PRICE PREDICTION MODEL

airbnb

DEVELOPED AND PRESENTED
BY VARSHINI CHELLAPILLA
FOR SPRINGBOARD

OVERVIEW

Airbnb: Airbnb is an online marketplace developed for homestay experiences, including but not limited to vacation homes, limited hotels, and short-term rentals. According to the company's website, there are six million active listings across the world with approximately 100,000 locations with active Airbnb listings.

New York, NY: New York is one of the largest cities in the world and is home to over eight million people. With a booming tourism industry, it is a center for entertainment, history and cultural experiences. In 2019, 6.6 million tourists visited New York City.

Situation At Hand: Airbnb contains nearly 40,000 listings in New York. The company allows hosts to set prices and list various features and descriptions of their listing. However, this is often a subjective approach. My intention was to build a machine learning model that could predict a listing's price based on its features, utilizing exploratory data analysis, supervised learning, statistical inference, Thus, the aim of this model is to provide a highly-qualified aid to hosts as they decide their prices.

DATA COLLECTION

The data used in this project was obtained from Inside Airbnb, an online project that provides data and advocacy about Airbnb's impact on residential communities. The data contains detailed information on each listing on the platform. It was compiled by InsideAirbnb on Sept. 1, 2021 and obtained for the purposes of this project shortly after.

ENVISIONED APPROACH

The questions posed at the beginning of the project are:

- What does the distribution of listing prices across New York look like?
- What are the factors that influence a listing's price?
- How does the price vary across different locations and types of dwellings?

The steps followed in this project to answer the above question and build a predictive model are: Data Wrangling, Exploratory Data Analysis, Feature Engineering and Data Preprocessing, Data Modeling, and Data Visualization and Presentation.

DATA WRANGLING

ORIGINAL DATASET

The original dataset obtained from InsideAirbnb contained 36,923 entries with 74 columns. The columns included information on the host, the location of the lisitng, features of the listing, availability, and the reviews given to listing by previous users on a scale of 0-5.

FACETS OF DATA WRANGLING

In order to properly carry out the various stages of data wrangling, the dataset was first observed from a bird's eye perspective before taking a detailed look at it column by column.

Columns were converted into the correct data types (boolean values, datetime objects, etc.). However, one of the major issues encountered was regarding the large number of missing values in the dataset as seen in *Figure 1*.

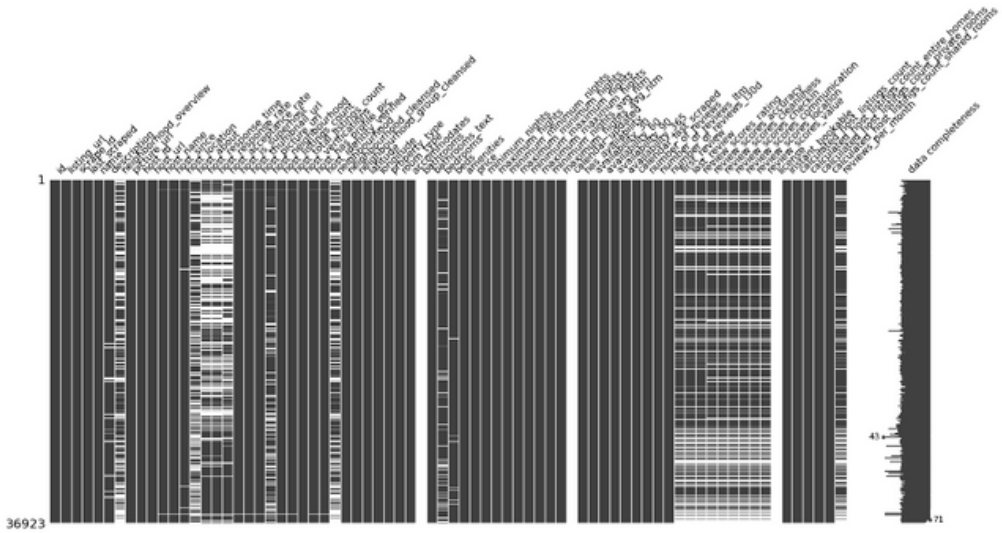


Figure 1: A matrix showcasing the missing data in the dataset and the overall data completeness.

Rows that contained missing values were either dropped or filled in with statistically inferred solutions. A column was removed if a significant portion of its data was missing. If the cells containing missing values make up for less than 5% of the data in the column, then those cells -- and their adjoining rows -- were dropped.

