University of Miami

School of Arts and Sciences

Statistical Learning with Applications

CSC 642 - Group Project

Project Report

DATA ANALYSIS OF AIRBNB NEW YORK 2019 OPEN DATA

Group Members: Sunity Sharma & Melvin Maria
Instructor's Name: Vanessa Aguiar

CONTENTS

ABSTRACT

This report presents an analysis of the New York City Airbnb dataset (2019) [1] using RStudio and R programming language. The study utilizes machine learning techniques such as regression and classification models to predict Airbnb prices based on various factors. The report covers data cleaning, variable identification, model application, and analysis of the best-performing model for this application. The statistical analysis aims to provide insights into the factors that influence Airbnb prices and their evolution over time. This study shows potential, and it is vital for understanding the dynamics of the Airbnb market and can inform pricing strategies for Airbnb hosts and rental businesses.

INTRODUCTION

Airbnb has become a popular platform for travelers seeking unique and affordable accommodations, and for homeowners looking to rent out their properties. In New York City, which is a very popular destination around the world there are numerous Airbnb listings, making it one of the largest markets in the United States. However, the prices of these listings vary widely depending on factors such as location, amenities, time of year, or maybe availability.

In this project, we aim to explore the factors that affect Airbnb prices in New York City and build predictive models to forecast the price of given listings throughout the city using statistical learning. We will use a dataset of over 45,000 Airbnb instances in New York City, which includes information such as location, room type, and number of reviews, yearly availability and so on.

To better assess our aim, we have come up with some questions about Regression and Classification which would be crucial for the analysis:

Regression Analysis Questions:

1. How do different factors, such as location, property type, and availability, impact the price of an Airbnb rental in New York?
2. Can we predict the price of an Airbnb rental in New York based on factors such as location, reviews, and availability? Will the results be significant?
3. Can we accurately predict the price of an Airbnb listing using a regression model, and which features are most important in determining price?

Classification Analysis Questions:

1. Can we accurately predict whether an Airbnb rental in New York is considered "affordable" based on factors such as location, amenities, and reviews?
2. Can we classify Airbnb rentals in New York as "good value" or "not worth the price" based on factors such as neighborhood, property type, and reviews?
3. Can we predict whether an Airbnb rental in New York is likely to receive a high volume of bookings based on factors such as availability, location, and property type?

STATE OF THE ART

There have been many studies and analyses performed on the Airbnb New York City dataset [1], each with their own unique approaches and findings. Some examples of previous studies on this dataset include:

- **Kalehbasti, Pouya Rezazadeh, Liubov Nikolenko, and Hoormazd Rezaei. "Airbnb price prediction using machine learning and sentiment analysis." arXiv preprint arXiv:1907.12665 (2019).** [2]
  The paper explores the prediction of Airbnb prices using a restricted set of features and machine learning techniques, including linear regression, tree-based models, SVR, and neural networks. The study analyzes the performance of these models based on their respective feature importance analyses and suggests that SVR performed the best in terms of accuracy. The paper also proposes future work, such as exploring other feature selection schemes and experimenting with neural network architectures.
- **Luo, Yuanhang, Xuanyu Zhou, and Yulian Zhou. "Predicting airbnb listing price across different cities." (2019)** [3]
  The paper conducts extensive feature extraction and engineering and experiments with various machine learning approaches for predicting Airbnb listing prices. The study shows that XGBoost and neural network models outperform other approaches and suggests the potential for a generalized model trained on datasets from multiple cities using neural networks. The paper highlights the importance of further research into transfer learning with neural networks and the need for improving neural network performance with extra feature extraction and hyper-parameter tuning in future work.
- **Y. Li, Q. Pan, T. Yang and L. Guo, "Reasonable price recommendation on Airbnb using multi-scale clustering," 2016 35th Chinese Control Conference (CCC), Chengdu, China, 2016, pp. 7038-7041, doi: 10.1109/ChiCC.2016.7554467.** [4]
  The paper proposes a multi-scale clustering approach to recommend reasonable prices for Airbnb listings, considering both spatial and temporal dimensions. The approach is based on a hierarchical clustering method that groups listings based on their characteristics and prices and provides price recommendations for each group.

Our analysis focused on exploring the impact of various factors, such as location, property type, availability, and amenities, on the price of Airbnb rentals in New York City. We used feature selection techniques, clustering analysis, and regression analysis to identify the most significant predictors of rental price. While our approach is not unique, our analysis provides a comprehensive exploration of the factors that impact Airbnb rental prices in New York City and offers insights that can be useful for hosts and travelers alike.

MATERIALS & METHODS

I. DATASET

The name of the dataset is "New York City Open Dataset (2019) – Kaggle [1]"

Website Link: https://www.kaggle.com/datasets/dgomonov/new-york-city-airbnb-open-data

It contains information about Airbnb listings in New York City, including variables such as Listing ID, Name, Neighborhood, Room Type, Price, Availability, etc.

It consists of around 48000 instances and will require data cleaning as it has a few missing values.

From the different variables that it possesses, it is a multivariate dataset with some of them being categorical in nature which gives us the ability to perform regression and classification and the following statistical interpretation.

id	name	host_id	host_name	neighbourhood_group	neighbourhood	latitude	longitude	room_type	price	minimum_nights	number_of_reviews	last_review	reviews_per_month	calculated_host_listings_count	availability_365
2539	Clean & quiet apt home by the park	2787	John	Brooklyn	Kensington	40.64749	-73.97237	Private room	149	1	9	2018-10-19	0.21	6	365
2595	Skylit Midtown Castle	2845	Jennifer	Manhattan	Midtown	40.75362	-73.98377	Entire home/apt	225	1	45	2019-05-21	0.38	2	355
3647	THE VILLAGE OF HARLEM....NEW YORK !	4632	Elisabeth	Manhattan	Harlem	40.80902	-73.94190	Private room	150	3	0	NaN	NaN	1	365
3831	Cozy Entire Floor of Brownstone	4869	LisaRoxanne	Brooklyn	Clinton Hill	40.68514	-73.95976	Entire home/apt	89	1	270	2019-07-05	4.64	1	194
5022	Entire Apt: Spacious Studio/Loft by central park	7192	Laura	Manhattan	East Harlem	40.79851	-73.94399	Entire home/apt	80	10	9	2018-11-19	0.10	1	0

Figure 1.: Initial data head

II. CLEANING THE DATA

Our data consisted of around 48000 instances and around up to 20000 missing but, it was mostly from columns which would not affect the analysis much, so we removed the columns "last review", "name", "ID", "Host ID" and some rows associated with the column "review per month" illustrated in figure 2.

	neighbourhood_group	neighbourhood	latitude	longitude	room_type	price	minimum_nights	number_of_reviews	reviews_per_month	calculated_host_listings_count	availability_365
0	Brooklyn	Kensington	40.64749	-73.97237	Private room	149	1	9	0.21	6	365
1	Manhattan	Midtown	40.75362	-73.98377	Entire home/apt	225	1	45	0.38	2	355
3	Brooklyn	Clinton Hill	40.68514	-73.95976	Entire home/apt	89	1	270	4.64	1	194
4	Manhattan	East Harlem	40.79851	-73.94399	Entire home/apt	80	10	9	0.10	1	0
5	Manhattan	Murray Hill	40.74767	-73.97500	Entire home/apt	200	3	74	0.59	1	129

Figure 2.: Data with no missing values

III. EXPLORATORY ANALYSIS

For classification, first an exploration through the different aspects of the data was made as it is shown in the figures below. This involved generating summary statistics for the different variables, creating visualizations such as histograms and box plots, and computing correlations between the variables.

