New York City Airbnb Price Prediction Using Data Analytics

Group Members:

- Lipika Baniya
- Prashant Kuikel
- Soniya Shrestha
- Karthik Gollakore

***Abstract*— This project is to assist Airbnb hosts in NYC in optimizing their pricing strategies by leveraging data analytics. Using features such as neighborhood, room type, and booking trends, a predictive pricing model is developed to help new hosts set competitive nightly rates.**

***Keywords*— Airbnb, Predictive Pricing, NYC, Data Analytics.**
***Link* – [NYC Airbnb Dataset]**

I. INTRODUCTION

Since 2008, guests and hosts have used Airbnb to expand on traveling possibilities and present a more unique, personalized way of experiencing the world. Today, Airbnb became a one-of-a-kind service that is used and recognized by the whole world. Data analysis on millions of listings provided through Airbnb is a crucial factor for the company. These millions of listings generate a lot of data – data that can be used for security, business decisions, understanding of customers' and providers' (hosts) behavior and performance on the platform.

II. DATA PREPARATION AND PREPROCESSING

This research was conducted using a dataset from the New York City Airbnb Open Data on Kaggle. The collection includes 16 columns and 48,896 rows of information on location, home type, and reviews, among other characteristics of Airbnb listings. These qualities are essential for analysis and predictive modeling. The majority of raw data, however, contains a number of problems that must be fixed before any analysis can begin, including outliers, missing values, and irrelevant features.

No.	Attribute - Description	Distinct Values	Example	Data Type	Purpose/Usage
1	ID	Unique Identifier	2539	Numeric	Serves as a unique identifier for each listing
2	name	Varies by Listing	Clean and Quiet apt home by the park	Text	Indicates the name of the listing, useful for branding and recognition
3	host_id	Unique Identifier for host	2787	Numeric	Identifies the host of the listing, useful for tracking host activity
4	host_name	Varies by Host	John	Text	The name of the host, provides a personal touch and accountability
5	neighbourhood_group	Varies (like manhattan and brooklyn)	Brooklyn	Categorical	Indicates the broader area for categorizing listing geographically
6	neighborhood	Varies (like Kensington , Chelsea)	Kensington	Text	Specifies the exact neighborhood, and helps users find listings in desired areas)
7	latitude	Geographic Coordinates	40.647	Numeric	Provides geographical location for mapping and search functions
8	longitude	Geographic Coordinates	-73.972	Numeric	Provides the geographical location for mapping and search functions
9	room_type	Private room, Entire home/apt	Entire home/apt, Private home	Categorical	Indicates type of accommodation, helps users filter options based on their needs
10	price	Numeric Values	149	Numeric	Shows price per night, essential for budgeting

					and comparison of listings
11	minimum_nights	Whole Number	1	Numeric	Indicates the minimum stay requirement, affects booking decisions
12	number_of_reviews	Whole Number	9	Numeric	Reflects the popularity and feedback of the listing, helps assess quality
13	last_review	Date	10/19/2018	Date	Indicates the date of the most recent review, useful for assessing current guest experiences
14	reviews_per_month	Numeric	0.21	Numeric	Indicates average reviews received monthly, useful for gauging engagement
15	calculated_host_listings_count	Whole Number	6	Numeric	Reflects how many listings the host has, useful for understanding host's experience
16	availability_365	Whole Number (0 - 365)	365	Numeric	Shows the number of days the listing is available for booking, important for potential guests

Table 1: Features selected from the dataset

A. *Data Cleaning and Preprocessing*

In the exploration of the dataset, the issue discovered was that the New York City Airbnb consists of 16 columns. Missing attributes is a negative aspect to our project as many machine learning algorithms do not process missing values. In the case where they do, however, missing values may cause complications in feature extraction and pattern recognition which will then translate into false predictions. So, we replaced the null values in review per month with 0.

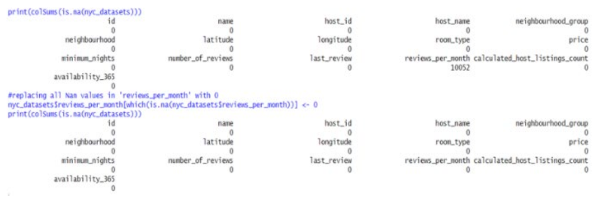

Figure 1: Handling Missing Values in Airbnb Dataset

We also checked some duplicate in the dataset Airbnb with the following figure:

```
> #Counting how many duplicate rows exist
> duplicate_count <- sum(duplicated(nyc_datasets))
> print(duplicate_count)
[1] 0
```

Figure 2: Checking duplicate

While handling the duplicate values in our Airbnb datasets, we did not find any duplicate rows as shown in the above output.

