**TABLE OF CONTENTS**

**I COVER PAGE------------------------------------------------------------------------------------------------------------2**

**II EXECUTIVE SUMMARY-----------------------------------------------------------------------------------------------3**

**III DATA DESCRIPTION---------------------------------------------------------------------------------------------------3**

**IV DATA PREPARATION AND EXPLORATION----------------------------------------------------------------------3**

**V RESEARCH QUESTION------------------------------------------------------------------------------------------------7**

**VI METHODOLOGY-------------------------------------------------------------------------------------------------------7**

**LINEAR REGRESSION**

**LASSO REGRESSION**

**NEURAL NETWORKS**

**XG BOOSTING**

**VII RESULTS AND FINDING-----------------------------------------------------------------------------------------------8**

**VIII CONCLUSION------------------------------------------------------------------------------------------------------------9**

**IX APPENDIX----------------------------------------------------------------------------------------------------------------9**

**I. COVER PAGE**

Project Title: **Predicting Prices of Airbnb Houses in New York City**

Team Members:

Venkata Srirangam (116269612)

Jay Sethia (116144780)

Keya Mankar (116238793)

Ankit Yadav (116240079)

**ORIGINAL WORK STATEMENT**

We the undersigned certify that the actual composition of this proposal was done by us and is original work.

**Typed Name Signature**

Venkata Srirangam

Jay Sethia

Keya Mankar

Ankit Yadav

**II. EXECUTIVE SUMMARY**

Airbnb has grown to be one of the most popular and affordable platforms for accommodation for travellers. It is difficult for new hosts to decide the optimal price for their listing that is competitive yet affordable for guests and profitable. The goal of this project was to predict the Airbnb house prices in New York City with minimum error and maximum reliability. We analysed about 48,000 listings in order to understand how to leverage listing attributes to help hosts understand trends in prices and accurately predict the optimal pricing for both, hosts and guests. After exploring machine learning along with data mining algorithms throughout the course, we've chosen linear regression with variable transformations as the best model for prediction. We performed feature selection to get insights into the variables that affect price the most and ran Linear, Neural Network, and Boosting models in order to determine the best solution that will accurately predict prices.

**III. DATA DESCRIPTION**

The data set was obtained from the [InsideAirbnb](http://insideairbnb.com/get-the-data.html). Our primary dataset is the result of a full join between listings and reviews grouped by house listings ids.

The data dictionary is defined in the Appendix.

**IV. DATA PREPARATION AND EXPLORATION**

The dataset has records of all the Airbnb Houses in New York City up till March 2019. The initial data set consisted of 106 variables and 48,000 rows. The data was split into 70% training and 30% test.

