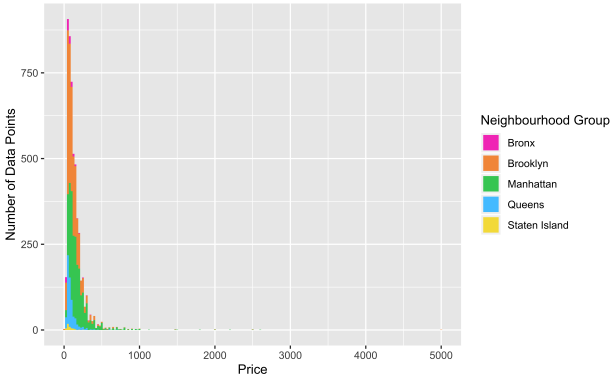

Figure 4.2: After Random Sampling

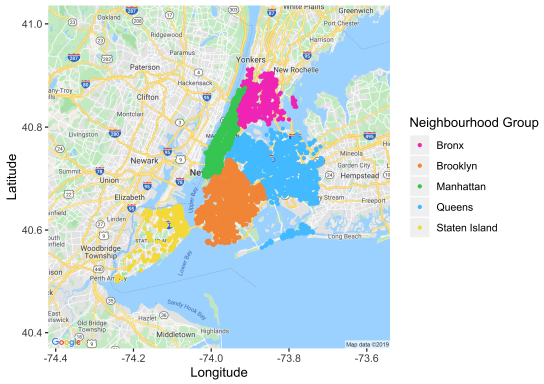

Figure 4.3: Before Random Sampling

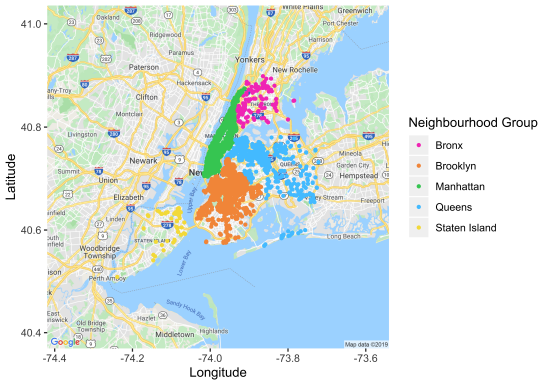

Figure 4.4: After Random Sampling

To ensure readability of the visuals in the rest of the paper, we exclude the few points with prices greater than $1,000. However, the regression analysis does include these points. A closer look at the price distribution can be found in figure (4.5) below:

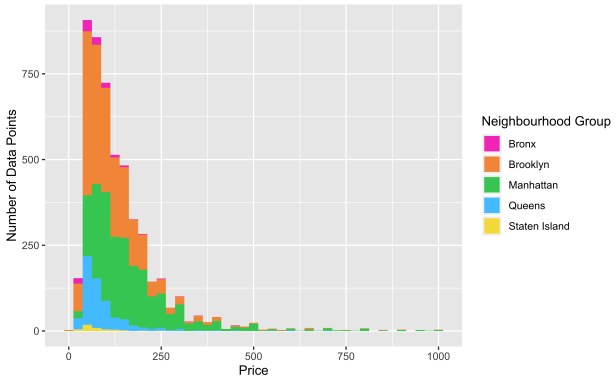

Figure 4.5: Histogram of Prices - Closer Look

The numerical-valued attributes can be directly represented by variables, but we needed to map each of the categorical values to dummy variables. These include neighborhood_group, neighborhood, and room_type. We dropped the id, name, host_id, host_name, and last_review columns because they have no categorical or numerical meaning for our analysis. We also cut out the latitude and longitude variables because location is already represented by the dummy variables.

4.2 Variables

Now that we have a reduced dataset, we would like to check for any potential multicollinearity among our independent variables. Multicollinearity is when one independent variable can predict another with a fair degree of accuracy. Strong interactions between variables can be an indicator of any potential multicollinearity. Multicollinearity can be measured in a multitude of ways, one

being correlation coefficients.

Below, in figure (4.6), the correlation matrix for all numerical variables is shown:

Figure 4.6: Correlation Plot By Neighbourhood Group

The plot shows no coefficients larger than 0.5, which indicates that none of our independent variables are correlated. As a result, we proceed by keeping all variables for further analysis.