We also handled the noisy data but after analysis with sort of codes we do not find any of the noisy data in our dataset.

Correcting data format issues

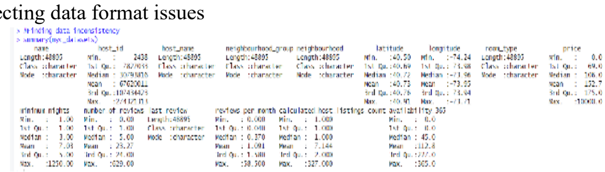

Figure 3: Identifying formatting

While checking data format issues in R, we did not find any inconsistencies.

For data merging, we had extracted one single file from Kaggle, so we did not need to process the merging of the data.We also decided to drop a columns that is not significant to our analysis i.e ID

B. *Data Exploration*

Once the dataset has been preprocessed, we visualize and analyze the data to decide on the train and test data split.

We identified outliers and noisy data based on the IQR and box plot. We noticed some weird data in price and minimum nights, so we created a box plot to view the data.

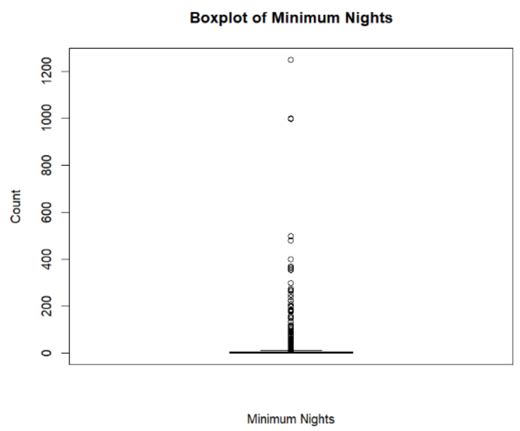

Figure 4: Boxplot of Minimum night with outliers and noisy data

From Figure 4 we can observe many outliers to make changes so now we removed all the outliers and noisy data that are above 11 minimum nights from our Airbnb dataset.

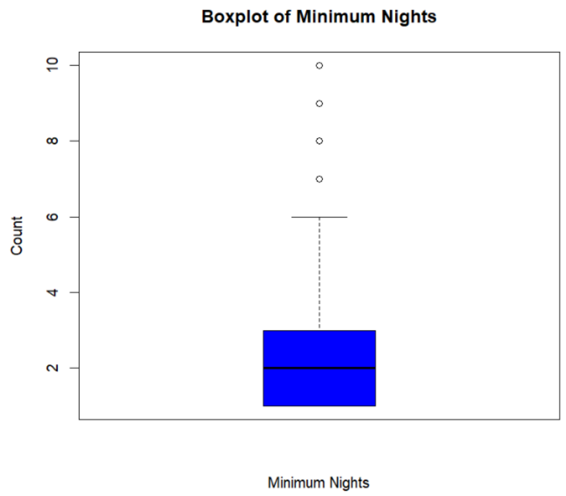

Figure 5: Minimum Night without outliers and noisy data

From Figure 5, many observations can be made. These are the observations from the plot.

i. The box plot suggests a slight skewness, meaning some Airbnb types might be slightly shorter or longer than the central range.
ii. Budget conscious travelers will probably find a temporary stay with a minimum stay of 2-4 nights as per the IQR and values of boxplot.

Similarly, we removed any price less than $80 and more than $5000 to create data consistency.

III. EXPLORATORY ANALYSIS

A. *Summary statistics and Data*

Variable	Mean	Median	Standard Deviation	Min	Max	Mode
Price-Target Variable	195.3295	150	185.866	81	4500	100

B. *Descriptive Statistics*

To understand the pricing and what factors affect the pricing we have addressed the following research questions along with an explanation of why we decided to address these questions

- What is the most common type of room among the guests? To help Airbnb hosts understand guest preferences and identify the room type with the highest demand.
- Who is the top performing host? To analyze host performance metrics and highlight the most successful hosts as examples for new or aspiring hosts.
- What is the distribution of prices across different room types? To assist new Airbnb hosts to understand pricing patterns for various room types.
- What is the distribution of prices in each neighborhood group? To help Airbnb hosts understand pricing trends within each neighborhood group.
- What is the average number of listings per host? This will help new Airbnb hosts understand the level of competition and identify whether multi-listing hosts dominate the platform or if single-listing hosts are more common in their area.
- What are the revenue trends across different neighborhood groups? - This will assist new Airbnb hosts in identifying which neighborhood groups generate the highest revenues, enabling them to select competitive pricing strategies and target the most lucrative areas.