From our initial exploration, we identified several important trends and patterns, including the relationship between the rental price and location, property type, and availability.

1. Different types of Airbnb Listings in New York

This plot generates the percentage of different types of Airbnb listings available in each neighborhood group of New York City. We can see that most Airbnb listings in New York City are for the **private room,** with Bronx and Queens having the highest percentages.

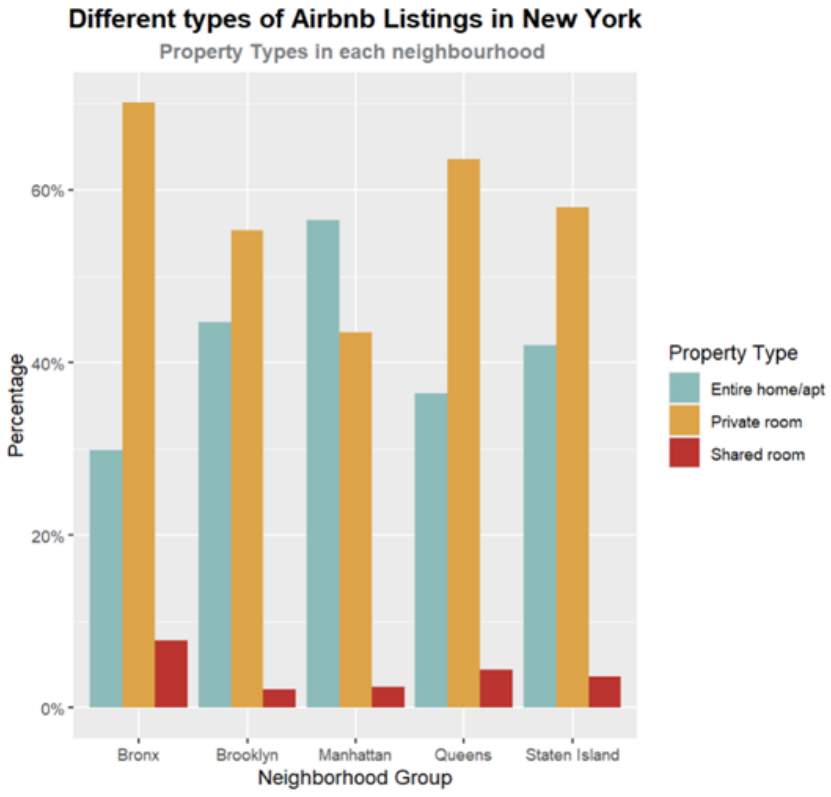

Figure 3: Different types of Airbnb Listings in New York

2. Comparison of Mean Price for each Neighborhood group

This is a bar chart of the **average price** of Airbnb listings in each **neighborhood group** of New York City. We can see that **Manhattan** has the **highest average Airbnb price**, followed by Brooklyn and Queens. The **Bronx** and **Staten Island** have significantly **lower average** Airbnb prices.

Figure 4: Comparison of Mean Price for each Neighborhood group

3. Comparison of Mean Price with all Room Types

The chart shows that the mean price of entire homes/apartments is the most (**163 USD**), followed by private rooms(81USD) and shared rooms(64 USD). The mean price of an entire home/apartment is **double** the mean price of a private room.

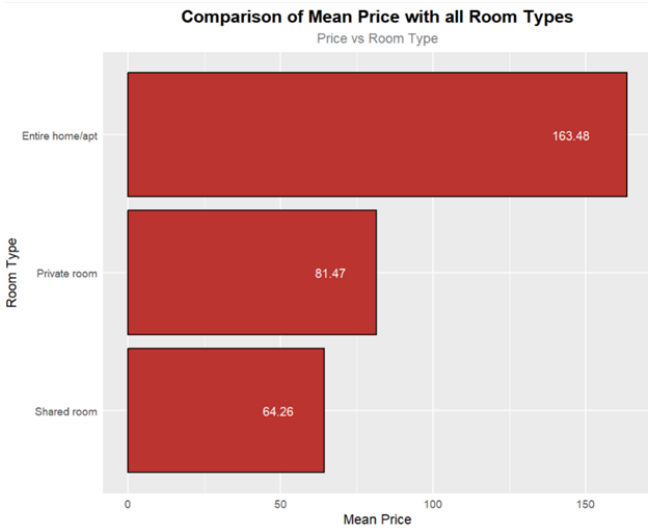

Figure 5: Comparison of Mean Price with all Room Types

4. Price distribution by neighborhood

This is a scatter plot that shows the mean price of listings in each neighborhood group, with Manhattan having the highest average price and Staten Island having the lowest.

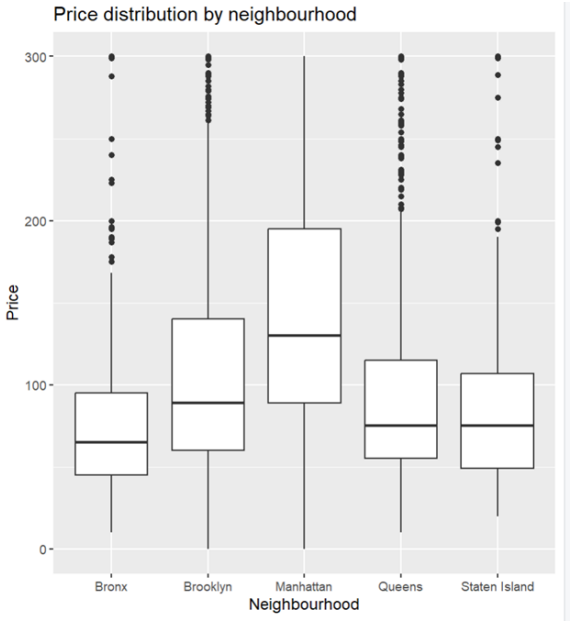

Figure 6: Price distribution by neighborhood

5. Price distribution by room type

This plot helps us compare the pricing for the type of property booked. Clearly, booking an entire home/apartment is more expensive, followed by Private room and then shared room.

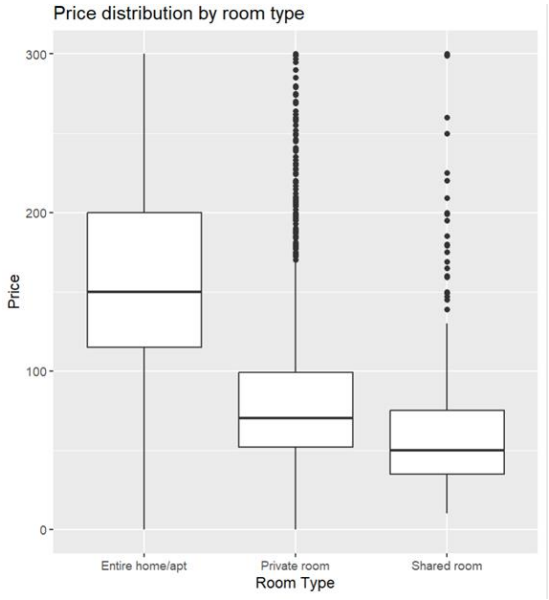

Figure 7: Price distribution by room type

6. Price distribution v/s Minimum nights (Scatter Plot)

This is a scatter plot of the price of Airbnb listings on the y-axis and the minimum number of nights required to book the listing on the x-axis. Each dot represents an Airbnb listing, and the color of the dot represents the neighborhood group of the listing. **No clear correlation** between the minimum number of nights required to book a listing and its price.

