Using Machine Learning to analyze Airbnb Data for New York City

Group 2C

Animesh Danayak, Dave Arno, Dhwani Mehta, Edwin Liu and Letian (William) Ma

BAX-401-002 Information, Insight, and Impact

TABLE OF CONTENTS

1. Executive Summary

Hosts on Airbnb are keen to find out how they should price their listings that would lead to higher profits for them in the long run, this is no different for hosts in New York City. As the fifth most popular city on Airbnb in the world, New York has over 50 thousand apartment listings and competition is fierce. In particular, the report aims to answer 3 key questions using New York City data from Inside Airbnb:

- How do rental properties vary across the neighborhoods in New York
- How do prices vary with neighborhoods, rental property types, and rental amenities
- How to predict Airbnb rental prices using features such as property type, number of beds, past reviews, etc

These are some extracts from our Exploratory Analysis to answer the pain points faced by hosts:

- New York City's listings have an average price tag of $128, with expensive neighborhoods listed as high as $750 per night in Fort Wadsworth and $550 in Westerleigh
- Uncommon properties such as boats and resorts are the most expensive property types, whereas the top three most common property types, apartments, houses and townhouses, are reasonably priced at under $200
- Williamsburg, Bedford-Stuyvesant, and Harlem are the top three neighborhoods with the most listings

We used Lasso Regression as our Machine Learning model to precisely predict listing prices of any property in NY City with a variation of $43. Using the outcomes from the model, we provided recommended prices of their properties to hosts along with other recommendations. Hosts can use our findings to enhance the attractiveness of their listings.

2. Introduction

What are the deciding factors when you book an Airbnb? Every year, hundreds of millions of travelers who travel with Airbnb worldwide have the same question in mind. The vast demand drives hundreds of thousands of new listings on Airbnb yearly, adding to the expansive 7 million total listings worldwide. As the fifth most popular city on Airbnb in the world, New York has over 50 thousand apartment listings and competition is fierce. With 72% of Airbnb hosts use their revenue to remain in their homes in New York, hosts are dire to find out how they can stand out among the competition, what are the features and amenities that would lead to higher prices, and whether they are likely to be successful by starting a new Airbnb business in a specific neighborhood.

In answering the pain points above, we started with an Exploratory Analysis of the Airbnb data we obtained from Inside Airbnb to have a general understanding of the Airbnb rental landscape in New York City. We then used machine learning to analyze Airbnb data for New York City to address the problems outlined in our Problem Statement. We used Lasso Regression (L1 regularization model) for predicting Prices. Using the outcomes from the model, we provided recommended prices of their properties to hosts along with other recommendations.

3. Problem Statement

Airbnb has drastically changed the tourism and hotel industry, now more so with the Pandemic and travel restrictions. People travel in smaller groups and avoid hotels. The preference has shifted to Airbnb apartments instead. Our project aims to answer 3 key questions:

1. How do rental properties vary across the neighborhoods in New York (Proximity to the subway stations)
2. How do prices vary with respect to neighborhoods, rental property types and rental amenities
3. How can we use machine learning to predict Airbnb rental prices using features such as property type, number of beds, past reviews, etc.

4. Data Characteristics

We obtained our dataset from Inside Airbnb: http://insideairbnb.com/get-the-data.html. The dataset that we are using in this report will be the *listings.csv* file from their New York City repository. *Listings.csv* contains 96 columns and 50,969 rows, and each row represents a listing in New York City on Airbnb.The data set contains attributes such as *price per day, property type, rental amenities, listing id, neighborhood, room type, and review scores.* This set of scraped data is extremely useful in helping us answer the problem statements above because of the size and completeness of the file, giving us plenty of attributes and features to work with.

We first explored the data in the Exploratory Data Analysis stage to grasp a better understanding and pull initial insights into the features that we have. We will present some of our findings with visualizations here.