Airbnb offers three types of listings,

- Entire home/apartment
- Private Room
- Shared Room

Room Type

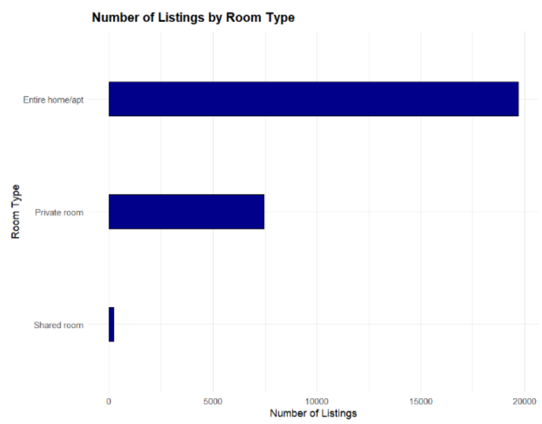

Figure 6: Number of Listings by Room Type

Horizontal Bar Plot for Room Type:

i. Entire home/apartment has the largest share of the listing types in New York City. This listing type is preferred by families, groups, or individuals seeking to stay in a completely private setting.
ii. Private rooms are the second most common listing type, appealing to solo travelers.
iii. Shared rooms represent the smallest share of listings. They are very cheap but offer minimal privacy, appealing to cost-conscious travelers.
iv. Based on this, we can conclude that New York City attracts tourists who would pay more to stay in private and independently owned accommodation.

Top Hosts

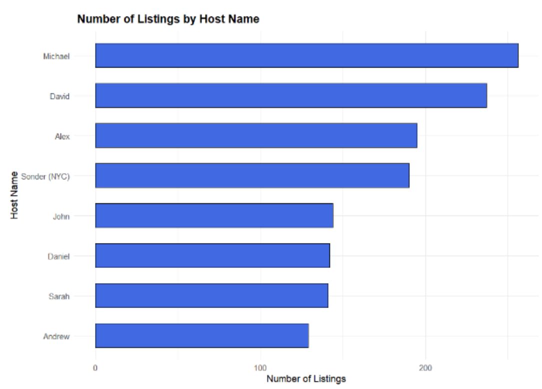

Figure 7: Number of Listings by Host Name

Distribution of Airbnb listings among hosts in New York City:

i. Michael leads as the top host, managing 256 properties, indicating a business scale operation.

ii. David follows closely with 237 listings, further reinforcing the presence of large-scale operators.

iii. Alex and Sonder manage 195 and 190 listings, respectively, highlighting significant contributions to the Airbnb market.

iv. John, Daniel, and Sarah have 144,142, and 141 Approx. listings each, suggesting moderately active hosting.

v. Andrew has the fewest listings, with 129 properties, indicating a likely individual host with a small-scale operation.

vi. The variation in host scale, from large operators like Michael to individual like Andrew, illustrates the diversity of the Airbnb market.

vii. The range gives travelers a wide selection of accommodation options based on their preferences.

Price Distribution Across Different Types

	room_type	Avg_Price	Median_Price	Min_Price	Max_Price	Total_Listings
	<chr>	<dbl>	<dbl>	<int>	<int>	<int>
1	Entire home/apt	215.	170	81	4500	19685
2	Private room	142.	105	81	3210	7446
3	Shared room	183.	116.	85	1800	202

Figure 8: Average Price by Room Type

Bar chart illustrating the price distribution across different Airbnb listing types:

i. Price range in Entire home/apartment has the highest price range because of larger space and amenities.

ii. The median price of Entire Home/Apartment is the highest at $170, followed by Shared Rooms at $116, and Private Rooms at $105.

iii. The Private Rooms are the most affordable option with a median price of $105, which is quite strange because in general we expect Private Rooms to be more expensive than Shared Room.

iv. Entire Home/Apartment listings, with 19685 listings, are the most popular accommodation type in New York City, while Shared Rooms remain the least popular with only 202 listings.

v. The minimum price is $81, and the maximum price is $4500, reflecting the range of accommodation options in New York City.

