Analysis of Airbnb in the Concrete Jungle

Team: Nonlinear

Rod Aryan	**Kathleena Inchoco**	**George Zhou**	**Hannah Park**	**Jon Ma**
ra2829	**ki2130**	**gz2214**	**hp2501**	**xm595**

Link: Dataset

1. Introduction

Airbnb is a popular website that allows people to list and rent short-term accommodations, such as apartments, houses, and rooms. Its success and rise to the top have also sparked debates about the company's effect on housing markets, urban planning, and the local economy. Our capstone group is interested in answering the following questions about New York City Airbnbs: 'What drives value for NYC Airbnbs ?'

Understanding what drives value for Airbnb rentals in New York City is interesting because it can help both customers and hosts make informed decisions. For customers, knowing what factors influence the price of a listing can help them search for the best value for their money. On the other hand, hosts can use this information to better market their listings and set competitive prices. Additionally, understanding how some units are able to charge more than others can also provide valuable insights. This could be due to a variety of factors, such as location or the overall quality of the listing. By understanding these factors, both customers and hosts can make more informed decisions about what to look for in a listing or how to differentiate their own listing from others.

2. Data Preparation

The dataset for this project was obtained from Kaggle New York City Airbnb Open Data [1]. The dataset is a comprehensive collection of information on Airbnb listings in New York City during the year 2019. It has 48,895 observations with 16 variables and is a mix of categorical and numeric values. This dataset is an excellent resource that offers a wealth of information that can be used to gain insights into the popularity and performance of Airbnb rentals in the city, as well as the factors that drive demand and pricing. Every row of the dataset represents a summary of information and metrics for Airbnb listings in New York City.

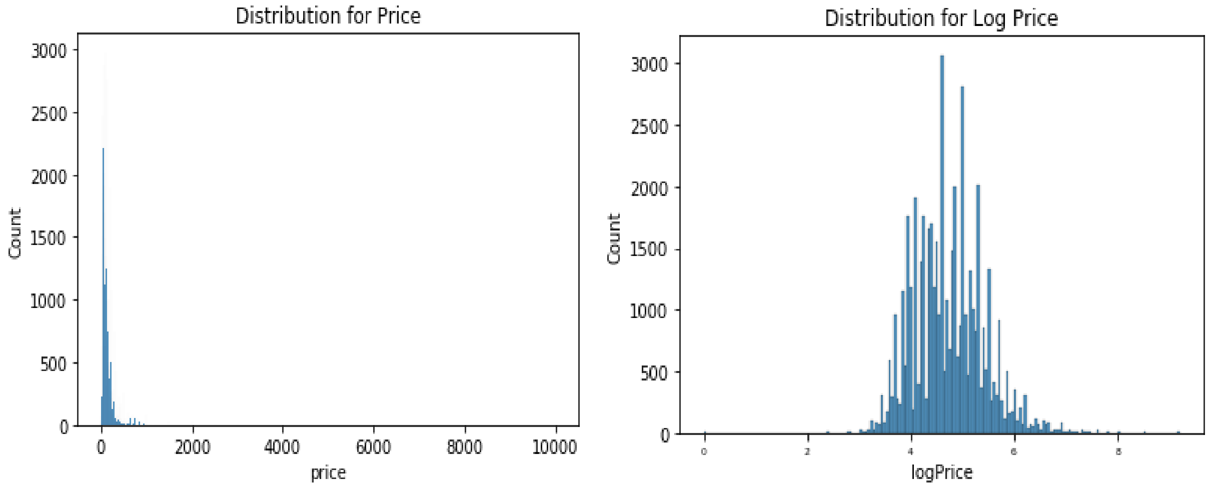
The columns of the dataset that contribute to our analysis are as follows:

1. *neighbourhood_group*: Categorical variable with 5 Boroughs (Manhattan, Brooklyn, Queens, Bronx, Staten Island).
2. *latitude*: Indicates latitude coordinates.
3. *longitude*: Indicates longitude coordinates.

4. *room_type*: Categorical variable with 3 room types (Entire home/apt, Private room, Shared room).
5. *price*: Numerical variable that indicates Airbnb price in dollars for a night. This variable was removed after creating logPrice variable
    a. *logPrice*: *price* variable transformed by log scaled.
6. *minimum_nights*: Numerical variable that indicates minimum night stays.
7. *number_of_reviews*: Numerical variable that indicates a number of reviews for each airbnb listing.
8. *reviews_per_month*: Numerical variable that indicates number of reviews per month.
9. *calculated_host_listings_count*: Numeric variable that indicates number of airbnb listings per host.
10. *availability_365*: Numerical variable that indicates number of days when listing is available for booking.

