Exploratory Data Analysis on New York City Airbnb Dataset

Mahmut Özmen, Germany

mahmut.oezmen@protonmail.com

https://www.linkedin.com/in/mahmut-ozmen

January 27, 2022

Citation:

Mahmut Özmen, *Exploratory Data Analysis on New York City Airbnb Dataset*, 2022

Abstract

Since 2008, Airbnb has been an online marketplace for arranging or selling lodging, primarily home stays and tourism experiences. New York City is the most populous city in the United States, as well as one of the most popular tourist and business destinations on the planet.

Most people considering renting out their homes on Airbnb have a few questions. How can I increase my rate, is location important, how should I price, and how can I get more customers? Customers, like hosts, have similar questions. They wonder if the price is reasonable and where they should look for better housing. I used the NYC Airbnb open dataset to analyze and make price predictions in order to answer these questions. In this work, we clarify such question using Python 3 and its necessary data analysis libraries as well as scikit-learn for price prediction.

Individuals and businesses alike benefit from the outcome of this project. Data that can be analyzed and used for a variety of purposes, including security, business decisions, gaining a better understanding of customers' and providers' (hosts) behavior and performance on the platform, guiding marketing initiatives, launching innovative new services, and much more.

This dataset consists of around 48,895 observations, and 16 columns. Each host is represented in each row. We observe both categorical and numerical values in this dataset. This dataset should be cleaned as a beginning because it may contain missing values. Besides, data visualization is necessary since we have 16 columns. We make predictions using the price, review, and location columns using machine learning techniques with regression models. Random forest regression outperforms the other regression models on this dataset.

Table of Contents

List of Figures

1 Introduction

Since 2008, Airbnb [1] has been an online marketplace for arranging or selling lodging, primarily home stays and tourism experiences. New York City is the most populous city in the United States, as well as one of the most popular tourist and business destinations on the planet.

Most people considering renting out their homes on Airbnb have a few questions. How can I increase my rate, is location important, how should I price, and how can I get more customers? Customers, like hosts, have similar questions. They wonder if the price is reasonable and where they should look for better housing. I used the NYC Airbnb open dataset [2] to analyze and make price predictions in order to answer these questions. In this work, we clarify such question using Python 3 and its necessary data analysis libraries as well as scikit-learn [3] for price prediction.

This dataset consists of around 48,895 observations, and 16 columns. Each host is represented in each row. We observe both categorical and numerical values in this dataset. This dataset should be cleaned as a beginning because it may contain missing values. Besides, data visualization is necessary since we have 16 columns. We make predictions using the price, review, and location columns using machine learning techniques with regression models. Random forest regression outperforms the other models.

1.1 Motivation

The project outcomes are used not only by individuals but also by companies. Data that can be analyzed and used for security, business decisions, understanding of customers' and providers' (hosts) behavior and performance on the platform, guiding marketing initiatives, implementation of innovative additional services and much more.

2 Data Cleaning and Wrangling

Aim of the data cleaning and wrangling steps are

- Ensuring that all features are of the correct data type
- To ensure that missing data are properly imputed
- To create additional potential useful features
- To prepare the dataset for exploratory data analysis (EDA) and statistical analysis

2.1 Data Type Correction

All data types are correctly placed in the dataset, so there is no need to place them.

2.2 Missing Value Imputation

Using the *isnull().sum()* method of *pandas* library, we can see the columns *name*, *host_name*, *last_review*, and *reviews_per_month* columns have missing values. Later, I have checked the correlation between rows and columns that contain missing values.

Dealing with Missing Data

The *name* and *host_name* columns are irrelevant with our further work. Hence, we can drop them.

As we have no review for the data, then there is no need for *last_review* column, and *reviews_per_month* column is 0 for these rows of the data. Therefore, we can replace *reviews_per_month* values with 0, and we remove the *last_review* column since it is not relevant with our framework. After deleting these columns, we have a total of 13 columns in this dataset.

Using the *isnull().sum()* method of *pandas* library, we can see the columns *name*, *host_name*, *last_review*, and *reviews_per_month* columns have missing values. Later, I have checked the correlation between rows and columns that contain missing values.

3 Data Story

This chapter provides insight about the Airbnb NYC dataset.

