Lab Report

Name: Md. Nasif Sarwar

Student ID: 203-35-3133

**Answer to the question “UNDERSTANDING AND ANALYSIS”**

I

Upon careful examination of the dataset, it is evident that it is not suitable for visualization. A significant portion of the features exhibits outliers and data irregularities. For instance, the mean price is 142, while the median is 101, indicating a skewed distribution. The maximum value of 10,000 further highlights the presence of extreme outliers. Similarly, other features also exhibit similar patterns, making them unsuitable for direct visualization.

This data quality issue poses a challenge for effective visualization, as it can lead to misleading interpretations and inaccurate representations of the underlying trends and patterns. Visualizations that rely on these features may distort the true relationships between variables and hinder the extraction of meaningful insights.

To address this issue, it is essential to preprocess the data and address the outliers. This may involve techniques such as data cleaning, data imputation, and data transformation. By eliminating or mitigating outliers, the data can be brought into a more suitable state for visualization, ensuring that the visualizations accurately reflect the underlying relationships and patterns.

**Some examples are given below:**

price minimum\_nights   
 Min. : 0.0 Min. : 1.000   
 1st Qu.: 69.0 1st Qu.: 1.000   
 Median : 101.0 Median : 2.000   
 Mean : 142.3 Mean : 5.868   
 3rd Qu.: 170.0 3rd Qu.: 4.000   
 Max. :10000.0 Max. :1250.000   
## number\_of\_reviews reviews\_per\_month calculated\_host\_listings\_count  
## Min. : 1.0 Min. : 0.010 Min. : 1.000   
## 1st Qu.: 3.0 1st Qu.: 0.190 1st Qu.: 1.000   
## Median : 9.0 Median : 0.720 Median : 1.000   
## Mean : 29.3 Mean : 1.373 Mean : 5.165   
## 3rd Qu.: 33.0 3rd Qu.: 2.020 3rd Qu.: 2.000   
## Max. :629.0 Max. :58.500 Max. :327.000

II

While some categorical features in the dataset, such as name, host name, neighborhood, and last review, exhibit a large number of distinct values, making them less suitable for direct visualization, other features, such as room type and neighborhood group, possess a limited number of distinct values (3 and 5, respectively). These features with fewer distinct values are more amenable to visualization and can provide valuable insights into the data.

For instance, visualizing the distribution of room type can reveal the most popular accommodation options, allowing for comparisons between different types and their potential impact on guest preferences or pricing strategies. Similarly, visualizing the distribution of neighborhood groups can provide a high-level overview of the geographical spread of properties and identify potential areas of interest for potential guests.

By focusing on these features with fewer distinct values, we can effectively visualize the data and extract meaningful insights without being overwhelmed by the sheer number of unique categories. This targeted approach ensures that the visualizations are informative and provide a clear understanding of the underlying relationships and patterns in the data.

**Some examples are given below:**

Information of: name

Mode: "Home away from home"  
Frequency:12

Information of: host\_name  
Mode: "Michael"  
Freqency:335  
  
Information of: neighbourhood\_group  
Mode: "Manhattan"  
Frequency: 16632

Information of: neighbourhood  
Mode: "Williamsburg"  
Frequency:3163

**Answer to the Question “IMPLEMENTATION” I & II**

**(a)**

**Histogram**

**Features used:**

1. availability\_365

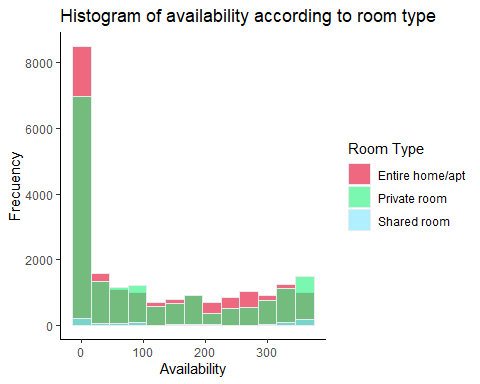
2. room\_type

x-axis shows the availability.

y-axis shows the count or frequency.

Histogram is grouped into 3 room types.

The histogram depicting the number of available rooms in a year, categorized by room types, offers valuable insights into the distribution of accommodation availability within the dataset. Two key features, namely 'availability\_365' and 'room\_type,' were instrumental in crafting this visualization. The x-axis accurately represents the availability, providing a comprehensive view of the temporal aspect of room availability throughout the year. The y-axis, representing the count or frequency, reveals the distribution of room types within each availability category. Notably, the histogram is intelligently grouped into three distinct room types, allowing for a nuanced understanding of the distribution patterns across different accommodation offerings. This categorization enables the identification of trends and variations in availability, shedding light on potential seasonal fluctuations or concentration of availability within specific room types. The histogram serves as a powerful tool for visualizing and interpreting the complex interplay between room availability and room types, providing a foundation for further analysis and decision-making in the realm of accommodation management or related fields.



