Predicting Airbnb Prices in New York City

Kyle Guanzon, Hannah Wen, and Charles Zhang

Introduction

Airbnb has revolutionized the hospitality industry by offering a platform that connects travelers with unique lodging options around the world. Unlike traditional hotels, Airbnb listings provide a diverse range of accommodations, from entire homes and apartments to private and shared rooms, catering to various budget ranges and preferences. This flexibility and variety, combined with the ability to stay in convenient and often residential locations, make Airbnb a popular choice for many travelers. Additionally, the platform offers opportunities for hosts to earn income by renting out their properties.

Our project focuses on understanding the factors that influence Airbnb rental prices in New York City, one of the most dynamic and competitive markets for short-term rentals. By analyzing a comprehensive dataset of Airbnb listings, we aim to uncover the key determinants of pricing and provide actionable insights for both hosts and potential guests. Our analysis employs a combination of statistical techniques and machine learning models to explore the complex interactions between various features and rental prices, ultimately seeking to enhance pricing strategies and optimize listing performance.

Exploratory Data Analysis

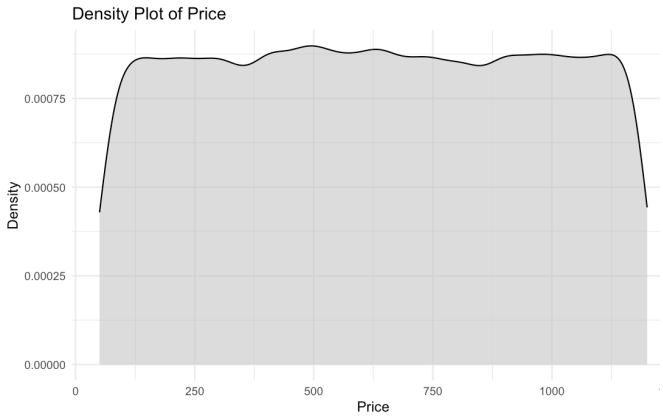


Figure 1

Data Reduction

The data was reduced from 102,776 observations and 26 columns to 84,655 and 18 columns. Columns were removed because they were either redundant, such as country code, or not able to analyze, such as the description. However, new columns were created from data from some of these columns. Inaccurate values, such as the year of the most recent review being larger than 2023, were removed as well as N/A values in columns we considered for EDA and modeling.

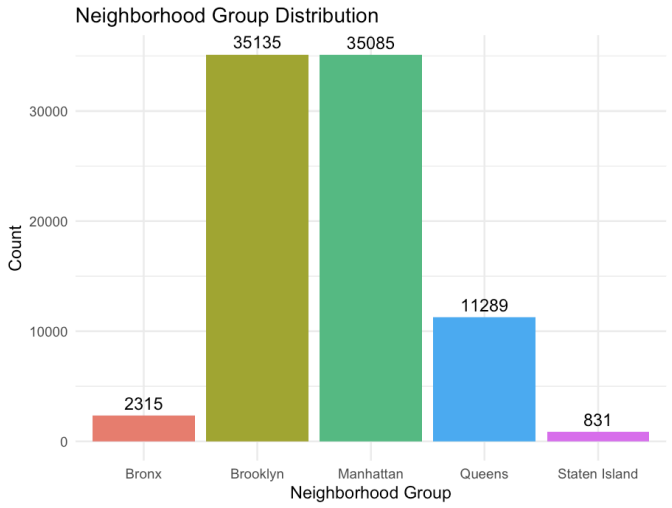


Figure 2

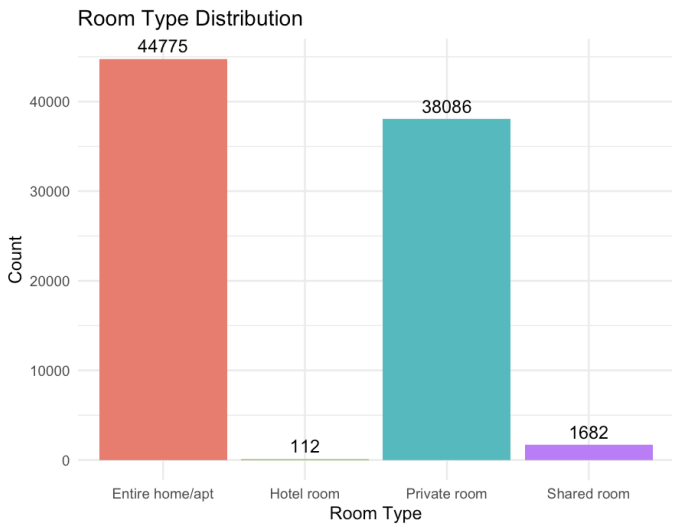


Figure 3

There are four different room types: entire homes and apartments, hotel rooms, private rooms, and shared rooms. Entire homes/apartments are the most popular type of room, representing 52.89% of listings. Private rooms are the second most popular, representing 44.99%. Entire homes/apartments and private rooms account for 97.88% of listings.

For Manhattan, Brooklyn, and Staten Island, entire homes and apartments are the most common room type while private rooms are the most popular for Queens and the Bronx. The variable ***cancellation_policy*** records the level of flexibility of the cancellation policy of the listing with the rankings: flexible, moderate, and strict. These classes are balanced in the dataset, each accounting for approximately 28,000 observations. It is also observed that the distributions of data for each class are also fairly uniform. For example, when comparing to the variable ***instant_bookable***, which records whether the listing allows instant bookings, each class has approximately 50% of its listings with and without this ability. Going further, the distribution of neighborhood groups is also fairly equal for these classes with approximately 41% of observations accounting for listings in Manhattan and Brooklyn each. Additionally, the distribution of room types did not change according to the cancellation policy, with approximately 53% and 45% of listings representing an entire home/apartment and private room, respectively. When it comes to price, the level cancellation policy did not appear to have an effect on the mean.