3.1 Host Distribution

We can see that the top ten hosts with the most listings have a good distribution. The first host has over 300 counts, while the second has around 230 as we can see in Figure 3.1.

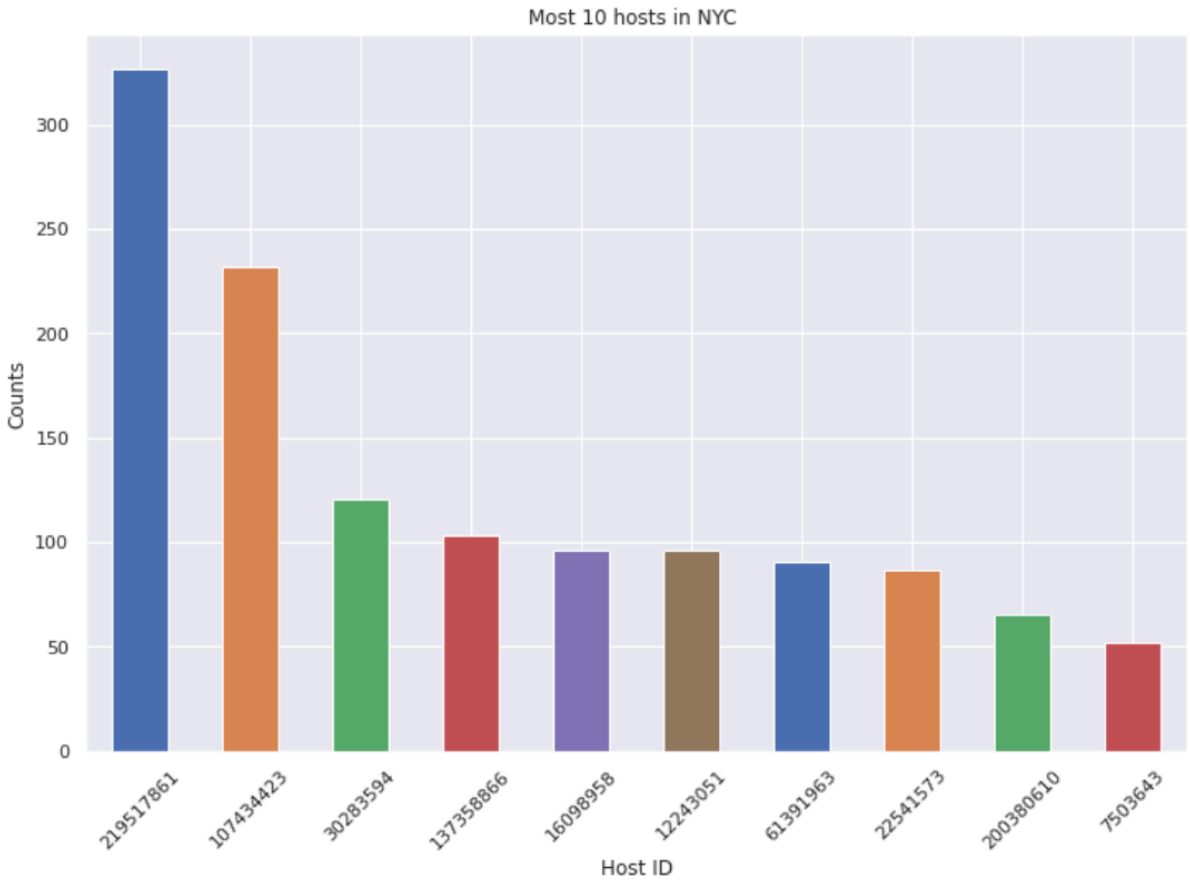

Figure 3.1: Host Distribution. Most ten hosts in NYC.

3.2 Price Distribution

Looking at the price distribution graph (Figure 3.2) we notice that most of the prices are distributed between $40 and $100 per day. The number of prices are very low

above $\$400$. Hence, we limited price distribution with $\$1000$ because prices above $\$1000$ are outliers. The mean price is $\$133$.

Figure 3.2: Price Distribution.

3.3 Neighbourhood Group

If we check the number of the distribution of neighborhood group on Figures 3.3, 3.4, we observe that there are around 21000 hosts in Manhattan, 20000 hosts in Brooklyn, 5500 hosts in Queens, 1000 hosts in Bronx, and 350 hosts in Staten Island. We conclude that most of the hosts are located in Manhattan and Brooklyn. Bronx and Staten Island have quite low number of hosts.

