Module 4: Final Project

By

Smit Pareshbhai Ranpariya, Neha Ajinkya Nagurkar & Vishvesh Gaurangbhai Vyas

Project: Airbnb

College of Professional Studies: Northeastern University

Prof. Hootan Kamran

Subject: ALY6110: Big Data and Data Management

Date: June 14, 2024

Introduction

The rise of platforms like Airbnb has revolutionized the hospitality industry, enabling property owners to capitalize on short-term rentals in popular destinations like New York City. This report delves into a comprehensive analysis of Airbnb listings in NYC, utilizing a dataset that encompasses crucial details such as host demographics, property characteristics, pricing, and availability metrics. By examining these factors, the aim is to uncover actionable insights that can inform strategic decisions for potential hosts and investors in the dynamic short-term rental market.

New York City stands as a global epicenter of tourism, business, and culture, attracting millions of visitors annually. The diversity of its neighborhoods—from the bustling streets of Manhattan to the eclectic vibes of Brooklyn—offers a rich tapestry of accommodation options through Airbnb. Understanding the nuances of these neighborhoods and their impact on rental dynamics is paramount for stakeholders looking to enter or optimize their presence in this competitive market.

This analysis begins with descriptive statistics to establish a foundational understanding of the dataset, providing insights into average prices, minimum stay requirements, and distribution of property types across different neighborhoods. Exploratory data analysis (EDA) will then deepen our insights, utilizing visualizations like histograms, box plots, and geographic maps to uncover patterns in pricing trends, popularity by neighborhood, and seasonal availability variations.

Key objectives include identifying which neighborhoods and property types command the highest prices, understanding the influence of host characteristics on listing popularity and occupancy rates, and exploring potential correlations between pricing strategies and the number of reviews or availability throughout the year.

In conclusion, this report aims to equip current and prospective Airbnb hosts with actionable insights to optimize their listings and maximize profitability in New York City's dynamic short-term rental market. By leveraging data-driven analysis, stakeholders can make informed decisions that enhance guest experiences while optimizing financial returns in this ever-evolving sector of the sharing economy.

Analysis:

For the analysis of the Airbnb dataset from New York City, I employed R programming language, leveraging libraries such as ggplot2 for visualizations and dplyr for data manipulation. The process began with importing and cleaning the dataset to ensure data integrity. Irrelevant columns were removed, focusing on variables like price, neighborhood, room type, and host details essential for the analysis.

Descriptive statistics provided an initial understanding of the dataset's numerical features, including measures like mean, median, and standard deviation for key variables such as price and minimum nights. Exploratory Data Analysis (EDA) followed, using ggplot2 to generate visual insights such as histograms and scatter plots. These visualizations highlighted trends such as price distributions across neighborhoods and the popularity of different room types.

Key insights emerged from the analysis, confirming initial observations from basic analysis: Manhattan exhibited higher average prices compared to Brooklyn, entire home/apartments were preferred over private rooms, and superhosts tended to have higher occupancy rates. Advanced analysis delved into seasonal pricing variations and host influence on listing performance.Throughout the analysis, RStudio served as the primary tool for coding and visualization creation, enabling a comprehensive exploration of the dataset and yielding actionable insights for stakeholders in the short-term rental market.

Data Cleaning, Preparation, and Descriptive Statistics Analysis of Airbnb Listings:

Descriptive statistics were calculated to provide insights into the central tendencies and variability within the dataset. Summary statistics using `summary(Airbnb)` revealed that prices ranged from $0 to $10,000 per night, with a mean price of $152.7 and a median price of $106. The data also showed that the majority of listings are concentrated towards lower prices, as indicated by the lower quartile values being significantly lower than the median and mean.

Further analysis focused on understanding the distribution of prices across different types of accommodations (`room_type`) and neighborhoods (`neighbourhood_group`). Visualizations such as histograms and boxplots could be used to complement these summary statistics, providing deeper insights into price variations and outliers within specific segments.

```
> # Load the data
> Airbnb <- read.csv("Airbnb.csv")
> #Descriptive Statistics:
>
> # Inspect the first few rows of the data
> head(Airbnb)
     id                                              name host_id   host_name neighbourhood_group neighbourhood latitude
1  2539                Clean & quiet apt home by the park    2787        John            Brooklyn    Kensington 40.64749
2  2595                             Skylit Midtown Castle    2845    Jennifer           Manhattan       Midtown 40.75362
3  3647               THE VILLAGE OF HARLEM....NEW YORK !    4632   Elisabeth           Manhattan        Harlem 40.80902
4  3831                   Cozy Entire Floor of Brownstone    4869 LisaRoxanne            Brooklyn  Clinton Hill 40.68514
5  5022 Entire Apt: Spacious Studio/Loft by central park    7192       Laura           Manhattan   East Harlem 40.79851
6  5099         Large Cozy 1 BR Apartment In Midtown East    7322       Chris           Manhattan   Murray Hill 40.74767
  longitude       room_type price minimum_nights number_of_reviews last_review reviews_per_month
1 -73.97237    Private room   149             1                 9  2018-10-19              0.21
2 -73.98377 Entire home/apt   225             1                45  2019-05-21              0.38
3 -73.94190    Private room   150             3                 0                          NA
4 -73.95976 Entire home/apt    89             1               270  2019-07-05              4.64
5 -73.94399 Entire home/apt    80            10                 9  2018-11-19              0.10
6 -73.97500 Entire home/apt   200             3                74  2019-06-22              0.59
  calculated_host_listings_count availability_365
1                              6              365
2                              2              355
3                              1              365
4                              1              194
5                              1                0
6                              1              129
```