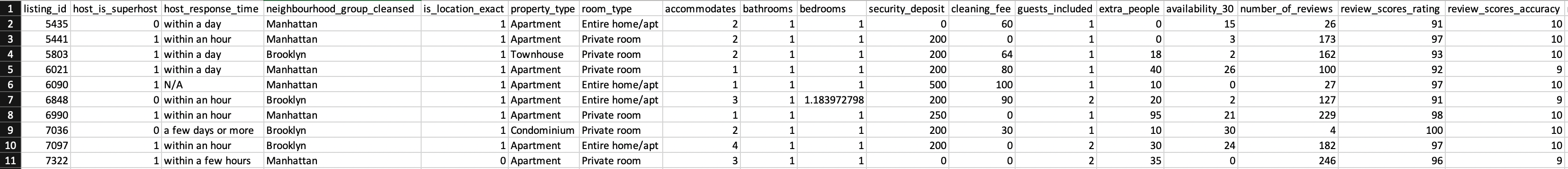


Table 1: - Snapshot of 1 - 18 variables

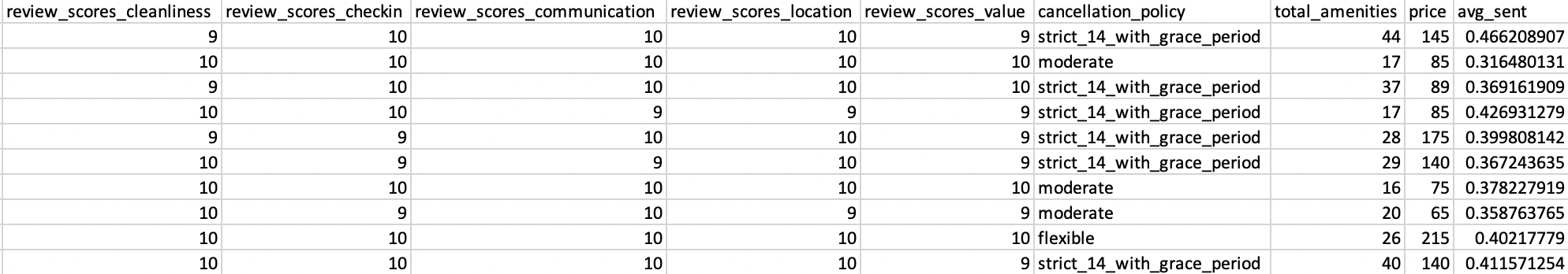


Table 2: - Snapshot of last 19 - 26 variables

We narrowed down the dataset to 26 variables using Backward and Forward Selection. This dataset will be used to run models on. The full dataset is being used for the visualizations done in our Exploratory Data Analysis.