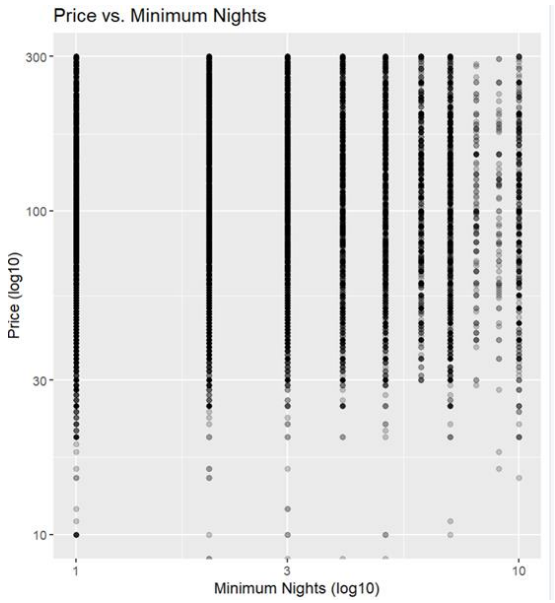

Figure 8: Price distribution v/s Minimum nights

7. Correlation Heat Map

This is a correlation matrix between the numeric variables in the Airbnb dataset. The darker shades of blue indicate a stronger negative correlation, while the darker shades of red indicate a stronger positive correlation. **Number of reviews and availability** shows **strong positive relation**. Between **availability and minimum nights**, we see **strong negative relation**.

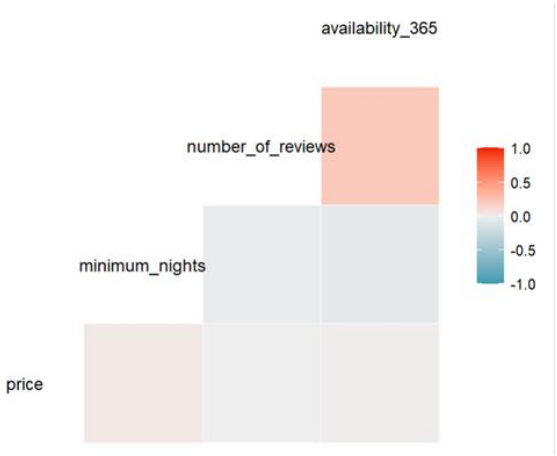

Figure 9: Correlation Heat Map

IV. FEATURE ENGINEERING

We perform some feature engineering to create new features that may be useful for our data analysis.

1. Feature – Times Square v/s Central Park

We have created a new feature for the distance to Times Square and Central Park. We observe the relation between distribution of bookings with respect to closeness to each of the destinations.

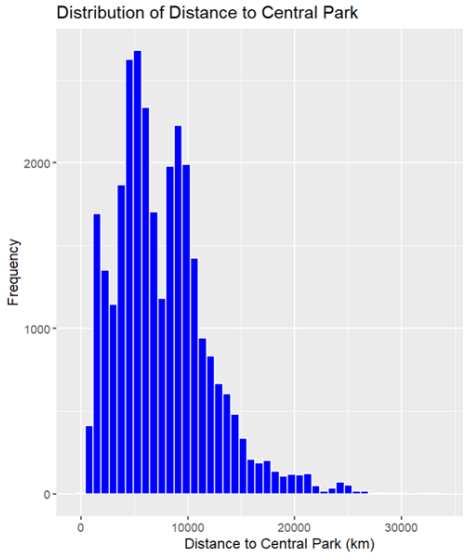

Figure 10

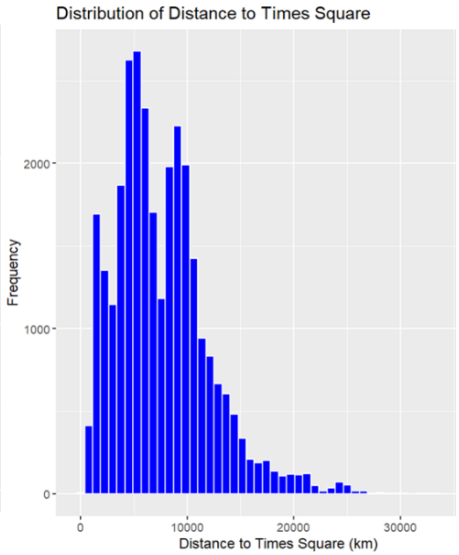

Figure 11

2. Feature - LaGuardia Airport v/s JFK Airport

We have created a new feature for distance to LaGuardia Airport and JFK Airport. We observe the relation between distribution of bookings with respect to closeness to each of the airports.

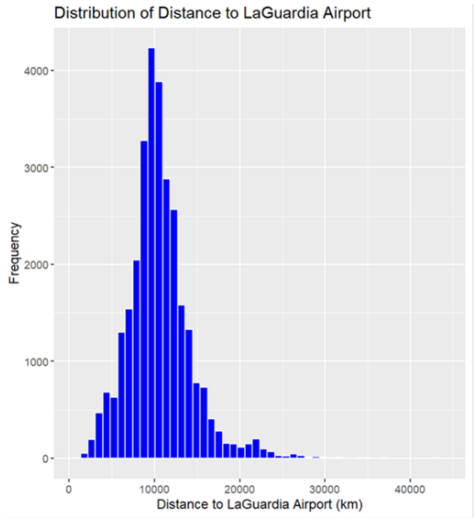

Figure 12

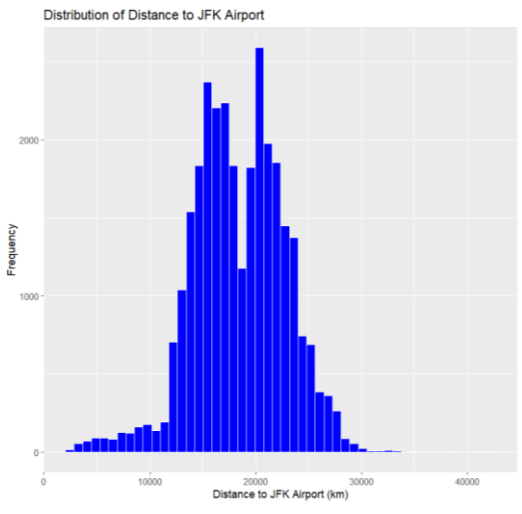

Figure 13

3. Analysis – Most Correlated Features with Price

We plot the most correlated features that impact pricing. We see that the new features i.e., distance from Times Square, distance from Central Park and distance from JFK Airports are in the top 10 most correlated features. Also, Private room and entire room, properties in Manhattan and Brooklyn are the most correlated with pricing.