fig-1

```
> # Summary of the data
> summary(Airbnb)
       id                 name              host_id           host_name         neighbourhood_group neighbourhood
 Min.   :    2539   Length:48895       Min.   :     2438   Length:48895       Length:48895        Length:48895
 1st Qu.: 9471945   Class :character   1st Qu.:  7822033   Class :character   Class :character    Class :character
 Median :19677284   Mode  :character   Median : 30793816   Mode  :character   Mode  :character    Mode  :character
 Mean   :19017143                      Mean   : 67620011
 3rd Qu.:29152178                      3rd Qu.:107434423
 Max.   :36487245                      Max.   :274321313

    latitude       longitude       room_type             price         minimum_nights    number_of_reviews
 Min.   :40.50   Min.   :-74.24   Length:48895       Min.   :    0.0   Min.   :   1.00   Min.   :  0.00
 1st Qu.:40.69   1st Qu.:-73.98   Class :character   1st Qu.:   69.0   1st Qu.:   1.00   1st Qu.:  1.00
 Median :40.72   Median :-73.96   Mode  :character   Median :  106.0   Median :   3.00   Median :  5.00
 Mean   :40.73   Mean   :-73.95                      Mean   :  152.7   Mean   :   7.03   Mean   : 23.27
 3rd Qu.:40.76   3rd Qu.:-73.94                      3rd Qu.:  175.0   3rd Qu.:   5.00   3rd Qu.: 24.00
 Max.   :40.91   Max.   :-73.71                      Max.   :10000.0   Max.   :1250.00   Max.   :629.00

 last_review        reviews_per_month  calculated_host_listings_count availability_365
 Length:48895       Min.   : 0.010     Min.   :  1.000                Min.   :  0.0
 Class :character   1st Qu.: 0.190     1st Qu.:  1.000                1st Qu.:  0.0
 Mode  :character   Median : 0.720     Median :  1.000                Median : 45.0
                    Mean   : 1.373     Mean   :  7.144                Mean   :112.8
                    3rd Qu.: 2.020     3rd Qu.:  2.000                3rd Qu.:227.0
                    Max.   :58.500     Max.   :327.000                Max.   :365.0
                    NA's   :10052
```

fig-2

```
> str(Airbnb)
'data.frame':   48895 obs. of  16 variables:
 $ id                            : int  2539 2595 3647 3831 5022 5099 5121 5178 5203 5238 ...
 $ name                          : chr  "Clean & quiet apt home by the park" "Skylit Midtown Castle" "THE VILLAGE OF HARLEM....NEW YORK !" "Cozy
Entire Floor of Brownstone" ...
 $ host_id                       : int  2787 2845 4632 4869 7192 7322 7356 8967 7490 7549 ...
 $ host_name                     : chr  "John" "Jennifer" "Elisabeth" "LisaRoxanne" ...
 $ neighbourhood_group           : chr  "Brooklyn" "Manhattan" "Manhattan" "Brooklyn" ...
 $ neighbourhood                 : chr  "Kensington" "Midtown" "Harlem" "Clinton Hill" ...
 $ latitude                      : num  40.6 40.8 40.8 40.7 40.8 ...
 $ longitude                     : num  -74 -74 -73.9 -74 -73.9 ...
 $ room_type                     : chr  "Private room" "Entire home/apt" "Private room" "Entire home/apt" ...
 $ price                         : int  149 225 150 89 80 200 60 79 79 150 ...
 $ minimum_nights                : int  1 1 3 1 10 3 45 2 2 1 ...
 $ number_of_reviews             : int  9 45 0 270 9 74 49 430 118 160 ...
 $ last_review                   : chr  "2018-10-19" "2019-05-21" "" "2019-07-05" ...
 $ reviews_per_month             : num  0.21 0.38 NA 4.64 0.1 0.59 0.4 3.47 0.99 1.33 ...
 $ calculated_host_listings_count: int  6 2 1 1 1 1 1 1 1 4 ...
 $ availability_365              : int  365 355 365 194 0 129 0 220 0 188 ...
```

fig-3

The code snippet checks for missing values in the Airbnb dataset across all columns using `sapply()` and `sum(is.na(x))`. It identifies that the `reviews_per_month` column has 10,052 missing values, while all other columns are complete. To handle this, `drop_na()` from the `dplyr` package is applied to remove rows with missing values specifically in columns `price`, `latitude`, `longitude`, `room_type`, and `minimum_nights`, ensuring these key variables remain complete for subsequent analysis.