Airbnb Price by Policy Level and Neighborhood Group

Neighborhood Group	Flexible	Moderate	Strict
Bronx	621.32	644.67	625.47
Brooklyn	630.67	626.42	625.50
Manhattan	622.20	622.36	624.85
Queens	634.64	628.79	628.58
Staten Island	646.77	606.32	617.51

Figure 4

The mean prices for flexible, moderate, and strict policies were $627.44, $625.25, and $625.50, respectively. However, among the neighborhood groups, the level of cancellation policy did change the mean prices despite each neighborhood group having approximately even numbers of each level. For example, the highest mean price for listings in the Bronx is $644.67, as seen in *Figure 4*, with moderate cancellation policies, while the highest mean price for listings in Staten Island is $646.77 with flexible cancellation policies. Additionally, each neighborhood group has a different range of means among the cancellation policy levels. For example, listings in Brooklyn have mean prices for each cancellation policy type ranging from $625.47-$630.37 while listings in Staten Island range from $605.70-$645.19.

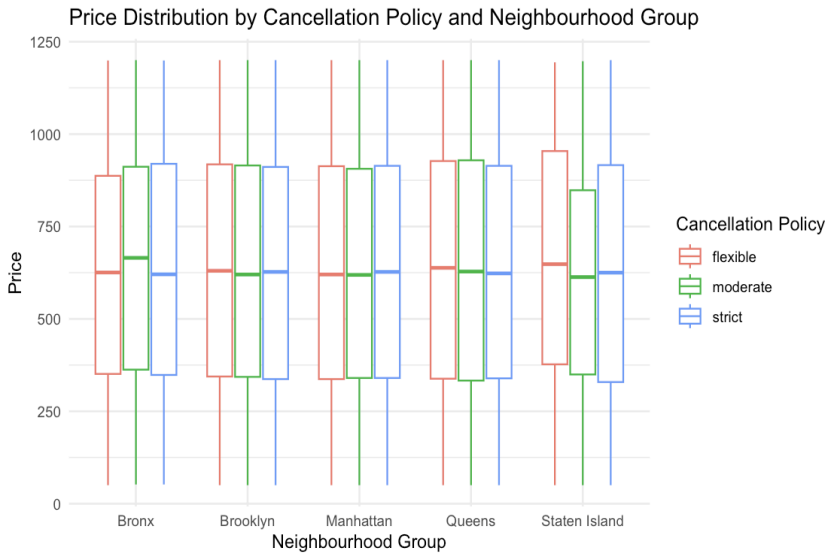


Figure 5

Looking at the boxplot graph of the distribution of price over cancellation policy and neighborhood group in *Figure 5*, it is observed that the distributions are very similar for Brooklyn, Manhattan and Queens, but there are notable differences between policy types for the Bronx and Staten Island.

When it came to the distribution of price over whether listings were instantly bookable or not, there did not appear to be a notable difference among the different neighborhood groups. The difference between mean prices depending on whether a listing was instantly bookable only ranged between 2-10 dollars and there was not a consistent pattern of instant bookings being more or less expensive.

Staten Island was the only neighborhood group with a difference in mean prices of over $10, with instant bookings having an average price of $633.76 and non-instant booking having an average price of $612.08.

There are several variables which cover the reviews of the listings as well as the amount of listings the host has. The average number of monthly reviews a listing gets is 1.374565 and the average number of total reviews is 32.196153. The average rating of a listing is 3.280365 and the average number of listings for a host is 7.028988. The median rating is 3 and the median number of listings for a host is 1. The amount of monthly reviews is moderately correlated with the total number of reviews with a Pearson correlation coefficient of 0.594685. However, the correlation between the number of reviews and rating as well as number of host listings is very weak. Ratings are split into whole integer scores of 1-5. Each score accounts for approximately 23% of the listings except for the score of 1, which represents 8.85% of listings. Looking at the distribution of ratings over each neighborhood group, it appears that the highest mean price differs according to the rating, which may indicate that the ratings should be treated as a categorical variable.

Service fee is a variable which keeps track of the service fee of each listing. It is very strongly correlated with the price because it is supposed to represent 10-14% of the price of the listing. The Pearson correlation coefficient is 0.9999909 which is a nearly perfect correlation. Because of this, the service fee will not be included in predictive modeling.

The ***minimum nights*** variable records the minimum number of nights required for a booking. The mean is 7.447 while the median is 3. The distribution of this variable changes depending on room type, with a median of 1 for hotel and shared rooms, but a median of 2 for private rooms and a median of 3 for entire homes and apartments. The neighborhood groups the Bronx, Queens, and Staten Island had a median of 2 while the groups Brooklyn and Manhattan had a median of 3, despite the fact that the groups have similar distribution of room types.

Mean Future Availability by Room Type

Room Type	Mean Future Availability
Entire home/apt	142.4597
Hotel room	218.7679
Private room	140.5521
Shared room	173.3639

Figure 6

Mean Future Availability by Neighborhood Group

Neighborhood Group	Mean Future Availability
Bronx	183.3158
Brooklyn	132.9930
Manhattan	139.6424
Queens	166.8812
Staten Island	201.4970

Figure 7

The column ***availability_365*** records the number of future nights the listing is available to be booked for. The mean is 142.3 and the median is 102. The first and third quartile are 7 and 267, respectively. Private rooms have the lowest mean future availability with a mean of approximately 141 nights available in the future while hotel rooms have the highest with a mean of approximately 219 nights available in the future, as reflected in *Figure 6*. Additionally, the mean future availability differs according to neighborhood groups, with Staten Island having the highest mean future availability of approximately 201 nights, as seen in *Figure 7*, and Brooklyn having the lowest mean future availability of approximately 133 nights.

Construction year represents the year the listing has been constructed or last remodeled. The minimum is 2003 and the maximum is 2022. Both the median and mean are 2012. The most common construction year is 2006, but all of the construction years have a relatively even distribution, representing between 4100-4400 listings each. The median construction year differs slightly depending on the neighborhood group. Brooklyn has the highest median year of 2013 while Staten Island has the lowest median of 2011. The remaining groups have a median construction year of 2012.