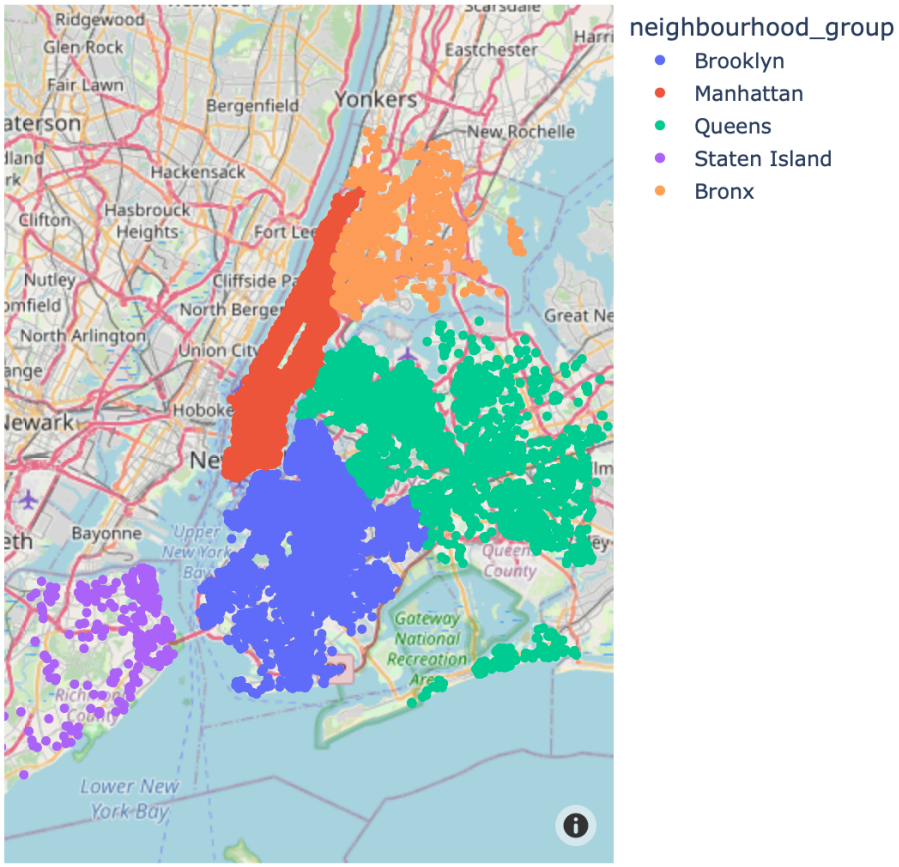

Figure 3.3: Distribution of the Data with respect to Neighbourhood Group on the NY map

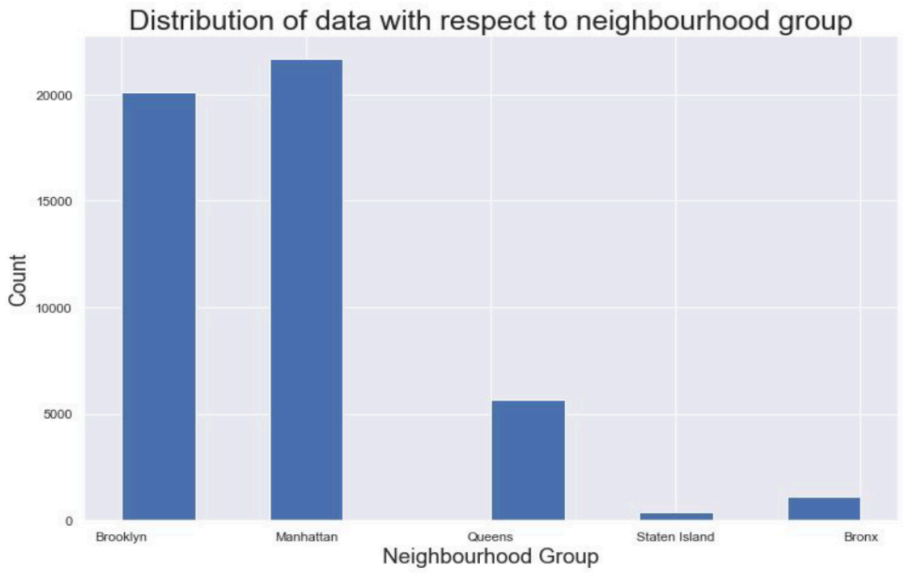

Figure 3.4: Distribution of the Data with respect to Neighbourhood Group

Neighbourhood Group and Price Distribution

The dots with red color indicate the apartment or rooms with higher price. We consider prices up to $\$500$ to get a good representation on the plot. We observe that Manhattan region contains more expensive apartments 3.5.