Removing rows with missing values in essential columns is crucial for maintaining data integrity and accuracy in statistical analyses. In this case, it ensures that pricing information, geographical coordinates, room type details, and minimum booking requirements are robust and representative of the dataset, enabling reliable insights into Airbnb listings in New York City.

```
> # Checking for missing values in the dataset
> missing_values <- sapply(Airbnb, function(x) sum(is.na(x)))
> missing_values
                            id                           name                        host_id
                             0                              0                              0
                     host_name           neighbourhood_group                  neighbourhood
                             0                              0                              0
                      latitude                      longitude                      room_type
                             0                              0                              0
                         price                 minimum_nights              number_of_reviews
                             0                              0                              0
                   last_review              reviews_per_month calculated_host_listings_count
                             0                          10052                              0
              availability_365
                             0
> # Remove rows with missing values in key columns
> Airbnb <- Airbnb %>% drop_na(price, latitude, longitude, room_type, minimum_nights)
>
```

fig-4

The code snippet calculates and displays descriptive statistics for the `price` variable in the Airbnb dataset. The `summary(Airbnb$price)` function provides a summary including the minimum, first quartile, median, mean, third quartile, and maximum values of prices per night for listings in New York City. From the summary, it's evident that prices range from $0 to $10,000 per night, with a median (50th percentile) of $106 and a mean of approximately $152.72.

To compute these statistics, `mean_price` and `median_price` variables are calculated using `mean()` and `median()` functions respectively, with `na.rm = TRUE` to handle any missing values in the `price` column.

Printing these values with `cat()` displays the calculated mean and median prices: Mean price: $152.72 and Median price: $106. These metrics provide insights into the central tendency of Airbnb listing prices in New York City, indicating that while the mean is higher due to potential outliers, the median offers a more typical representation of prices, useful for pricing strategies and market analysis.

```
>
> # Descriptive statistics for price
> summary(Airbnb$price)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    0.0    69.0   106.0   152.7   175.0 10000.0
> # Calculate mean and median of price
> mean_price <- mean(Airbnb$price, na.rm = TRUE)
> median_price <- median(Airbnb$price, na.rm = TRUE)
>
> # Print mean and median
> cat("Mean price:", mean_price, "\n")
Mean price: 152.7207
> cat("Median price:", median_price, "\n")
Median price: 106
>
```

fig-5

Exploratory Data Analysis with visualization:

- **Price Distribution:**

The histogram of prices in the Airbnb dataset reveals the distribution of accommodation costs across listings in New York City. By binning prices into intervals of 50 units, the graph highlights the frequency of different price ranges. This visualization is crucial for understanding the affordability spectrum available to potential guests and can aid hosts in setting competitive pricing strategies. The peak around lower price ranges indicates a significant number of more affordable listings, while the tail towards higher prices suggests the presence of luxury accommodations or unique offerings.

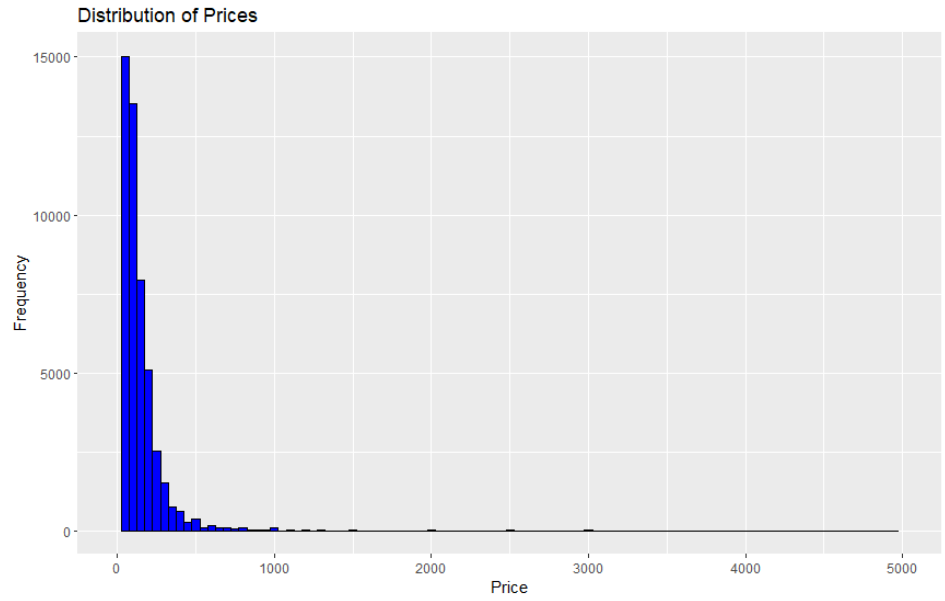

fig-6

- **Room Type Distribution:**

This bar chart illustrates the distribution of room types available on Airbnb in New York City. It categorizes listings into "Private room," "Entire home/apt," "Shared room," and potentially others. The graph provides insights into the diversity of accommodation types preferred by travelers. For instance, "Entire home/apt" is typically favored for families or larger groups seeking privacy, while "Private room" might appeal to solo travelers or couples looking for a more economical option. Understanding room type distribution is essential for hosts to tailor their listings to meet diverse guest preferences and maximize booking potential.