4.3 A Closer Look

Prices to stay at an Airbnb in New York can vary wildly by borough. In figure (4.7) below, a map of Airbnb prices by location can be found:

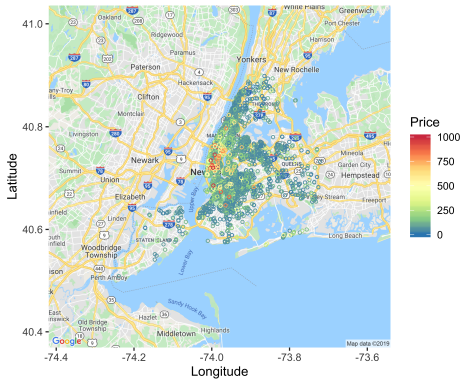

Figure 4.7: Prices By Location

As you can see, Manhattan is clearly the most expensive borough. Considering the density of theaters, restaurants, and other visitor attractions in Manhattan, it doesn't come as too much of a

shock that it leads the pack in terms of Airbnb price. In our appendix (8), we have a graph breaking out price to each of the five boroughs.

Similarly, we can look at New York in terms of availability. Below, figure (4.8) shows the frequency of available rental properties throughout the city.

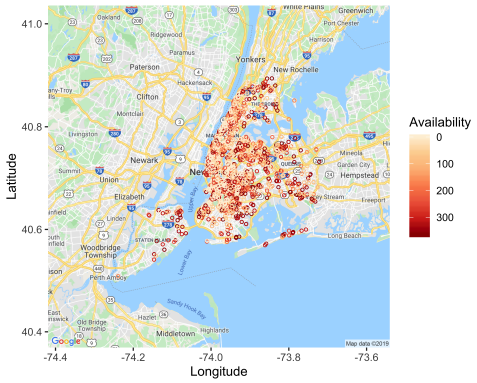

Figure 4.8: Availability By Location

The figure shows that Manhattan and Brooklyn are the most dense areas in terms of openings. These are the two most densely populated areas of the city, so it shouldn't come as a surprise that they lead the way in terms of Airbnb rental property openings.

Furthermore, we can break the city out by the type of rental openings. Specifically, we look at openings where the entire apartment/house can be rented, openings where a private room can be rented, and openings where a shared room can be rented. The figures below summarize our findings:

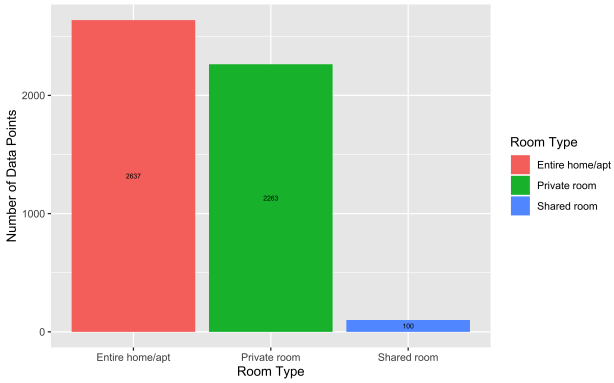

Figure 4.9: Counts of Room Types

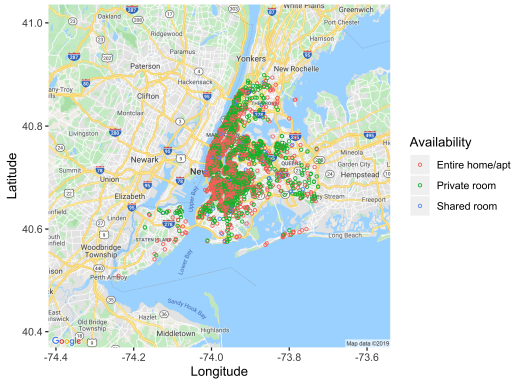

Figure 4.10: Map of Room Types

Very few property owners rent out a shared room, which is likely due to the fact that most people would prefer to have their own space when possible. Manhattan has a higher concentration of entire properties available for rent, which would make sense when considering the high cost of a very small space in the city. There simply isn't room for someone to live in their home and also rent out space to a guest.