Looking to investigate the string type columns house_rules and NAME, new variables were created from string manipulation. The ***house_rules*** variable was manipulated to create new variables by searching for certain text in each observation. The variable ***no_smoking*** noted whether there were the words "smoke" or "smoking" in each listing's rules representing whether the listing prohibited smoking or did not mention it. Approximately 24% of listings mentioned smoking in their house rules. The mean price for listings which mentioned smoking in their rules is $619.12 while the mean price for listings which did not mention smoking is $628.21. The distribution of the no smoking variable does not differ according to the level of cancellation policy but does according to neighborhood groups. Staten Island has the least amount of mentions of smoking with 17.8% of listings while Queens and the Bronx had the most with approximately 21% of listings.

Another variable ***no_rules*** is another boolean column which measures whether the listing had any house rules described or not. 52.76% of listings did not have any house rules listed while 47.24% of listings did. This distribution also did not differ according to cancellation policy level but did for neighborhood groups. The Bronx has the highest percentage of listings without house rules with approximately 59% of listings while Brooklyn and Manhattan have the lowest with approximately 52% of listings. The mean price for listings with house rules is $623.38 and $628.46 for listings without house rules.

A variable was created from manipulating the NAME column which contains the listing description to note the ***number of bedrooms*** an Airbnb has. It is noted that not every description will mention the number of bedrooms. The 3 most popular number of bedrooms are 1, 2, and 3 representing 92%, 5.6%, and 1.7% of listings, respectively. There is very little correlation between the number of bedrooms and the price, but it is noted that the mean prices varied wildly according to the number of bedrooms, indicating a non-linear relationship. Additionally, the correlation can also be explained by 1 bedroom representing a very large majority of listings.

Figure 8 presents a correlation matrix. It is observed that there are very little linear relationships between predictors and price. In general, between predictors there are also little linear relationships, with the most significant one being a moderate positive relationship between number of reviews and reviews per month.

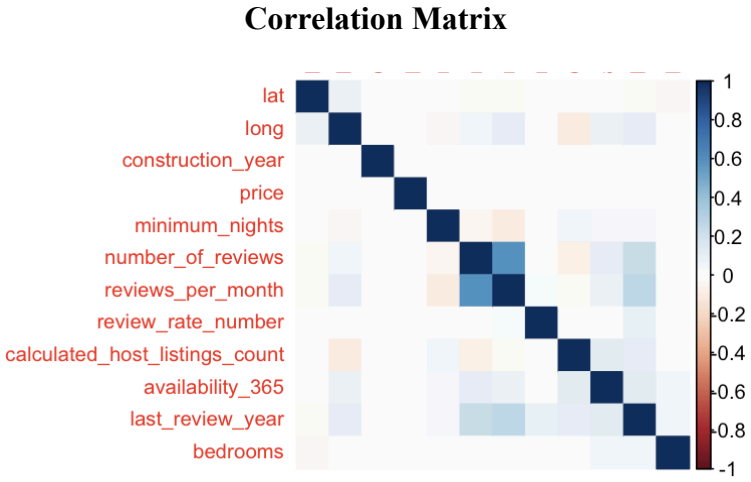


Figure 8

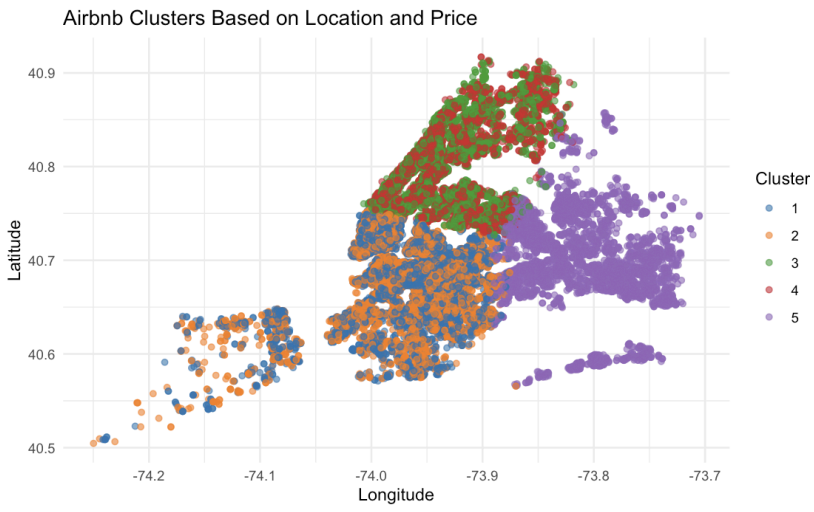


Figure 9

Clustering

K-means clustering was applied with k = 5 to partition the dataset into 5 clusters based on geographical coordinates (latitude and longitude) and price, as illustrated in *Figure 9*. By clustering data points, we were able to isolate high, low, and variable pricing zones across the city's boroughs. High-price clusters were predominantly located in Manhattan and certain parts of Brooklyn, indicating areas with higher priced listings on average. In contrast, more moderate and varied pricing was observed in Queens, the Bronx, and Staten Island, suggesting a mix of budget options and residential areas. The clustering indicates that while there is variation in price in many locations, location may be an important factor in pricing.

Methodology

Preparing for predictive modeling, a train and test data set were created. Using the Airbnb dataset, the service fee, review rate number, house rules, last review date, neighborhood group, and description name columns were removed to create the data used for modeling. Because variables were created from the house rules and description columns which were free text strings, they were removed from the training set. Service fee had a very high correlation with price and neighborhood group had a high correlation with neighborhood, so they were removed due to multicollinearity issues. Last review date and the review rating number were removed because new altered versions of these columns were created as variables. This leaves 18 columns in the data set. The training data set is made up of 80% randomly chosen observations of the Airbnb data set while the testing set is made up of the remaining 20%. A dummy variable dataset which one hot encoded the categorical data was created as well.