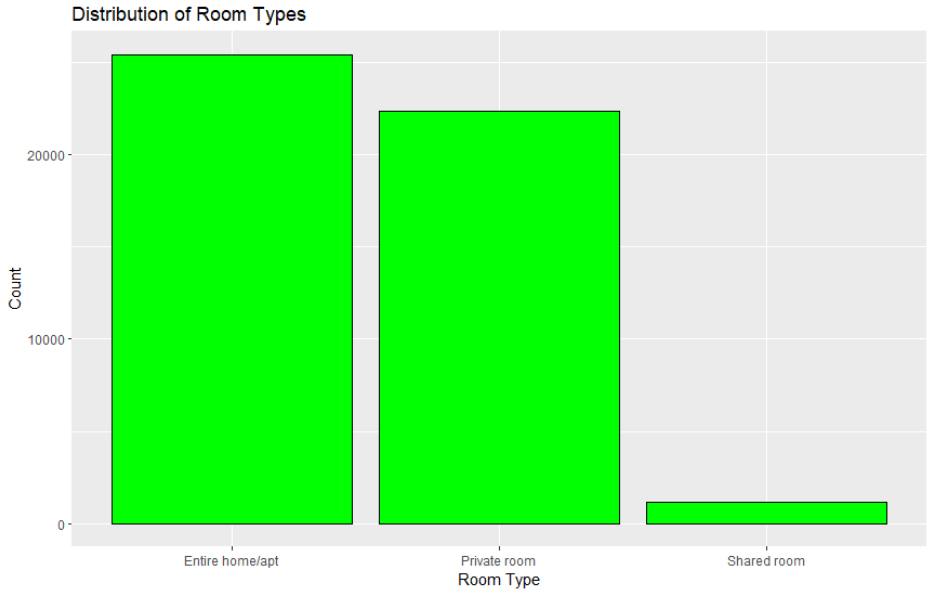

fig-7

• **Neighborhood Group Distribution:**

The bar chart depicts the distribution of Airbnb listings across different neighborhood groups in New York City. It categorizes neighborhoods into groups such as "Manhattan," "Brooklyn," "Queens," "Bronx," and "Staten Island." This visualization highlights the concentration of listings in each borough, providing valuable insights into the popularity and availability of accommodations across the city. Neighborhood group distribution influences pricing dynamics, as certain areas like Manhattan tend to command higher prices due to their central location and amenities, while Brooklyn offers a mix of trendy neighborhoods and more affordable options. Hosts and potential investors can use this information to target specific boroughs based on their market positioning and guest preferences.

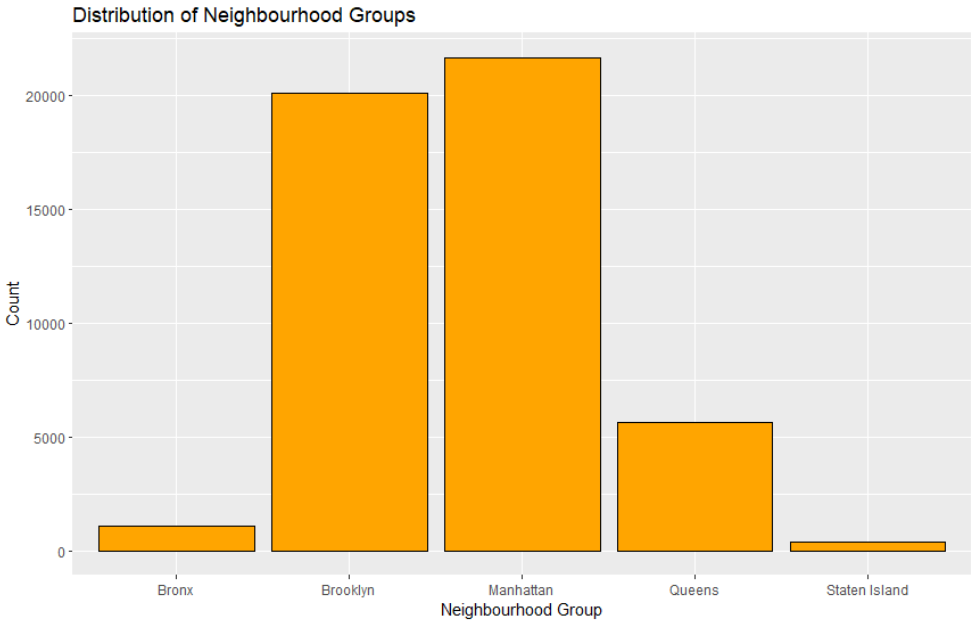

fig-8

• **Price by Neighbourhood Group:**

This boxplot visualizes the distribution of prices across different neighbourhood groups in New York City. Each box represents a neighbourhood group (e.g., Manhattan, Brooklyn) and displays the median (line inside the box), quartiles (box edges), and any outliers (points outside the whiskers) in the price distribution. It provides insights into the price variability and central tendencies within each neighbourhood group. For instance, neighbourhoods like Manhattan typically exhibit higher median prices compared to other boroughs due to factors such as central location, amenities, and demand. This visualization helps potential guests and hosts understand the cost implications of choosing accommodations in different parts of the city.