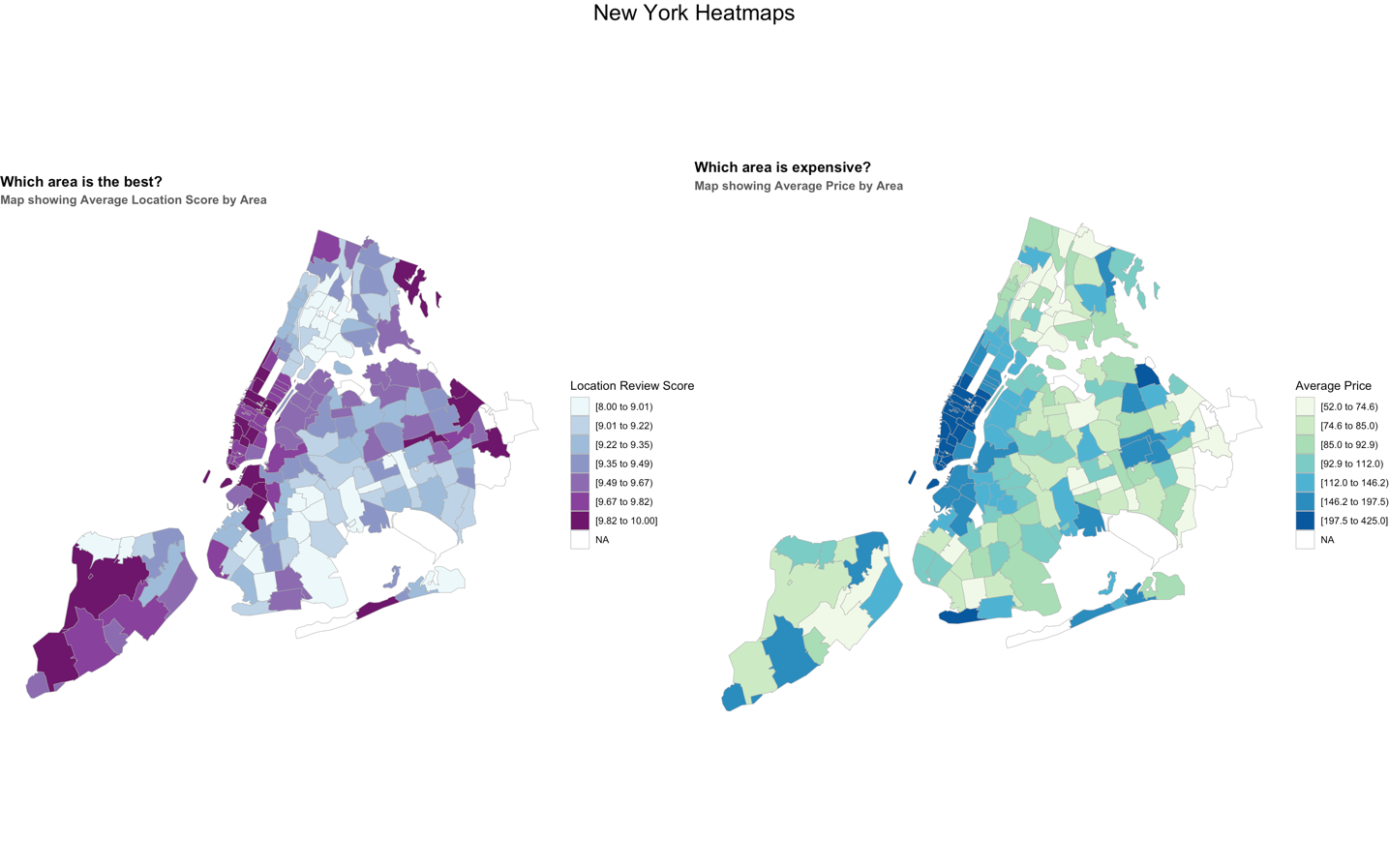


Figure 1: - New York Reviews and Price Heatmaps

The above Heat Maps helped us understand the distributions across the city neighbourhoods for reviews and price.