Linear regression, lasso regression, PCA, random forest, bagging, and boosting models were utilized as regression methods for predicting the price. A linear regression with price as the predicted variable and all of the variables was performed. Lasso regression was trained on the one hot encoded dataset. The random forest models were tested with the ranger function in R due to the size of the data. Two random forest models were tested with the default $m = p/3 \approx 5$ and a bagging was performed with $m = p = 16$. Both models had 500 trees and returned the total decrease in node impurity that results from splits over that variable, averaged over all trees. Additionally, a random forest model was trained on PCA transformed data. Finally, boosting was performed with xgboost on the one hot encoded data.

Main Results

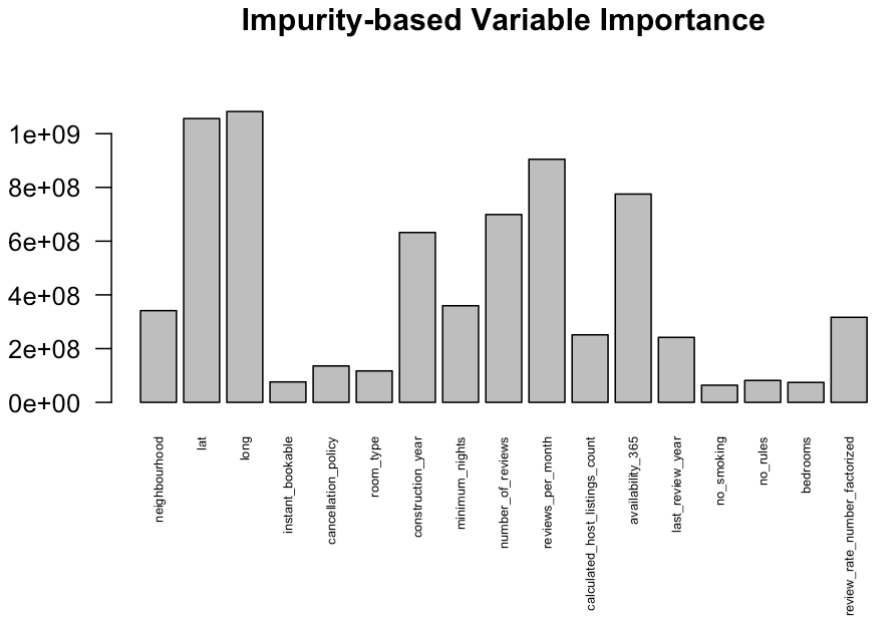


Figure 10

The linear regression model performed very poorly. The MSE of the model is 109,750.6 while The variance of the prices in the test data set is 109,512.4. The R^2 is 0.03993643%. The lasso model had a test MSE of 109,481.9 and an R^2 of 0.029551%. This indicates that linear models capture very little of the variance in the prices, further confirming suspicions that there are not many linear relationships in the data set. The random forest model with m = 5 performed significantly better, with a test MSE of 75,265.98, training MSE of 73,905.07, and an R^2 of 39.15638%. The random forest bagging model performed only slightly better than the previous model, with a test MSE of 70,338.94, training MSE of 71,696.71, and an R^2 of 40.70773%.

Figure 10 illustrates the node impurity of each predictor for a random forest model with m = 17 and 1000 trees. It is observed from the node impurity score graph of the bagging model with 1000 trees that the 5 most important features were latitude, longitude, number of reviews per month, number of nights available to book in the future, and the total number of reviews. The least important variables are instant bookable, room type, mention of smoking in house rules, and the number of bedrooms.

The bagging model trained on the PCA transformed data performed significantly worse, with a training MSE of 93,895.26 and a test MSE of 92,786.34. The R^2 is 15.96899%.

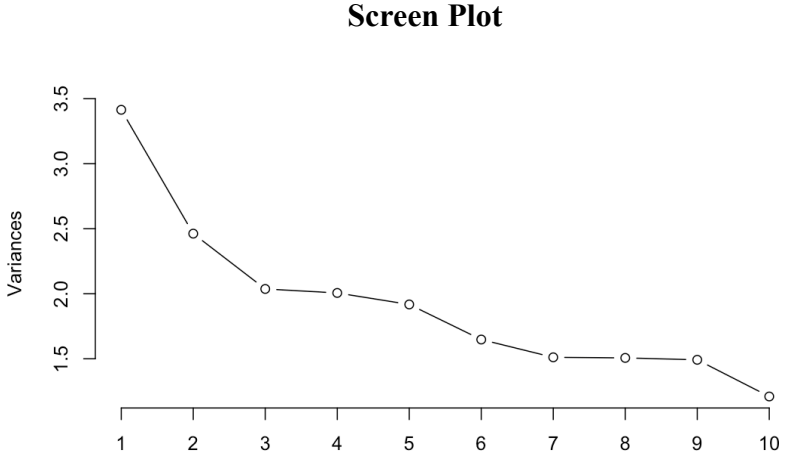


Figure 11

Utilizing XGBoost, a model was trained on the one hot encoded data. Models were trained with a maximum depth ranging from 6 to 100, learning rate of 0.1 and 0.01, and 50 and 100 rounds. The best model was trained with a max depth of 75, 100 rounds, and a learning rate of 0.1. The model performed fairly well in comparison to other models, with a test MSE of 65,734.07 and an R^2 of 40.0938%. However, the model overfit with a training MSE of only 142.3765. A model tested with 100 rounds had a slightly lower test MSE of 65,409.67, but overfit significantly more with a training MSE of only 0.2531277.

Location appeared to be an important determinant of listing pricing, shown from feature importance graphs emphasizing the importance of latitude and longitude. These metrics being more important than neighborhood indicate that very specific location details are more valuable in predicting the price than general classes.