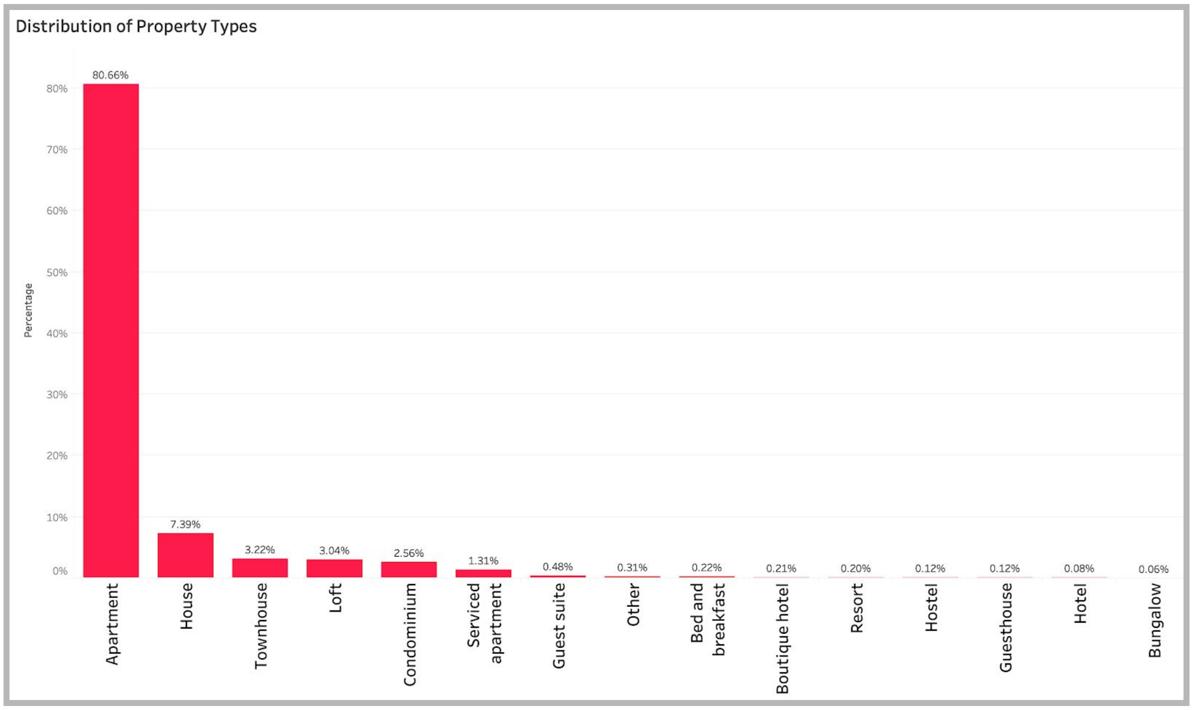


Figure 1 shows that 80% of listings are apartments, followed by houses and townhouses.