During the data exploration, we found that the price of Airbnb listings had a skewed distribution, with a few very high-priced listings pulling the average price up (Figure 1). Thus, the *price* feature was recalculated using a log function to generate the *logPrice* variable. Taking the log of the price can help to transform the data and make it more normally distributed, which can be helpful for some machine learning models that assume a normal distribution of the input data. The original price column was then dropped.

Figure 1. Distribution of Price and Distribution of Log Price



Since machine learning models in python require all input and output variables to be numeric, our categorical variables are encoded as dummy variables so that we can fit and evaluate models we build in the later analysis. Additionally, encoding the categorical variables can allow for a more fine-grained analysis of the data, as the encoded values can be treated as continuous variables rather than just categorical labels.

Missing values in *reviews_per_month* and *number_of_reviews* features are replaced with 0.

Note that for all of our analyses we set the seed of each random state to *save_state* = 18212178.

3. Inference Question

Question:
Is there a relationship between the number of reviews for a unit and the borough the unit is located in, or the room type?

Approach:
Based on the figure of the distribution of the number of reviews, the distribution of the number of reviews for each borough is not normally distributed (Figure 2), so we will use the Kruskal-Wallis test. It was also possible that the room type for each unit could play a role in the number of reviews, so we did additional Kruskal-Wallis tests while keeping room type constant across different boroughs. Before conducting the Kruskal-Wallis test, we did a power analysis to assess if there was enough sample size for adequate statistical power of 0.8. All tests had enough power given the sample size. The null hypothesis for each Kruskal-Wallis test is that the median number of reviews for each neighborhood group are the same, and the alternative hypothesis for each test is that there is a difference in median number of reviews.

Analysis:
Table 1 has the p-values for each Kruskal-Wallis test performed. The significance cut-off was at α=0.05. For each test, even if keeping the room type constant across different boroughs, all results are significant, where there is a relationship between boroughs and number of reviews.

Figure 2. Number of Reviews Distribution by Neighbourhood

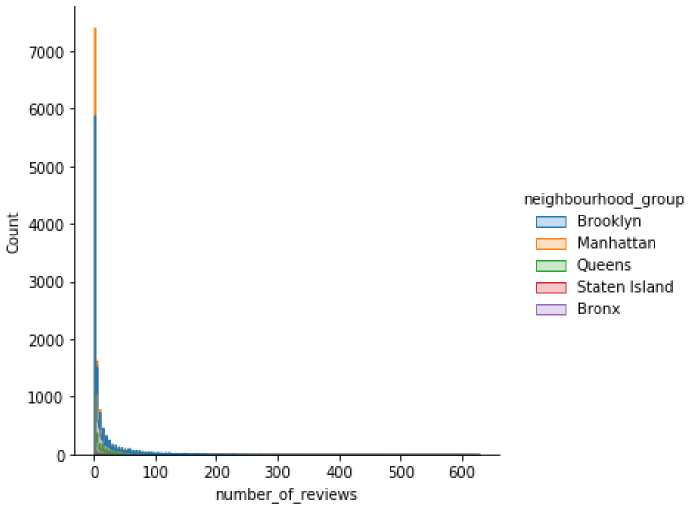


Table 1. P-Values For Median Number of Reviews by Borough/Room Type from Kruskal-Wallis Test

Among Boroughs: 4.85e-65	Among Boroughs for private rooms: 8.19e-17	Among Boroughs for shared rooms: 0.0004	Among Boroughs for Entire home/apt: 3.61e-134

Figure 3. Median Number of Reviews by Borough and Room Type

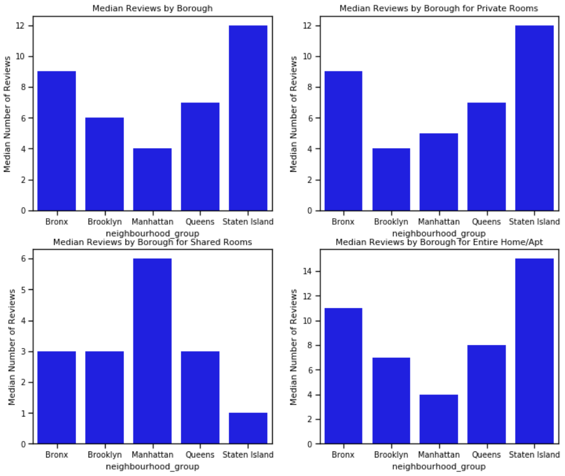
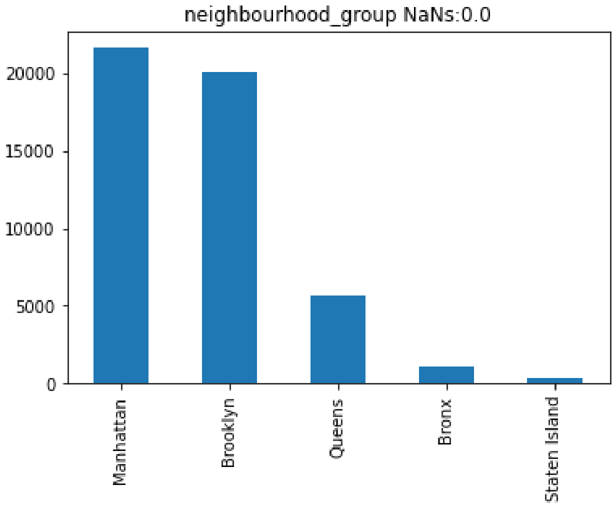