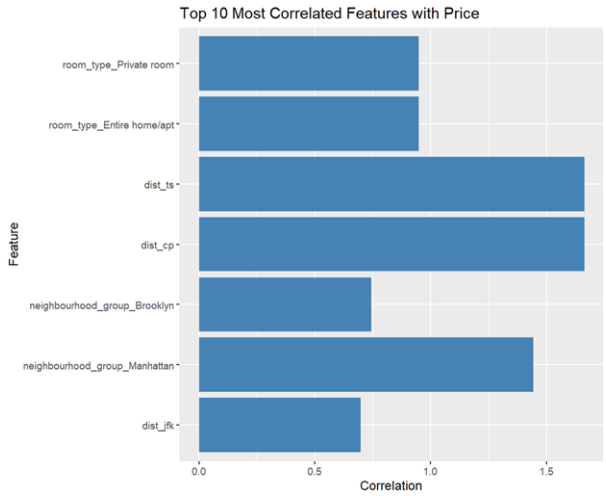

Figure 14: Most Correlated Features with Price

4. Analysis – Price Distribution for AirBnb Rentals

We plot the price distribution for Airbnb Rentals and observe the following trends:

- Most listings between 40 USD to 100 USD were booked
- Minimum pricing starts from 40USD
- Ranges to 300 USD and above
- Frequency decreases as the price increases.

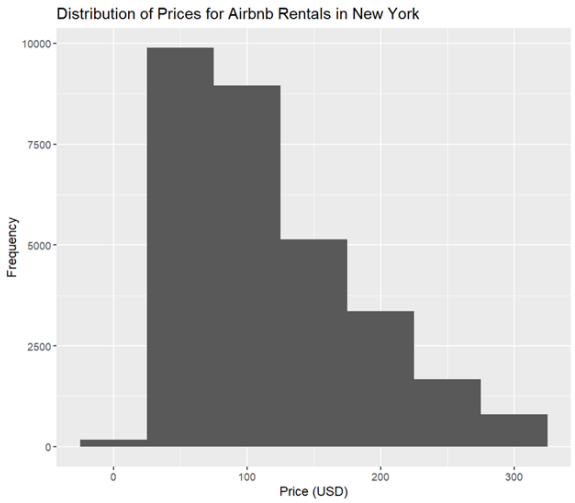

Figure 15: Price Distribution for AirBnb Rentals

5. Analysis – Price by Room Type

We plot the Price v/s Room type with respect to each Neighborhood Group and observe the following trends.

- Booking an entire home/apt in Manhattan is the most expensive.
- Booking an entire home/apt is the most expensive.
- Manhattan is the most expensive neighborhood.

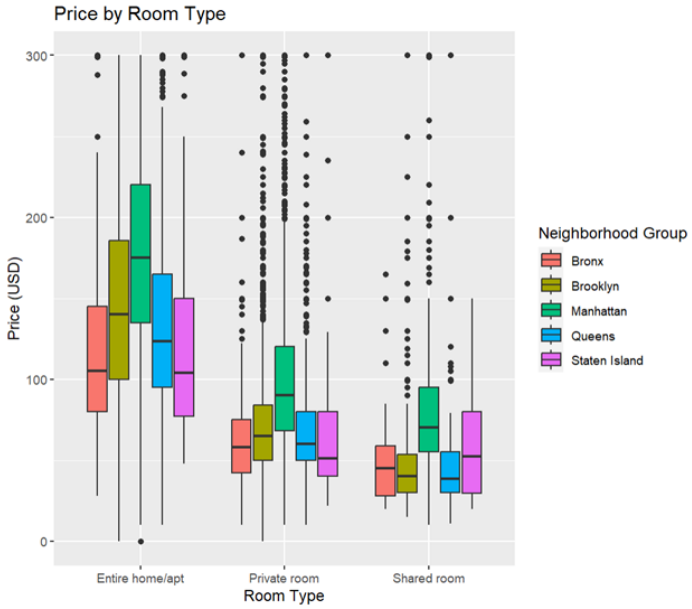

Figure 16: Analysis – Price by Room Type

V. DATA TRANSFORMATION

After cleaning the data, we transformed the data using dummy variables for the categorical variables in the dataset to work with a fully numerical values dataset which resulted in a dataset of 234 variables (illustrated in figure 3) in total due to fact that there are many neighborhoods in New York.

	latitude	longitude	price	minimum_nights	number_of_reviews	reviews_per_month	calculated_host_listings_count	availability_365	neighbourhood_group_Bronx	neighbourhood_group_Brooklyn	...	neighbourhood_Williamsbridge	nei
0	40.64749	-73.97237	149	1	9	0.21	6	365	0	1	...	0	
1	40.75362	-73.98377	225	1	45	0.38	2	355	0	0	...	0	
3	40.68514	-73.95976	89	1	270	4.64	1	194	0	1	...	0	
4	40.79851	-73.94399	80	10	9	0.10	1	0	0	0	...	0	
5	40.74767	-73.97500	200	3	74	0.59	1	129	0	0	...	0	

5 rows × 234 columns

Figure 17: Cleaned Data

VI. PRE-PROCESSING

Still the process of preprocessing the data for analysis we loaded the data into RStudio, and before dividing the data into testing and training we performed an outlier's value check and from around 38000 to around 27000, price was one of the examples illustrated in figure 18 and 19, and the same was done for all the initial purely numerical variables from the beginning.