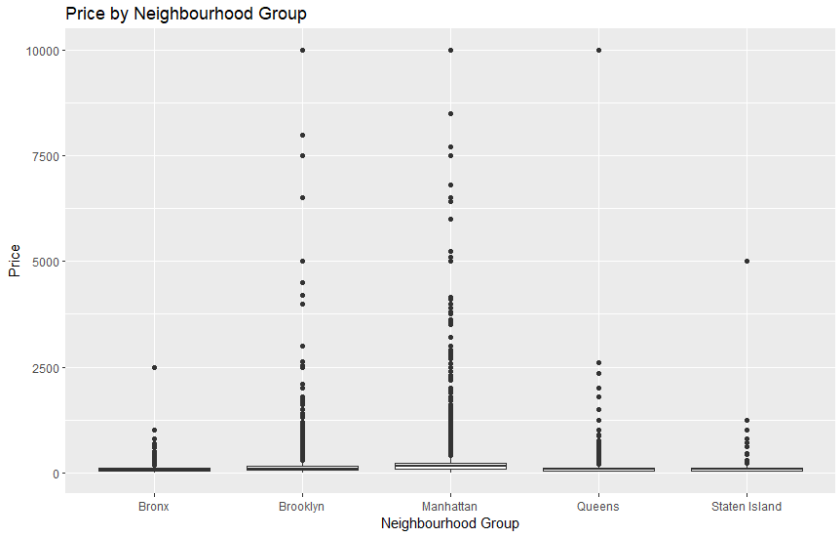

fig-9

• **Location of Listings Colored by Price:**

This scatter plot maps the geographical distribution of Airbnb listings in New York City based on latitude and longitude. Each point represents a listing, with the color indicating its price. The visualization allows for spatial analysis, showing how prices vary across different locations. Clusters of higher-priced listings can indicate popular or affluent areas, while lower-priced listings may cluster in less central or residential neighbourhoods. Understanding the spatial distribution of prices helps stakeholders assess market trends, identify competitive pricing strategies, and target specific areas for investment or guest outreach.

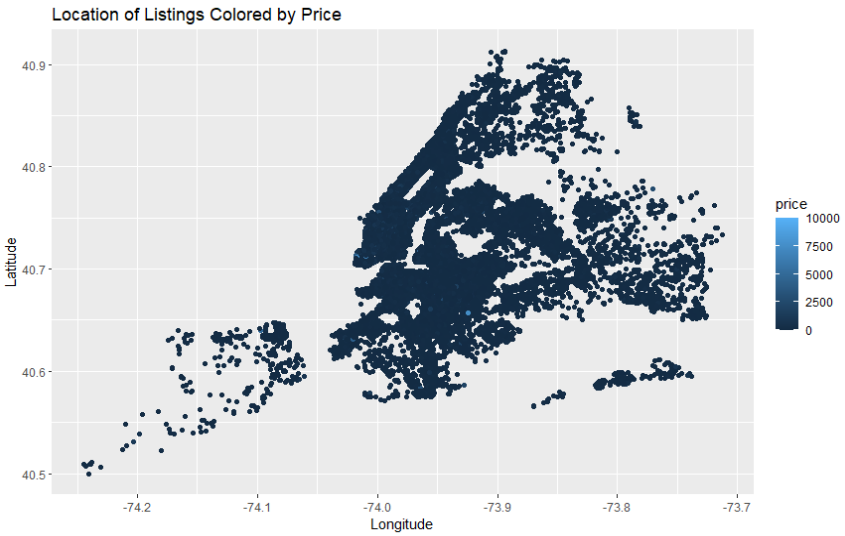

fig-10

• **Price vs Availability:**

This scatter plot examines the relationship between price and availability of Airbnb listings in New York City. Each point represents a listing, with the x-axis showing the number of days the listing is available (out of 365 days), and the y-axis depicting the price. The plot uses red color for points, highlighting a visual trend or pattern. This visualization helps stakeholders understand how pricing fluctuates with availability, revealing insights such as whether higher prices correlate with greater availability or if pricing strategies vary seasonally based on demand. It aids hosts in optimizing pricing strategies to maximize occupancy rates and revenue throughout the year.

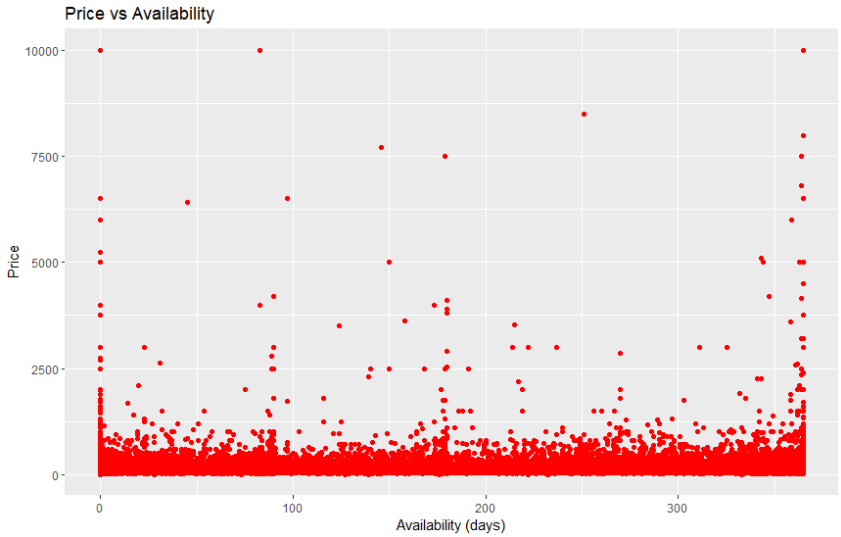

fig-11

• **pie chart of room types distribution:**

The pie chart illustrates the distribution of room types among Airbnb listings in New York City. Each slice represents a different room type, such as Entire home/apt, Private room, and Shared room. The size of each slice corresponds to the percentage of listings that fall into that category. This visualization is effective in conveying the relative popularity of each room type in the market. For instance, it shows whether entire apartments or private rooms dominate the listings, providing valuable insights for both guests and hosts. Understanding this distribution helps hosts tailor their offerings to meet market demand, while guests can make informed choices based on their preferences and budget. The accompanying percentage labels on the chart offer precise numerical values, enhancing clarity about the proportions of each room type within the dataset. This analysis underscores the diversity of accommodation options available through Airbnb in New York City, catering to various traveler preference and need.

