AIRBNB CASE STUDY

- RAJDIPA & MUNNA

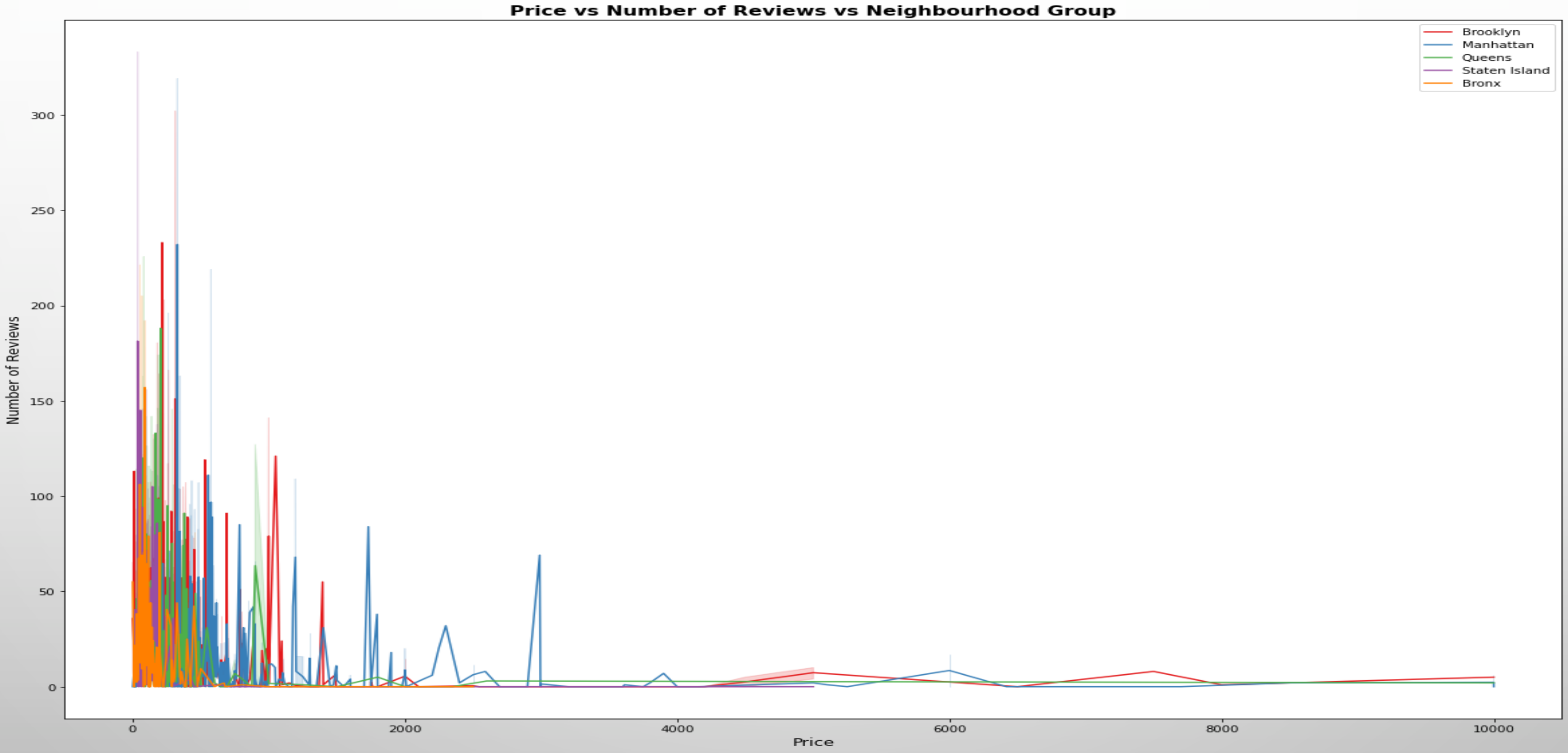
OBJECTIVE

Airbnb has seen a major decline in revenue. Now that the restrictions have started lifting and people have started to travel more, Airbnb wants to make sure that it is fully prepared for this change.