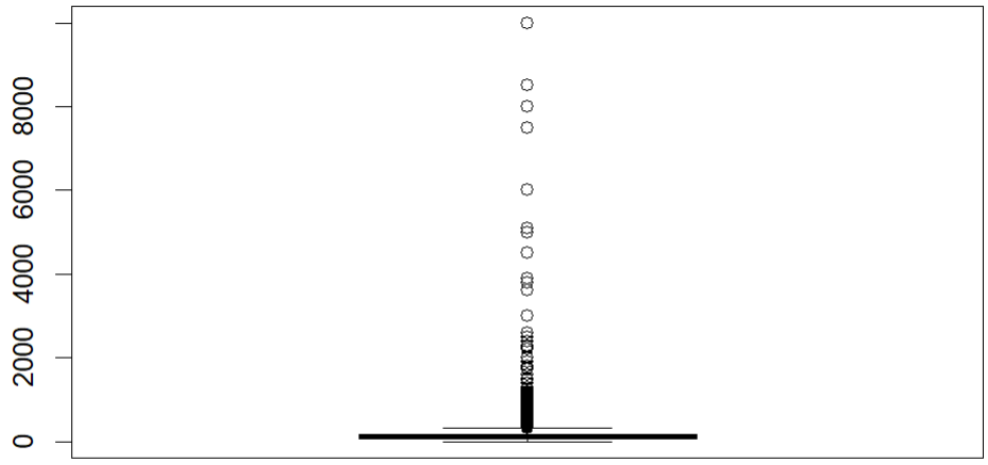

Figure 18: Boxplot for Price before check

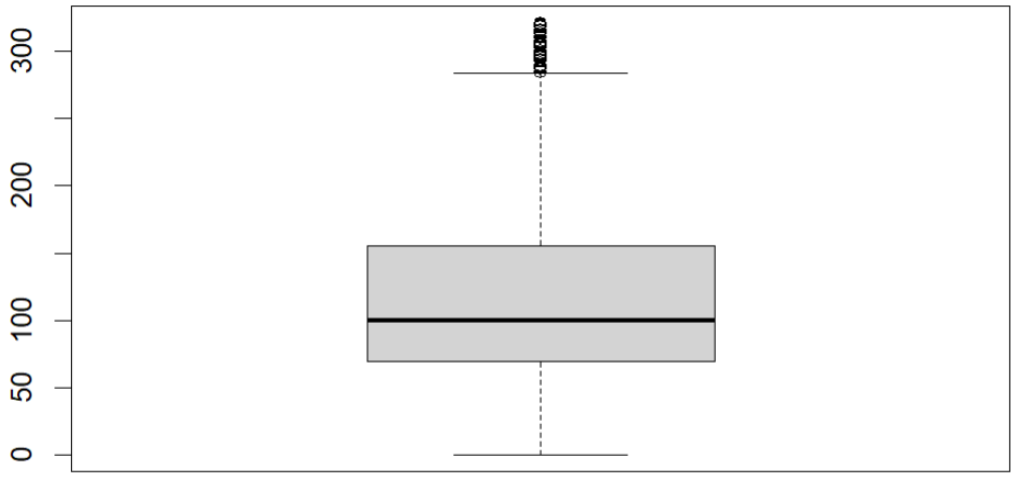

Figure 19: Boxplot for Price after check

After the outliers check we divided the data into training set and testing set to perform the different models. The following division was set as 80% for training and 20% for testing which resulted in 21978 instances for training and 5495 instances for testing.

VII. MODELS AND EVALUATION

The applied models consisted of:

- **Multiple Linear Regression**

A statistical method that models the relationship between multiple independent variables and a dependent variable. We have used multiple linear regression to model the relationship between various factors (such as location, amenities, price, etc.) and the price of a listing in the Airbnb dataset.

- **LASSO**

A regression analysis method that uses regularization to select important features and reduce overfitting in high-dimensional datasets. We have used LASSO to select the most important features among a large set of potential predictors in order to predict the price of a listing in the Airbnb dataset.

- **Partial Least Squares (PLS)**

A multivariate regression method that identifies linear combinations of predictor variables that explain the maximum amount of variance in the dependent variable. We have used PLS to identify a set of linear combinations of predictor variables that explain the maximum amount of variance in the dependent variable (price) in the Airbnb dataset. This has helped us to identify which combinations of factors are most important in determining the price of a listing.

- **Decision Tree (Regression Tree)**

A non-parametric supervised learning method that models a decision-making process by recursively partitioning the data into subsets based on the values of predictor variables. We have used a decision tree to predict the price of a listing based on the values of various features (such as location, amenities, and number of bedrooms) in the Airbnb dataset.

- **Random Forest**

A machine learning algorithm that combines multiple decision trees to reduce overfitting and improve predictive accuracy. Random forest could be used to combine multiple decision trees to create a more accurate prediction of the price of a listing. This could help to reduce overfitting and improve the accuracy of the model.

- **Boosting**

A machine learning technique that combines multiple weak models to create a strong model by iteratively adjusting the weights of misclassified instances. Boosting could be used to iteratively adjust the weights of misclassified listings in order to improve the accuracy of the model. This could help to create a more accurate prediction of the price of a listing.

For Multiple Linear Regression, Random Forest and Boosting, their implementation was made in a default way, especially for Random Forest and Boosting as Cross Validation would make the implementation become computationally expensive. For LASSO, PLS and Decision Tree Cross Validation was implemented.

To evaluate the model's metrics, we used Mean Squared Error (MSE), Coefficient of Determination or Proportion of Variance(R2) and Adjusted R2.

Models Setup

Models	Parameters	Metrics
Multiple Linear Regression	Default	MSE, R2, Adjusted R2
LASSO	k-fold CV, Best Lambda, alpha=1	MSE, R2, Adjusted R2
PLS	k- fold CV, scale=True, ncomp=Optimal	MSE, R2, Adjusted R2
Regression Tree	k-fold CV, best	MSE, R2, Adjusted R2
Random Forest	mtry=p/3, Importance=True	MSE, R2, Adjusted R2
Boosting	Gaussian, n.trees=5000, interaction.depth=4	MSE, R2, Adjusted R2

Figure 20: Models with evalutaion technique used.

RESULTS

All the results for the generated regression models will be associated with dataset from the pre-process for regression analysis.

I. MULTIPLE REGRESSION MODEL

In the figure below, we observe the detailed summary of the multiple regression model generated.

```
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 44.2 on 21755 degrees of freedom
Multiple R-squared:  0.5529,    Adjusted R-squared:  0.5484
F-statistic: 121.2 on 222 and 21755 DF,  p-value: < 2.2e-16
```

Figure 21: Multiple Linear Regression model

The statistical analysis reveals a significant model at F(222, 21775)= 121.2, with a p-value less than 0.05. This indicates that the model is statistically significant and can be used to predict the outcome variable. Additionally, the independent variables explain 55.3% of the variance, suggesting that they are important predictors of the outcome variable. The correlation coefficient (R) between the variables is 0.744, indicating a strong and positive relationship between them. The mean squared error (MSE) is 2009.64, which measures the average squared difference between the predicted and actual values and is used as a measure of the model's accuracy. These findings provide valuable insights into the relationship between the variables and can aid in making informed decisions based on the model's predictions.

II. LASSO REGRESSION MODEL

In the figures below, we observe the detailed LASSO graphs after the cross validation for the best lambda and of the coefficients being forced to 0.

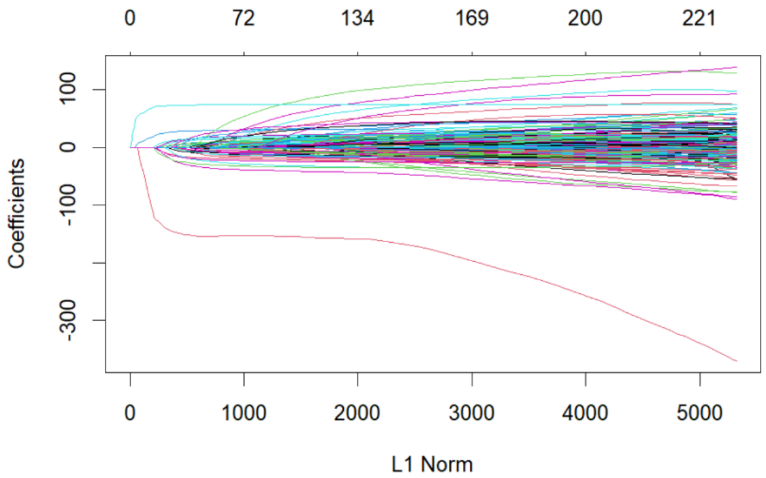

Figure 22.: Lasso

Using Lasso regression and 10-fold cross-validation, the best lambda value was found to be 0.1421658. The model had an R2 of 0.54 and an adjusted R2 of 0.52, indicating that 54% of the variance in the dependent variable can be explained by the independent variables. Additionally, the mean squared error (MSE) was calculated to be 2007.05, indicating that the model has good predictive accuracy. These results suggest that the Lasso regression model is a good fit for the data and can be used to make accurate predictions.