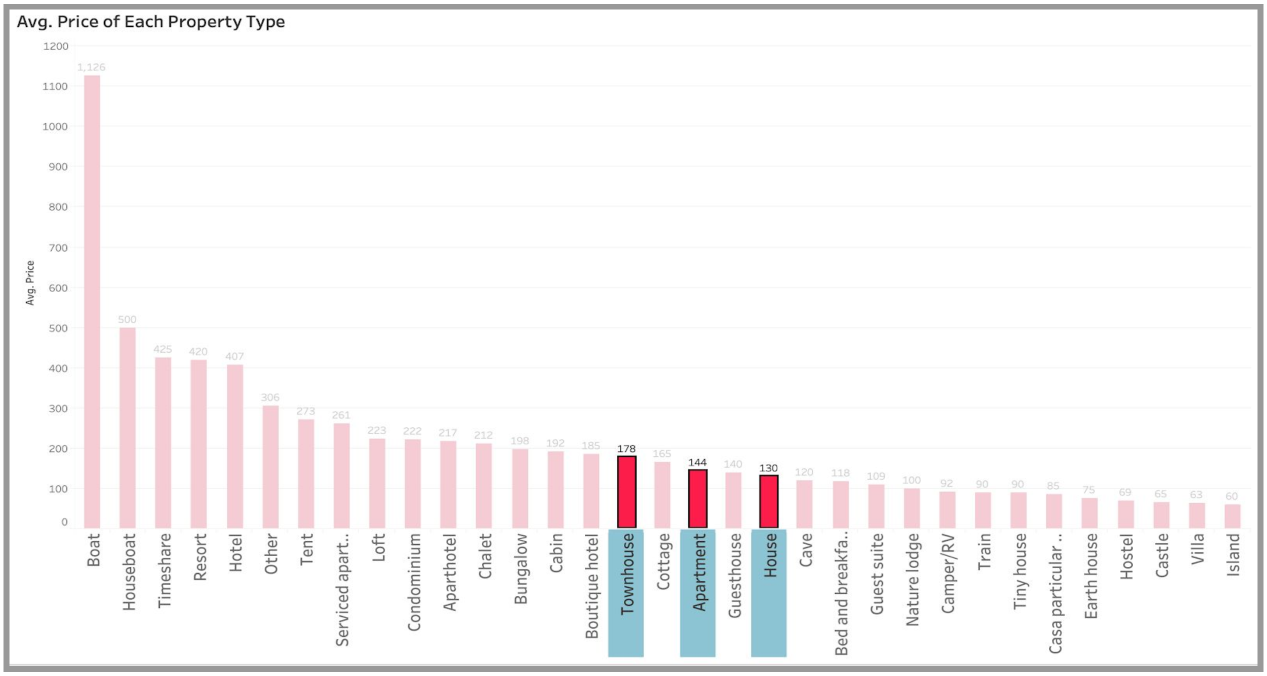


Figure 2 shows that the most expensive property types are uncommon properties such as boats and resorts, whereas the top three most common property types shown in figure 1 are reasonably priced at under $200.

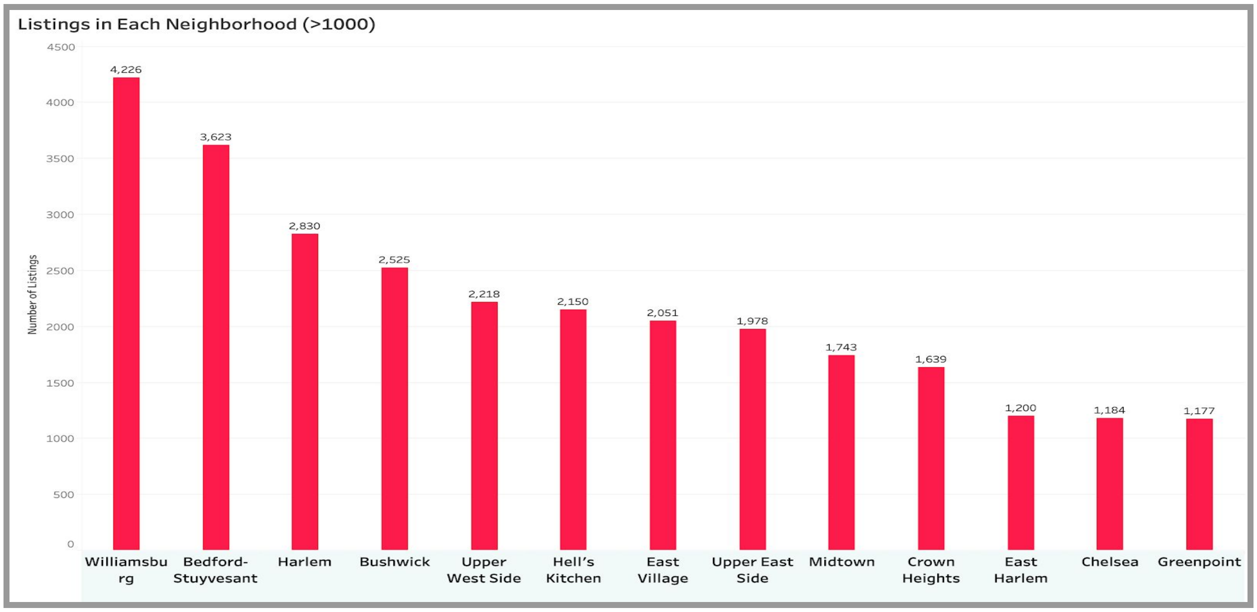


Figure 3 depicts that the neighborhood with the highest density of listings in New York (only neighborhoods with more than 1000 units are shown). Williamsburg, Bedford-Stuyvesant, and Harlem are the top three neighborhoods with the most listings