Figure 4. Number of Units by Borough



We see that other than for units that are shared rooms, Staten Island has the highest median number of reviews (Figure 3). This is interesting since typically Airbnb users will feel more confident with booking units that are reviewed more thoroughly. This could suggest that Staten Island Airbnb properties are more “vetted” compared to units in the other boroughs. It is also interesting to note that though the median number of reviews is the highest in Staten Island, Staten Island has the fewest number of units (Figure 4).

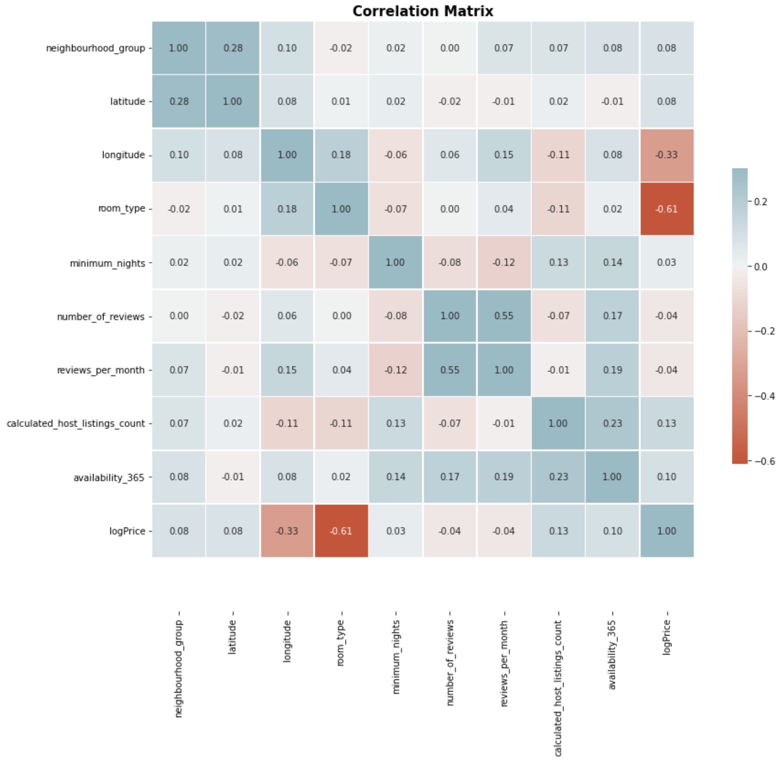
4. Prediction Question

Question:
Can we create a model that can help us predict the price for a unit in NYC?

Approach:
To answer this question, we developed 3 regression models, ridge, lasso, and linear regression. The predictors for the models were transformed/feature engineered for better performance. Since the price variable is heavily skewed right, we will be predicting the log price value, which is relatively normally distributed. To address overfitting and cross-validation concerns, the data will be split into 80% training (n=39,116) and 20% testing data (n=9,779) for the 3 models, as well as the evaluation on the root mean square error and the coefficient of determination value. For ridge and lasso regression, hyperparameter tuning was utilized for the best performance.

Analysis:
Before carrying out the necessary steps for our 3 regression models, we plotted a correlation matrix between the features to get an idea of their relationship. A correlation matrix helped us identify if there were confounding variables in our dataset (Figure 5). Confounding variables are correlated with both the predictor variables and the target variable, and they can cause problems in prediction models if they are not accounted for.

Figure 5. Correlation Matrix



After running a simple automated confounding variables test to analyze the correlation matrix, we found that ['number_of_reviews', 'reviews_per_month', 'calculated_host_listings_count', 'availability_365'] all came back as confounding. After another analysis (calculating the p-values of each coefficient), the variables outside of "reviews per month" that showed as confounding were not actually negatively affecting the model. Therefore, we continued with creating our models, accounting for the confounding variable "reviews per month". By identifying these issues in the correlation matrix, we took steps to address them before building our prediction model. Since the variable 'room_type ' is considered a categorical variable, we created additional dummy variables for each room type (private rooms, shared rooms, and entire home/apt).