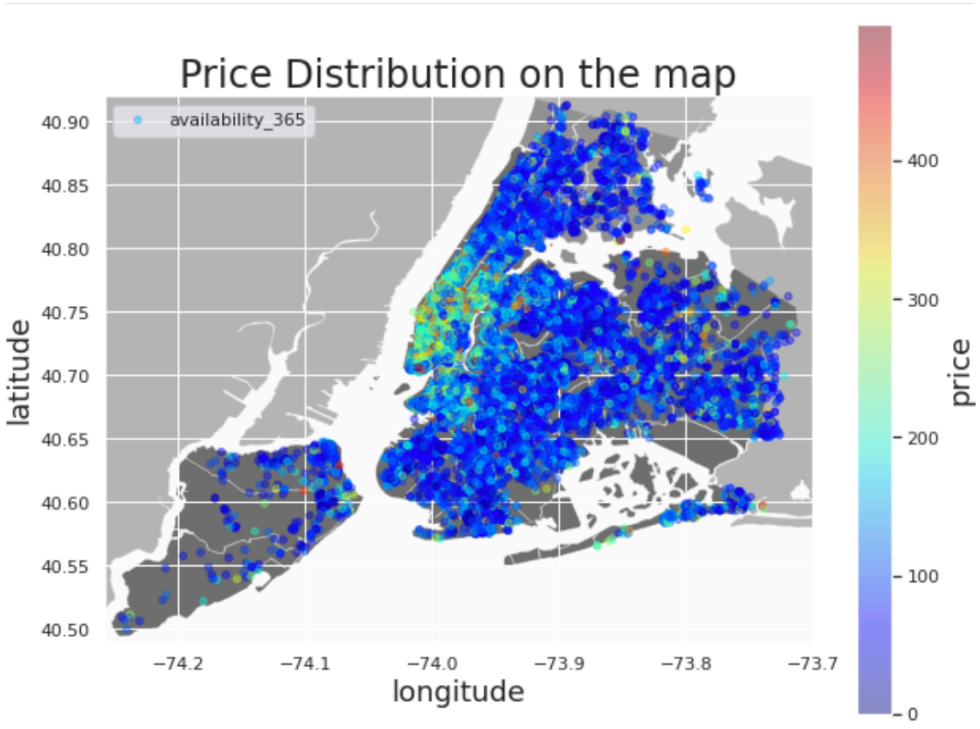

Figure 3.5: Distribution of the Data with respect to Neighbourhood Group

We concentrate on the average and range of price distribution between neighborhood groups 3.6. We become aware that the range and price distribution in Manhattan is higher than others. Queens and Staten Island appear to have very similar distributions. Bronx is the cheapest neighborhood.