Some columns were cleaned and divided into neater categories. The column containing information on property type originally contained over 50 different types that occasionally overlapped with each other. Many of them had relatively few listings under them. Therefore, intuitively, I categorized them into eight overall types. Similarly, the bathrooms column and the type of room column were given the same treatment.

Additionally, columns that were redundant were removed, as were columns that were deemed irrelevant (i.e., text columns that required NLP skills not in the scope of this project).

EXPLORATORY DATA ANALYSIS

AIM

The goal of the exploratory data analysis process was to determine the factors that affect the base price of a rental unit in New York. To do so, the various relationships between different features and their influence on the price of a listing were studied and recorded.

PRICE DISTRIBUTION IN NEW YORK, NY

The price of listings in the given dataset varied greatly. The minimum price for a listing was $10 while the maximum stood at $10,000. While most of the prices were in the lower hundreds, there were many outliers that skewed the data to the right.

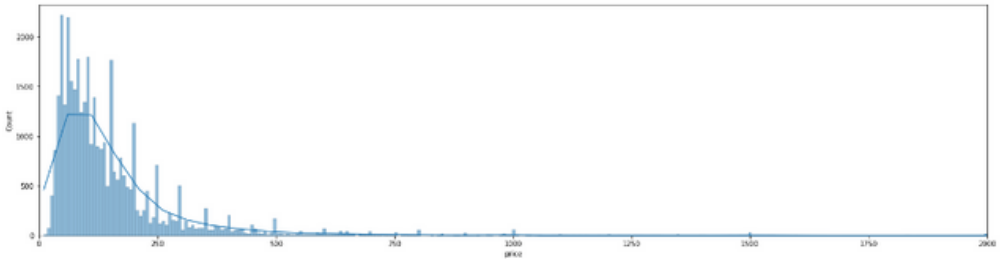


Figure 2: Data distribution of the **price** column using a histplot in the Seaborn package.

FACTORS AFFECTING PRICE

Using a correlation matrix to determine linear relationships and the predictive power score package to determine non-linear relationships, the following columns were observed to have a relationship with the price column:

- Amenities offered
- Number of people accommodated
- Number of bedrooms and beds
- Number of bathrooms

The most common number of bedrooms appears to be 1 and the most common bathrooms appears to be between 1 and 2. The most common number of beds appears to be between 0 and 5. The most common number of people a listing can accommodate is less than 3.

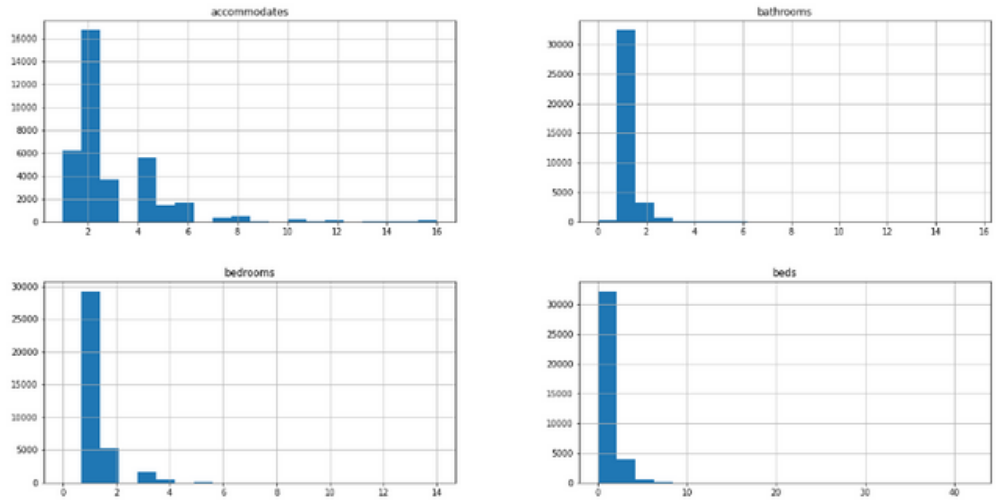


Figure 3: A histogram of the Amenities, Bedrooms, Bathrooms, and Beds columns.

Among the various amenities offered by different listings, the top 10 most common amenities were: wifi, long-term stays allowed, heating, kitchen, essentials, smoke alarm, air conditioning, hangers, carbon monoxide alarm, hair dryer. Additionally, because of the known parking issues in New York, I decided to also consider free street parking as an amenity to look into. With each of these options, a null hypothesis was taken as the provision of the amenity not having an effect on the price. The t-statistic test was performed was performed on each and the p-value was calculated to determine if the null hypothesis was rejected or accepted. The results of the analysis were the following columns having a proven relationship with price: long term stays allowed, provision of a kitchen, air conditioning, hangers, and free street parking.