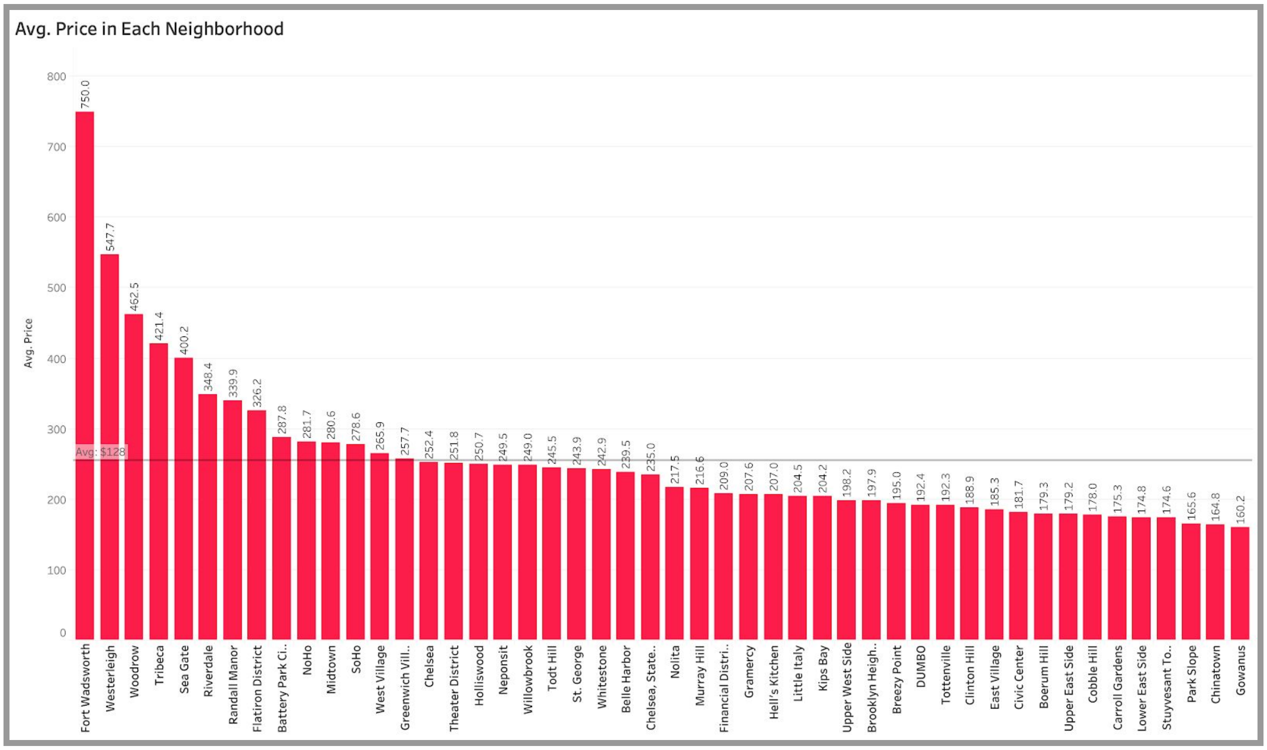


Figure 4 shows the average price of listings in each neighborhood. New York City's listings have an average price tag of $128, with expensive neighborhoods listed as high as $750 per night in Fort Wadsworth and $550 in Westerleigh.

5. Feature Engineering

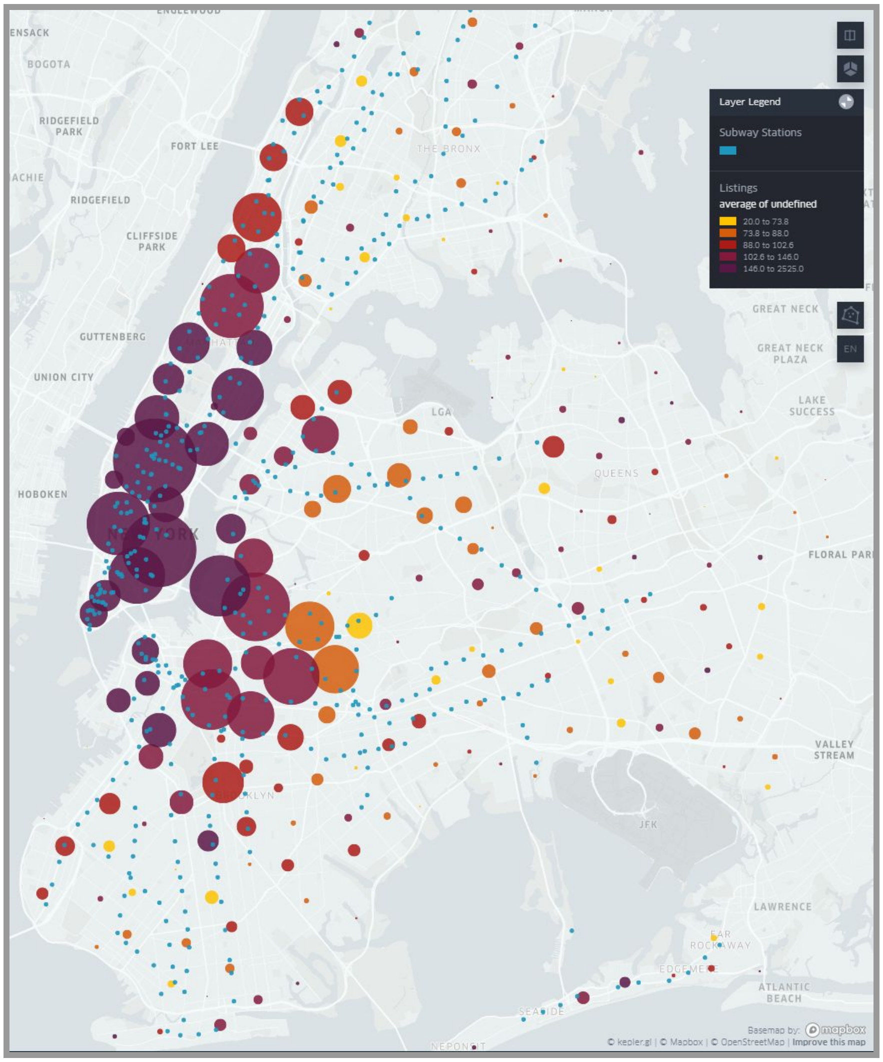
Next we brainstormed potential features we could add to our models. We decided that finding the distance between each listing and the nearest NYC subway station would be a meaningful attribute in the value of listings. Luckily our data include the latitude and longitude of each listing which gave us a good starting point. Next we found the coordinates of all subway stations in the city from NYC Open Data, a public site dedicated to providing public information. After cleaning both datasets, we programmed a function to calculate the distances from the closest subway station for each listing. To accomplish this we relied on the haversine formula in order to take into account the earth's curvature.