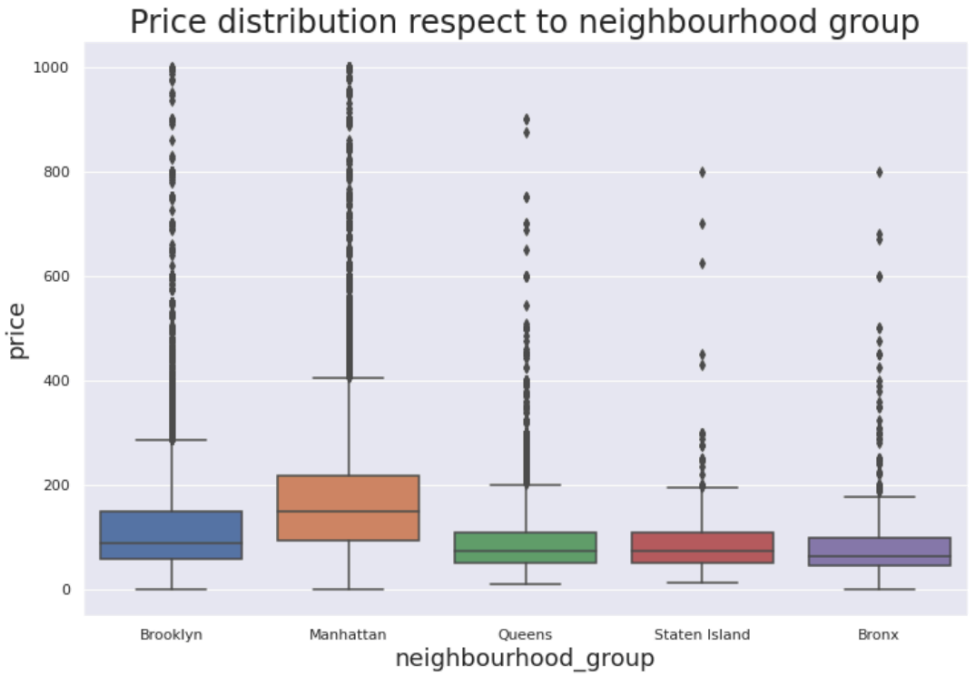

Figure 3.6: Price Distribution with respect to Neighbourhood Group

We are curious whether the data sets are similar because their host counts are very close to each other. Hence, we employed t-Test for average of prices between Manhattan and Brooklyn neighborhood groups to check if they are significantly similar. The null hypothesis in these datasets was that both the data sets are significantly similar. The alternative hypothesis was that both the data sets are significantly different. When we take into account of the significance level to be at 5%, then to accept null hypothesis, our p-value should be more than the chosen level of significance. In the result of the t-Test, the p-value is 0. Hence, we ignore null hypothesis. The t-Test correctly highlights that the mean of each dataset are different and are statistically different from each other.

3.4 Room Types

When we look at the different room types available in New York City (Figure 3.7), we notice that there are quite a few. In a house or apartment, the total number of rooms is around 25000, private rooms are around 23000, and shared rooms are around 1160. The number of private rooms and the total number of rooms in the house/apartment are very close to each other, and the number of shared rooms is very low. The highest number is for the entire home apartment, followed by private rooms, and shared rooms are the least preferred.

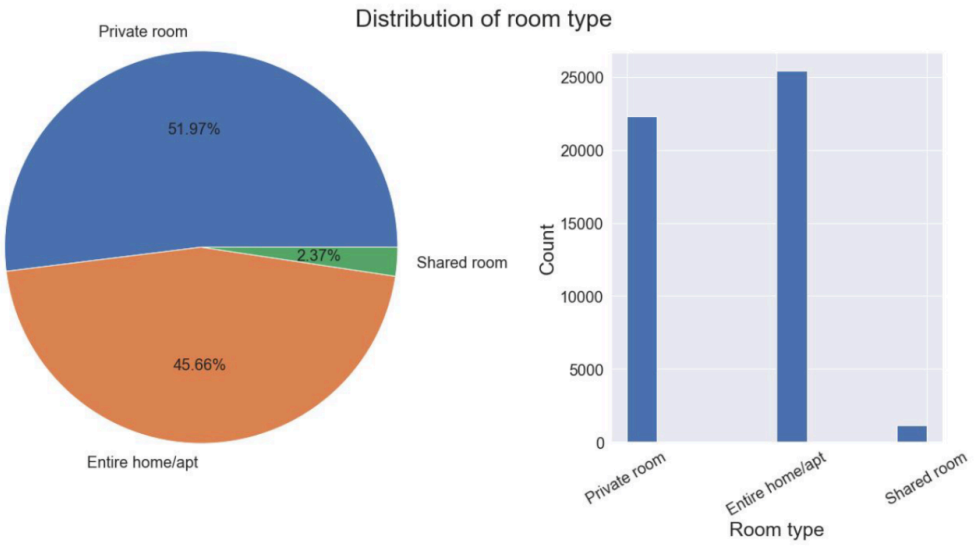

Figure 3.7: Distribution of Room Types

Room Type and Neighbourhood Group Distribution

When we look at the distribution of room types (Figure 3.8), we can see that Manhattan and Brooklyn have a lot of listings for each one. That was to be expected, given that Manhattan and Brooklyn are two of the most popular tourist destinations and thus have the most listing availability. While Manhattan is first in the categories of entire home/apt and shared room, it is second in the category of private room.