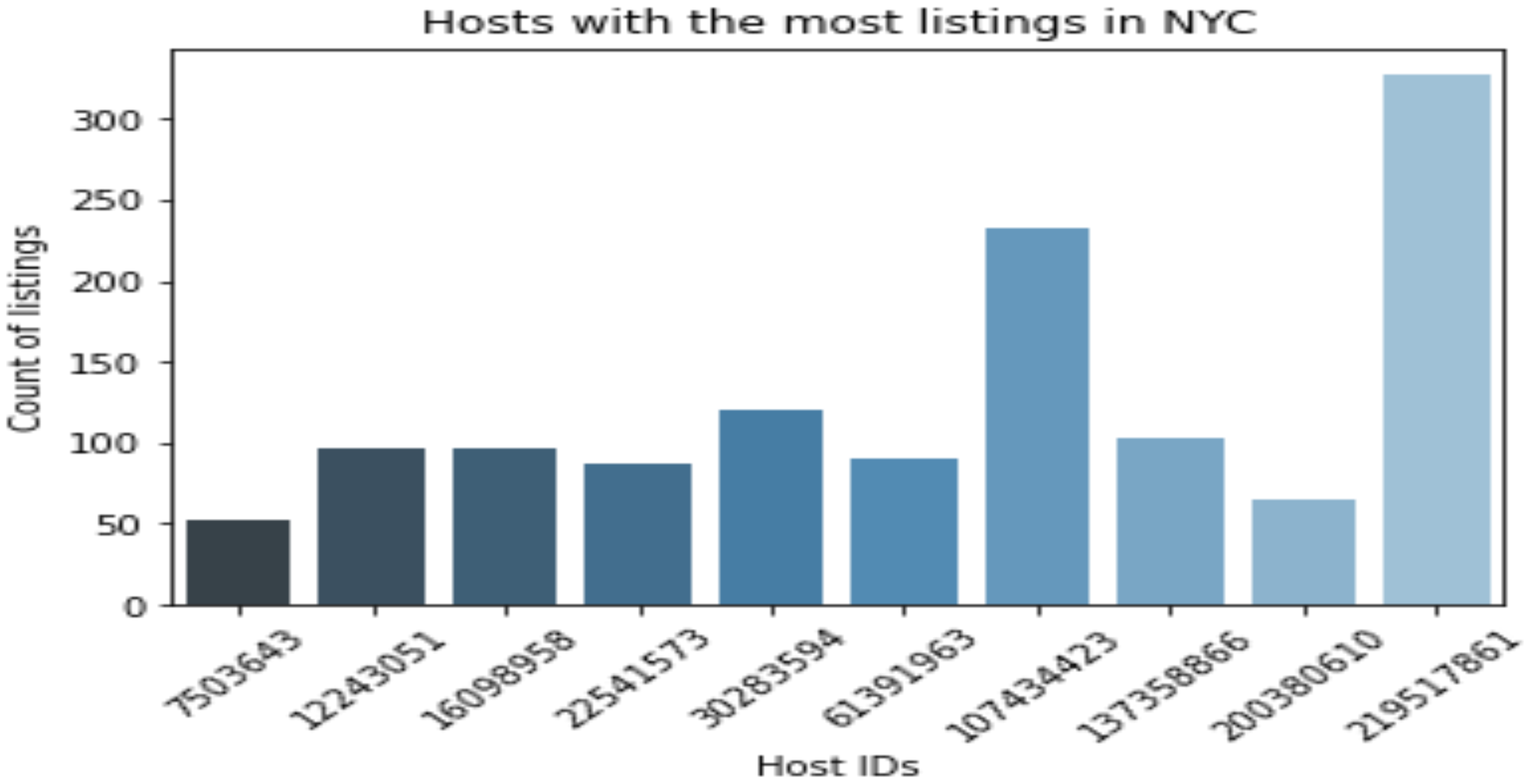
The different leaders at Airbnb want to understand some important insights based on various attributes in the dataset so as to increase the revenue such as -

- Which type of hosts to acquire more and where?
- The categorisation of customers based on their preferences.
  - What are the neighbourhoods they need to target?
  - What is the pricing ranges preferred by customers?
  - The various kinds of properties that exist w.r.t. customer preferences.
  - Adjustments in the existing properties to make it more customer-oriented.
- What are the most popular localities and properties in New York currently?
- How to get unpopular properties more traction? and so on...

GRAPH IS ABOUT PRICE VS NUMBER OF REVIEWS BASED ON NEIGHBOURHOOD GROUP. IT SHOWS US THE LOWEST PRICES HAVE HIGHER REVIEWS THAN THE HIGHER PRICES. IT SHOWS NEGATIVE CORRELATION BETWEEN PRICE AND NUMBER OF REVIEWS. ALSO MANHATTAN, BROOKLYN AND QUEENS AREAS HAVE HIGHER REVIEWS THAN OTHERS.