Price Distribution across different Neighborhood Groups

	neighbourhood_group	Avg_Price	Median_Price	Min_Price	Max_Price	Total_Listings
	<chr>	<dbl>	<dbl>	<int>	<int>	<int>
1	Bronx	159.	120	81	2500	345
2	Brooklyn	173.	140	81	4500	10230
3	Manhattan	219.	169	81	4160	14346
4	Queens	155.	122.	81	2600	2262
5	Staten Island	160.	120	83	1000	150

Figure 9: Average price by Neighborhood Group

The above visualization depicts the Airbnb listings into major neighborhood groups in New York City.

i. The metrics such as average price, median price, total listings, and price range including maximum and minimum price are used to analyze the pricing trends for each neighborhood group.

ii. Manhattan has the highest average and median price with the highest number of listing due its central location and high demand.

iii. Staten Island is more affordable than Manhattan and Brooklyn, but the demand is very limited. It may be due to the lowest proportion of quieter travelers and less preferred to stay in sub-urban style.

iv. Queens is the third most popular destination for budget friendly travelers which constitutes 2262 listings but more reasonable pricing than Staten Island.

v. Travelers with higher budgets likely prefer Manhattan or Brooklyn for proximity to attractions.

Listings per Host

	host_type	count
	<chr>	<int>
1	Multi-Host	6703
2	Single-Host	20630

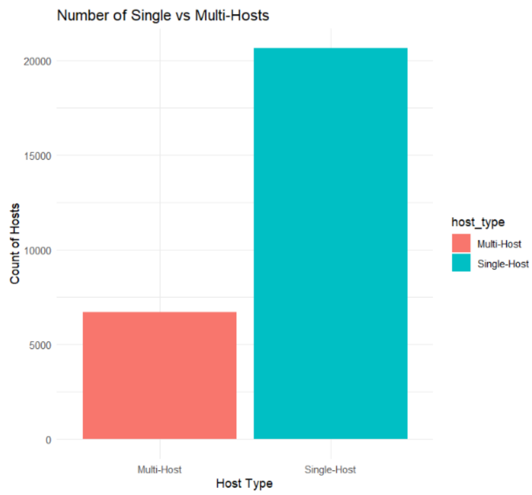

Figure 10: Listing per host

Hosts with multiple listings are more likely to be running a business and are unlikely to be living on the property.

The above visualization shows the number of single versus multi hosts for Airbnb listings by comparing the count of hosts:

i. The single hosts represent the majority of listings, with more than 20,000 listings. These are the hosts who manage a single property and a significant proportion of Airbnb's host base.
ii. Multi hosts comprise a smaller group of hosts.
iii. Managing multiple properties is less popular on Airbnb.

Revenue Trends Across Different Neighborhood Groups

	neighbourhood_group	total_revenue
	<chr>	<int>
1	Bronx	122208
2	Brooklyn	5100341
3	Manhattan	8680197
4	Queens	866500
5	Staten Island	65614

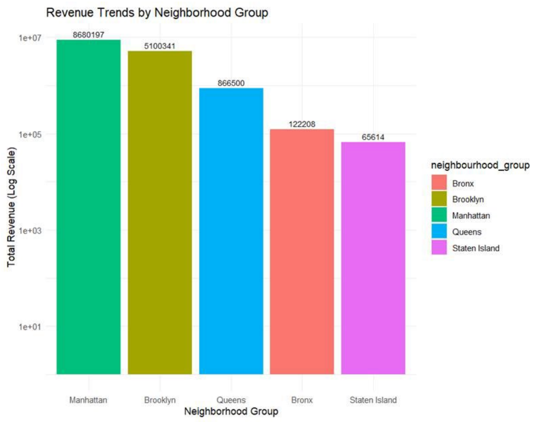

Figure 11: Revenue Trends by Neighborhood Group

The Chart displays the revenue trends in the neighborhood group for Airbnb listings in New York City using a log scale.

i. Manhattan has the highest total revenue among all neighborhood groups, being a prime tourist destination with fuller facilities and luxury accommodations.
ii. Brooklyn stands at the second highest total revenue being a prime tourist destination but cheaper than Manhattan.
iii. Queens generated the third highest total revenue appeals to budget conscious customers. Brooklyn stands at the second highest total revenue being a prime tourist destination but cheaper than Manhattan.
iv. The Bronx generates the second lowest total revenue 5100341, indicating significantly less demand compared to other groups and making it less popular.
v. Staten Island has the total lowest revenue in the neighborhood group attributed to its suburban style and limited attractions for tourists.