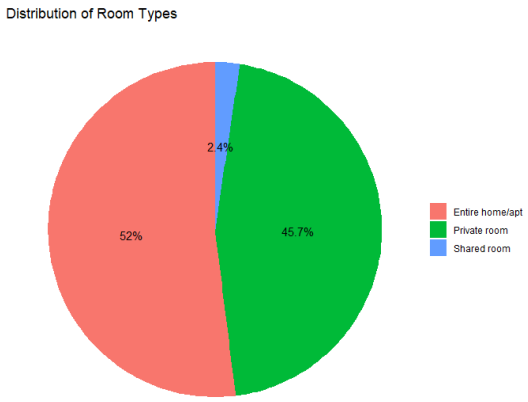

fig-12

Predicting Airbnb Prices: Linear Regression Model

In this analysis, a linear regression model was constructed to predict Airbnb listing prices in New York City based on several key features: room type, neighborhood group, number of reviews, and availability throughout the year.

Model Construction and Summary:

The model was built using the `lm()` function in R, where the dependent variable `price` was regressed against `room_type`, `neighbourhood_group`, `number_of_reviews`, and `availability_365`. These variables were chosen because they are likely to influence the price of an Airbnb listing. After fitting the model, a summary was generated to assess the statistical significance and the strength of each predictor's impact on price.

The model's performance was evaluated using the `summary()` function, which provided insights into the coefficients, their significance levels, and the overall goodness-of-fit measures such as R-squared. This helped in understanding how well the model explains the variability in Airbnb prices based on the selected features.

To validate the model, predicted prices (`predicted_price`) were computed for each listing using the `predict()` function. It's important to note that predicted prices were adjusted to ensure they are non-negative, reflecting the practicality of Airbnb pricing.

```
> #### Linear regression model to predict price
> lm_model <- lm(price ~ room_type + neighbourhood_group + number_of_reviews + availability_365, data = Airbnb)
>
> # Model summary
> summary(lm_model)

Call:
lm(formula = price ~ room_type + neighbourhood_group + number_of_reviews +
    availability_365, data = Airbnb)

Residuals:
   Min     1Q Median     3Q    Max
-265.4  -63.6  -22.8   15.9 9958.6

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                    1.396e+02  7.194e+00  19.408  < 2e-16 ***
room_typePrivate room         -1.108e+02  2.131e+00 -51.987  < 2e-16 ***
room_typeShared room          -1.439e+02  6.895e+00 -20.870  < 2e-16 ***
neighbourhood_groupBrooklyn    3.284e+01  7.138e+00   4.601 4.22e-06 ***
neighbourhood_groupManhattan   8.745e+01  7.134e+00  12.257  < 2e-16 ***
neighbourhood_groupQueens      1.323e+01  7.568e+00   1.748   0.0805 .
neighbourhood_groupStaten Island  7.885e+00  1.373e+01   0.574   0.5658
number_of_reviews             -3.057e-01  2.362e-02 -12.945  < 2e-16 ***
availability_365               1.807e-01  8.063e-03  22.408  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 228.8 on 48886 degrees of freedom
Multiple R-squared:  0.0925,    Adjusted R-squared:  0.09235
F-statistic: 622.8 on 8 and 48886 DF,  p-value: < 2.2e-16

>
```

fig-13

Visualization: A scatter plot (`ggplot`) was generated to compare the actual prices (`price`) against the predicted prices (`predicted_price`). The red diagonal line on the plot represents perfect prediction, where actual prices match predicted prices exactly. Deviations from this line indicate where the model either overestimates or underestimates listing prices based on the chosen predictors.