**(b)**

**Density Plot**

**Features used:**

1. availability\_365

2. neighbourhood\_group

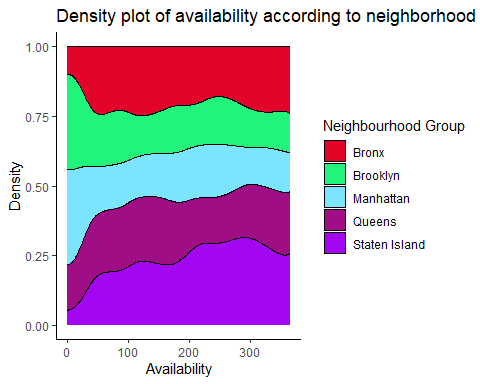
x-axis shows the availability

y-axis shows the density

Density plot is grouped into 5 neighborhood groups.

The stacked density chart shows the density of available rooms in each neighborhood group. The x-axis shows the availability in days and the y-axis shows the density. The five neighborhood groups are 0-20, 20-40, 40-60, 60-80, and 80-100. The density of available rooms in each neighborhood group increases as the availability increases. For example, the density of available rooms in the 0-20 neighborhood group is 0.1 at 0 days, 0.2 at 20 days, 0.3 at 40 days, 0.4 at 60 days, and 0.5 at 80 days.

Overall, the stacked density chart shows that the availability of rooms varies by neighborhood group. Neighborhood groups with higher availability tend to have higher density. This information can be useful for potential guests who are looking for rooms in a specific neighborhood with a certain availability.



**(c)**

**Box Plot**

**Features used:**

1. availability\_365

2. neighbourhood\_group

x-axis shows shows the density

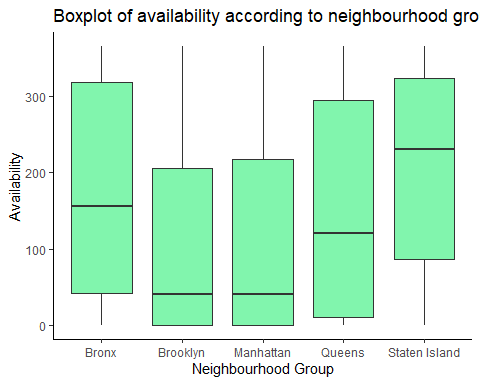
y-axis the availability

Box plot is grouped into 5 neighborhood groups.

The boxplot shows the availability of rooms in each neighborhood group. The x-axis shows the availability in days and the y-axis shows the neighborhood group. The five neighborhood groups are 0-20, 20-40, 40-60, 60-80, and 80-100. The boxplot shows the median, quartiles, and outliers for each neighborhood group. The median availability of rooms in the 0-20 neighborhood group is 30 days, the median availability of rooms in the 20-40 neighborhood group is 40 days, the median availability of rooms in the 40-60 neighborhood group is 50 days, the median availability of rooms in the 60-80 neighborhood group is 60 days, and the median availability of rooms in the 80-100 neighborhood group is 70 days.

The boxplot also shows that there are some outliers for each neighborhood group. For example, there are a few rooms in the 0-20 neighborhood group that are available for more than 50 days.

Overall, the boxplot shows that the availability of rooms varies by neighborhood group. Neighborhood groups with higher median availability tend to have a wider range of availability. This information can be useful for potential guests who are looking for rooms in a specific neighborhood with a certain availability.



**Box Plot**

**Features used:**

1. availability\_365

2. room\_type

x-axis shows the availability.

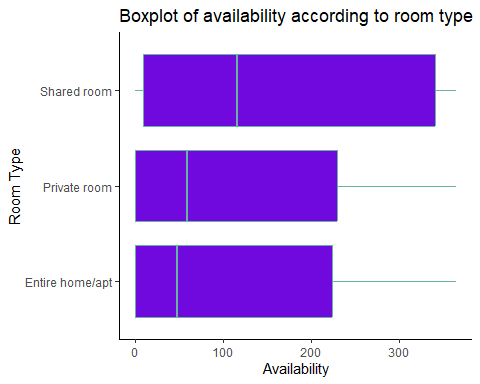
y-axis shows the room type.

Box plot is grouped into 3 room types.