Discussion

The lasso and linear models performed poorly most likely due to the lack of linear interactions between variables, especially between price and all other variables. The bagging model on the PCA data likely performed poorly due to this, as well. Because of the lack of linearity, the PCA data could not identify linear combinations as well to perform dimensionality reduction. Utilizing the random forests method, we were able to achieve our best models with an R^2 of approximately 40%. Bagging consistently performed slightly better than the random forest model with m = 5, possible because it provides more information by considering all features while splitting. While the test MSE in boosting was lower than those of the random forest models, neither of the random forest models overfit like the boosting model did. While training the boosting models, increasing the maximum depth increased the complexity of the tree and increased the extent to which the models overfit. However, it was observed that the test MSE consistently got lower as the model overfit more. This may be because the training and test data set do not differ much. This could be likely because of the complex relationships between predictors which require more complex models, even if the difference between test and train MSE grows larger. While the random forests performed the best, it is difficult to interpret which does not provide much insight on the relationships between variables. Additionally, utilizing trees made it so that handling the categorical data is much easier than in other methods. Because the data set did not contain any information about the occupancy or physical state listings, we were unable to capture a lot of the variance in price. While the number of bedrooms columns was created from the listing description, it is not a very reliable variable. It was one of the least important variables in the model, likely due to the fact that the vast majority of listings were given 1 as the number of bedrooms. Even with ensuring that the correct number of bedrooms was pulled, not every description contained the number of bedrooms, a critical factor of pricing. Additionally, there are several complex details such as luxurious renovations or well decorated rooms which were not included in the dataset but also can play a large part in pricing. The lack of this information explains why our best model is

still fairly weak. Moving forward, still missing this information, an in-depth analysis of the description and house rules may provide more accurate information on the number of bedrooms or the occupancy of the listings may improve model accuracy.

Conclusion

In our report, we explored the factors influencing Airbnb pricing across New York City neighborhoods through statistical and machine learning techniques. Both the analysis and modeling confirmed the significant impact of geographic location and room type on listing prices, while also highlighting the challenges posed by the non-linear nature of the data. Linear models such as linear and lasso regression captured minimal variance, underscoring the complex interactions within the dataset. In contrast, tree-based methods like random forests and boosting demonstrated superior performance, with our best model being the bagging random forest model notably achieving an R^2 of approximately 40%.

It is acknowledged that a model with an R^2 of 40% is relatively weak, and with a test MSE of 70,670.64, there is still a lot of variance to be captured by the model. However, we believe that our models have been limited by our dataset, which lacks information on several crucial aspects of listing pricing such as occupancy, square footage, and the availability of certain amenities. While attempts to draw information about these features were made through creating new columns such as number of bedrooms, this information was not confirmed and could not be extracted from every observation.

Moving forward, working with a dataset with more features on the qualities of listings would provide more information on the complex relationships between variables. Additionally, string manipulation of listing descriptions and house rules focusing on descriptions which mention the location or convenience of a listing may also provide more insight since both the random forest and boosting models demonstrated that location is an important variable in predicting price. Finally, further experimenting with the hyper-parameters of the bagging and boosting models may improve predictive performances.

Appendix

Mean Prices of Each Rating Depending on Neighborhood Group

Neighborhood Group	1	2	3	4	5
Bronx	623.24	649.29	621.06	631.13	623.86
Brooklyn	627.84	628.21	626.94	624.74	630.05
Manhattan	630.44	630.07	619.97	627.93	611.90
Queens	629.34	626.74	638.92	620.69	636.33
Staten Island	837.54	592.22	639.02	662.79	561.33

Appendix 1

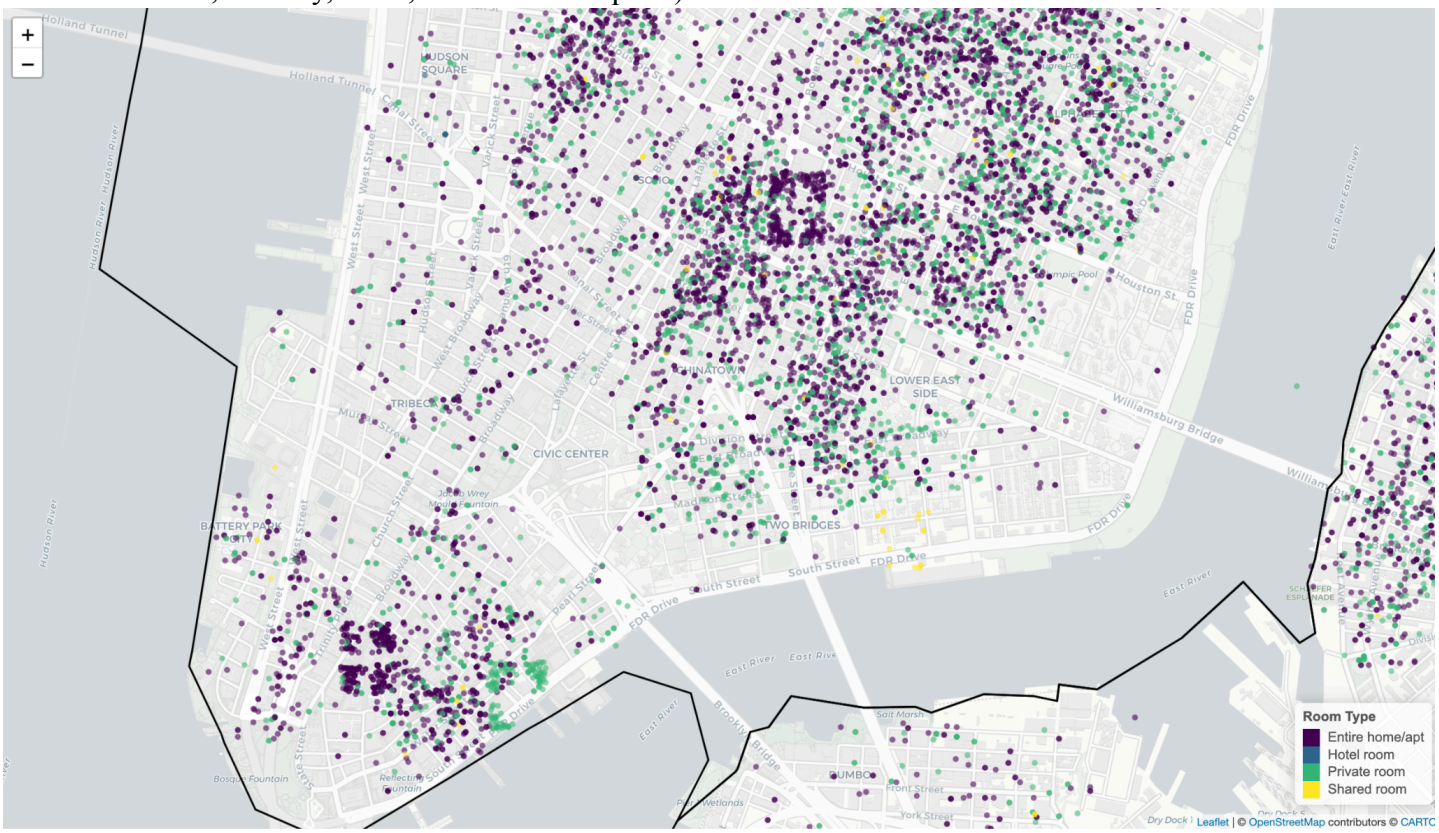
Mean Prices According to Number of Bedrooms