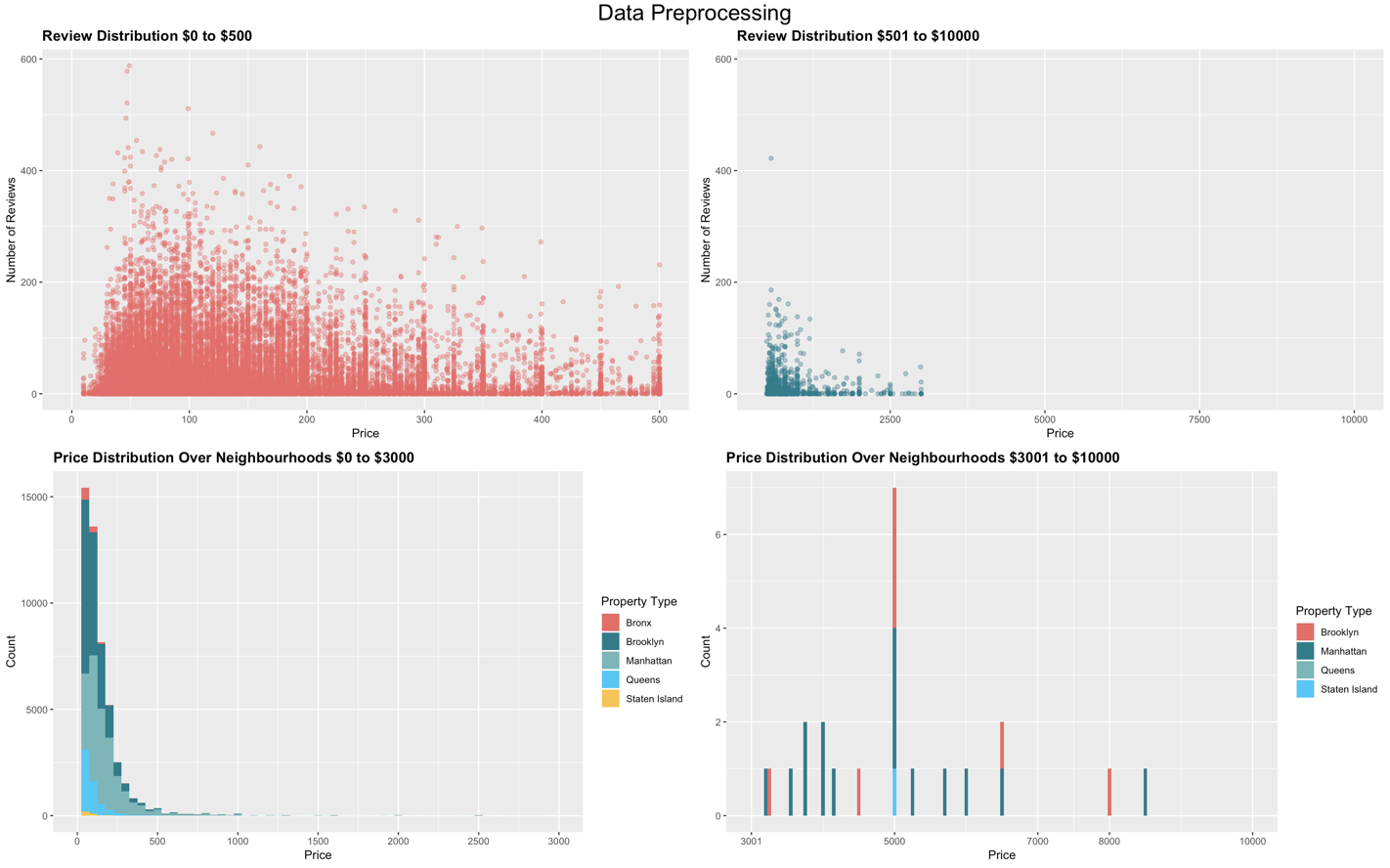


Figure 2: - Reviews and Price Distribution

Analysed the range of prices of the properties listed and removed houses with prices above $3,000 as there were only 30 such records but severely impacted predictions.

About 10,000 listings had no reviews and removing these observations would prove fatal. After learning about other features for these listings, we assigned median value for these listings, which was shared by 80% of the listings.