Pricing monthly trends

	month	avg_price
	<ord>	<dbl>
1	Jan	181.
2	Feb	182.
3	Mar	183.
4	Apr	188.
5	May	186.
6	Jun	184.
7	Jul	183.
8	Aug	179.
9	Sep	180.
10	Oct	184.
11	Nov	185.
12	Dec	207.

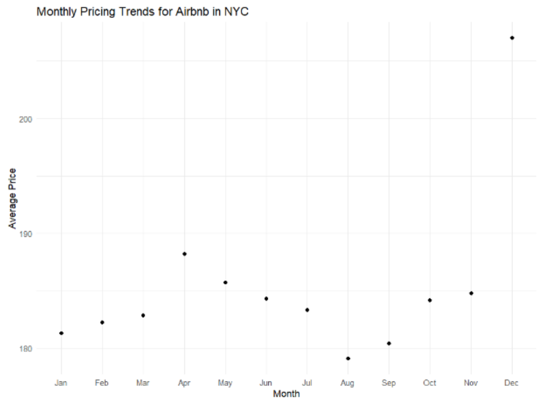

Figure 12: Pricing Monthly trends

The above scatter plot presents data related to monthly pricing trends for Airbnb in NYC. It shows clear seasonal variations in the average monthly price.

i. Prices are highest in December at $207, which seems due to the holiday season.
ii. In contrast, the lowest monthly price fall in August at $179 may be due to hot weather in NYC.
iii. The average price remains relatively stable in other months, ranging from $180 to $195.
iv. Hosts can adjust their pricing to maximize profits in peak price seasons like December whereas make discount strategies during low-demand months to attract more bookings.

Revenue Potential vs Price for Airbnb in NYC

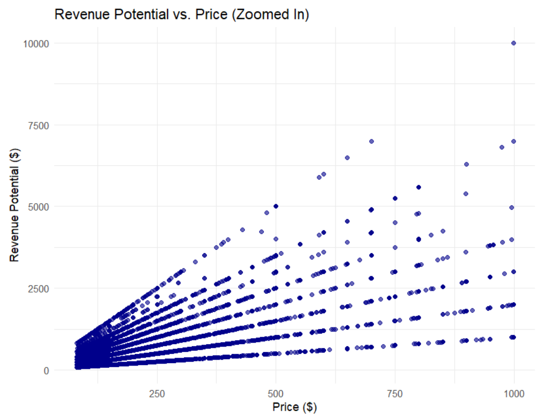

Figure 13: Revenue Potential vs Price

The Chart displays the Revenue Potential vs Price for Airbnb in New York City

i. The scatterplot shows a positive relationship between price and revenue potential. As the price of a listing increases, its revenue potential also increases, though the spread becomes wider at higher price points.
ii. It implies that fewer expensive prices have higher revenue assuming minimum nights are met.
iii. The major listings are concentrated in the lower price range below $250 that suggests higher Airbnb properties in NYC are operated by mid-range budget customers.

iv. In some cases, a smaller number of properties exceeds $500 in price and contribute significantly to revenue.

Assumptions:

i. Empty rows for column last review mean that the Airbnb has never been booked before

ii. We are taking the last_review to find price trends during the holiday season, we are expecting the customer to write the review within 11 days of their stay.

iii. To improve our dataset quality and ensure accurate analysis, we removed outliers by excluding listings with nightly rates below $80 and above $5000. As, these extreme values can distort statistical measures and hinder meaningful insights.

IV. METHODOLOGY

A. *Proposed Work*

This section details the specific algorithms and machine learning models that were trained on the dataset to predict the pricing trends in NYC.

i. Multiple Linear Regression

Multiple linear regression is a linear regression with two or more independent variables, also called the explanatory variables, and one dependent variable, also called the response variable. Multiple linear regression helps make predictions and understand relationships between one or more independent variables.