$$d = 2r \arcsin\left(\sqrt{\sin^2\left(\frac{\varphi_2 - \varphi_1}{2}\right) + \cos(\varphi_1)\cos(\varphi_2)\sin^2\left(\frac{\lambda_2 - \lambda_1}{2}\right)}\right)$$

where

- φ_1, φ_2 are the latitude of point 1 and latitude of point 2 (in radians),
- λ_1, λ_2 are the longitude of point 1 and longitude of point 2 (in radians).

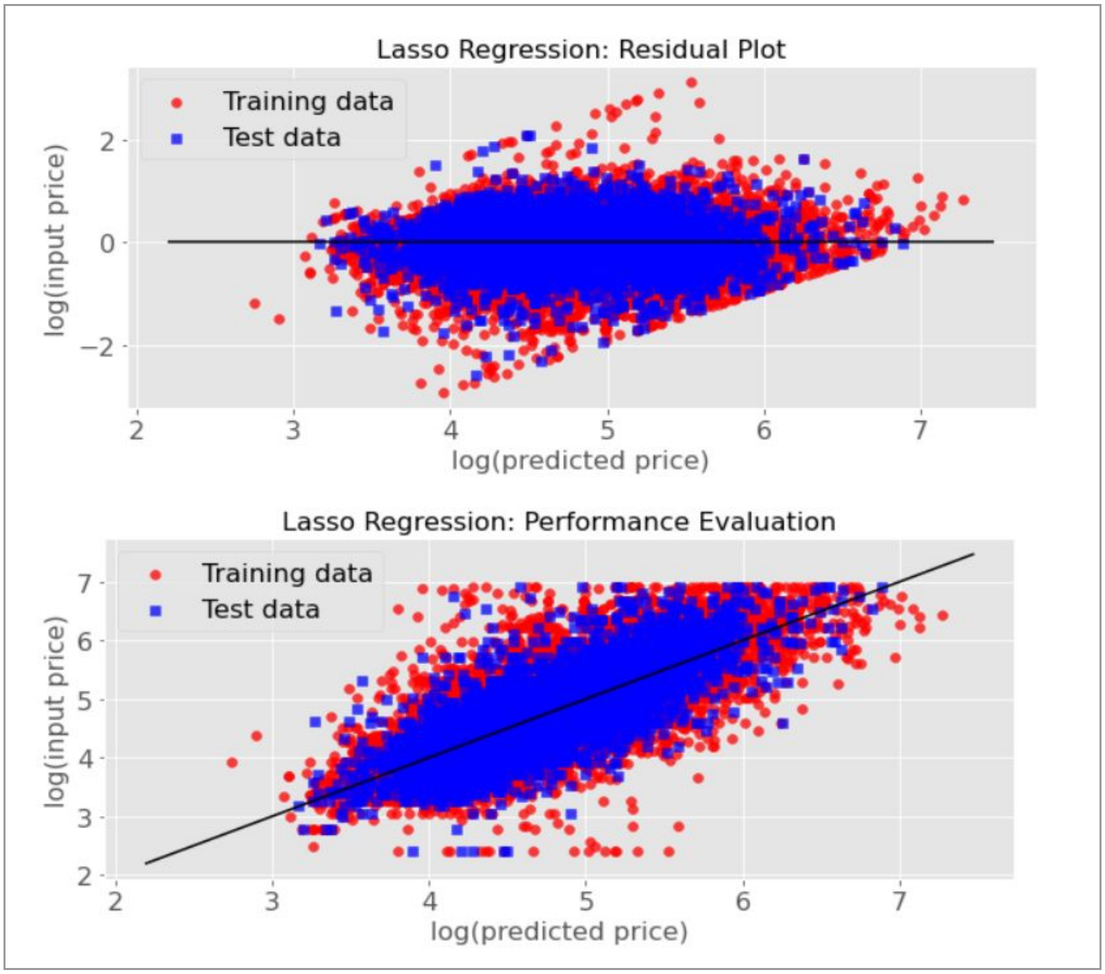
Using this we were able to produce the exact distances to the closest subway stations for all 50,969 listings.