Table 2. Evaluation on the Prediction Models

Model Type	**RMSE**	**COD**
Linear Regression	0.499	0.494
Ridge Regression, *α=0.0001	0.503	0.483
Lasso Regression, *α=0.0001	0.509	0.469

***Note: In Table 2, the α value is the tuning parameter, not the significance level.**

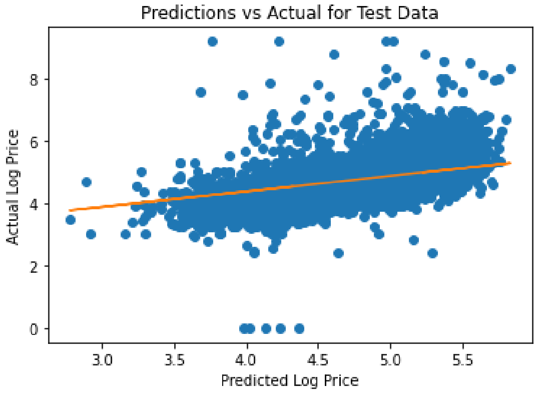
We see that the linear regression performs the best with a root mean squared error of 0.499 and a COD score of 0.494. Below, Table 3 shows the coefficient betas for the predictors for the linear regression.

Table 3: Linear Regression Coefficient Beta Values for Log Price

Log Minimum Nights	-0.0924
Log Availability	0.0544
Log Number of Reviews	-0.0586
Latitude	1529.2751
Longitude	-2783.5881
Interaction between Longitude and Latitude	-37.6629
Room Type: Private Room	-0.7859
Room Type: Shared Room	-1.1908

Positive coefficients suggest positive indicators for log price (drives the price up), while negative coefficients mean negative indicators (drives the price down). We see that predictors related to location (longitude, latitude, and interaction between longitude and latitude) have relatively large coefficient magnitudes which will heavily influence the predicted log price. Private rooms and shared rooms are negative indicators of log price. This is reasonable as it is most likely more expensive to rent out a whole unit than a single room. It is interesting to note that the number of reviews has a negative correlation with price. One possible explanation is that the number of reviews may be positively correlated with the number of overall visitors, such that visitors are more likely to book cheaper units.

Figure 6. Predicted vs Actual for Test Data



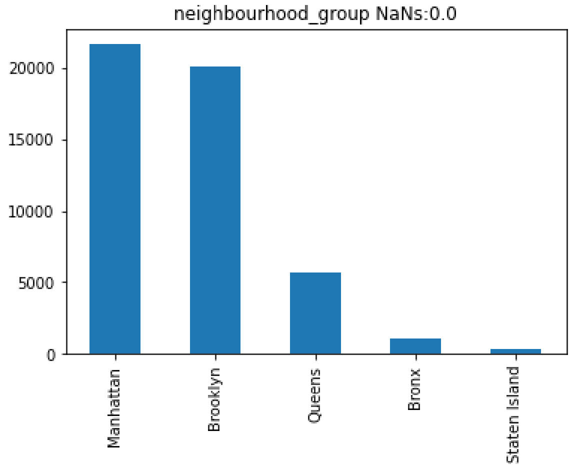
When plotting against the actual log price by predicted log price by linear regression, there is a positive correlation, such that our model under-predicts a log price that is supposed to be much higher (Figure 6). This may explain why lasso and ridge regression in our case performed worse than linear regression since regularization adds bias, which would cause even more underfitting.

5. Classification Question

Question:
Can we classify whether a unit is in Manhattan or not given other features?

Approach:
In the data exploration, we found about 45% of the Airbnb listings are located in Manhattan, 40% in Brooklyn, and the rest 15% of the listings are located in the other three boroughs. (Figure 7)

Figure 7. Neighbourhood_group Variable Analysis



Therefore, we created a binary variable *Manhattan* in order to reflect 1 for Manhattan and 0 for not Manhattan and used this variable as our target variable.

We took out the neighborhood group, latitude, longitude, and neighborhood columns because these reveal the location of the Airbnb unit. We kept columns such as *room_type, minimum_nights, number_of_reviews, reviews_per_month, calculated_host_listings_count, availability_365*, and *logPrice*. In order to encode categorical variables to numeric, we also created dummy variables for the room_type column.

Most Airbnb units were located in Manhattan with fewer units located in the Bronx, Queens, Brooklyn, and Staten Island. In order to account for this difference in our dataset, we used the bootstrap method to resample observations with replacements from our dataset to get an equal number of Manhattan and non-Manhattan Airbnb units in our analysis. For each classification model, we used an 80/20 training/test split and sampled our data 1,000 times to get a mean AUC score across each classification model in order to reduce bias and evaluate the model's performance. First, we used K-Means clustering and then continued with three other classification methods: logistic regression, decision tree, and naive bayes.