Equation

$$y = b_0 + b_1*x_1$$

ii. Tools Used

For the implementation of the model, it was coded in R Studio. The data exploration was done using R.

iii. Research Question

How different variables/factors (room type, neighborhood etc) impact the pricing in NYC Airbnb.

V. RESULTS AND FINDINGS

A. *Model Performance Measure*

Model 1: Target Value – Price

Independent Variable – Room Type + Minimum nights + Availability 365 +Neighborhood Group + Number of reviews + Listing type

```
Call:
lm(formula = price ~ room_type + minimum_nights + availability_365 +
    neighbourhood_group + number_of_reviews + listing_type, data = nyc_datasets)

Residuals:
   Min     1Q Median     3Q    Max
-235.2  -73.1  -32.6   20.2 4210.3

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                   161.369156  10.070106  16.025  < 2e-16 ***
room_typePrivate room         -88.329175   2.509621 -35.196  < 2e-16 ***
room_typeShared room          -58.261245  12.559972  -4.639 3.52e-06 ***
minimum_nights                 -3.433131   0.611837  -5.611 2.03e-08 ***
availability_365                0.267548   0.009238  28.962  < 2e-16 ***
neighbourhood_groupBrooklyn    36.687492   9.727145   3.772 0.000163 ***
neighbourhood_groupManhattan   89.161823   9.684636   9.207  < 2e-16 ***
neighbourhood_groupQueens       7.741322  10.241239   0.756 0.449717
neighbourhood_groupStaten Island -4.315840 17.321677 -0.249 0.803240
number_of_reviews              -0.507604   0.024173 -20.999  < 2e-16 ***
listing_typeSingle-Listing     -9.589629   2.661802  -3.603 0.000316 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 177.1 on 27322 degrees of freedom
Multiple R-squared:  0.09256,   Adjusted R-squared:  0.09222
F-statistic: 278.7 on 10 and 27322 DF,  p-value: < 2.2e-16
```

Figure 14: Model 1

From this analysis, the model shows that room type, availability of listings, and location in Manhattan or Brooklyn are the most significant predictors of price. Conversely, Queens and Staten Island add little value in explaining the variance in price.

i. The intercept is 161.37, that represents the baseline price when all predictors are at zero.

ii. Private room and Shared room are most significant variable with $p < 0.001$ in the room type where price are reduced by $88.39 and 58.26 respectively compared to the entire homes/apartment. Private rooms are less expensive accommodation.

iii. Manhattan and Brooklyn neighborhoods group have strong impact on price compared to the Bronx and increased the price by $89.16 and $36.86. They are highly significant variable with $p < 0.001$.

iv. For every additional day a property is available, the price will increase by $0.18 that have moderate positive impact.

v. Single listing type have negative impact that reduce the price by $9.59, implying single listings are cheaper compared to other listing types.

vi. Minimum nights and number of reviews have a slight negative effect on price. It implies that longer minimum stays will lower the price by $1.63 and listing with more review reduce the price by $0.32 with high significant.

vii. Neighborhood groups such as Queens and Staten Island have minor positive effect price that can raise the price by $8.73 and $13.26 but not statistically significant.

viii. The residual standard error predicts the difference in predicted price from actual price that is about $177.17, which is quite high.

ix. Model 1 explains the 9.26% of the variance in the price, with the residual error of 177.09.

x. The F-statistic 278.7 and p-value < 0.001suggests that the overall model is statistically significant and independent factors together influence the price.

Based on the Model 1; we picked the most significant 3 variables

Model 2: Target Value – Price

Independent variable: Neighborhood Group + Room Type + Availability 365

```
Call:
lm(formula = price ~ availability_365 + neighbourhood_group +
    room_type, data = nyc_datasets)

Residuals:
   Min     1Q Median     3Q    Max
-215.1  -73.2  -35.4   19.3 4246.1

Coefficients:
                                   Estimate Std. Error t value Pr(>|t|)
(Intercept)                      138.107093   9.753956  14.159  < 2e-16 ***
availability_365                   0.233558   0.008737  26.732  < 2e-16 ***
neighbourhood_groupBrooklyn       30.501566   9.801504   3.112 0.001861 **
neighbourhood_groupManhattan      85.142063   9.761455   8.722  < 2e-16 ***
neighbourhood_groupQueens          6.962300  10.326895   0.674 0.500196
neighbourhood_groupStaten Island  -3.424676  17.466445  -0.196 0.844555
room_typePrivate room            -83.695945   2.448556 -34.182  < 2e-16 ***
room_typeShared room             -45.166923  12.638229  -3.574 0.000352 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 178.6 on 27325 degrees of freedom
Multiple R-squared:  0.07711,   Adjusted R-squared:  0.07688
F-statistic: 326.2 on 7 and 27325 DF,  p-value: < 2.2e-16
```