In regards to property type, the most popular type was rental unit at a whopping 27,449. It was immediately followed by residential homes at 3,042 (as is visualized in the following page). An ANOVA test was created to establish whether a relationship within the property type and listing price exists. It does! Additionally, a similar ANOVA test was used to find the F-statistic and identify the distribution of room types across the city. The statistical test proved that the room type of the listing has an effect on its price. Doing a similar test on the bathroom type offered by each listing (private or shared) proved that it, too, has an effect on the price of a listing.

Lastly, it was observed that most reviews seemed to lie between 4 and 5, particularly above 4.8. Most of the listings with reviews are observed to be priced between $0 and $1000. With a few exceptions, listings with higher prices usually had higher ratings but the same could not be said for the reverse relationship. Nearly 10,000 listings, however, weren't given any reviews. Using a t-statistic test, it was also noted that the inclusion of reviews for a listing tended to have an effect on the price of the listing as well. However, interestingly, the mean price of a listing without reviews was higher than that with reviews.

Other observed relationships include minor details with host information and availability. The median price of a listing was found to be higher if it was instantly bookable, had availability and if the host was an Airbnb superhost, a special program for Airbnb's most experienced hosts.

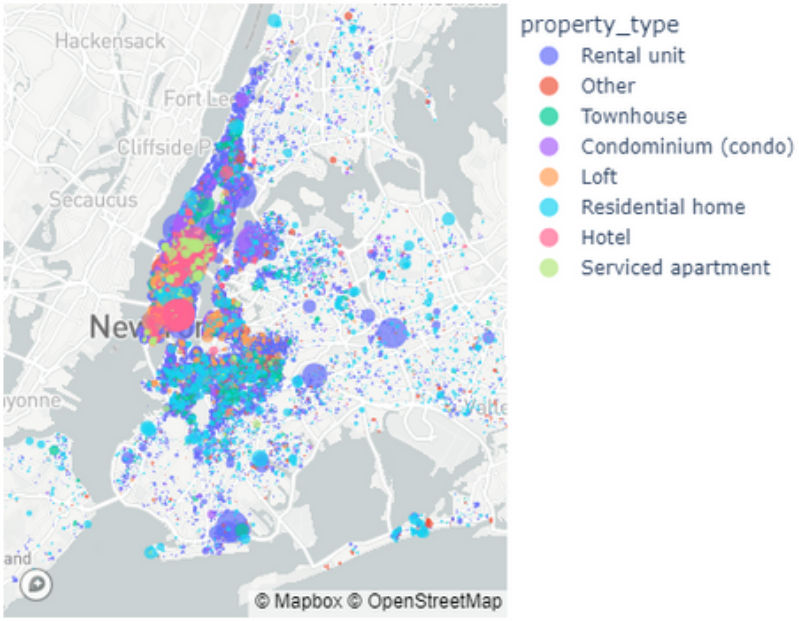


Figure 4: A map of all the listings in New York, NY color-coordinated based on the property type of the listing.