Analysis:
Before exploring more complex classification models, we attempted to use K-Means clustering on the New York City Airbnb open data [2]. Alas, the results were not very successful. K-Means is a centroid-based algorithm that tries to find clusters by assigning points to the nearest cluster center, but the

distribution of Airbnb listings in New York City is likely to have complex, overlapping clusters with a variety of shapes and densities. As a result, K-Means struggled to accurately identify the clusters in the data. To try and improve the results, we used the "elbow method" as a means of parameter tuning to find the optimal number of clusters. This method recommended using 4 clusters, which was lower than our original prediction of 5 (given the number of boroughs). However, the results were not satisfactory. This suggests that K-Means may not be the best method for clustering the New York City Airbnb data, and we needed to explore alternative clustering algorithms. The centers of the clustering algorithm are added in the appendix [2]. Below are figures showing the K-means clustering, plotted on each listing's longitude and latitudinal coordinates (Figure 8):

Figure 8. K-Means Clustering Scatterplots

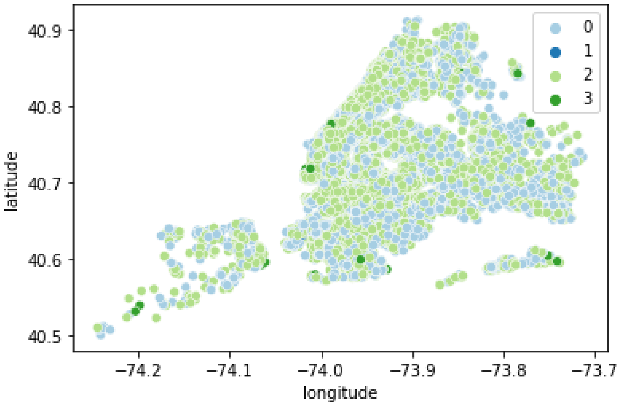
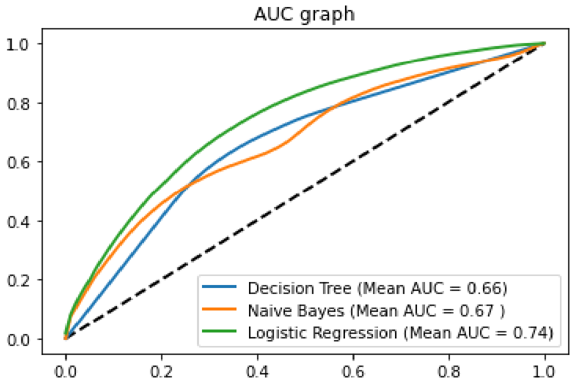


Figure 9. Mean AUC Curves of Classification Models

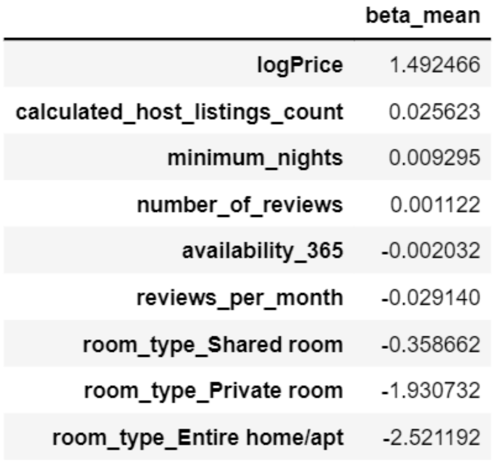


However, across all of the classification methods, the logistic regression performed the best at classifying the NYC Airbnb units with a mean AUC score of 0.74 (Figure 9).

Analyzing the beta coefficients for feature selection [4], we found that the most important positively correlated features to classify the units were *logPrice, calculated_host_listings_count* and *minimum_nights* (Table 4). This makes sense because listings in Manhattan tend to be pricier than listings in Brooklyn, Queens, the Bronx, and Staten Island. Also, hosts that are able to let more than one listing will likely optimize in the most popular area which is Manhattan.

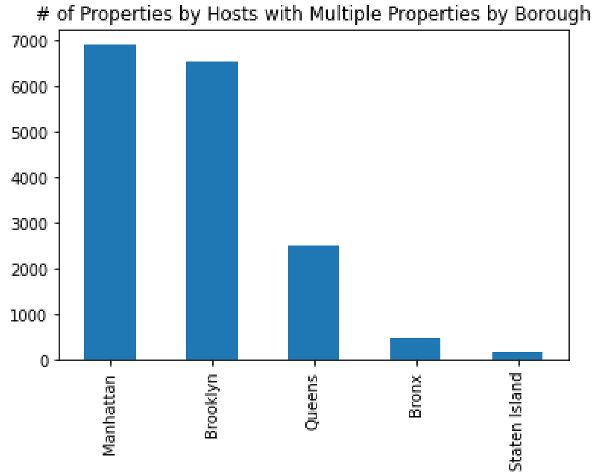
These statistics corroborate a common phrase that is important to real estate: location, location, location. The data supports what we understand intuitively about real estate in New York City. Manhattan remains a coveted borough in terms of housing and those visiting will pay a premium to have the opportunity to live in Manhattan. Similarly, hosts follow this demand to capture the supply of housing options in Manhattan, reflected by the positive beta coefficient for *calculated_host_listings_count* (Figure 10).