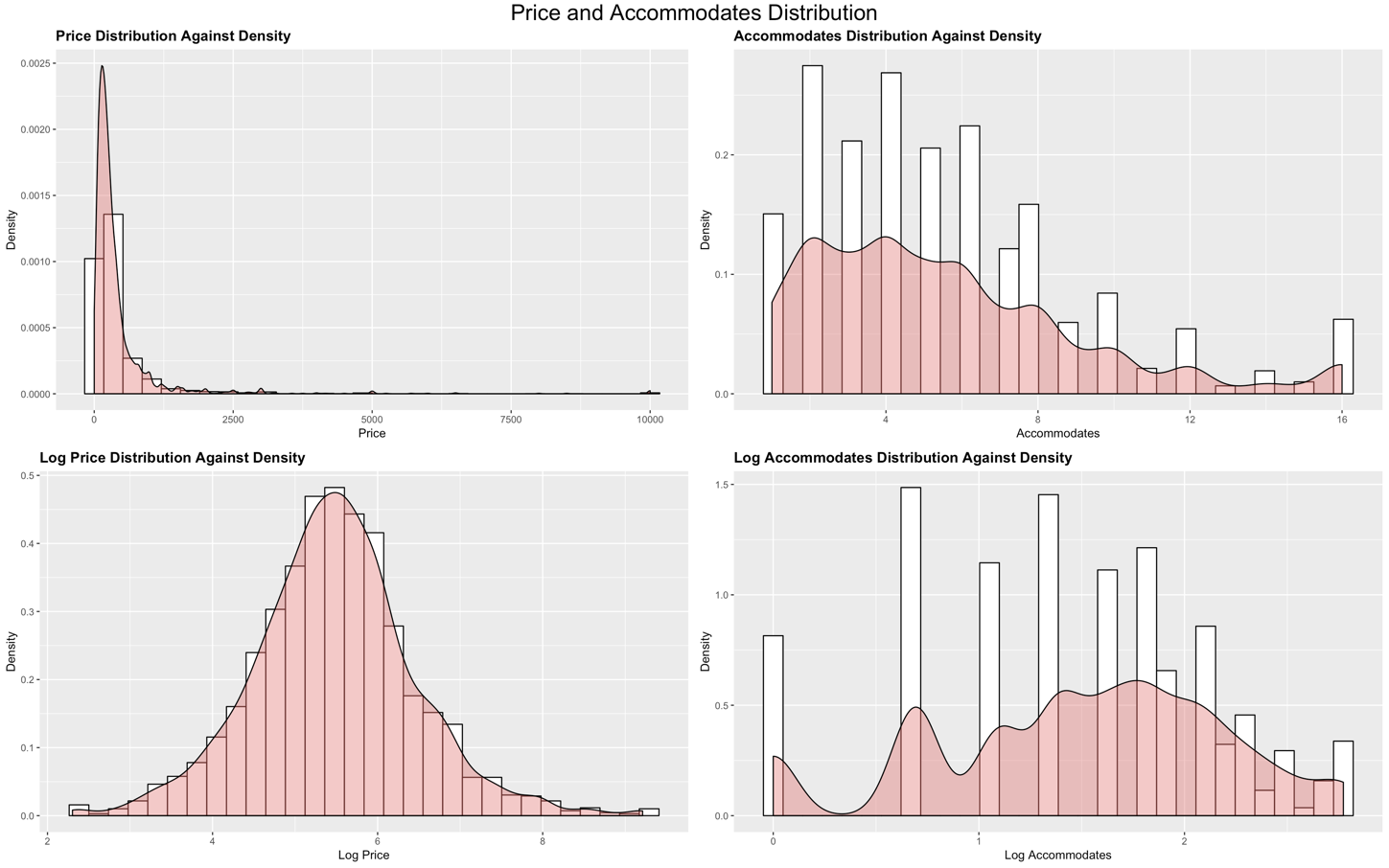


Figure 3: - Logging Price and Accommodates columns.

The Data showed an inherent right skewness for the price and accommodates columns, log normalizing these columns helped us make better predictions.