Figure 15: Model 2

i. The intercept is 138.11 that represents the baseline price when all predictors (room type, neighborhood group, and availability) are set to zero.

ii. For each additional day of availability in a year, the price increases by $0.23 that implies a very small positive but highly significant effect.

iii. Neighborhood groups such as Brooklyn and Manhattan have strong positive effects on price that increase the price by $30.50 and $85.14 representing listing in this group are more expensive. They are highly significant with a p-value less than 0.001.

iv. Private room and Shared room have also a large negative impact on price that reduce the rooms by $83.69 and $45.17 compared to entire home/apartment with highly significant.

v. Queens have very small positive effects that increase price by \$6.96 whereas Staten Island reduces the price by \$3.42 but not statistically significant.
vi. Model 2 explains 7.71% of the variance in the price, with the residual error of 178.58.
vii. In conclusion, the whole model is statistically significant, overall p-value is less than 0.001, meaning that neighborhood groups and room type collectively influence prices.

B. *Brief Summary of Results*

Comparison of Regression Models

Basis of Comparison: The two regression models were compared based on the following criteria:

i. Explained Variance (R2): Measures the proportion of variability in the dependent variable (price) explained by the model's predictors.
ii. Residual Standard Error (RSE): Indicates the average deviation of the actual prices from the predicted prices.
iii. Model Simplicity: Evaluates the complexity of the model based on the number of predictors included.
iv. Significance of Predictors: Assesses which predictors significantly contribute to the model.

Comparison Results

Model 1:

i. R2=0.09256, Adjusted R2=0.09222
ii. Residual Standard Error = 177.1
iii. Includes predictors such as room_type, minimum_nights, availability_365, neighbourhood_group, number_of_reviews, and listing_type.

Explains slightly more variance in price compared to Model 2 but is more complex.

Model 2:

i. R2=0.07711, Adjusted R2=0.07688
ii. Residual Standard Error = 178.6
iii. Includes only room_type, availability_365, and neighbourhood_group as predictors.

Simpler model but slightly less effective in explaining price variability.

C. *Recommendation*

Based on the comparison:

i. Model 1 is better suited if the goal is to maximize explanatory power and understand the impact of additional variables like number_of_reviews and listing_type.
ii. Model 2 is preferred if simplicity and ease of interpretation are prioritized.

Thus, Model 1 is recommended for this project because it explains more variance and provides additional insights into the factors influencing Airbnb pricing in New York City.

VI. CONCLUSION AND RECOMMENDATION

In summary, our goal was to create a Predictive Price Modelling tool where enter all the relevant factors such as location of the listing, listing properties, available amenities etc. and the Machine Learning Model will recommend the Price for the listing.

This machine learning model will help a new host can predict the optimal nightly prices based on factors like neighborhood, room type, and demand. This analysis offers insights into the new hosts on competitive pricing strategies and provides guidelines for how these factors can influence price settings.

Hosts should also focus on optimizing amenities, prioritizing those that significantly impact pricing, to increase their listing's appeal. Regularly updating prices to reflect market conditions and competitor rates is essential for maintaining competitiveness. Additionally, integrating this tool into Airbnb's host dashboard could provide a seamless and accessible experience, benefiting a wider audience of hosts.

REFERENCES

Inside Airbnb. (n.d.). *About Inside Airbnb*. Retrieved from https://insideairbnb.com/

Karkala, D. (n.d.). *Airbnb price modeling: A machine learning case study*. Retrieved from https://www.deepakkarkala.com/docs/articles/machine_learning/airbnb_price_modeling/about/index.html

DataCamp. (n.d.). *Simple linear regression tutorial*. Retrieved from https://www.datacamp.com/tutorial/simple-linear-regression

Inside Airbnb. (n.d.). *New York City Airbnb listings and data*. Retrieved from https://insideairbnb.com/new-york-city/

Gomonov, D. (n.d.). *New York City Airbnb Open Data* [Dataset]. Kaggle. https://www.kaggle.com/datasets/dgomonov/new-york-city-airbnb-open-data