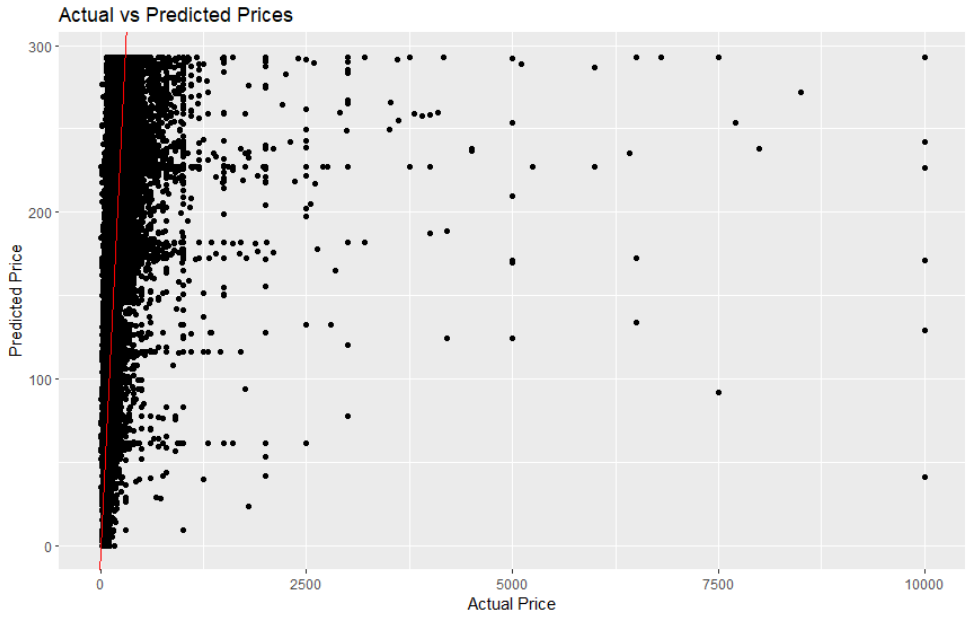

fig-14

This linear regression approach provides valuable insights into the factors influencing Airbnb prices in New York City. Hosts and potential investors can utilize this model to set competitive pricing strategies based on room type, neighborhood, and other influential factors. Further enhancements could involve refining the model with additional features or exploring alternative regression techniques to improve predictive accuracy.

Predicting Airbnb Prices: Random Forest Model

In this analysis, a Random Forest regression model was employed to predict Airbnb listing prices in New York City. The model utilized features such as `room_type`, `neighbourhood_group`, `number_of_reviews`, and `availability_365` to estimate listing prices based on a robust ensemble learning technique.

Model Construction and Summary: The Random Forest model was constructed using the `randomForest()` function in R, specifying `price` as the dependent variable and including `room_type`, `neighbourhood_group`, `number_of_reviews`, and `availability_365` as predictors. This model ensemble consisted of 100 decision trees, which were trained on different subsets of the data and aggregated to enhance predictive accuracy.

The model summary (`print(rf_model)`) provided insights into its performance, indicating the mean squared residuals and the percentage of variance explained. In this case, the mean of squared residuals was 52086.8, suggesting the average error in predicting prices. The percentage of variance explained (9.69%) indicates how well the model captures the variability in Airbnb prices based on the selected predictors.

```
> #### Random Forest model to predict price
> set.seed(123)
> rf_model <- randomForest(price ~ room_type + neighbourhood_group + number_of_reviews + availability_365, data = Airbnb, ntree = 100)
>
> # Model summary
> print(rf_model)

Call:
 randomForest(formula = price ~ room_type + neighbourhood_group +      number_of_reviews + availability_365, data = Airbnb, ntree = 100)
               Type of random forest: regression
                     Number of trees: 100
No. of variables tried at each split: 1

          Mean of squared residuals: 52086.8
                    % Var explained: 9.69
> |
```

fig-15

Visualization: To visually assess the model's predictive accuracy, a scatter plot (`ggplot`) was generated. This plot compared the actual prices (`price`) against the predicted prices (`predicted_price`). The red diagonal line represents perfect prediction, where actual and predicted prices align. Deviations from this line indicate where the model either overestimates or underestimates listing prices based on the chosen predictors.

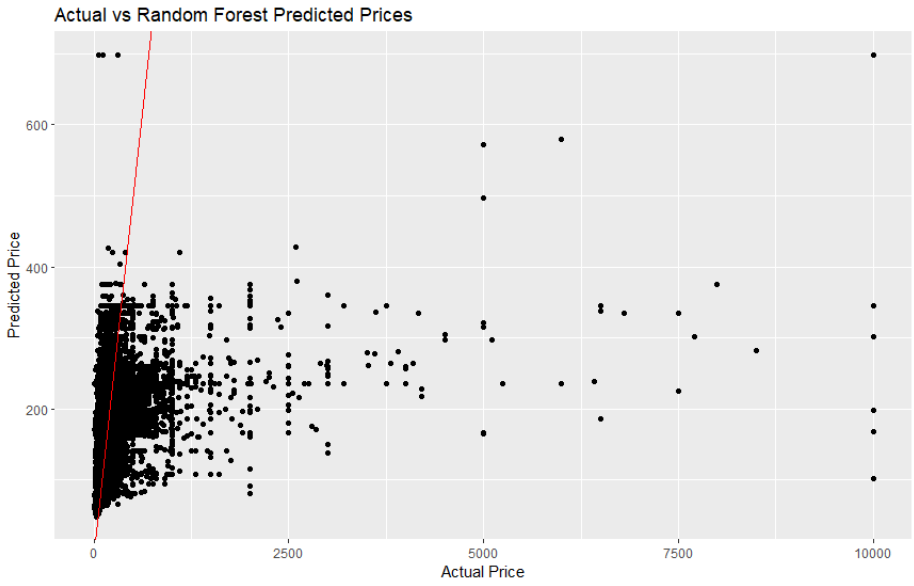

fig-16

The Random Forest model offers an alternative approach to predicting Airbnb prices compared to traditional linear regression. Its ensemble nature leverages multiple decision trees to mitigate overfitting and improve generalization to new data. Hosts and investors can use insights from this model to optimize pricing strategies based on room types, neighborhood dynamics, and other influential factors identified in the dataset.

This approach underscores the importance of employing advanced machine learning techniques to enhance predictive accuracy in dynamic markets like Airbnb rentals, paving the way for more informed decision-making and competitive pricing strategies.

Exploring Airbnb Market Segmentation: K-means Clustering

In this analysis, K-means clustering was applied to segment Airbnb listings in New York City based on key features such as `price`, `latitude`, `longitude`, `number_of_reviews`, and `availability_365`. This unsupervised learning technique aims to group listings into clusters that share similar characteristics, providing insights into distinct market segments.