1	2	3	4	5	6	7	8	9	14
626.84	618.34	619.61	619.52	596.82	660.08	574.05	473.5	691	530.5

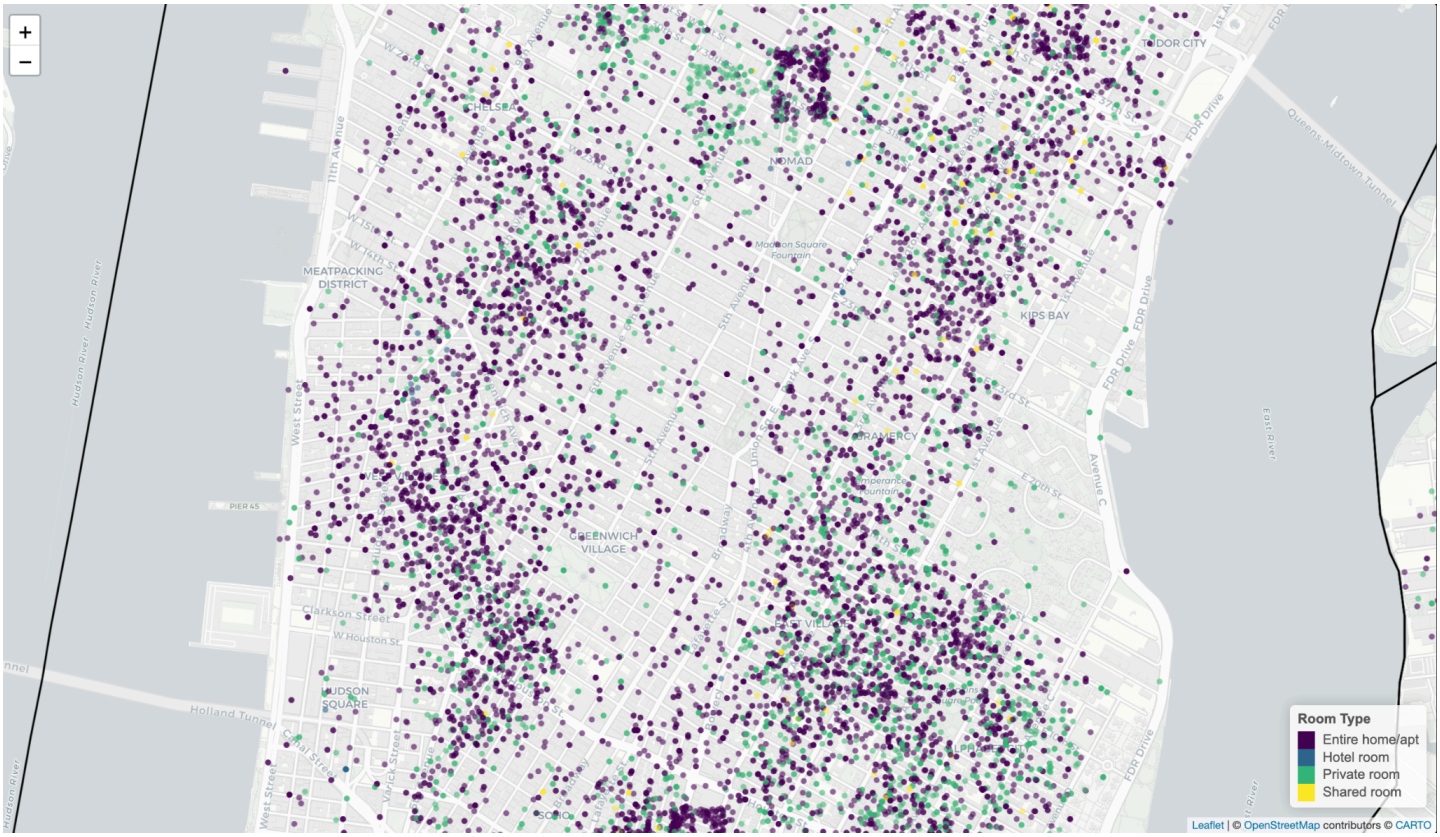
Appendix 2

Map Visualizations

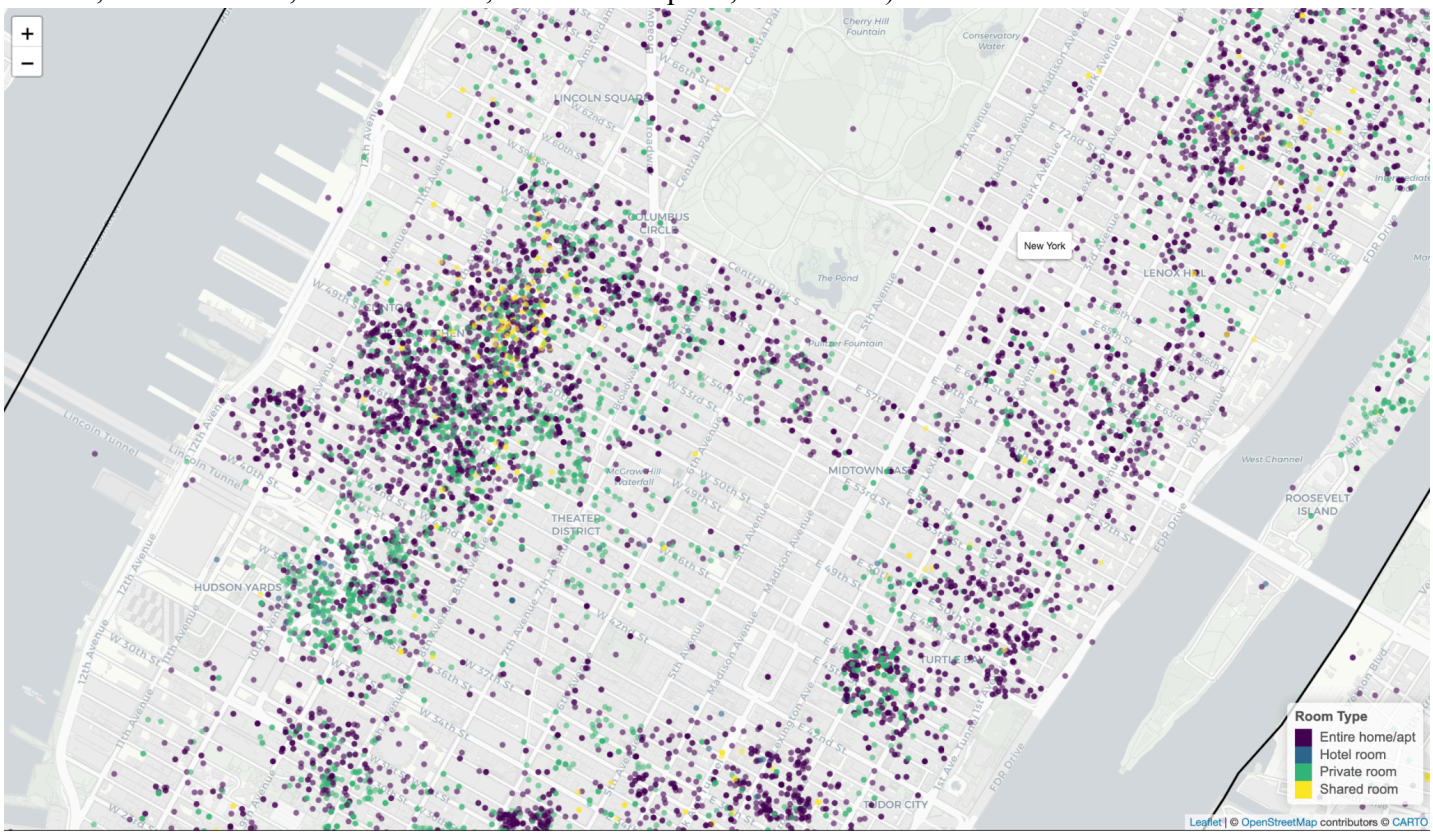
South - The Borough of New York (encompassing Financial District, Battery Park City, Tribeca, Chinatown, Lower