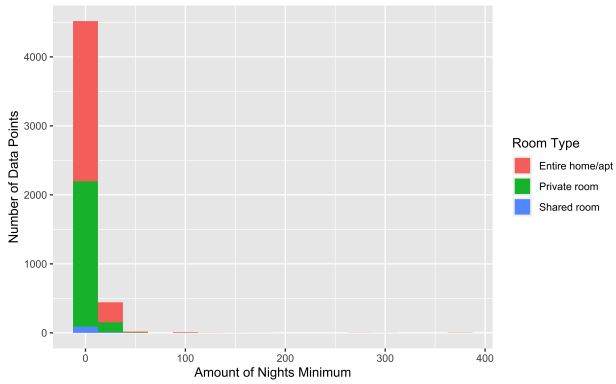

Figure 4.11: Histogram of Minimum Nights By Room Type

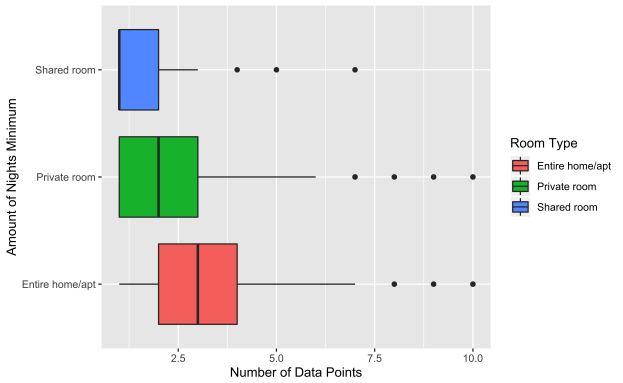

Figure 4.12: Box Plot of Minimum Nights By Room Type

5 Regression Analysis

We break up our analysis into two sections. The first is the creation of a base model that includes all variables, and the second is the determination of a best model through regression analysis. The R implementations and outputs of the models from both of these parts can be referenced in the appendix (8).

5.1 Part 1 Analysis: Base Model

We began our analysis with a model containing all of our variables. We will refer to this model as our base model and use it as a benchmark for comparison to other models tested.

Independent tests were run on this model to identify relationships between all independent variables. We define the set of all independent variables as V where $V = \{$ number_of_reviews, minimum_nights, reviews_per_month, calculated_host_listings_count, availability_365, shared_room_dummy, staten_dummy, brooklyn_dummy, queens_dummy, bronx_dummy, private_room_dummy $\}$. Our base model is the following:

$$\text{Base Model: } price = \beta_0 + \beta_i X_i \;\; \forall i \in V \tag{5.1}$$

The summary of this base model in figure (5.1) reveals that all variables tested are significant except for reviews_per_month and calculated_host_listings_count. All other variables have p-values significantly less than 0.05. The analysis can also be found on page 14 of the appendix (8).

```
lm(formula = price ~ number_of_reviews + minimum_nights + reviews_per_month +
    calculated_host_listings_count + availability_365 + shared_room_dummy +
    staten_dummy + brooklyn_dummy + queens_dummy + bronx_dummy +
    private_room_dummy, data = project_data_sample)

Residuals:
   Min     1Q Median     3Q    Max
-191.7  -53.5  -15.1   19.1 4876.5

Coefficients:
                               Estimate Std. Error t value Pr(>|t|)
(Intercept)                    213.86270    4.22492  50.619  < 2e-16 ***
number_of_reviews               -0.20184    0.04956  -4.072 4.73e-05 ***
minimum_nights                  -0.64761    0.13913  -4.655 3.33e-06 ***
reviews_per_month               -1.40665    1.41241  -0.996    0.319
calculated_host_listings_count  -0.03833    0.08366  -0.458    0.647
availability_365                 0.19493    0.01632  11.946  < 2e-16 ***
shared_room_dummy             -135.87966   14.41519  -9.426  < 2e-16 ***
staten_dummy                  -107.97891   21.32911  -5.063 4.29e-07 ***
brooklyn_dummy                 -46.55858    4.36708 -10.661  < 2e-16 ***
queens_dummy                   -66.63348    6.77516  -9.835  < 2e-16 ***
bronx_dummy                    -84.93433   13.99659  -6.068 1.39e-09 ***
private_room_dummy            -108.38674    4.11035 -26.369  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 140.9 on 4988 degrees of freedom
Multiple R-squared:  0.1877, Adjusted R-squared:  0.1859
F-statistic: 104.8 on 11 and 4988 DF,  p-value: < 2.2e-16
```

Figure 5.1: Base Model Summary

We also ran a VIF test to look for any potential multicollinearity. All of the VIF factors hovered around 1, meaning that once again no significant multicollinearity was detected. This further indicates that the variables in our model are independent of one another. Additionally, we performed outlier tests including Cooks Distance measure (CooksD) and Studentized Deleted Residuals (SDR). These outlier results can be referenced in the appendix (8).