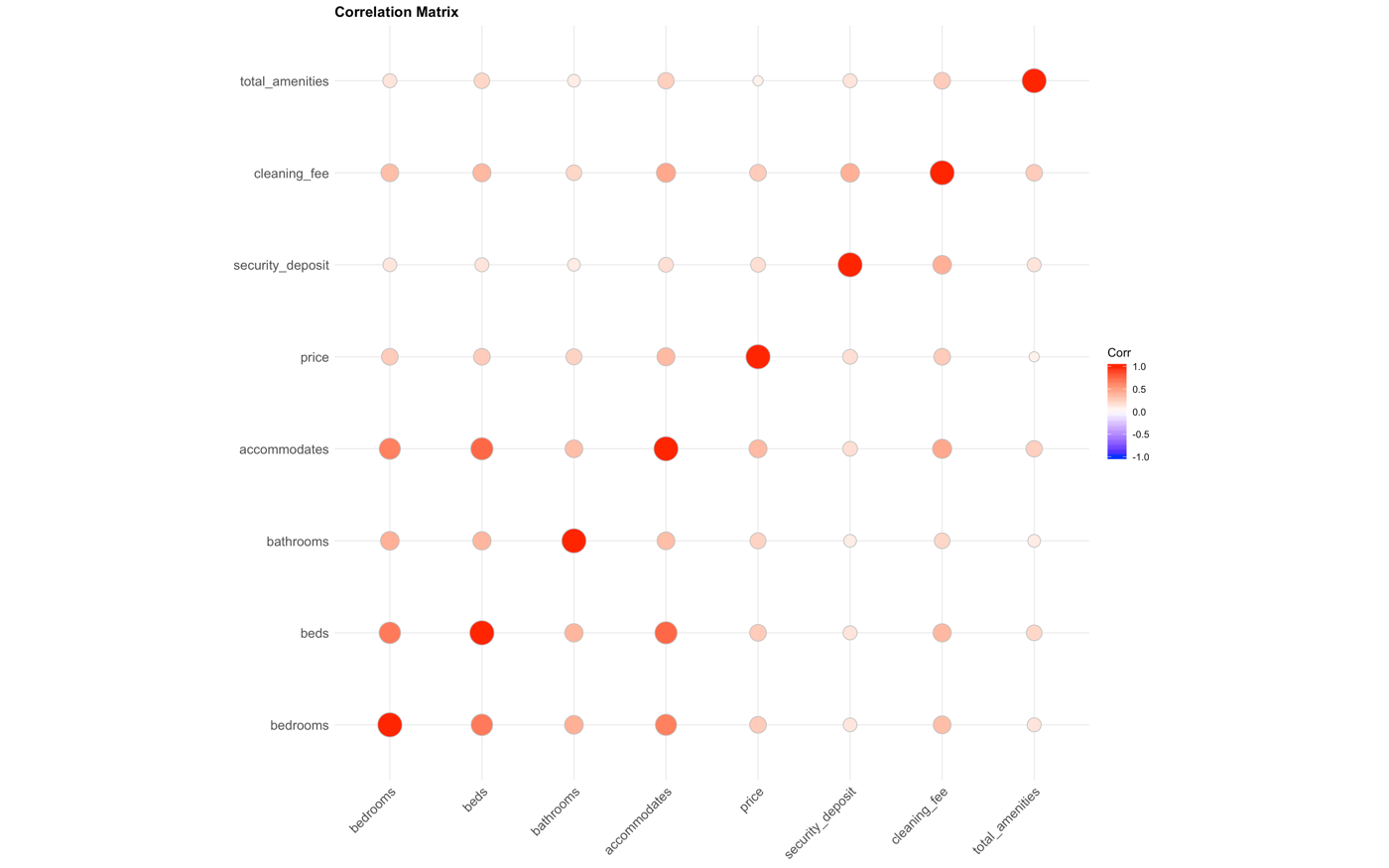


Figure 4: - Correlation Matrix for variable selection

Further we went on to explore the collinearity between variables and removed highly correlated variables that in most cases provided redundant data.