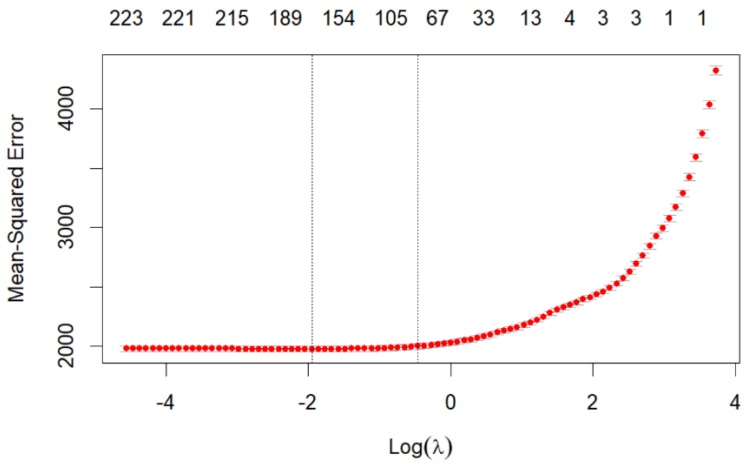

Figure 23.: MSE vs Best Lambda

III. DIMENSION REDUCTION - PARTIAL LEAST SQUARES (PLS)

In the figures below, we observe the detailed PLS summaries for cross validation and for the optimal parameter found after cross validation.

```
library(pls)

#a) PLS with cross-validation to optimize M
training_data_filtered <- training_data[, -nearZeroVar(training_data)]
set.seed(1)
pls.cv = plsr(price ~ ., data = training_data_filtered, scale = TRUE, validation = "CV")
summary(pls.cv)
```

```
Data:   X dimension: 21978 15
    Y dimension: 21978 1
Fit method: kernelpls
Number of components considered: 15

VALIDATION: RMSEP
Cross-validated using 10 random segments.
       (Intercept)  1 comps  2 comps  3 comps  4 comps  5 comps
CV           65.77    49.05     48.3    47.97     47.7    47.43
adjCV        65.77    49.05     48.3    47.97     47.7    47.43
       6 comps  7 comps  8 comps  9 comps  10 comps  11 comps
CV       47.16    46.96    46.93    46.92     46.91     46.91
adjCV    47.16    46.96    46.92    46.92     46.91     46.91
       12 comps  13 comps  14 comps  15 comps
CV        46.91     46.91     46.91     46.91
adjCV     46.91     46.91     46.91     46.91

TRAINING: % variance explained
       1 comps  2 comps  3 comps  4 comps  5 comps  6 comps  7 comps
X        15.88    30.63    40.24    48.28    53.74    57.39    60.81
price    44.41    46.10    46.85    47.47    48.07    48.65    49.08
       8 comps  9 comps  10 comps  11 comps  12 comps  13 comps
X        66.57    73.31      76.2     81.12     86.55     91.06
price    49.16    49.18      49.2     49.20     49.20     49.20
       14 comps  15 comps
X         93.63     100.0
price     49.20      49.2
```

Figure 24: PLS CV Summary

Using 10-fold cross-validation, the optimal number of components for Partial Least Squares (PLS) regression was found to be 9. The model had an R2 of 0.492 and an adjusted R2 of 0.490, indicating that 49.2% of the variance in the dependent variable can be explained by the independent variables. Additionally, the mean squared error (MSE) was calculated to be 2210.67, indicating that the model has good predictive accuracy. These results suggest that the PLS regression model with 9 components is a good fit for the data and can be used to make accurate predictions.

```
# Plot the validation curve and extract the optimal number of components (M)
validationplot(pls.cv, val.type = "MSEP")
```

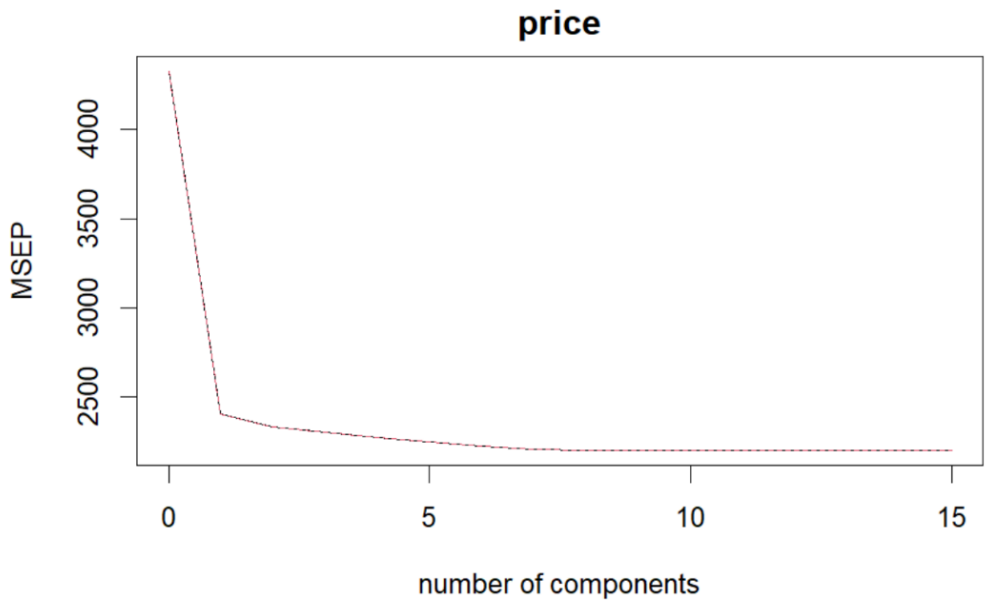

Figure 25.: MSEP vs best component

```
# Refit the model using all the data and the optimal M
pls.fit = plsr(price ~ ., data = training_data_filtered, scale = TRUE, ncomp = optM)
summary(pls.fit)
```

```
Data:   X dimension: 21978 15
    Y dimension: 21978 1
Fit method: kernelpls
Number of components considered: 9
TRAINING: % variance explained
       1 comps  2 comps  3 comps  4 comps  5 comps  6 comps  7 comps
X        15.88    30.63    40.24    48.28    53.74    57.39    60.81
price    44.41    46.10    46.85    47.47    48.07    48.65    49.08
       8 comps  9 comps
X        66.57    73.31
price    49.16    49.18
```

Figure 26: Refit the model for optimal M

IV. DECISION TREE (REGRESSION TREE)

In the figures below, we observe the detailed Regression Tree summary, the tree using the best parameter after cross validation and the graph of actual values and predicted values.