Every room type puts Queens in third place. There are only a few listings on Staten Island. It's at the bottom of the list for each room type. As a result, we can say that Staten Island is New York City's least visited destination.

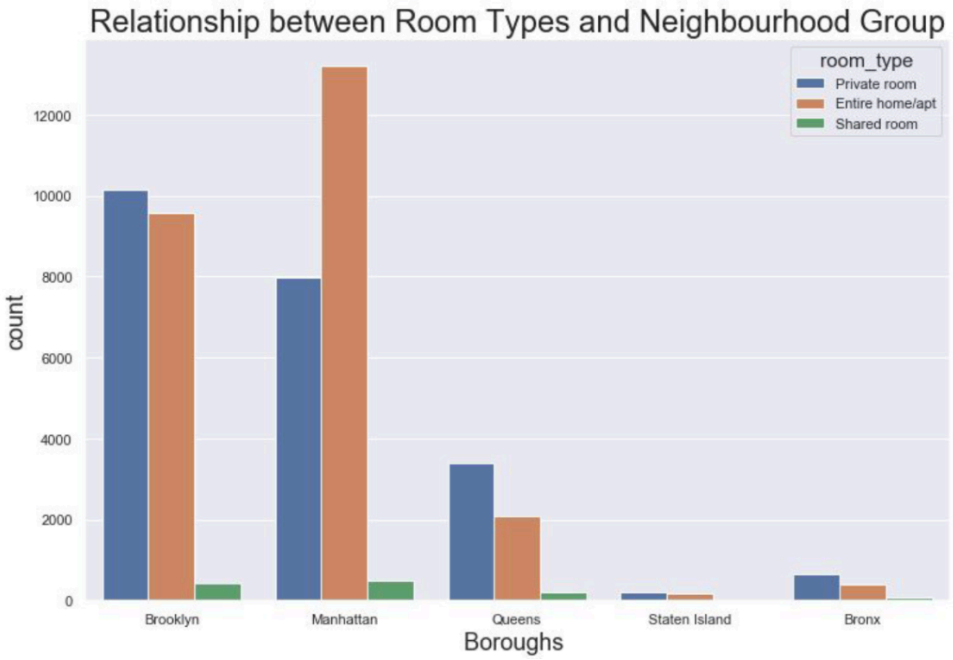

Figure 3.8: Relationship between Room Types and Neighbourhood Groups

Room Type and Price Distribution

The average price for an entire home/apt is around $\$150$, making it the most expensive room type. It also has the greatest range. The average cost of a private room is $\$75$, while the average cost of a shared room is $\$50$. To get a good representation of the violin graft, we have considered prices up to $\$600$. As we can see, their means are not that dissimilar, and their ranges are very similar when we look at the violin graph. Hence, we focus on their correlation.

In both the private and shared rooms, we use z-Test. In these datasets, the null hypothesis was that private rooms and shared rooms are significantly similar (Figure 3.9). The alternative hypothesis was that the two data sets differed significantly. To accept the null hypothesis, we believe the significance level should be greater than 5%. Therefore, our p-value should be greater than the chosen significance level. Because the p-value$= 0$ in my test result is less than 0.05 (5%), we can disregard the null hypothesis. Both datasets are statistically significantly different, according to the z-Test.