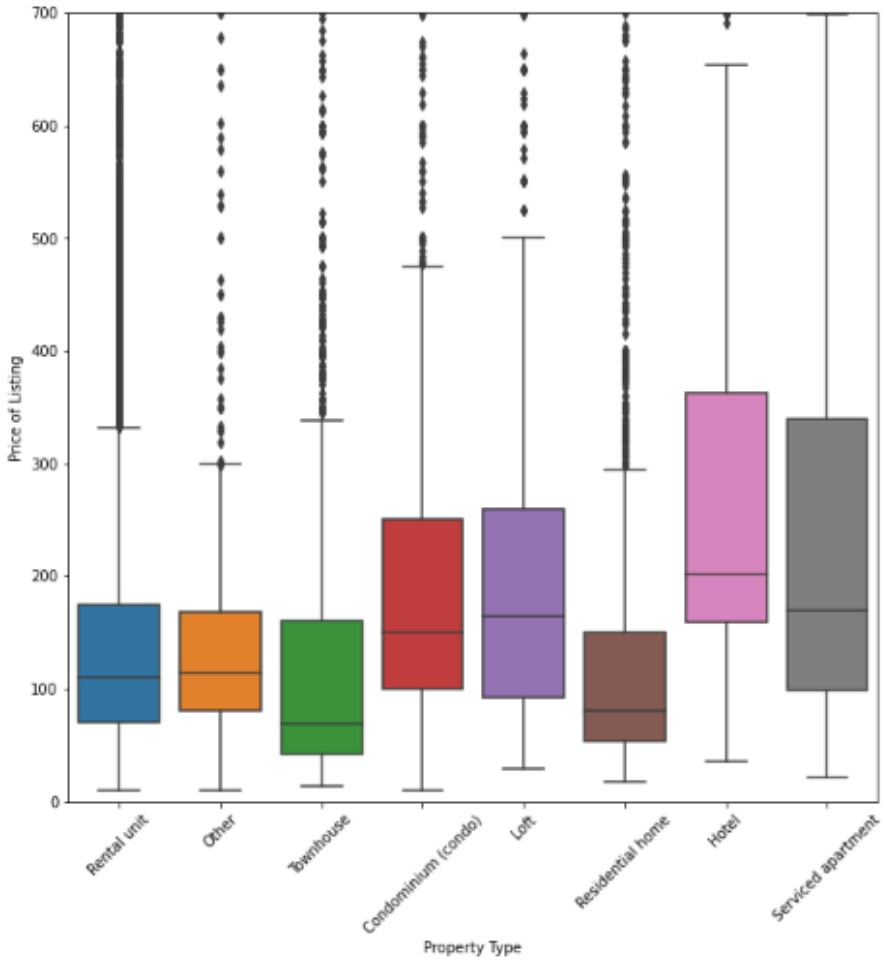


Figure 5: A boxplot of all the listing in New York by property type.

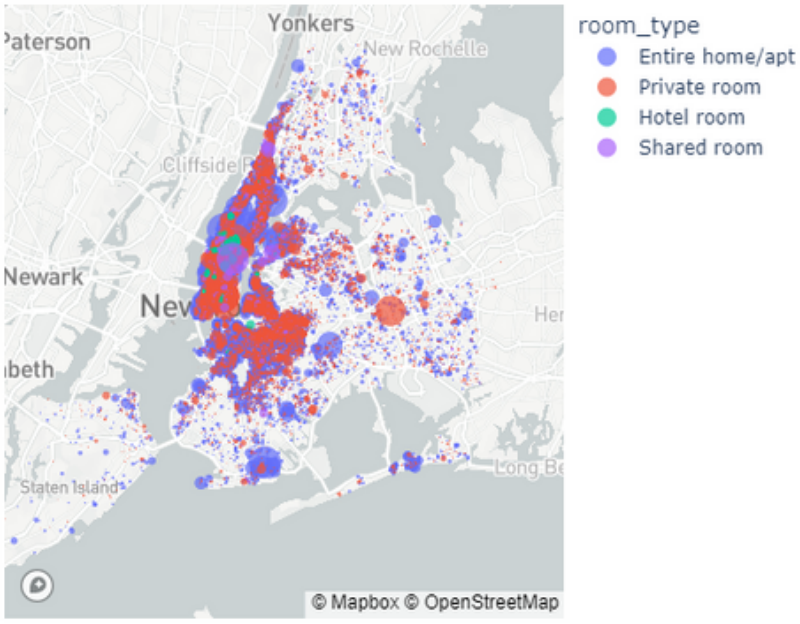


Figure 6: A map of all the listings in New York, NY color-coordinated based on the room type of the listing.