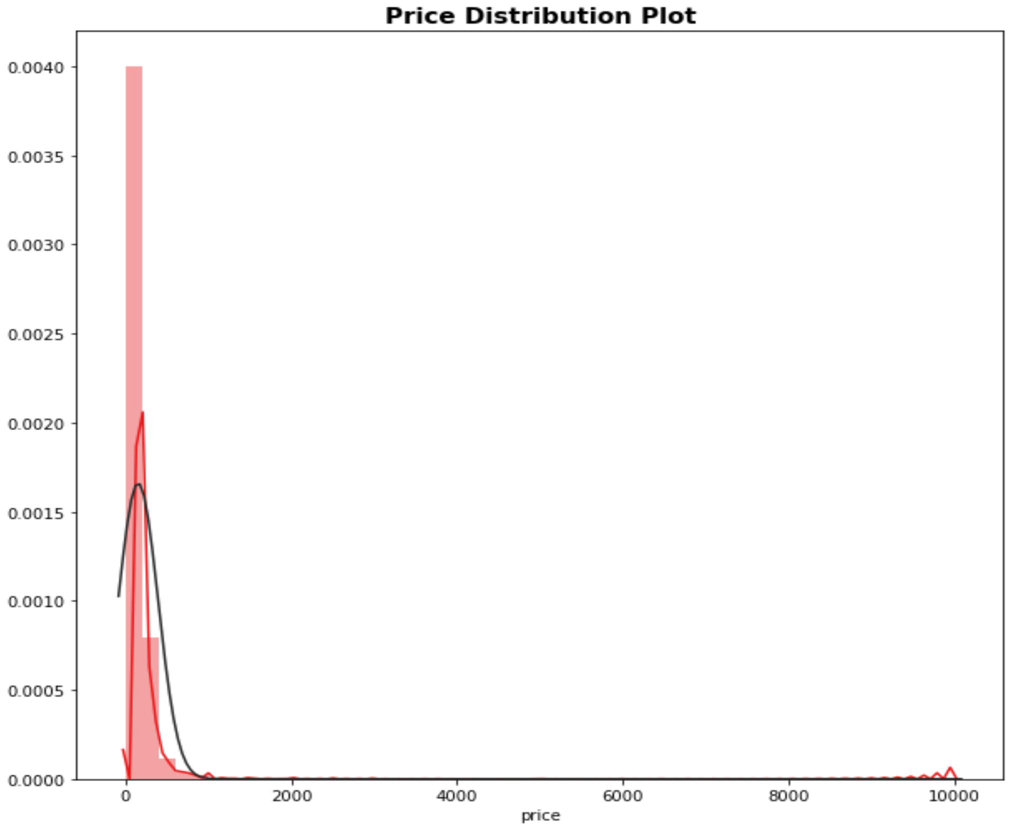


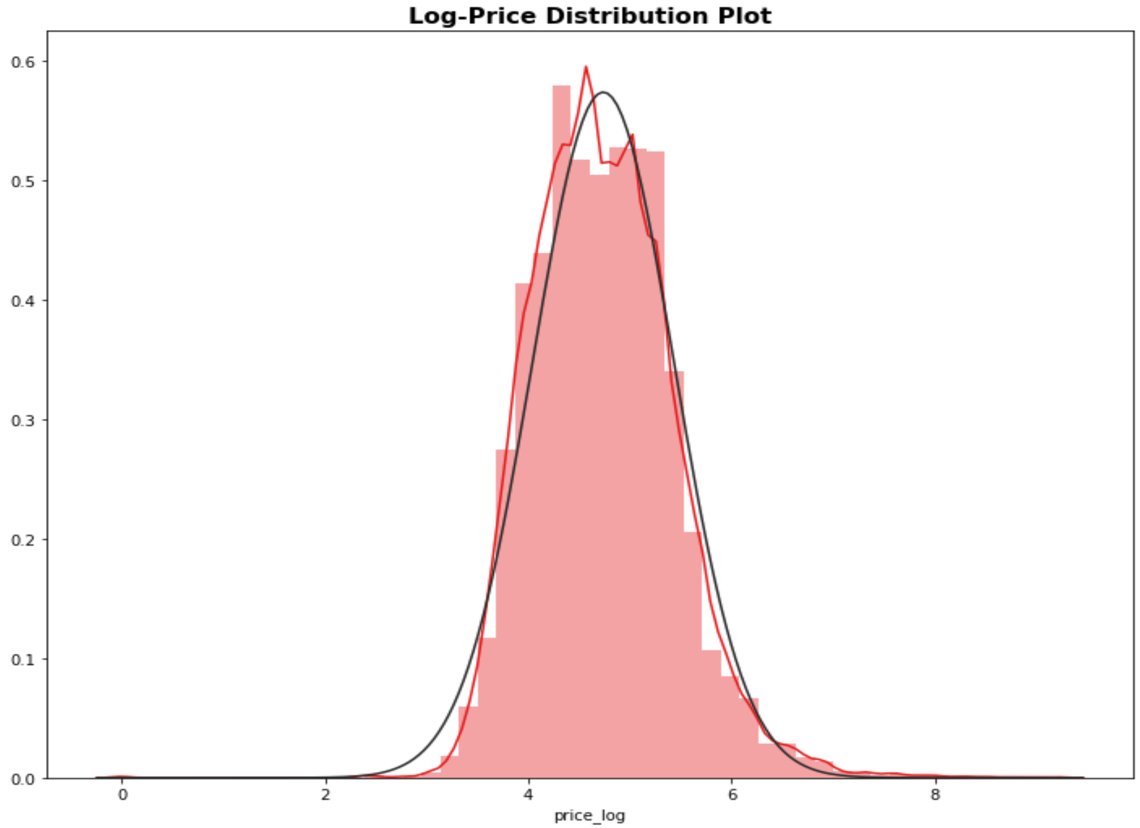
we can see that there is a good distribution between top 10 hosts with the most listings. First host has more than 300+ listings.