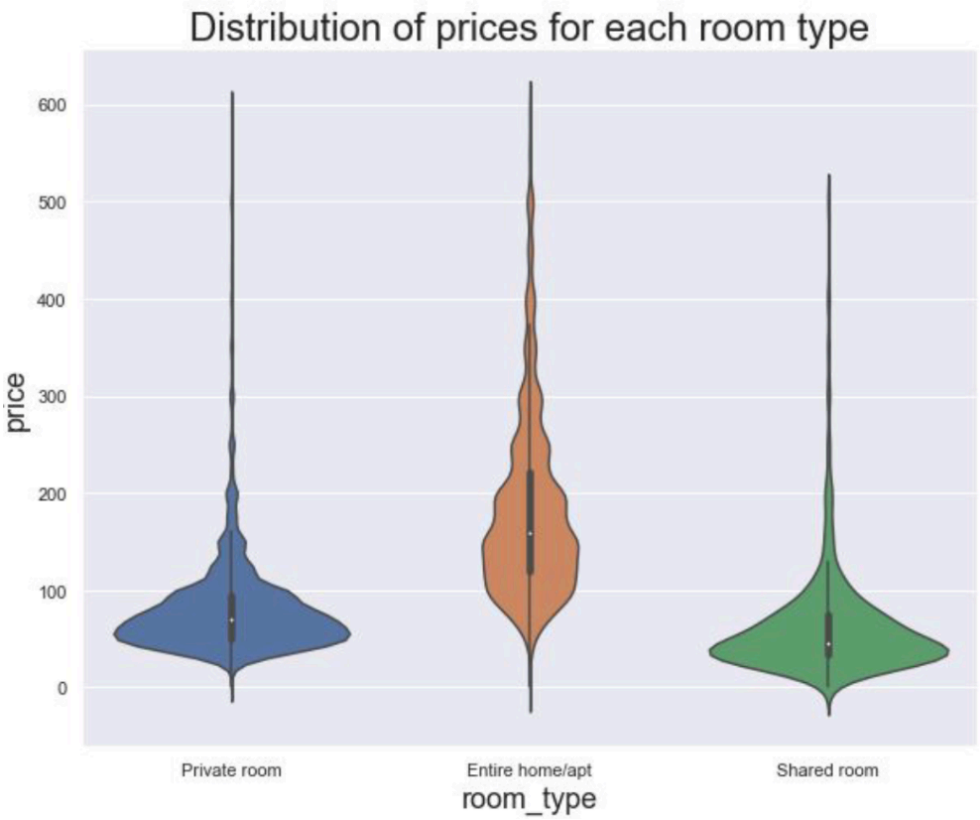

Figure 3.9: Price Distribution between Room Types

4 Modelling

4.1 Data Preparation for Modelling

To begin, we determined the upper limit of price columns in order to eliminate outliers. The price distribution's upper limit is $\$334$. we raise the prices to $\$334$. we still had 45918 rows when we checked the data shape. As a result, we lost around 3000 data points. It's not a big deal because we still have enough data to figure out which machine learning model performs best. The types of the three columns 'neighbourhood', 'neighbourhood group', and 'room type' are objects in the data set. They were converted to a numerical value by me. Because conversion created columns for each category in the object columns, we had 237 columns.

We made feature columns out of all of the columns except the price. Price is the column that will be used to estimate prices. After that, We divided the data into training (75%) and testing (25%) samples.

4.2 Regression Models

To begin, we used regression models, and computed R^2 and RMSE (Root Mean Squared Error) (Figure 4.1). The R^2 statistic indicates how close the data are to the fitted regression line. The higher the R-squared, the better the model fits your data in general. The RMSE is the square root of the residuals' variance. It shows how well the model fits the data in terms of absolute fit–how close the observed data points are to the model's predicted values.

With default parameters, we used Linear Regression. The R2 value is 0.52 and the RMSE value is 46. I used Lasso Regression after linear regression with various regularization (alpha) parameters of 1, 0.1, and 0.01. The best result was alpha=0.01. The R2 value is 0.52 and the RMSE value is 46.6. We visualize them in order to compare their relationships.

When we look at the plot and the code, the regularization parameter's default value (given by α) is 1 in Lasso regression. Only 11 of the 237 features in the data set are used in this way (non zero value of the coefficient). When $\alpha = 0.1$, non-zero features, training and test score goes up. Table is in the next page.

For $\alpha = 1$, we observe that most of the coefficients are zero or nearly zero, which is not the case for $\alpha = 0.1$. Reudcing the α further, non-zero features increases. Training and test scores are similar to basic linear regression cases.

Besides, I employed Ridge Regression, ElasticNet Regression, Decision Tree Regressor, Random Forest Regressor, Gradient Boosting Regressor, Linear SVM, Nonlinear SVM.