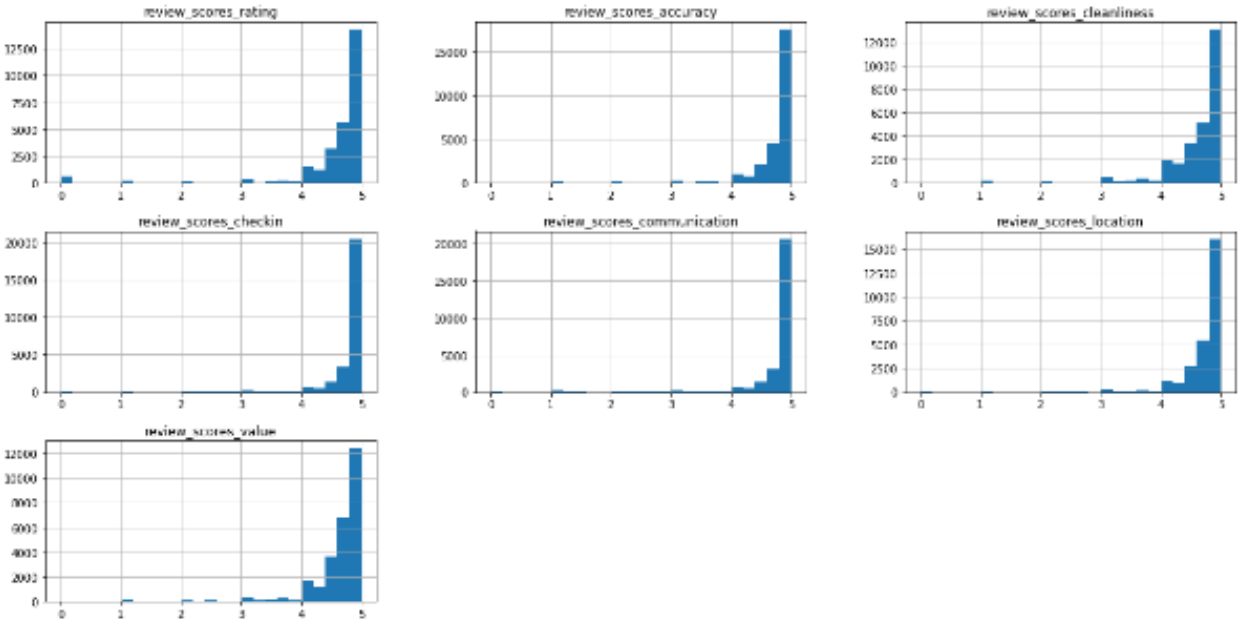


Figure 7: Histograms of the seven reviews columns. Most of the reviews lie between 4 and 5, particularly above 4.8.

LOCATION & PRICE OF A LISTING

BOROUGH

Manhattan had the highest number of listings at 16,182, while Staten Island had the lowest with 313. In terms of average prices of listings, however, while Manhattan once again claimed the highest value at $214, Bronx had the lowest at $104.

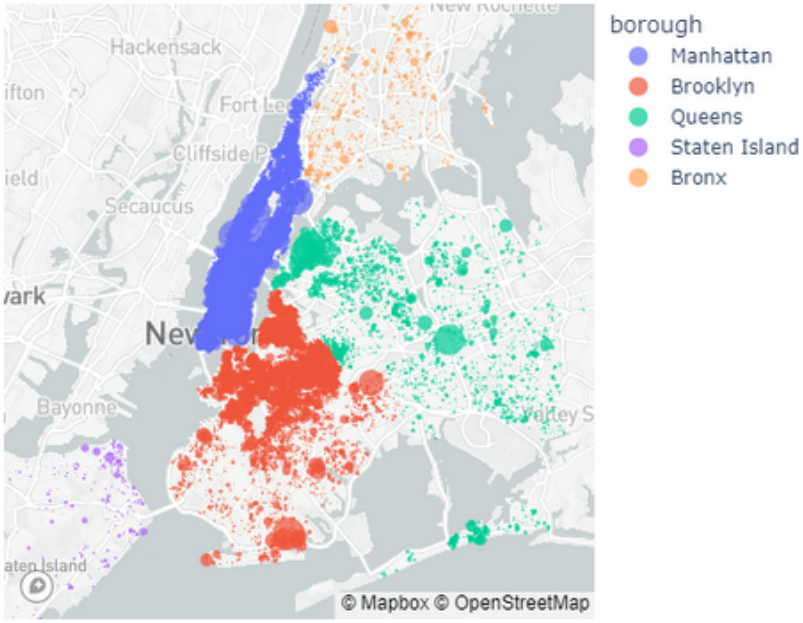
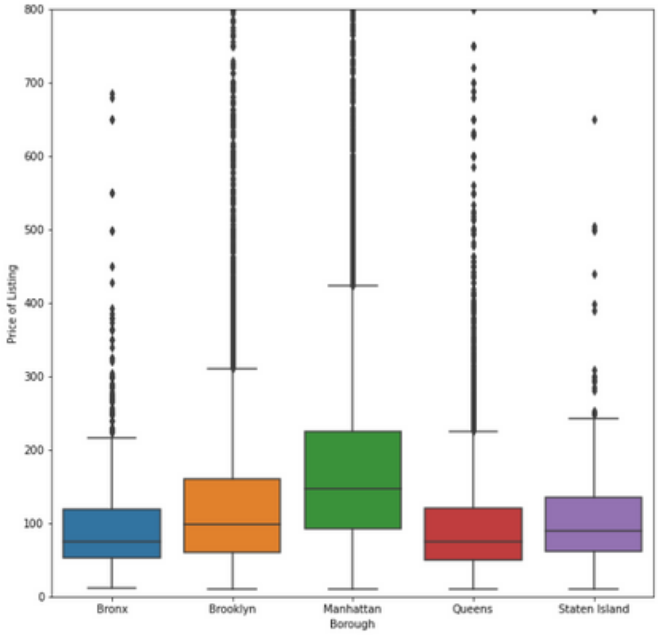


Figure 8: A map of all the listings in New York, NY color-coordinated based on the borough they are located in.