The map above illustrates the relationships between NYC subway stations and listings on Airbnb. The listings are represented in varying sizes of clusters, with the number of listings represented by the size of clusters and the average price of listings represented by color. Clusters in yellow are listings with the lowest average price while clusters in purple are listings with the highest average price.

6. Model Selection and Results

For predicting Price, we used Lasso Regression (linear regression model with L1 regularization) using the scikit-learn in Python, the hyperparameters were tuned using a 10-fold cross-validation on the training data set. We ran 10,000 iterations on the Lasso Regularization. The Lasso Regression tries to optimize a cost function which in turn reduces underfitting of the model.



The results on the response variable (Price) from the training data set and the predictions made on the test data set is summarized in the table below:

Model	Train RMSE	Test RMSE	Train R^2	Test R^2	Train Accuracy	Test Accuracy
Lasso Regression	$82	$79	0.64	0.64	69%	69%

7. Recommendations and Managerial Implications

The model uses distance from a prominent landmark in NY City (Subway Station) to determine how it affects rental price. This can be extrapolated to include different locations anywhere in the world. This would help property owners determine the correct price for their property and also allow Airbnb to predict their possible revenues by different locations.

A shortcoming of this model is that it does not address the causality between all the variables. Also, certain external factors like customer preferences, host ratings and inherent local traits could also play a role in rental prices.

There also exist biases in the data selection. For example: we chose NY City and the Subway station as the landmark. Using another city or landmark could perhaps reduce or increase the chances of correctly predicting the rental amount.

We could improve the prediction accuracy by perhaps reducing the number of features used to determine the price. We could also look at other alternative models like gradient boosting trees, k-nearest neighbors and random forest regression.

8. Conclusion

Our project aims to identify rental prices within different neighborhoods in NY city. We used a dataset of Airbnb listings in NY city for the past year. Our data had different metrics like location of property and its distance from the subway station, size of the property, list of amenities that it offers, etc. Of the 100 metrics, we shortlisted ~50 which were related to pricing of a property by using regression analysis. We ran Lasso Regression to predict the price of a property in NY city with highest test accuracy. Our model has an error rate of $43, that is, it allows you to precisely predict the rental price of any property in NY City with a variation of less than the cost of a decent meal for two!

9. References

1. Peshin, A., Gupta, S., Agrawal, A. (2019). *Analyzing Airbnb Rentals Dataset*. GitHub. https://github.com/Ankit-Peshin/airbnb#readme
2. Atta-Fynn, R., Zien, C. (2019). *Analysis and Machine Learning Modeling of New York City Airbnb Data*. NYC Data Science Academy. https://nycdatascience.com/blog/student-works/analysis-and-machine-learning-modeling-of-new-york-city-airbnb-data/
3. Dgomonov. (2019). *New York City Airbnb Open Data*. Kaggle. https://www.kaggle.com/dgomonov/new-york-city-airbnb-open-data
4. Cox, M. (2020). *Get the Data*. Inside Airbnb. http://insideairbnb.com/get-the-data.html
5. Heath, A. (2015). *Here's how much New Yorkers make renting their apartments on Airbnb*. Business Insider. https://www.businessinsider.com/how-much-new-yorkers-make-from-airbnb-2015-12
6. Deane, S. (2020). *2020 Airbnb Statistics: Usage, Demographics, and Revenue Growth*. Stratos Jet Charters, Inc. https://www.stratosjets.com/blog/airbnb-statistics/
7. Metropolitan Transportation Authority (MTA). (2010). *City Subway Stations. [Data File]*. NYC OpenData. https://data.cityofnewyork.us/Transportation/Subway-Stations/arq3-7z49

10. Appendix

- This is where you add your data descriptions, analysis (e.g. model selection criteria, other possible ways to solve the problem), extra tables and figures. Please do not treat this section as the forgotten child. It is just as important and hence connect it to the report. Again, less jargon and no orphan notations. Number and label each appendix, if you have more than one.