The above distribution graph shows that there is a right-skewed distribution on price. This means there is a positive skewness. Log transformation will be used to make this feature less skewed. This will help to make easier interpretation and better statistical analysis

Since division by zero is a problem, log+1 transformation would be better.





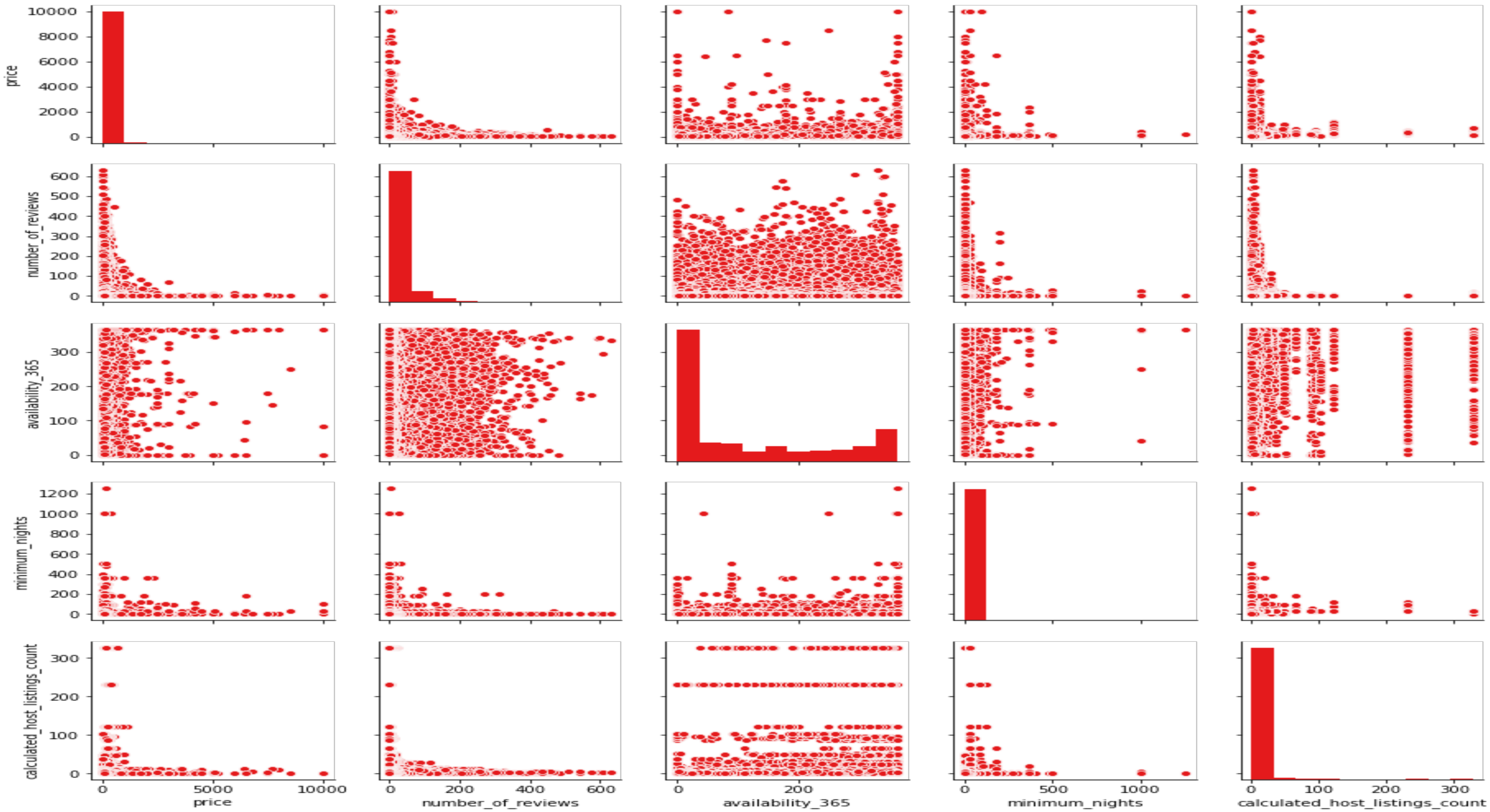
MULTICOLLINEARITY

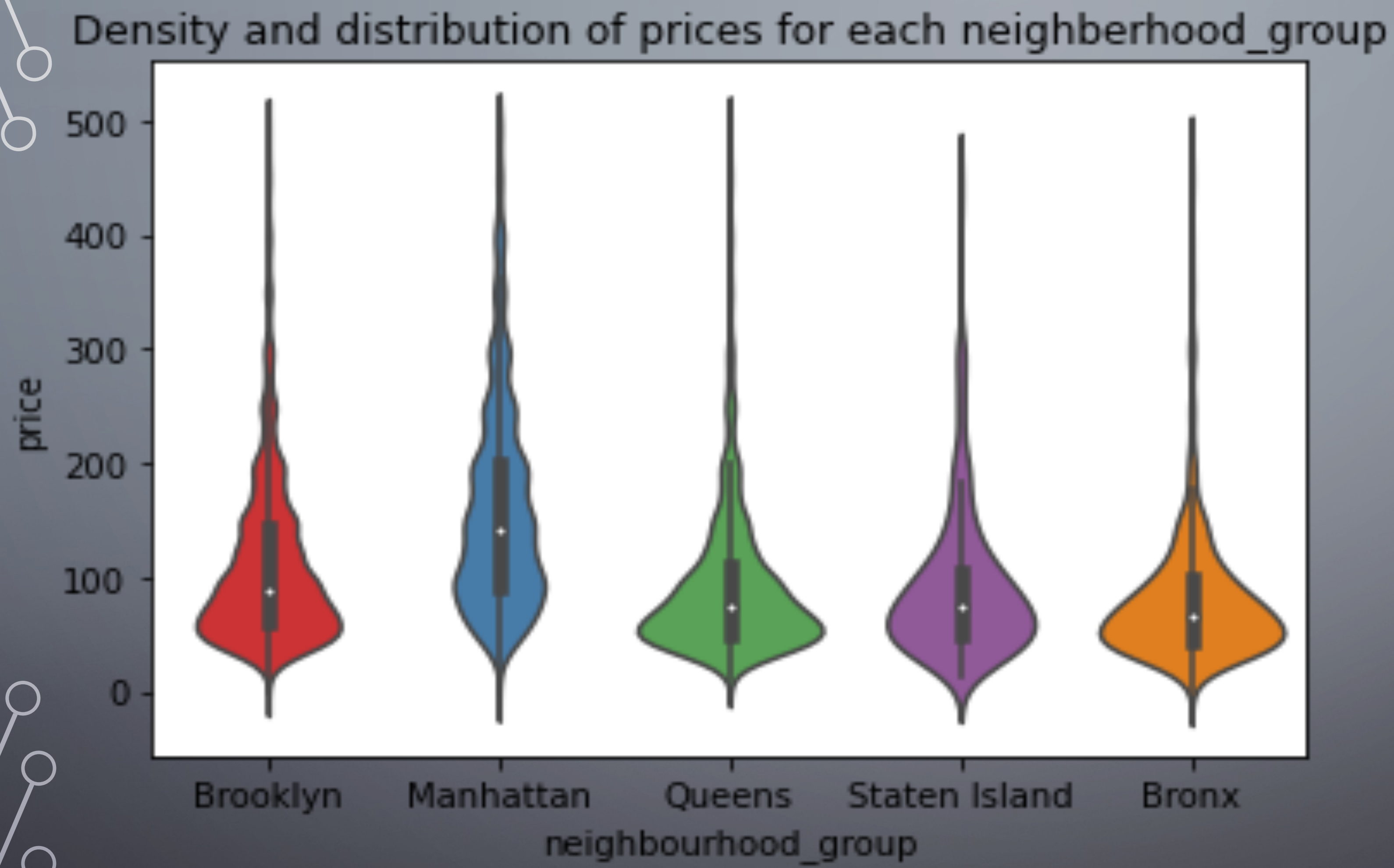
- Multicollinearity will help to measure the relationship between explanatory variables in multiple regression. If there is multicollinearity occurs, these highly related input variables should be eliminated from the model.
- In this kernel, multicollinearity will be control with Eigen vector values results.

```
array([1.88077542, 1.94050008, 1.634899 , 0.23788366,
0.32659394, 0.41850772, 1.19230211, 1.05330098,
0.63654962, 0.80614149, 0.87254598])
```