East Side, Bowery, Soho, and Hudson Square)



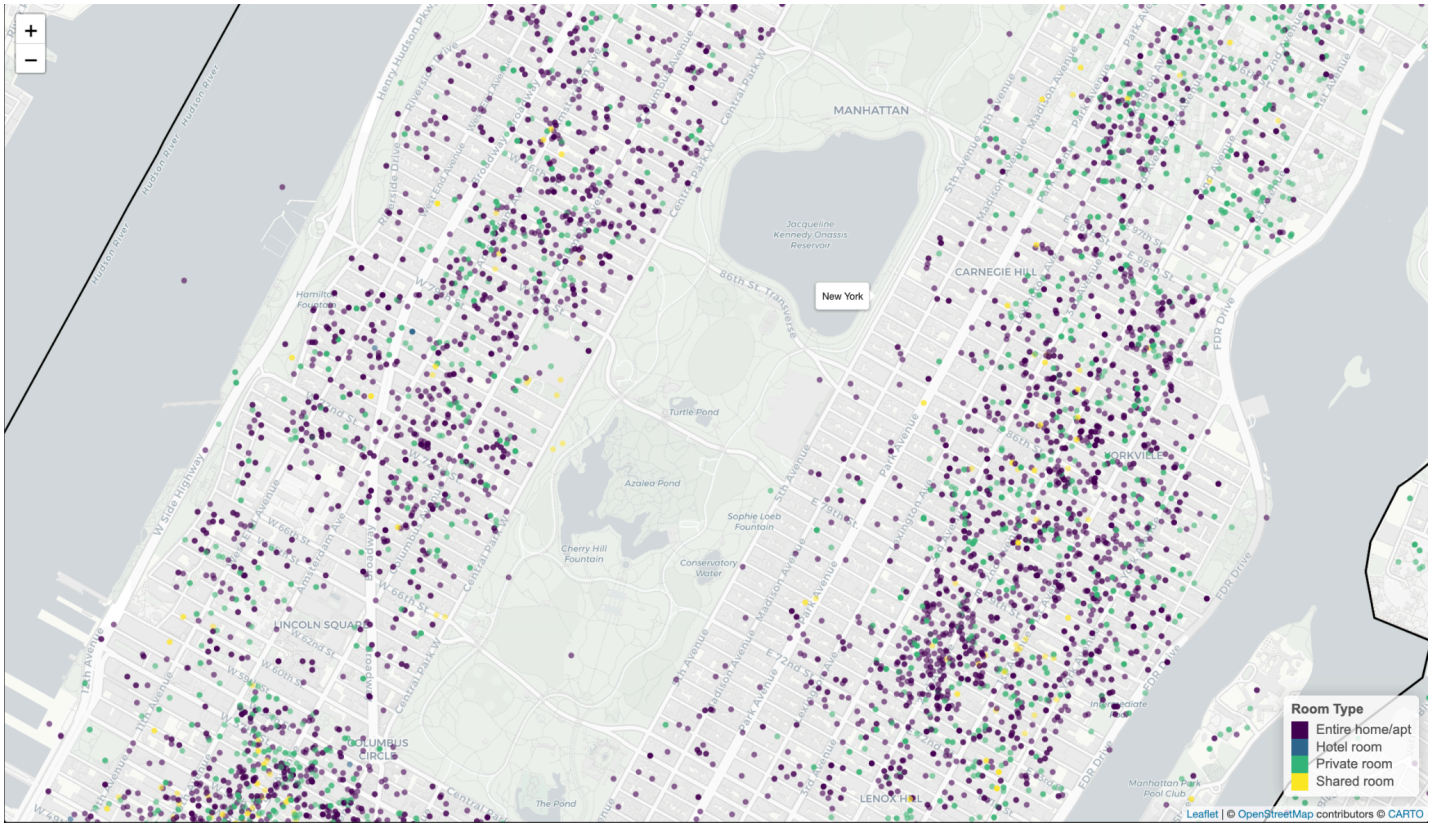
Central South - The Borough of New York (encompassing West Village, Meatpacking District, Soho, Nolita, Greenwich Village, Flatiron, Gramercy, Bedford-Stuyvesant, Chelsea, Kips Bay, and Hudson Square)