The normal probability plots of residuals can be seen in the figure below:

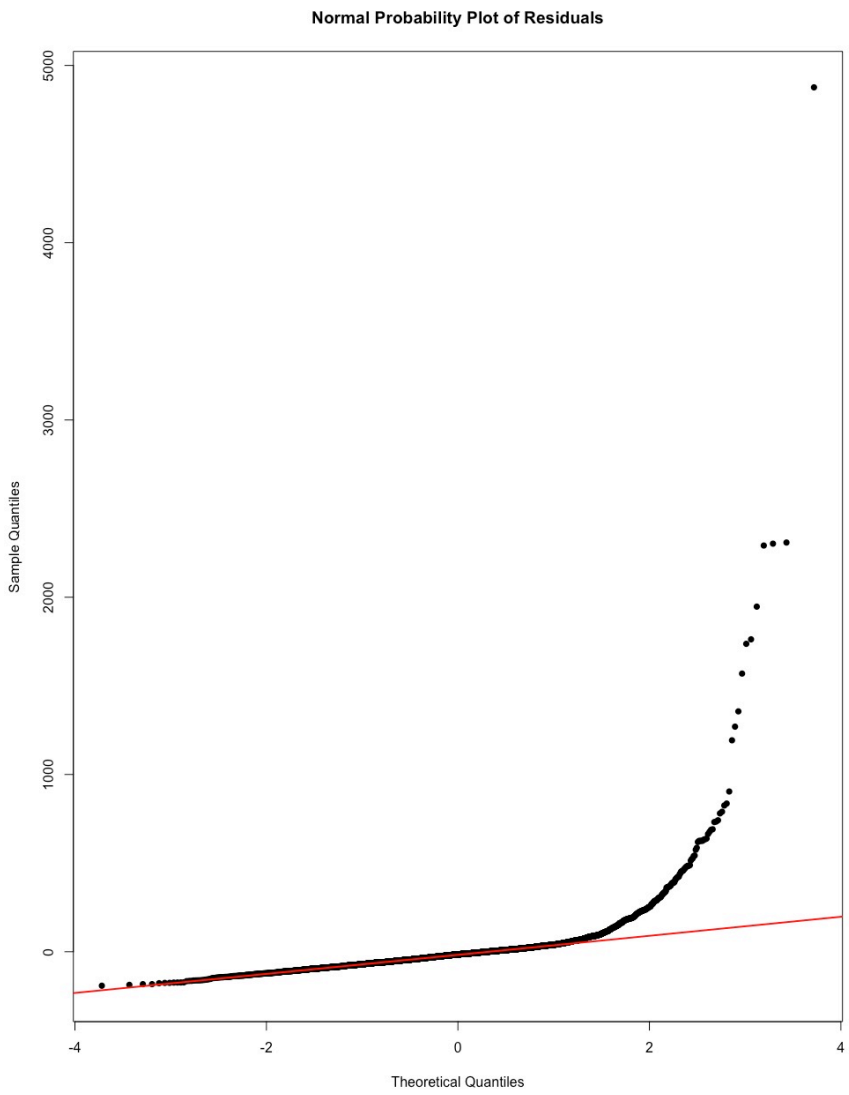

Figure 5.2: Normal Probability Plots of Residuals

The plot of residual error vs fitted values:

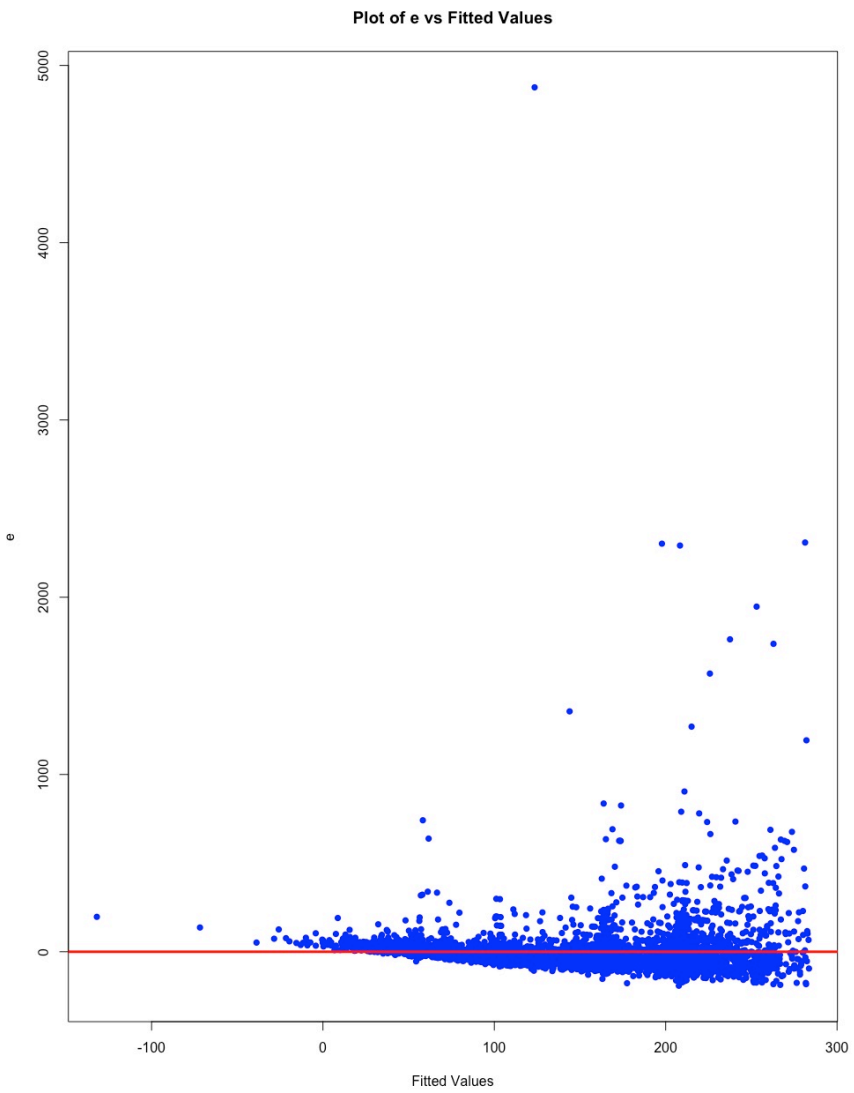

Figure 5.3: Error vs. Fitted Values

5.2 Part 2 Analysis: Best Model

In order to determine the best model, we performed stepwise elimination in both directions. The lowest AIC corresponds to the best model. This resulted in the following best model:

We define the set variables for the best model as V_{best} where V_{best} = { private_room_dummy, availability_365, shared_room_dummy, brooklyn_dummy, bronx_dummy, number_of_reviews, minimum_nights, queens_dummy, staten_dummy }. Our best model is the following:

$$\text{Best Model: } price = \beta_0 + \beta_i X_i \;\; \forall i \in V_{\text{best}} \tag{5.2}$$

The summary of this best model in figure (5.4) reveals that all variables tested are significant. All variables have p-values significantly less than 0.05. The analysis can also be found on page 26 of the appendix (8).

```
lm(formula = price ~ private_room_dummy + availability_365 +
    shared_room_dummy + brooklyn_dummy + bronx_dummy + number_of_reviews +
    minimum_nights + queens_dummy + staten_dummy)

Residuals:
   Min     1Q Median     3Q    Max
-192.3  -53.6  -15.0   19.1 4878.4

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)         212.55929    4.05706  52.392  < 2e-16 ***
private_room_dummy -108.39142    4.10207 -26.424  < 2e-16 ***
availability_365      0.19200    0.01592  12.061  < 2e-16 ***
shared_room_dummy  -135.72689   14.41210  -9.418  < 2e-16 ***
brooklyn_dummy      -46.21379    4.34315 -10.641  < 2e-16 ***
bronx_dummy         -85.16244   13.97069  -6.096 1.17e-09 ***
number_of_reviews    -0.22470    0.04266  -5.267 1.44e-07 ***
minimum_nights       -0.63347    0.13795  -4.592 4.50e-06 ***
queens_dummy        -67.19520    6.71403 -10.008  < 2e-16 ***
staten_dummy       -107.82482   21.31308  -5.059 4.36e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 140.9 on 4990 degrees of freedom
Multiple R-squared:  0.1875,    Adjusted R-squared:  0.186
F-statistic: 127.9 on 9 and 4990 DF,  p-value: < 2.2e-16
```