Table 4. Logistic Regression Mean Betas for Manhattan



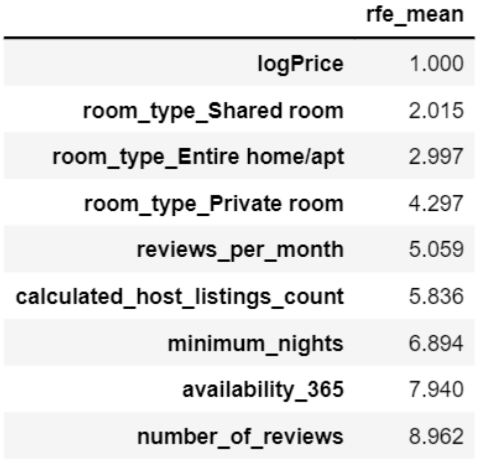
	beta_mean
logPrice	1.492466
calculated_host_listings_count	0.025623
minimum_nights	0.009295
number_of_reviews	0.001122
availability_365	-0.002032
reviews_per_month	-0.029140
room_type_Shared room	-0.358662
room_type_Private room	-1.930732
room_type_Entire home/apt	-2.521192

Figure 10. Count of Multiple Listings by Borough



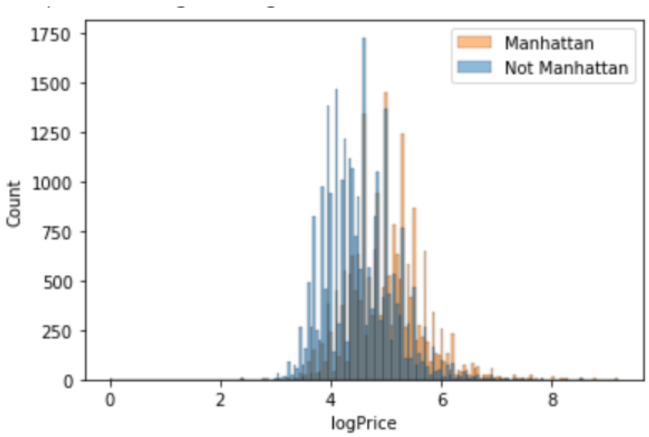
Using Recursive Feature Elimination (RFE) for feature selection [4], the most important features to classify the units in Manhattan or not Manhattan were *logPrice, room_type_Shared room, room_type_Entire home/apt* and *room_type_Private room* (Table 5). RFE ranks the n features ordinally from 1 to n. Score of 1 means the feature is the most important while a score of n means the feature is least important. What we can see is that again, logPrice is the most important feature in classifying rentals in Manhattan or not. This makes sense because as we can see from Figure 11, units in Manhattan are on average more expensive.

Table 5. Mean RFE Scores



	rfe_mean
logPrice	1.000
room_type_Shared room	2.015
room_type_Entire home/apt	2.997
room_type_Private room	4.297
reviews_per_month	5.059
calculated_host_listings_count	5.836
minimum_nights	6.894
availability_365	7.940
number_of_reviews	8.962

Figure 11. LogPrice Distribution by Manhattan and Not Manhattan



Decision tree classification was performed with the independent variable of Manhattan (binary 1 or 0) to predict whether a listing is in Manhattan or not given other features. Even though the Decision Tree algorithm can handle different data types, the Decision Tree Classifier from ScikitLearn [3] doesn't

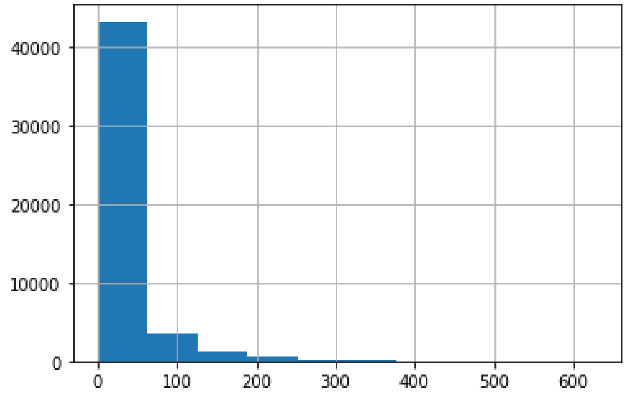
support categorical data. Thus, categorical features were encoded to dummy variables before training the model.
A decision tree is prone to overfitting unless a model has very few splittable features. Although bagging on the decision tree was done to reduce the possibility of overfitting, the model's mean AUC score was 0.66 (Figure 9). Since an AUC score between 0.5 to 0.7 is considered as poorly classifying observations into classes, this suggests that the decision tree model may not necessarily be the best model to answer our question. Moreover, the tree visualization was so complex and it was difficult to interpret due to the number of nodes it created.