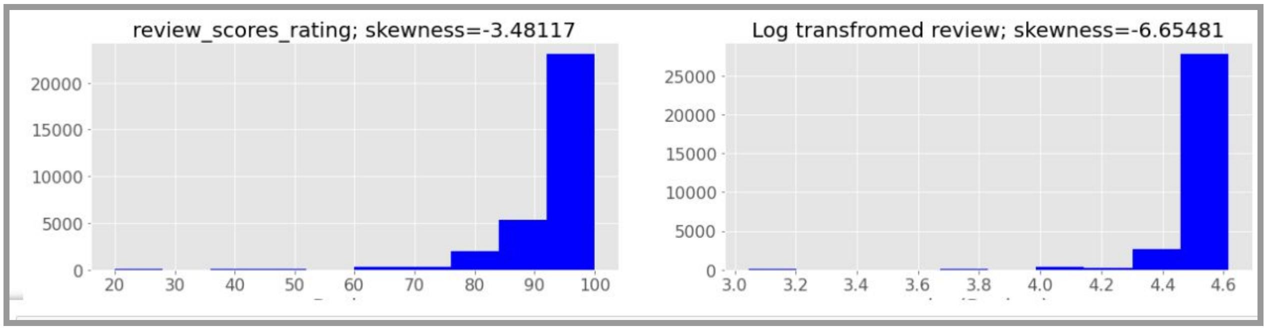
1. Interim Model Development steps and results:

2. Data Sources: See references section

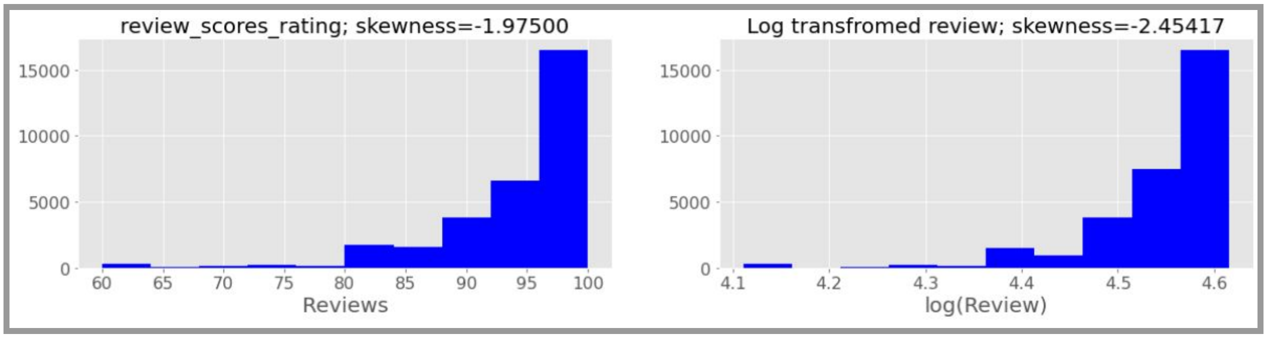
3. Code Base: Attached

4. Tableau File: Attached

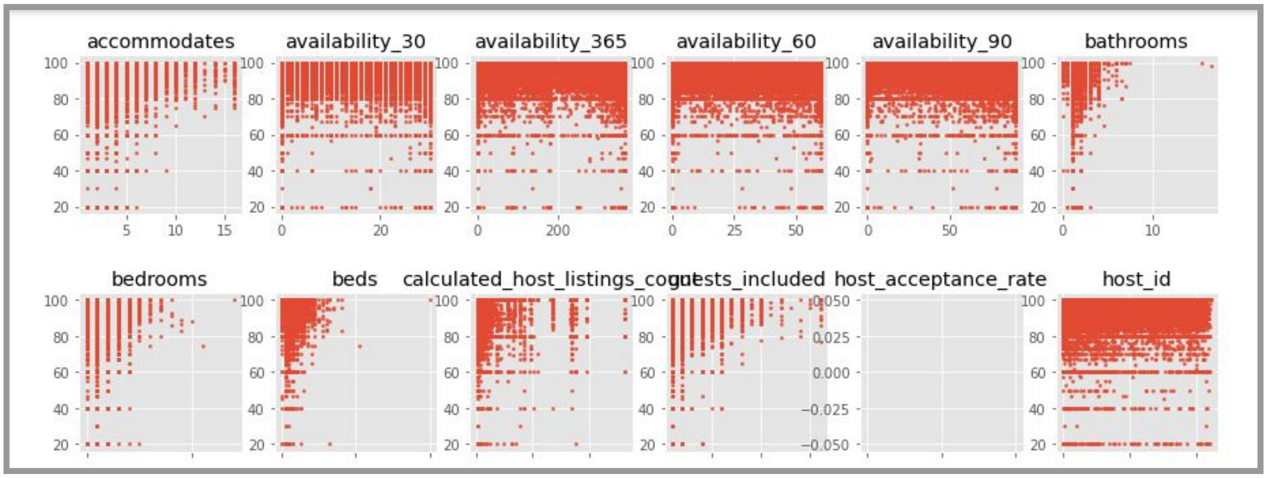
We started out with having Reviews as the response variable but EDA on the Reviews data showed heavy bias in the reviews:

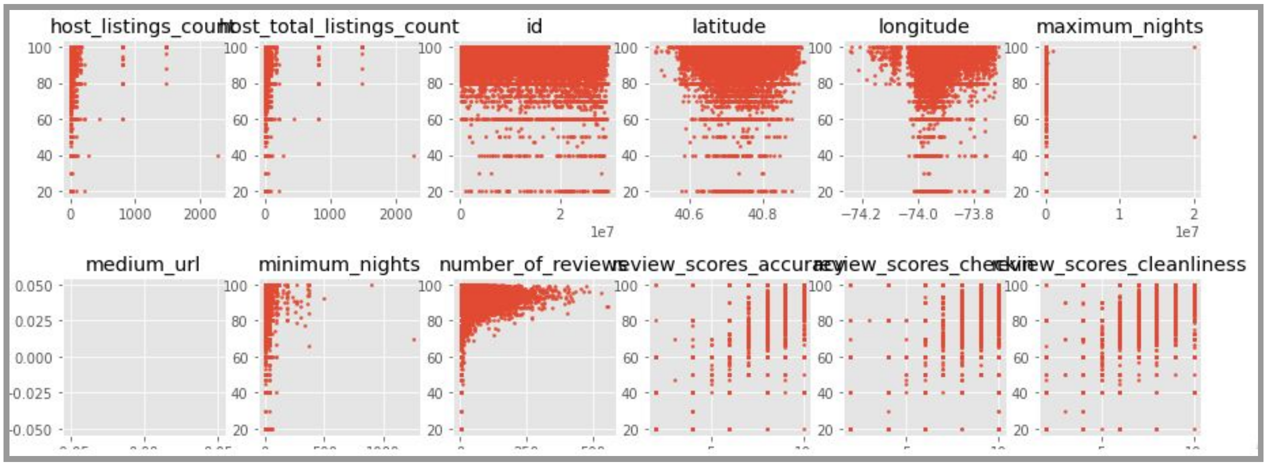


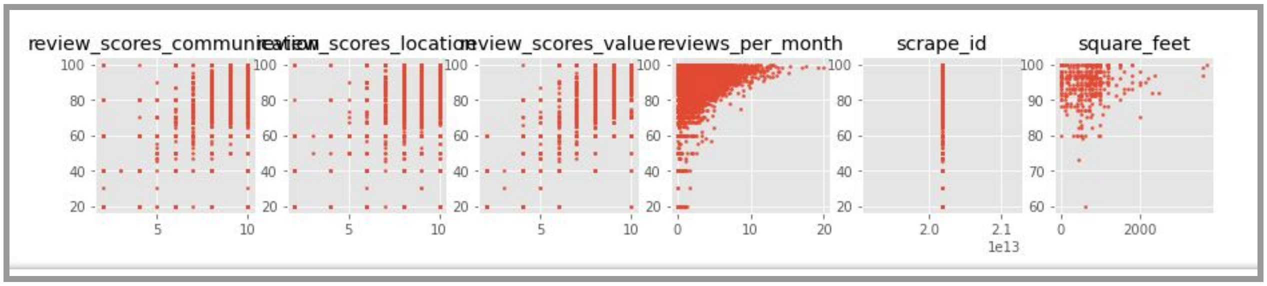
We tried removing the outliers to no avail:



Distribution of numerical variables wrt Reviews:







We realized that the biased nature of Airbnb’s reviews stems from the fact that most people connect with their hosts while staying at properties and don’t want to give negative reviews (as to in a hotel/motel setting where customers are more critical). This was heavily criticized and Airbnb changed their reviews policy in the late 2019. More about the policy changes here.