Manhattan has the largest range of variability of prices while Bronx has the smallest. However, a large majority of the listings in each borough are listed at a price of less than $1,000.

The most concerning aspect of the spread of the data points when grouped by location were the outliers. In order to understand the outliers better, we plotted ECDF curves which showed us the effect that outliers have on teh distribution of the data. All five boroughs' ECDF plots showed a curve at ~0.9.

Figure 9: A boxplot of the listings in New York based on the borough they are located in.

An ANOVA test was used to confirm that the borough in which a listing is located has an effect on the price of a listing.

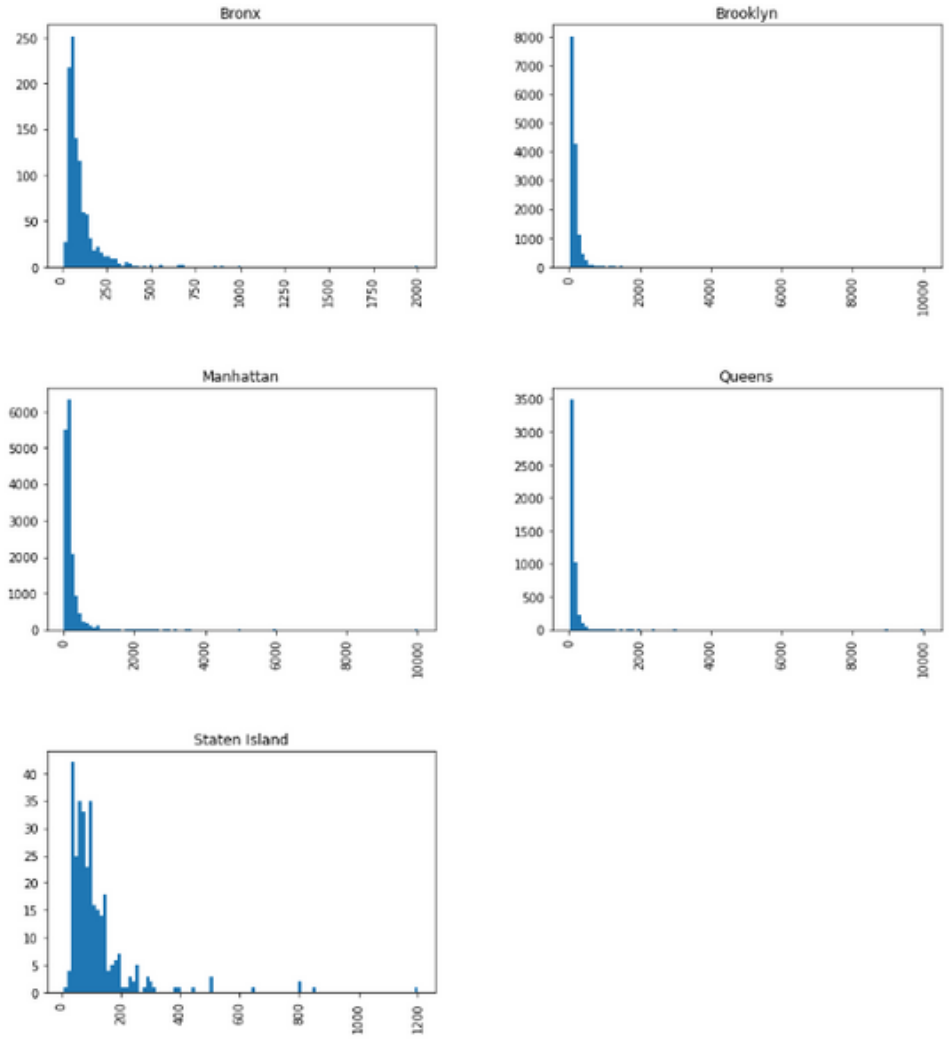


Figure 10: Histograms of the price of listings divided by the borough, aiming to show the price distribution in each.

NEIGHBORHOOD

There are a total of 223 neighborhoods in all five boroughs. Graniteville, located in Staten Island, has the least expensive mean price at $45.5 and Fort Wardsworth, also located in Staten Island, has the most expensive mean price at $800.0. However, it should be noted that Fort Wardsworth only contains 1 listing. It is also interesting to note that the neighborhood with the most expensive mean price for listings is not located in the borough with the most expensive mean price for listings.