We also performed Naive Bayes Classification. Longitude, latitude, as well as neighborhood and neightborhood_group are highly correlated to our dependable variable, since our targeted variable is regard to whether the location of the Airbnb is on Manhattan Island or not. That's the motivation behind dropping these variables, since they all represent the same type of geographic information.

In addition, slightly different from previous classification, a new feature called wighted_pop was generated for the purpose of the fact that we notice how some listings were reviewed many times but the actual minimum number of nights stayed were quite low, as you can see in Figure 12. In order to scale up the variance in such important information, it would be easier for the machine to learn from the small but critical changes.

Furthermore, bayesian networks are often useful for expressing the probability of events given one or many conditions, especially via direct acrylic graphs. For our analysis, we chose to utilize such a powerful tool to classify whether a listing is in Manhattan (1) or not (0). A Naive Bayes Classifier was applied on the features mentioned above because a naive Bayesian network assumes the independence between predictor variables, which fits our goal to solely explore the insights of how features selected could directly help us classify whether a listing is in Manhattan. Multiple versions of the Naive Bayes Classifier were tested and evaluated based on the characteristics of our dependent and independent variables, including Gaussian Naive Bayes, Categorical Naive Bayes, and Multinomial Naive Bayes. To differentiate them, we examined the distributions of each independent variable. Since the majority of the selected features are continuously distributed before/after some feature transformation, a Gaussian Naive Bayes Classifier won the competition as the best Naive Bayes Classifier for our final analysis.

Figure 12. No. of Reviews Distribution



6. Conclusion

In analyzing the New York City Airbnb Open Data dataset, our prediction models showed that location as well as the type of room available, were strong predictors of the price of a rental. These models aimed to predict the price of listings as a target feature and took into consideration various factors. In addition to our prediction models, we also developed classification models with the goal of predicting whether a listing was located in Manhattan or not. We found that logistic regression performed the best in our classification analysis. We also conclude that the location of a listing is the primary factor in determining its value. Airbnb rentals in Manhattan were consistently the most expensive, likely due to the high demand for rentals in this popular and tourist-friendly neighborhood.

Throughout the project, we found that the available data has some limitations as it doesn't provide Airbnb unit ratings, amenities, or post-COVID data. Not having rating values can be a limitation in analyzing this Airbnb dataset because ratings can provide important information about the quality of the listings and the experiences of previous guests. Ratings can also be used as a proxy for the overall popularity of a listing. Without this information, it may be more difficult to accurately assess the attractiveness or demand for different listings. Not having a list of amenities can also be a limitation in analyzing this dataset. Amenities can be a significant factor in determining the appeal of a listing, and the availability of certain amenities can differentiate one listing from another. Without this information, it may be more difficult to accurately assess the features and benefits of different listings. Finally, not having post-COVID data can also be a limitation in analyzing this dataset. The COVID-19 pandemic has had a significant impact on the hospitality industry, and it is likely that this has affected the demand and popularity of Airbnb listings in New York City. Without data on how the pandemic has impacted the Airbnb market in the city, it may be difficult to accurately assess the current state of the market and make accurate predictions about the future.

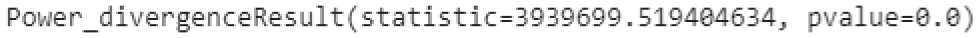
There are a few ways that the limitations of the dataset, such as the lack of post-COVID data and the absence of amenities information, could potentially be addressed. One solution would be to obtain more recent data that includes post-COVID information. This could provide a more accurate picture of the current state of the Airbnb market in New York City and help to address the limitations of the existing dataset. Another option would be to supplement the existing dataset with external data sources that provide information on amenities or other relevant factors. For example, data on the availability of public transportation or the proximity of restaurants and other attractions could potentially be used to enrich the existing dataset. Additionally, data from other cities may be able to provide some insights into the factors that drive value for Airbnb rentals. By comparing data from multiple cities, it may be possible to identify patterns and trends that could be relevant to the New York City market.

Extra Credit

For our Extra Credit research, we found that owners who had multiple listings were more likely to have their listings in Manhattan, as this is where they could command the highest prices. We divided our data into two groups: “is_Manhattan” and “not_Manhattan”. The group “is_Manhattan” contains observations of *calculated_host_listings_count* that were in Manhattan, while the group “not_Manhattan” contains observations of *calculated_host_listings_count* that were in Manhattan. We conducted a chi-squared test to see if there was a statistically significant difference in the distribution of the categories of each group in *calculated_host_listings_count* in Manhattan or not Manhattan.