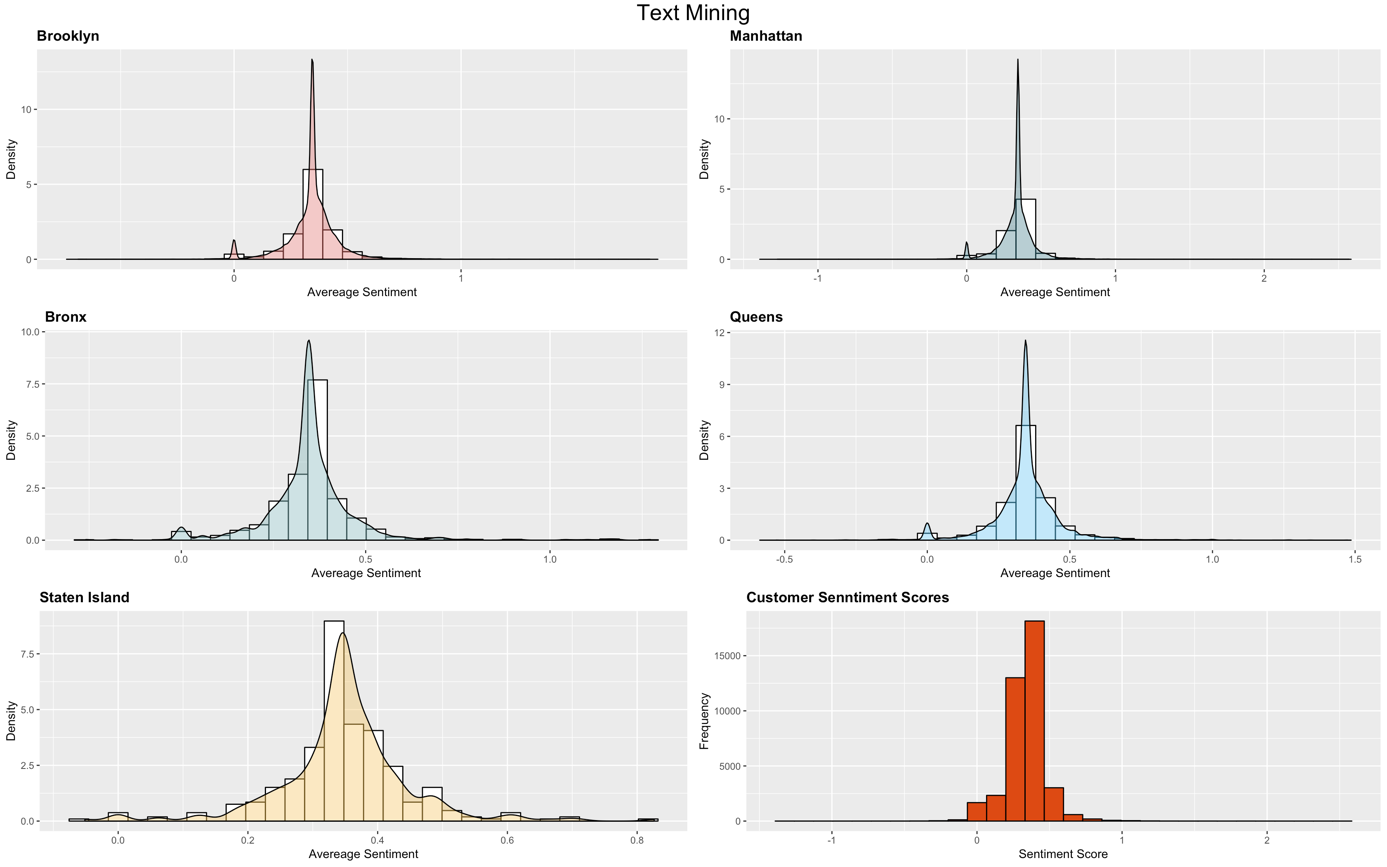


Figure 5: - Reviews Distribution Neighbourhood wise

Finally, we used Text Mining on the Reviews Dataset to calculate the sentiment scores of the listings in the city. We can see that Airbnb guests tend to leave positive reviews from the distributions in the above graphs.

**V. RESEARCH QUESTION:**

New York City is one of the top 3 cities in terms of number of Airbnb hosts. In the recent years, there has been an exponential increase of number of hosts in New York. This resulted in diversity in prices of houses in New York City area. The question we are attempting to answer is identification of factor and attributes that affect house prices, their relationship and behaviour.

**VI. METHODOLOGY**

The machine learning algorithms we used for performing the regression analysis are:

1) Linear Regression  
Based on what we learnt from exploratory analysis and extensive data pre-processing, we performed linear regression using 26 variables. After running an initial model and observing the residuals vs. fitted plot, we believed variables could be transformed to better explain relationship.  
We transformed the dependant variable and one predictor variable to log values and performed linear regression. The resultant model could explain the relationship much better and yielded a much better R2.

2) Lasso Regression  
Performed feature selection on 26 features, and predicted using minimum lambda at 0.1, which yielded results similar to linear regression with variable transformation, albeit with a lower R2 value.

3) Neural Networks  
Ran deep learning model using Keras, with 2 hidden layers having 30 and 15 nodes respectively. We ran the model for 200 epochs which plateaued at a MAE of 86.10.

4) XG Boosting

After evaluating optimal number of iterations before which Performed Extreme Gradient Boosting with 2000 iterations and found results to be similar to linear regression.