On the other hand, Bedford-Stuyvesant in Brooklyn contains the highest number of listings while eight neighborhoods contain only 1 listing.

PRE-PROCESSING AND TRAINING

AIM

The preprocessing and training stage of the project involves steps to improve the quality of the dataset and select the features needed during the upcoming modeling of the data.

FEATURE SCALING

Feature scaling is the process of normalizing a range of independent variables or features of data. This was performed using **scikit-learn**'s StandardScaler. Any additional missing values were also filled in using statistical inference.

TARGET VARIABLE

Price was identified as the target variable of our model. However, since it contained many outliers, I removed outliers using quantile-based outlier detection. If a value was above the 95th quantile, then it was considered an outlier. With this, nearly 2,000 rows were removed. While some outliers were still present, it made sense to leave them in so as to not disturb the data too much.

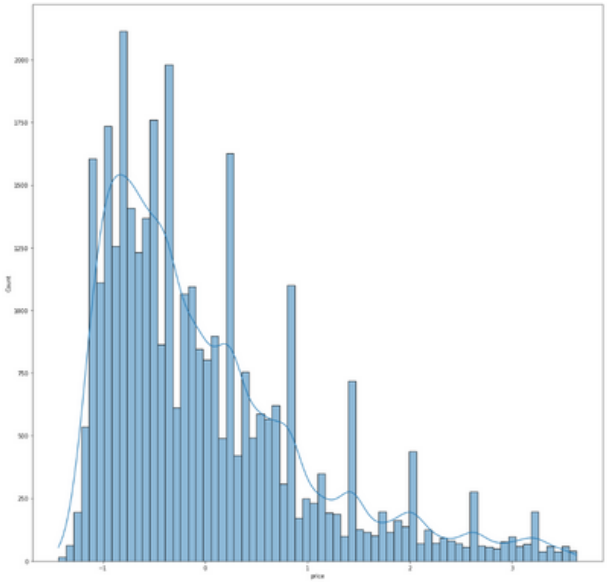


Figure 11: Data distribution of the **price** column.

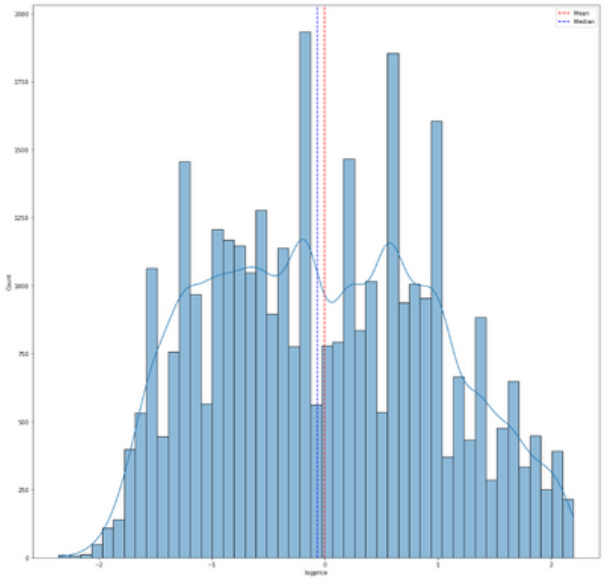


Figure 12: Data distribution of the **logprice** column.

Additionally, as shown in *Figure 11* and *Figure 12*, the data distribution of the price column was extremely right-skewed. The "bumps" on the right curve were caused because most of the values ended up being on the intervals of $150, $200, $250, and so on. In order to make the target variable resemble a Gaussian (normal) distribution more, **scikit-learn**'s PowerTransformer was used to transform the column into a logarithmic feature. This removed the skewing in the data distribution.

CATEGORICAL ENCODING

Categorical columns were split using the **pandas**' get_dummies method to create binary columns for each category: 1 for the category being present, 0 for it being absent. This was applied on Host Response Time, Boroughs, Property Types, Room Types, and Bathroom Types.

For Neighborhoods, instead of creating 200+ additional features, it was decided that the top 20 neighborhoods would only be used. Similar treatment was used on methods of Host Verification, using only top 5 most common methods.

For Amenities, only the amenities that had a proven relationship with the price of a listing based on statistical testing conducted int he EDA section were chosen to act as the basis of categorical encoding for this particular column.

DATA MODELING

Various machine learning methods were applied to our prepared and preprocessed dataset in order to predict our target variable.