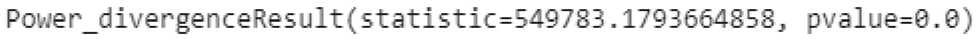
The following shows the chi-squared statistics and the p-values for the following groups:

is_Manhattan:



```
Power_divergenceResult(statistic=3939699.519404634, pvalue=0.0)
```

not_Manhattan:



```
Power_divergenceResult(statistic=549783.1793664858, pvalue=0.0)
```

We can see that the results are statistically significant for each, showing that the distribution of *calculated_host_listings_count* varies significantly in Manhattan and not in Manhattan. By understanding the factors that drive the value of Airbnb rentals in New York City, such as *calculated_host_listings_count,* hosts can make more informed decisions about how to make the most of their efforts to attract customers to their listings. Specifically, by having more listings in Manhattan which is more popular to visitors, this will likely increase the occupancy rate of their host listings.

Appendix:

[1] *New york city airbnb open data. (n.d.). Retrieved December 19, 2022, from* *https://www.kaggle.com/datasets/dgomonov/new-york-city-airbnb-open-data*

[2] Clustering Group

	room_type_Entire home/apt	room_type_Private room	room_type_Shared room	price	number_of_reviews	availability_365
K-Means Euclidean Cluster 1	5.26289528e-01	4.41214994e-01	3.24954781e-02	1.47088443e+02	3.50455311e+01	2.81236263e+02
K-Means Euclidean Cluster 2	5.06827435e-01	4.73410586e-01	1.97619793e-02	1.25147291e+02	1.77460382e+01	2.60830254e+01
K-Means Euclidean Cluster 3	8.60465116e-01	1.39534884e-01	-3.46944695e-18	5.84586047e+03	1.95348837e+00	1.90186047e+02
K-Means Euclidean Cluster 4	8.44769404e-01	1.46231721e-01	8.99887514e-03	9.69273341e+02	1.05793026e+01	1.84896513e+02
K-Means Euclidean Cluster 1	5.06827435e-01	4.73410586e-01	1.97619793e-02	1.25147291e+02	1.77460382e+01	2.60830254e+01

[3] *Sklearn. Tree. Decisiontreeclassifier*. (n.d.). Scikit-Learn. Retrieved December 19, 2022, from https://scikit-learn/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html

[4] P. Tan, M. Steinbach, and V. Kumar. (2005) Introduction to Data Mining. Addison Wesley https://towardsdatascience.com/a-look-into-feature-importance-in-logistic-regression-model-a4aa970f9b0f

[5] Airbnb. Why Reviews Matter. 29 Nov. 2019, https://www.airbnb.com/resources/hosting-homes/a/why-reviews-matter-41

What People Contributed:

George

I worked on primarily the prediction and inference aspects of the project. This includes writing the first and final draft for those sections, and the coding. I also helped with brainstorming ideas for the other sections and helped with coding difficulties. I was responsible for constructing the inference and prediction analyses related to the relationship between number of reviews, location, and the price of the units.

Rod

I helped outline the different questions for the project, separating them by inference/classification/prediction and highlighting the best statistical and machine learning methods for answering them. I also wrote a portion of the code for the classification questions, with one specific example being applying different clustering algorithms to the dataset and comparing their performances. Throughout the coding process, I provided assistance to other portions of the code for the capstone project which needed help. Finally, in terms of writing the report, I created the outline used for our report as well as aided in ensuring that each section of the report met the guidelines of the rubric. Overall, I contributed a large amount of writing and editing to each section of the report.

Hannah

I carefully read through capstone project rubric and sample reports and assisted team members to make sure we meet all the requirements listed in the instructions. I wrote the decision tree portion of code for the classification question. Throughout the coding process, I provided assistance by coordinating times our team can work together on the project either in person or through zoom. For writing the report, I helped outline the report so that it can have clear structure. I wrote data preparation and the decision tree analysis in the classification section.

Kathleena

I helped develop the narrative of the project and keep the team focused on staying on track in our progress to delivering the report. I also promoted iterations of the report that included critiquing our output, understanding what was important to keep our analysis moving forward, and encouraging others to drop parts of the report that were peripheral toward our goal. I was responsible for the logistic regression analysis and feature selection analysis as part of the classification model for the report. I also edited and revised the written sections for the prediction and inference analyses of the report.

Jon

Contributed in: general idea generation, project structural outline, coding/result generation, classification, visualization, analysis, extra credit and final report writing. I worked primarily on the naive bayes classification portion of the project.