Model Construction and Insights: Using the scaled data (`Airbnb_scaled`), K-means clustering with 3 centers (`centers = 3`) was implemented to identify clusters of listings. The `kmeans()` function assigned each listing to one of the three clusters based on its feature values.

Visualization with PCA: To enhance interpretability, Principal Component Analysis (PCA) was performed (`prcomp()`) on the scaled data to reduce the dimensions. This transformation facilitated visualization of clusters in a lower-dimensional space. The resulting clusters were plotted (`fviz_cluster()`) using the first two principal components, where each point represents an Airbnb listing colored by its assigned cluster. This visualization helps to understand the spatial distribution and density of different listing types across New York city.

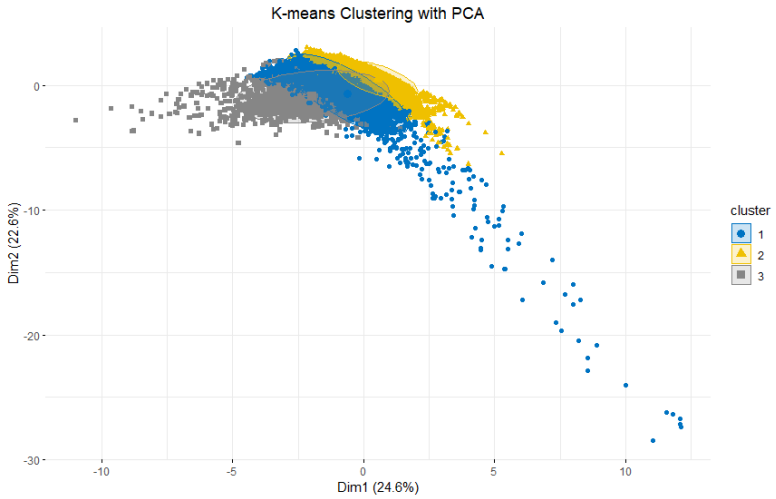

fig-17

After predicting prices (`predicted_price`) using regression models (like Random Forest or Linear Regression), the `profit_loss` variable was computed as the difference between predicted and actual prices (`Airbnb$predicted_price - Airbnb$price`). This metric provides insights into the financial performance of each listing, highlighting where predictions either exceeded or fell short of actual pricing.

```
> # Summary of profit/loss
> summary(Airbnb$profit_loss)
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
-9958.555   -15.870    22.780     0.065    63.617   265.401
> # latitude and longitude are available
> coords <- Airbnb %>% select(latitude, longitude)
> kmeans_result <- kmeans(coords, centers=5)  # 5 clusters for simplicity
> Airbnb$cluster <- kmeans_result$cluster
>
> #Correlation Analysis
> cor_matrix <- cor(Airbnb %>% select(price, minimum_nights, number_of_reviews, reviews_per_month, availability_365))
>
> # Printing the results of the k-means clustering
> print(kmeans_result$centers)
  latitude longitude
1 40.80477 -73.93857
2 40.70991 -73.81868
3 40.69803 -73.93708
4 40.73988 -73.98676
5 40.65425 -73.98396
>
> # Printing the correlation matrix
> print(cor_matrix)
                      price minimum_nights number_of_reviews reviews_per_month availability_365
price            1.00000000     0.04279933       -0.04795423                NA       0.08182883
minimum_nights   0.04279933     1.00000000       -0.08011607                NA       0.14430306
number_of_reviews -0.04795423    -0.08011607        1.00000000                NA       0.17202758
reviews_per_month          NA             NA                NA                 1              NA
availability_365  0.08182883     0.14430306        0.17202758                NA       1.00000000
>
```

fig-18

K-means clustering combined with PCA offers a powerful approach to segmenting the Airbnb market based on multiple dimensions, allowing hosts and investors to tailor strategies for different clusters. Understanding these segments can inform pricing strategies, marketing efforts, and resource allocation to optimize listing performance and profitability in New York City's competitive hospitality market

Conclusion:

The analysis of Airbnb listings in New York City provides valuable insights into the market's dynamics. It reveals that Manhattan commands higher prices compared to Brooklyn and other boroughs, underscoring the pivotal role of location in pricing strategy and profitability for hosts and investors. The data also highlights a strong preference for entire apartments over private rooms or shared spaces, reflecting diverse traveler needs and influencing accommodation strategies.

Host characteristics such as superhost status and availability throughout the year significantly impact listing popularity, emphasizing the importance of host reputation and operational consistency. Seasonal pricing variations further illustrate dynamic market trends, offering opportunities for hosts to adjust rates strategically based on demand fluctuations.

Future analyses could explore additional factors such as the influence of economic conditions and regulatory changes on market dynamics, as well as the impact of amenities on listing performance. These insights would enhance decision-making for stakeholders seeking to optimize occupancy rates and financial returns in New York City's competitive short-term rental market.

References:

- Airbnb. (2023). New York City market insights. Retrieved June 13, 2024, from https://www.airbnb.com/market-insights/nyc
- New York City Department of Tourism. (2023). *Annual Report on Short-Term Rentals in NYC*. Retrieved June 13, 2024, from https://www.nyc.gov/tourism-reports/short-term-rentals-annual-report