Figure 5.4: Best Model Summary

The normal probability plots of residuals can be seen in the figure below:

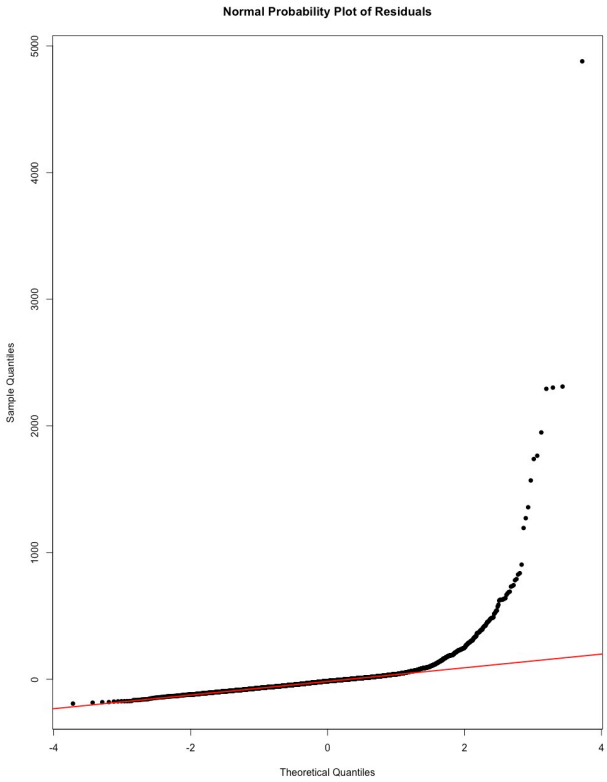

Figure 5.5: Normal Probability Plots of Residuals

Looking at the coefficient values and the p-values from the model, the location with the most significant correlation is clearly Manhattan. The data for Manhattan shows a positive correlation between demand and price, indicating that prices tend to be the highest in that borough. The trend for the Bronx and Staten Island locations is that demand decreases as price increases, which represents the fact that people will not stay in these locations if the rental property is not a relatively cheap option. In reality, this is an accurate indicator of how consumers will choose a place to stay, and also of the quality of the average rental property at each location.

6 Conclusion

Airbnb is a business that is successful because of its versatility of locations and services that are provided to guests as well as the ease of being a host. These two sides of the business have the same goal, price. Whether it be to search out the lower price as a consumer or to make the most of a location as a host, Airbnb has a variety of variables that adjust price, and throughout the report we have discussed the variables' interaction with price using regression and data analysis.

Our analysis has real-world appeal to both ends of the Airbnb market. It appeals to potential renters because they want to determine which Airbnb rental property is the best deal in a given

city. Quantifying the factors that increase the price of a destination can help renters choose a destination that aligns with their top preferences while also not breaking the bank. After all, having an affordable yet also comfortable and convenient place to stay for a vacation or work trip can almost singlehandedly make a trip successful.

Our findings also have appeal to potential hosts. While hosts seek to maximize profit, they can struggle to know what price range will achieve that goal. By understanding which factors influence the optimal price of renting their space, hosts can set a price for their property which maximizes their returns.

Our final recommendation to a potential client would be fairly simple. Airbnb users can have wildly different needs and desires, so it is hard to give a lot of concrete advice. Instead, we recommend that users make an informed decision using data and their personal desires before taking a trip to New York. As an example, if you are seeking to be very near the tourist attractions in the city but willing to spend money, Manhattan is probably the place for you to stay. The other neighbourhoods each have positives and negatives as well, and any potential New York tourist should understand them all before choosing their temporary home away from home.

7 References

Dataset: https://www.kaggle.com/dgomonov/new-york-city-airbnb-open-data