TRAIN/TEST SPLIT

In order to avoid overfitting on the same dataset when we train and test our dataset, we will use **scikit-learn**'s **train_test_split** to split our data into separate datasets for training and testing. We will split them on a 70:30 basis.

MODELS

The target variable **logprice** is a continuously varying variable. As such, regression algorithms from supervised learning were utilized. These models were Linear Regression, Ridge Regression, k-Nearest Neighbor (kNN), and Random Forest.

In order to quantify how well a regression models fits the dataset, R-squared was the evaluation metric specifically chosen because the target variable is a logarithmic value and R-squared can tell us how well our model can predict it in percentage terms.

After, the k-fold Cross Validation method was used to evaluate our models and choose the best model to move forward with.

MODEL SELECTION

The following models were attempted before choosing the model with the best performance and accuracy: Linear Regression, Ridge Regression, k-Nearest Neighbor, Random Forest, and Extra Trees Regressor. Of the five, Random Forest had the best average cross validation R-Squared score for the model. Thus, it was chosen.

Additionally, in order to improve the model's performance, GridSearchCV was used to identify the best values for the number of estimators, the maximum number of features and the maximum depth for the trees. With the optimal parameters, predictions were made.

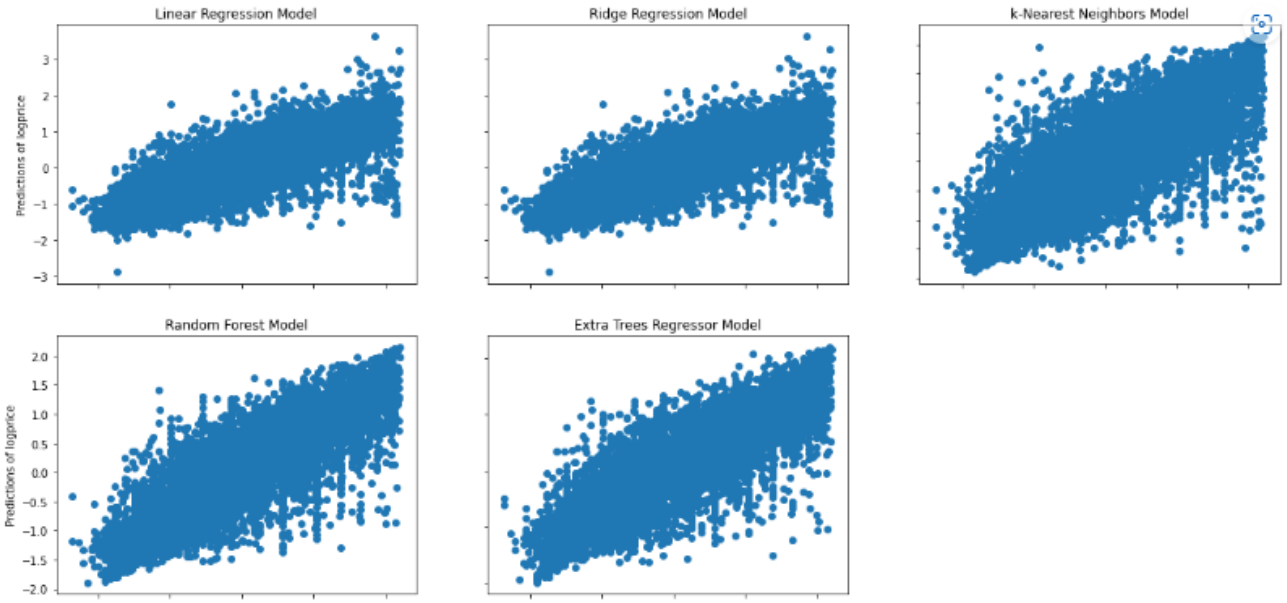


Figure 13: Scatter plot of the predictions from each baseline model.

CONCLUSION

It was concluded that, as it has the highest average cross validation R-squared score, the best performing model appears to be Random Forest.

Then, GridSearchCV was used to predict the best parameters for the model. These parameters were:

- max_depth: 10
- max_features: auto
- n_estimators: 150

FUTURE DIRECTION

There are many ways to further expand upon and improve upon the model as it exists right now. For starters, adding specific features regarding the distance of the listings from various locations and including various locations and experiences in the city would provide more specific information. Additionally, the inclusion of more amenities provided int eh listing during the prediction process would be helpful as well. A last recommendation I would add would be to use location-based data more aptly. This can take place in the form of thorough geospatial eploratory data analysis and inclusion during the modeling phase.