Using 10-fold cross-validation, the Decision Tree model had an R2 of 0.487 and an adjusted R2 of 0.465, indicating that 48.7% of the variance in the dependent variable can be explained by the independent variables. Additionally, the mean squared error (MSE) was calculated to be 2212.73, indicating that the model has reasonable predictive accuracy. These results suggest that the Decision Tree model is a good fit for the data and can be used to make predictions with moderate accuracy.

```
Regression tree:
tree(formula = price ~ ., data = training_data)
Variables actually used in tree construction:
[1] "room_type_Entire.home.apt"     "neighbourhood_group_Manhattan"
[3] "latitude"                      "longitude"
Number of terminal nodes:  6
Residual mean deviance:  2221 = 48790000 / 21970
Distribution of residuals:
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
-175.600  -27.360   -7.357    0.000   19.680  251.600
```

Figure 27: Regression Tree Summary

```
# prune.tree(): function to prune to be used in case we wanted to prune the tree
prune.airbnb=prune.tree(tree.airbnb,best=6)
plot(prune.airbnb)
text(prune.airbnb,pretty=0,cex=0.75)
```

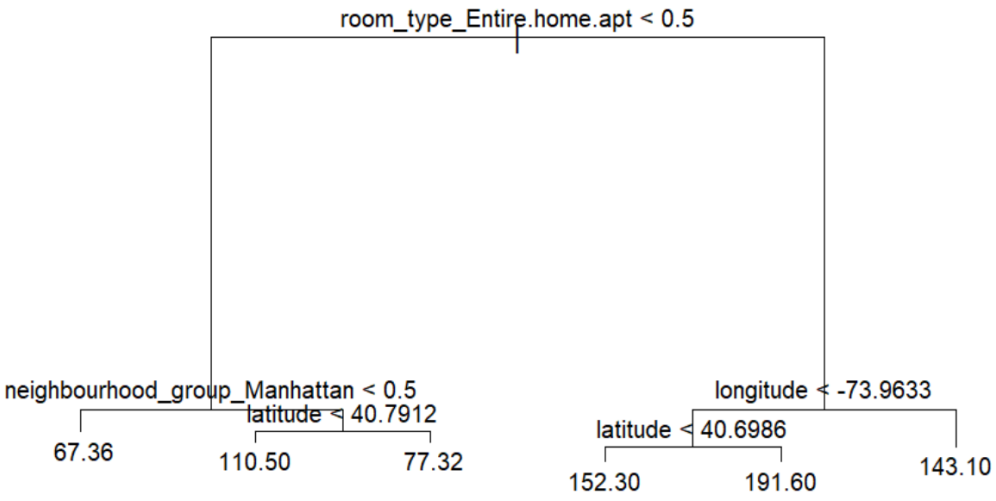

Figure 28: Decision Tree Regression

```
# Predicting based on CV results (i.e., use the unpruned tree)
yhat=predict(tree.airbnb,testing_data)
plot(yhat,Y.test)
abline(0,1)
```

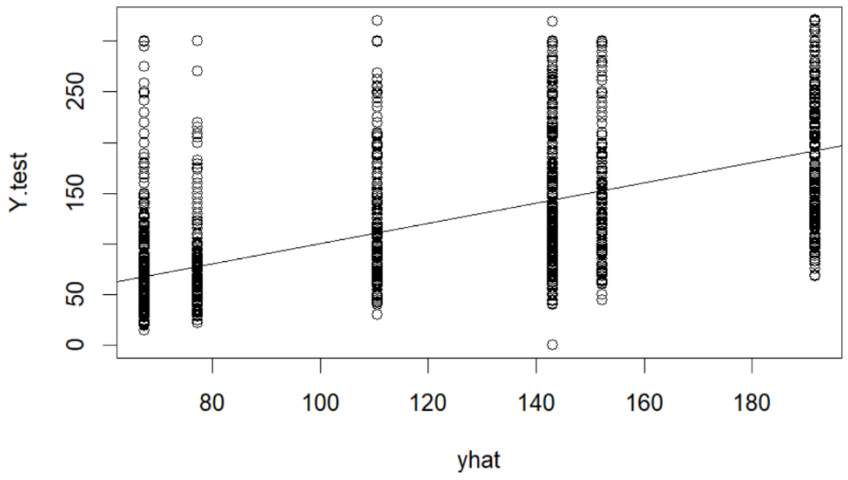

Figure 29: Actual vs Predicted and MSE for Regression Tree

V. RANDOM FOREST (REGRESSION)

In the figure below, we observe the important variables that have the most impact in the MSE of the predicted price.

The Random Forest model had an R2 of 0.575 and an adjusted R2 of 0.556, indicating that 57.5% of the variance in the dependent variable can be explained by the independent variables. Additionally, the mean squared error (MSE) was calculated to be 1838.12, indicating that the model has good predictive accuracy. These results suggest that the Random Forest model is a good fit for the data and can be used to make accurate predictions.

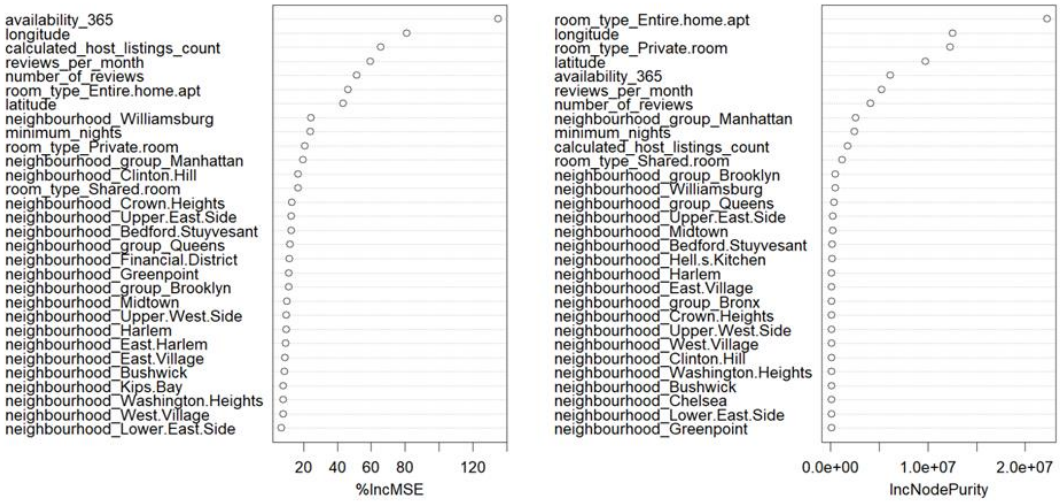

Figure 30: Variable Importance Graph

VI. GRADIENT BOOSTING (REGRESSION)

In the figure below, we observe the important variables and their relative importance on the predicted price.

The Gradient Boosting model had an R2 of 0.547 and an adjusted R2 of 0.527, indicating that 54.7% of the variance in the dependent variable can be explained by the independent variables. Additionally, the mean squared error (MSE) was calculated to be 1957.02, indicating that the model has reasonable predictive accuracy. These results suggest that the Gradient Boosting model is a good fit for the data and can be used to make predictions with moderate accuracy.

	var <chr>	rel.inf <dbl>
room_type_Entire.home.apt	room_type_Entire.home.apt	33.010253400
longitude	longitude	16.652173356
latitude	latitude	13.931319443
availability_365	availability_365	7.607061373
reviews_per_month	reviews_per_month	7.006827173
number_of_reviews	number_of_reviews	4.218330975
minimum_nights	minimum_nights	2.050346805
neighbourhood_group_Manhattan	neighbourhood_group_Manhattan	1.626343030
calculated_host_listings_count	calculated_host_listings_count	1.293361632
room_type_Private.room	room_type_Private.room	0.915764007
neighbourhood_Williamsburg	neighbourhood_Williamsburg	0.486439703
neighbourhood_Midtown	neighbourhood_Midtown	0.483198144
neighbourhood_East.Village	neighbourhood_East.Village	0.430282707
neighbourhood_Arverne	neighbourhood_Arverne	0.401882110
room_type_Shared.room	room_type_Shared.room	0.370164929
neighbourhood_Windsor.Terrace	neighbourhood_Windsor.Terrace	0.349282069
neighbourhood_Brooklyn.Heights	neighbourhood_Brooklyn.Heights	0.293165929
neighbourhood_Theater.District	neighbourhood_Theater.District	0.291713172
neighbourhood_Upper.West.Side	neighbourhood_Upper.West.Side	0.270750892
neighbourhood_SoHo	neighbourhood_SoHo	0.267575317
neighbourhood_Upper.East.Side	neighbourhood_Upper.East.Side	0.264055448
neighbourhood_West.Village	neighbourhood_West.Village	0.254707183
neighbourhood_Two.Bridges	neighbourhood_Two.Bridges	0.246634917
neighbourhood_Lower.East.Side	neighbourhood_Lower.East.Side	0.238518654
neighbourhood_Chinatown	neighbourhood_Chinatown	0.224439905

1-25 of 233 rows Previous 1 2 3 4 5 6 ... 10 Next

Figure 31: Relative Influence of each variable on price.