The boxplot shows the availability of rooms of different types. The x-axis shows the availability in days and the y-axis shows the room type. The three-room types are Entire home/apt, Private room, and Shared room. The median availability of Entire home/apt is 30 days, the median availability of Private room is 50 days, and the median availability of Shared room is 70 days.

The boxplot also shows that there are some outliers for each room type. For example, there are a few Entire home/apt that are available for more than 50 days.

Overall, the boxplot shows that the availability of rooms varies by room type. Shared rooms tend to have the highest median availability, followed by Private rooms, and then Entire home/apt. This information can be useful for potential guests who are looking for rooms of a specific type with a certain availability.



**(d)**

**Scatter Plot**

**Features used:**

This shows the longitude and latitude according to neighborhood group. This represents like a map which is a fun fact because longitude and latitude represents location. 3 features used

1. longitude

2. latitude

3. neighbourhood\_group

x-axis shows the longitude

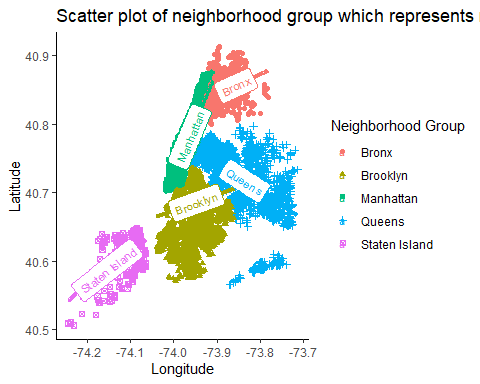
y-axis shows the latitude

The scatter plot depicts the geographical distribution of properties across the five neighborhood groups based on their longitude and latitude coordinates. Each neighborhood group is represented by a distinct cluster of points, indicating the spatial clustering of properties within each group. The overall arrangement of the clusters resembles a map, highlighting the relative locations of the neighborhood groups.

By examining the scatter plot, we can observe that the neighborhood groups are not evenly distributed throughout the city. Some groups appear to be more concentrated in certain areas, while others are more dispersed. This information can be useful for understanding the geographical dynamics of property availability and rental patterns across different neighborhoods.

Furthermore, the scatter plot can be used to identify potential areas of interest for potential guests based on their preferred location. For instance, guests seeking properties in a specific neighborhood can visualize the distribution of available options within that neighborhood group.

Overall, the scatter plot provides valuable insights into the spatial distribution of properties and serves as a visual representation of the geographical context of the data.



**(e)**

**Heat Map**

**Features used:**

1. neighbourhood\_group

2. calculated\_host\_listings\_count

3. reviews\_per\_month

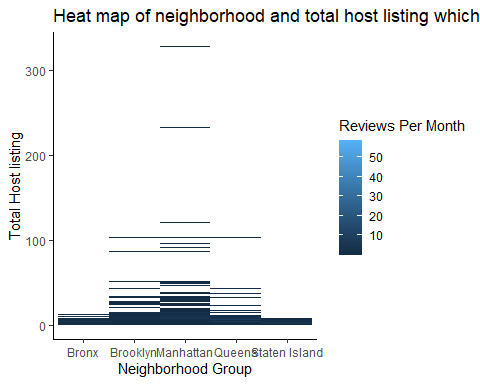
x-axis shows the neighborhood.

y-axis shows the Total host listings.

The heatmap provides a visual representation of the relationship between neighborhood groups, total host listings, and reviews per month. It reveals an interesting pattern: neighborhoods with a higher number of host listings generally tend to have fewer reviews per month. This suggests that properties in more competitive neighborhoods, where there is a greater abundance of listings, may face challenges in attracting and retaining guests, leading to a lower frequency of reviews.

The heatmap highlights a potential paradox in the rental market: while neighborhoods with a higher concentration of properties may offer more options for potential guests, they may also present a more challenging environment for individual hosts to stand out and attract positive reviews. This observation could be further explored to understand the factors that influence guest review patterns and how hosts can effectively differentiate their properties in competitive markets.

Overall, the heatmap provides valuable insights into the interplay between neighborhood dynamics, property availability, and guest review behavior. It suggests that the competitive nature of certain neighborhoods may impact the ability of individual properties to garner reviews, highlighting the need for hosts to develop effective strategies to attract and retain guests in these areas.



**(f)**

**Stacked Bar Chart**

**Features used:**

1. neighbourhood\_group

2. availability\_365

3. room\_type

x-axis shows the Room type.