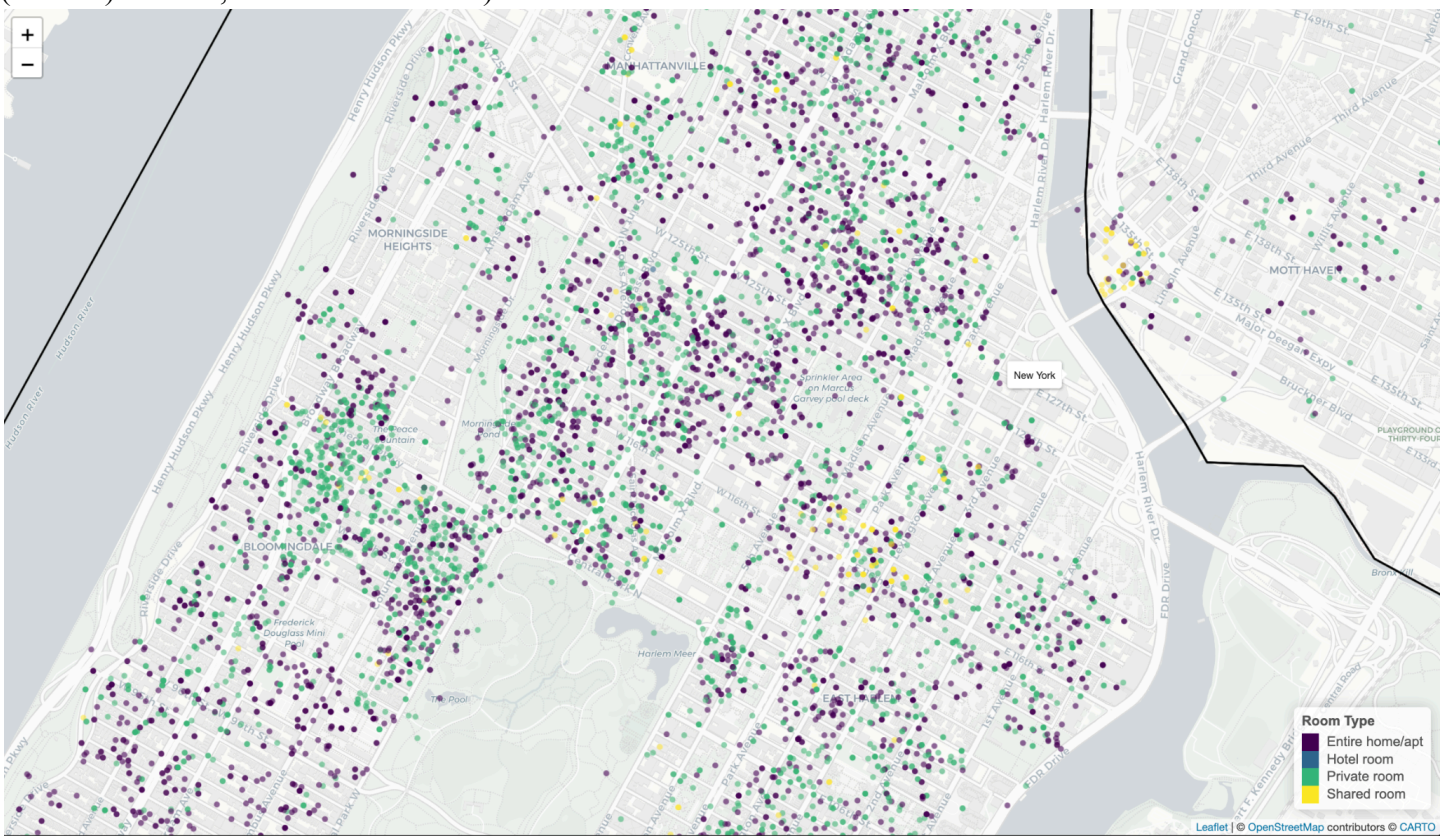
Central South - The Borough of New York (encompassing Hudson Yards, Tudor City, Turtle Bay, Theater District, Hells Kitchen, Midtown East, Columbus Square, Lenox Hill)



Central - The Borough of New York (encompassing Lenox Hill, Lincoln Square, Upper East Side, Yorkville, and Upper West Side)



North-Central - The Borough of New York (encompassing East Harlem, Bloomingdale, Morningside Heights, (Central) Harlem, and Manhattanville)



North - The Borough of New York (encompassing Sugar Hill, Hamilton Heights, and Manhattanville)