- None one of the eigenvalues of the correlation matrix is close to zero. It means that there is no multicollinearity exists in the data.

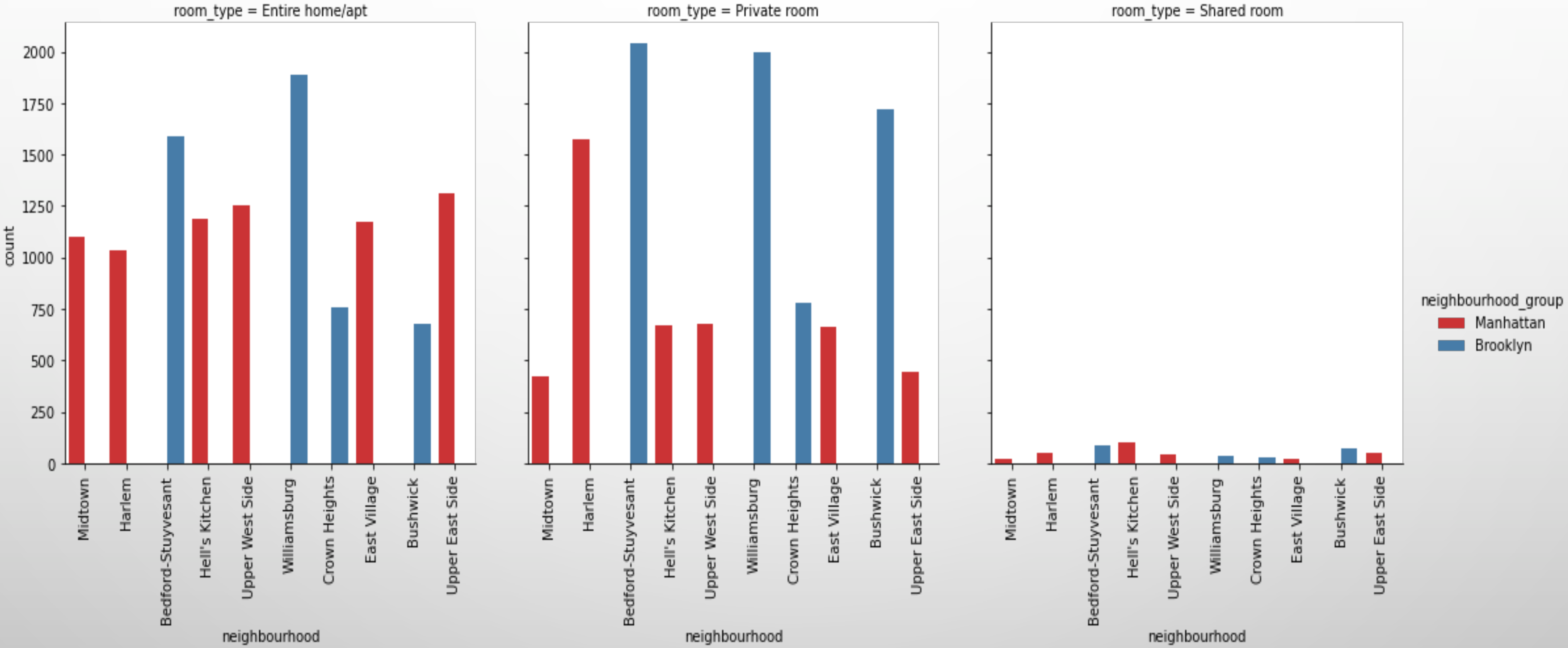
Visualizing Train Data Set



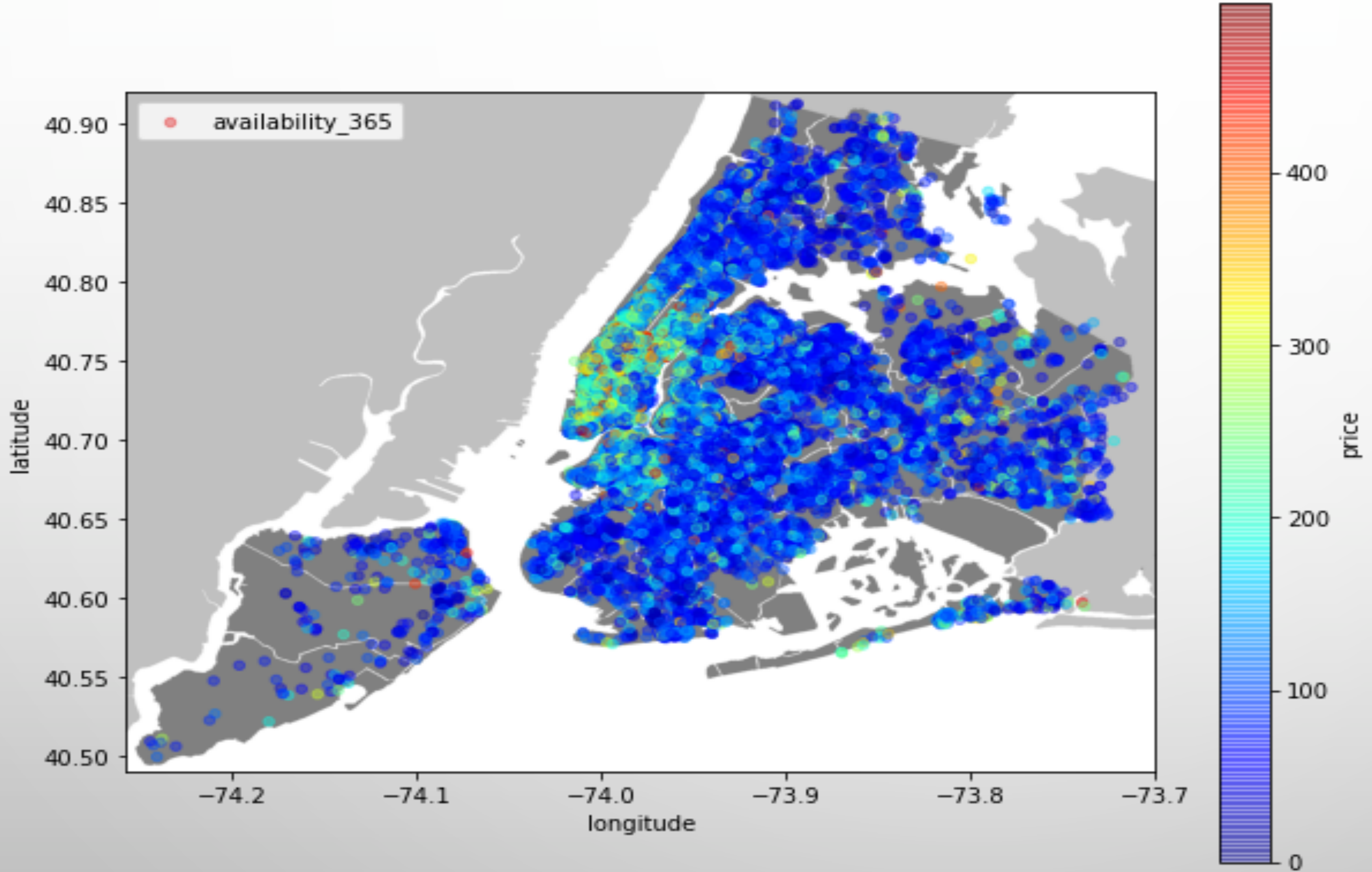


WITH A STATISTICAL TABLE AND A VIOLIN PLOT WE CAN DEFINITELY OBSERVE A COUPLE OF THINGS ABOUT DISTRIBUTION OF PRICES FOR AIRBNB IN NYC BOROUGHS. FIRST, WE CAN STATE THAT MANHATTAN HAS THE HIGHEST RANGE OF PRICES FOR THE LISTINGS WITH 150 *PRICE SAVER* OBSERVATION WITH $150 PRICE AS AVERAGE OBSERVATION, FOLLOWED BY BROOKLYN WITH $90 PER NIGHT. QUEENS AND STATEN ISLAND APPEAR TO HAVE VERY SIMILAR DISTRIBUTIONS, BRONX IS THE CHEAPEST OF THEM ALL. THIS DISTRIBUTION AND DENSITY OF PRICES WERE COMPLETELY EXPECTED; FOR EXAMPLE, AS IT IS NO SECRET THAT MANHATTAN IS ONE OF THE MOST EXPENSIVE PLACES IN THE WORLD TO LIVE IN, WHERE BRONX ON OTHER HAND APPEARS TO HAVE LOWER STANDARDS OF LIVING.