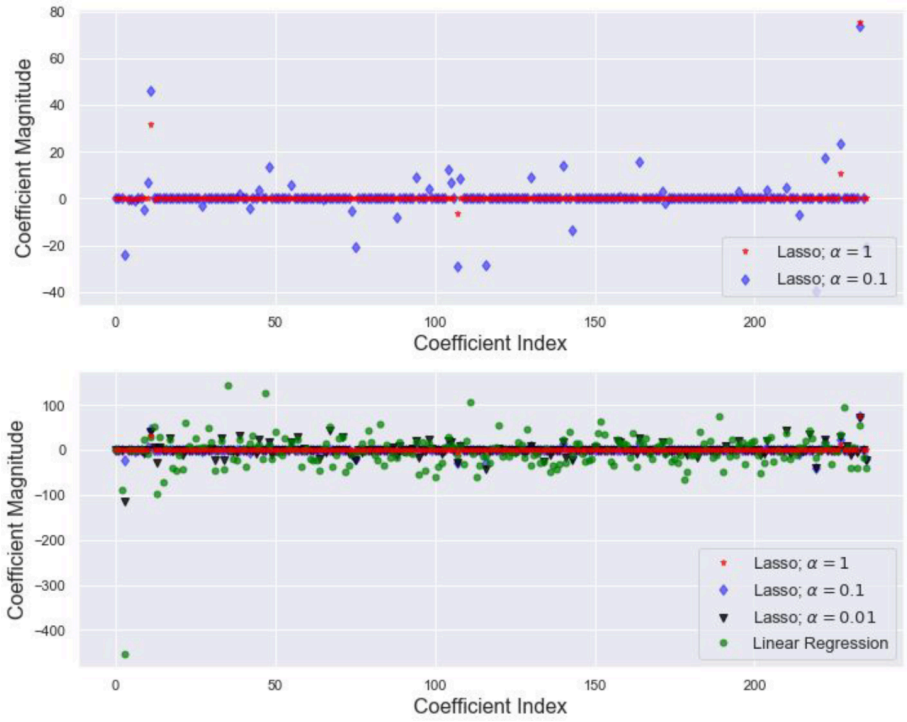

Figure 4.1: Linear and Lasso Regression.

Model Name	Best Parameter	R^2	RSME
Linear Regression	default	0.526	46.71
Lasso Regression	$\alpha = 0.001$	0.524	46.69
Ridge Regression	$\alpha = 1$	0.524	46.69
ElasticNet Regression	$\alpha = 0.001$	0.523	46.76
Decision Tree Regressor	default	0.16	61.9
Random Forest Regressor	default	0.58	43.88
Gradient Boosting Regressor	default	0.56	44.85
Linear SVM	default	0.0002	135
Nonlinear SVM	default	0.04	70

When we look at the table we can see that the Random Forest Regressor model has the highest R^2 value$= 0.58$ and the lowest RMSE$= 4385$. After we found the best model for the NYC Airbnb dataset, we used the best model for all samples without removing outliers. When we apply the model for all samples R^2 is 0.23 and RMSE is 218.

5 Results

5.1 Conclusion

This Airbnb dataset is a large dataset with a lot of columns. As a result, we were able to conduct extensive data analysis. First, we looked into hosts who could provide me with Airbnb host listing distributions. The top host has over three hundred listings, according to my research. I then proceed to price analysis. we noticed that the majority of the prices are in the $\$40$-$\$100$ per day range, and rice counts are extremely low above $\$400$. The average cost of my services is $\$133$. After that, we looked at the price distribution and neighborhood group. We can deduce that Brooklyn and Manhattan are the most expensive and well-traveled boroughs in New York City. Finally, we looked into room types and how they relate to price.

The entire home/apt has the highest number, followed by the private room, and the shared room has the lowest number. The average price for an entire home/apt is around $\$150$, making it the most expensive room type. It also has the greatest range. The average cost of a private room is $\$75$, while the average cost of a shared room is $\$50$. When it came to regression models, I found that a random forest regressor produced the best results. The most variance is found in random forest regression, which has a percent of variance of 58. The more variance that the regression model accounts for, the closer the data points fall to the fitted regression line.

5.2 Limitations

NYC Airbnb dataset does not include some essential information, e.g., room numbers, bedroom numbers, how many people can live in a house, and measures of houses. We believe that these features may affect the result of price prediction.

6 References

[1] *Airbnb*. URL: http://airbnb.com.

[2] *How is Airbnb really being used in and affecting the neighbourhoods of your city?* URL: http://insideairbnb.com.

[3] **F. Pedregosa et al.** “Scikit-learn: Machine Learning in Python”. In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.