VII. OVERALL PERFORMANCE COMPARISON OF REGRESSORS

Model	MSE	R2	Adjusted R2
Multiple Linear Regression	2009.64	0.553	0.548
Shrinkage Methods (LASSO)	2007.05	0.535	0.515
Dimension Reduction (PLS)	2210.67	0.492	0.490
Decision Tree	2212.73	0.487	0.465
Random Forest	**1838.12**	**0.575**	**0.556**
Gradient Boosting	1957.02	0.547	0.527

Figure 32: Model Comparison

DISCUSSION

Regression

Looking at our models, we could now answer the research questions that we came up with:

1. How do different factors, such as location, property type, and availability, impact the price of an Airbnb rental in New York?

   According to figure 31, location, property type and availability have high percentages of relative influence when prediction the prices.

2. Can we predict the price of an Airbnb rental in New York based on factors such as location, reviews, and availability? Will the results be significant?
   It was possible to predict the results using the following factors which resulted in a significant model as it is seen in figure 21, $F(222,21755) = 121.2$, $p<0.05$, $R2= 0.553= 55.3\%$ of variance explained by the predictors and $R=0.74$, suggesting a strong and positive relationship between variables.

3. Can we accurately predict the price of an Airbnb listing using a regression model, and which features are most important in determining price?
   According to figure 30, availability, location (longitude and latitude) and reviews per month were among the top three predictors for the prices. Furthermore, the Room type entire home apartment was also an honorable mention when doing the prediction, and in terms of accuracy, random forest has a model that has probably around 60% of proportion of variance meaning that it may have a good chance of being accurate with some prices given the following application.

CONCLUSION

With all the generated results it is possible to carefully conduct a comparison between the different regression models applied. **Random forest** was best regressor at this application since it produced the least overall mean squared error, in comparison to the others as it is showed in Figure 32.

Our analysis of the New York Airbnb dataset [1] revealed several important insights and patterns that can help to improve our understanding of the factors that impact the rental prices of Airbnb properties in New York. Some of the key findings include:

- The most important factors that impact the rental price of an Airbnb property in New York are location, property type, and amenities.
- The most expensive neighborhoods for Airbnb rentals in New York are Manhattan, Brooklyn, and Queens.
- The most popular neighborhoods for Airbnb rentals in New York are Williamsburg, Bedford-Stuyvesant, and Harlem.
- Rentals listed as "Entire home/apt" are generally more expensive than those listed as "Private room" or "Shared room".
- The availability of the rental may have a slight impact on the rental price, but other factors are likely to have a stronger impact.

And, some of the answers to relevant Airbnb queries include:

1. **What is the status of Airbnb market in NYC?**
    - **Manhattan** and **Brooklyn** account for over **85%** of the Airbnb listings in NYC, with each having more than 20,000 listings.
    - **Entire home/apartment** listings make up **52%** of the total listings, while private rooms make up 45%, and shared rooms only account for 2%. Brooklyn has the most private rooms, while Manhattan has the most entire homes/apartments.
    - The **median price** for a listing in Manhattan is approximately **$150**, significantly higher than other neighborhood groups where the median price is less than $100.
    - The **top 10 most expensive neighborhoods** in **Manhattan** have prices ranging from **$200 to $300**, with Tribeca being the most expensive. On the other hand, the top 10 least expensive neighborhoods in Manhattan range from $70 to $130, with Washington Heights being the cheapest.
    - **Availability** is generally **lower** in **Brooklyn** and **Manhattan** compared to other locations.

2. **Which factors affect the price?**
    - **Location** plays a significant role in determining the price of an Airbnb. For instance, Manhattan has significantly higher prices compared to other locations.
    - The **type of room** also affects the price, with entire home/apartments being more expensive than private and shared rooms. While private rooms are slightly more expensive than shared rooms.
    - There was no clear trend observed in the relationship between the **minimum nights** and **availability of listings** with their prices.

3. **Can we predict the price?**
    - We used five models (Multiple Linear Regression, Lasso Regression, Partial Least Squares (PLS), Decision Tree (Regression Tree), Random Forest and Gradient Boosting to predict the price, and **Random Forest** outperformed the other models.
    - The model suggests that the most crucial features for predicting the price are the **room type**, **location** and **availability**.
    - **Room types** which included the **entire apartment** also had a high relative importance for the prediction.
    - Model and Results are significant.

FUTURE WORK

There are several avenues for future work to further improve the analysis of Airbnb listing prices. One potential area of focus is the identification of methods to detect if listing prices are artificially inflated. Additionally, scaling up the project to include multiple regions across the United States and various types of Airbnb listings could provide more comprehensive insights into pricing trends. Another promising approach is the utilization of natural language processing techniques to analyze reviews and determine their sentiment for improved analysis. Time series analysis could also be incorporated to gain a more accurate understanding of how Airbnb prices change over time. Finally, exploring the use of advanced neural networks for improved accuracy, even if it may require more computational resources, could provide valuable insights. A recent study has also suggested including image quality as a feature, which could improve revenue, using deep learning and supervised learning analyses on an Airbnb panel dataset. Overall, these future directions have the potential to significantly enhance the analysis of Airbnb pricing and ultimately benefit both hosts and guests.

References

[1] Dgomonv, "Kaggle - New York City Airbnb Open Data," 2019. [Online]. Available: https://www.kaggle.com/datasets/dgomonov/new-york-city-airbnb-open-data. [Accessed January 2023].

[2] L. Nikolenko, P. R. Kalehbasti and H. Rezaei, "Airbnb price prediction using machine learning and sentiment analysis.," 2019. [Online]. Available: https://arxiv.org/abs/1907.12665.

[3] Y. Li, Q. Pan, T. Yang and L. Guo, "Reasonable price recommendation on Airbnb using Multi-Scale clustering," *35th Chinese Control Conference (CCC), Chengdu, China, 2016,* no. 10.1109/ChiCC.2016.7554467., pp. 7038-7041, 27 July 2016.

[4] Y. Luo, X. Zhou and Y. Zhou, "Predicting Airbnb Listing Price Across Different Cities - Stanford," *International Cross-Domain Conference for Machine Learning and Knowledge Extraction,* Vols. Springer, Cham, 2021, pp. 173-184, 14 December 2019.