NEIGHBORHOOD AND ROOM TYPE

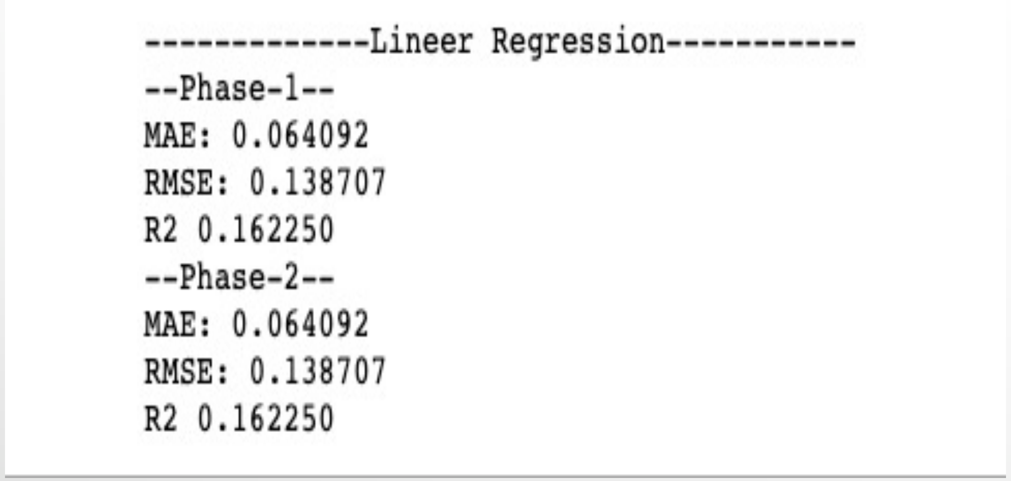


PRICE VS LOCATION IN NYC



After scaling our image the best we can, we observe that we end up with a very immersive heatmap. Using latitude and longitude points were able to visualize all NYC listings. Also, we added a color-coded range for each point on the map based on the price of the listing. However, it is important to note that we had to drop some extremely high values as they are treated as outliers for our analysis.

MODEL FORMATION USING LINEAR REGRESSION



```
-------------Lineer Regression-----------
--Phase-1--
MAE: 0.064092
RMSE: 0.138707
R2 0.162250
--Phase-2--
MAE: 0.064092
RMSE: 0.138707
R2 0.162250
```

The results show that the model have similar prediction results. Phase 1 and 2 have a great difference for the metric. All metric values are increased in Phase 2 it means, the prediction error value is higher in that Phase and model explainability are very low the variability of the response data around mean.

The MAE value of 0 indicates no error on the model. In other words, there is a perfect prediction. The above results show that all predictions have great error especially in phase 2.

RMSE gives an idea of how much error the system typically makes in its predictions. The above results show that all models with each phase have significant errors.

R2 represents the proportion of the variance for a dependent variable that's explained by an independent variable. 4. The above results show that, in phase 1, 13.8% of data fit the regression model and in phase 2, 13.8% of data fit the regression model.

CONCLUSION

Summarizing our findings, suggesting other features 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; we found that our top host has 327 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. For our data exploration purposes, it also would be nice to have couple additional features, such as positive and negative numeric (0-5 stars) reviews or 0-5 star average review for each listing; addition of these features would help to determine the best-reviewed hosts for NYC along with 'number_of_review' column that is provided. Overall, we discovered a very good number of interesting relationships between features and explained each step of the process. This data analytics is very much mimicked on a higher level on Airbnb Data/Machine Learning team for better business decisions, control over the platform, marketing initiatives, implementation of new features and much more. Therefore, I hope this kernel helps everyone!

APPENDIX

- Appendix A: Introduction and Importing Libraries
- Appendix B: Data Extraction
- Appendix C: Data Filtration
- Appendix D: Finding Unique Values
- Appendix E: Exploratory Data Analysis
    1. Multicollinearity
    2. Correlation Matrix
- Appendix F: Probability Graph
- Appendix G: Logarithmic Graph to showcase Skewness
- Appendix H: Linear Regression Modelling
    1. RFE
    2. Scalar MinMax
    3. R-squared value
- Appendix I: Conclusion

THANK YOU