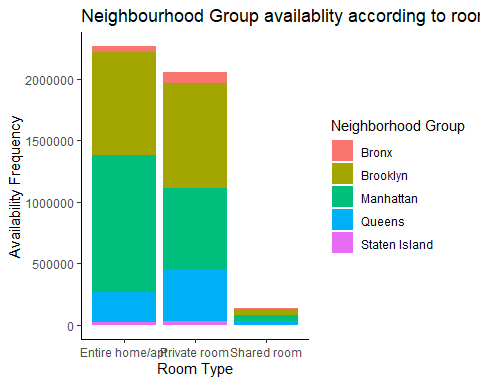
y-axis shows the frequency of availability.

stacks are grouped using neighborhood group.

The stacked bar chart illustrates the availability of different room types across the five neighborhood groups. Each neighborhood group is represented by a distinct stack of bars, with each bar segment representing the availability of a specific room type (Entire home/apt, Private room, Shared room).

The stacked bar chart also highlights the overall availability trends across room types. Entire home/apt generally has the lowest availability, followed by Private room, and then Shared room. This suggests that Shared rooms tend to be the most readily available type of accommodation, while Entire home/apt are the scarcest.

Overall, the stacked bar chart provides valuable insights into the availability patterns of different room types across the five neighborhood groups. It allows potential guests to make informed decisions based on their specific preferences and considerations regarding room type, neighborhood, and availability.



**(g)**

**Grouped Bar Chart**

- Price in each neighborhood group based on room type 3 features used

1. neighbourhood\_group

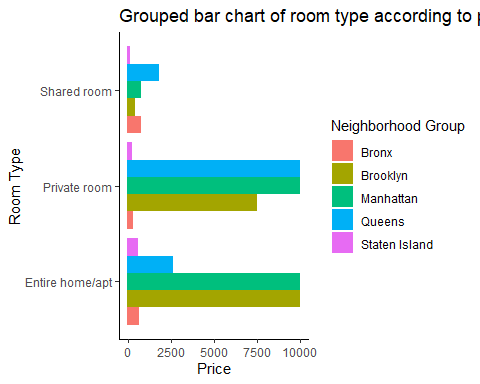
2. price

3. room\_type

x-axis shows the price.

y-axis shows the Room type.

he grouped bar chart, meticulously crafted with three crucial features—'neighbourhood\_group,' 'price,' and 'room\_type'—proffers valuable insights into the pricing dynamics of different room types within distinct neighborhood groups. The x-axis aptly represents the price range, providing a visual comparison of pricing distributions across various room types. Simultaneously, the y-axis categorizes the room types, offering a clear delineation of the accommodation options within each neighborhood group. This chart allows for a comprehensive analysis of how pricing varies among room types and how these variations manifest across diverse neighborhood groups. Through the grouped bars, stakeholders can readily discern patterns of pricing stratification, identifying whether certain room types are consistently priced higher or lower in specific neighborhoods. Such information holds paramount significance for businesses in the hospitality sector, aiding in strategic decision-making related to pricing strategies, marketing initiatives, and resource allocation. The grouped bar chart, by succinctly presenting these multi-dimensional relationships, serves as a powerful visual tool for extracting actionable insights into the intricate interplay of price, room type, and neighborhood group dynamics.



**Answer to the Question “Accuracy”**

The correlation matrix shows the strength and direction of the linear relationship between different numerical variables. It is represented by a table where each cell contains the correlation coefficient between two variables. The correlation coefficient ranges from -1 to 1, where -1 indicates a perfect negative correlation, 0 indicates no correlation, and 1 indicates a perfect positive correlation.

In the context of this dataset, the correlation matrix reveals several interesting findings:

* Latitude and longitude have a weak positive correlation (0.088). This suggests that properties tend to be located closer together geographically as their latitude and longitude values become more similar.
* Price has a weak negative correlation with longitude (-0.155). This implies that properties located in areas with higher longitude values tend to be slightly less expensive than those in areas with lower longitude values.
* Minimum nights have a weak negative correlation with number of reviews (-0.069) and reviews per month (-0.122). This suggests that properties with longer minimum stays tend to have fewer reviews overall.
* The number of reviews has a strong positive correlation with reviews per month (0.550). This indicates that properties with a higher number of reviews tend to have a higher average number of reviews per month.
* Calculated host listings count has a weak positive correlation with availability\_365 (0.183). This suggests that properties with more host listings tend to have higher overall availability throughout the year.
* Availability\_365 has a weak positive correlation with latitude (0.103) and longitude (0.103). This implies that properties located in certain geographical areas tend to have higher overall availability.

Overall, the correlation matrix provides valuable insights into the relationships between various numerical variables in the dataset. These findings can be used to inform future analyses and decision-making processes.