**VII. RESULTS AND FINDING**

We discovered that cancellation policy, host response time, property type, cleaning fee security deposit, review ratings, total amenities and sentiment score have the most influence on prices. We conclude that on average, our models under predict houses by about $50. Results have been summarized and tabulated below:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Model** | RMSE | MAE | MAPE | AE | Adjusted R2 |
| Linear Regression | 126.45 | 59.34 | 47.88 | -1.44 | 0.3630 |
| XGBoost | 126.46 | 59.33 | 47.81 | -17.43 | 0.3860 |
| **Linear Regression (Logged)** | **126.76** | **50.51** | **32.15** | **-1.42** | **0.6248** |
| XGBoost (Logged) | 126.75 | 50.54 | 32.20 | -17.53 | 0.4150 |
| Lasso Regression | 126.41 | 59.25 | 47.74 | -1.40 | 0.3860 |
| Neural Net (Keras) | 159.83 | 86.10 | 84.70 | -3.30 | ------- |

Table 3: - Key Performance Indicators for Data Mining Models

**VIII. CONCLUSION**

We believe linear regression with variable transformation is the best model to predict prices with respectable error margin to advise hosts, guests and investors. Hosts can be reassured that if their houses haven't been under-priced. Guests can be helped find a good deal that gets them the most bang for their buck. There also exist investors who are looking for long term property investments and host Airbnb guests. Such investors can be advised to identify houses that can yield highest returns.

**IX. APPENDIX:**

|  |  |  |  |
| --- | --- | --- | --- |
| **#** | **Variable** | **Description** | **Type** |
| *1* | Price | Price of the house listed identified by id | *Numerical* |
| 2 | bedrooms | Number of bedrooms in the listed Airbnb house. | Numerical |
| 3 | accommodates | Limit of guests per Airbnb house. | Numerical |
| 4 | host\_response\_rate | Likelihood of a host responding to enquiries and requests. | Numerical |
| 5 | security\_deposit | Amount of security deposit to be paid by guests to book property. | Numerical |
| 6 | guests\_included | Number of Guests that can stay at the house | Numerical |
| 7 | Review score rating | Review rating given by Customers to the house | Numerical (out of 100) |
| 8 | number\_of\_reviews | Number of reviews a property has accumulated to date. | Numerical |
| 9 | cancellation\_policy | Cancellation policy set by the host of the airbnb(strict, flexible or moderate) | Categorical |
| 10 | neighbourhood\_group\_cleansed | Neighborhood where the house is situated | Categorical |
| 11 | Host is superhost | Whether the Host is a Super Host, A Super Host has a proven history of great service | Boolean (T/F) |
| 12 | room\_type | The type of room | Categorical - Private Room, Entire Room |
| 13 | bathrooms | The number of bathrooms in the house | Numerical |
| 14 | host\_identity\_verified | Host Identity has been verified or not | Boolean (T/F) |
| 15 | property\_type | The type of the property like apartment or condo etc | Categorical |
| 16 | cleaning\_fee | The cleaning fee if imposed by the airbnb house owner | Numeric |
| 17 | extra\_people | The number of people beyond the ideal room capacity | Numeric |
| 18 | total\_amenities | The number of amenities the house offers | Numeric |
| 19 | review\_scores\_rating | The average rating of consumers for that particular house | Numeric |
| 20 | review\_scores\_accuracy | The average of the accuracy of the total ratings for a particular house | Numeric |
| 21 | review\_scores\_cleanliness | The average cleanliness rating for a particular house | Numeric |
| 22 | review\_scores\_checkin | Average of reviews for ease of checking in | Numeric |
| 23 | review\_scores\_communication | The average rating for the communication of host | Numeric |
| 24 | review\_scores\_location | The average rating awarded for the location of the airbnb house | Numeric |
| 25 | availability\_30 | Number of days not booked or free to book in the next 30 days | Numeric |
| 26 | review\_scores\_value | Average of scores for evaluation of experience had for price paid | Numeric |

**References:**