DIC Project Phase1

Teammate 1: Mahammad Thufail

Teammate 2: Sai Shirini Prathigadapa

Teammate 3: Rekha Anvitha Inturi

Problem Statement:

Title: NYC Airbnb Analysis

Problem Statement:

Exploring the relationship between Airbnb listing attributes such as location, room type, minimum nights, and availability, with listing prices. This analysis aims to uncover significant factors influencing pricing variations across different neighborhoods and property types, offering valuable insights for both guests seeking optimal value and hosts seeking to optimize their pricing strategies.

I. Background of Problem:

The advent of online platforms like Airbnb has revolutionized the hospitality industry, offering travelers a diverse range of accommodation options and enabling individuals to monetize their properties. However, determining the optimal price for an Airbnb listing can be challenging for hosts, as it involves considerations such as location, property type, and market demand.

a. Objectives:

1. **Identify Key Pricing Factors:** Analyze the Airbnb dataset to discern the primary factors influencing listing prices, including neighborhood, room type, and minimum nights.

2. **Understand Geographic Variations:** Explore how listing prices vary across different neighborhoods or geographical regions, shedding light on local market dynamics and demand patterns.

3. **Assess Room Type Impact:** Investigate the impact of room types (e.g., entire home/apartment, private room, shared room) on listing prices to understand preferences and price sensitivities.

4. **Optimize Minimum Nights:** Evaluate the relationship between minimum nights required for booking and listing prices, providing insights into pricing strategies for varying lengths of stay.

5. **Predictive Insights:** Develop predictive models to forecast listing prices based on relevant features, facilitating data-driven pricing decisions for hosts.

6. **Enhance Guest Experience:** Utilize findings to enhance the overall guest experience by ensuring fair pricing, thereby promoting customer satisfaction and loyalty.

By addressing these objectives, this analysis aims to provide actionable insights for both Airbnb hosts and guests, ultimately contributing to a more efficient and transparent marketplace for short-term rentals.

b. Significance of Problem:

1. **Economic Impact:** Optimising Airbnb listing prices is not only crucial for individual hosts but also contributes to the broader economy. By accurately pricing their listings, hosts can maximize their revenue potential, thereby stimulating economic activity in the local communities where Airbnb operates. Additionally, a well-functioning Airbnb marketplace fosters entrepreneurship by enabling individuals to generate income from underutilized assets, thus bolstering economic growth and job creation.

2. **Market Efficiency:** Understanding the factors that influence Airbnb listing prices enhances market efficiency by facilitating better allocation of resources. Hosts equipped with insights into pricing determinants can adjust their rates according to demand fluctuations, leading to a more balanced supply and demand equilibrium. Moreover, by promoting transparency and competition, an efficiently priced Airbnb market encourages innovation and fosters a more dynamic and responsive marketplace.

3. **Consumer Empowerment**: Transparent pricing in the Airbnb ecosystem empowers consumers by enabling them to make informed decisions based on their preferences and budget constraints. By gaining insights into pricing factors such as location, room type, and minimum nights, guests can identify listings that offer the best value for their money. This empowerment fosters trust and satisfaction among consumers, driving repeat bookings and positive word-of-mouth

recommendations. Additionally, transparent pricing encourages healthy competition among hosts, incentivizing them to offer competitive rates and high-quality accommodations to attract guests.

In summary, addressing the optimization of Airbnb listing prices is significant not only for its direct economic impact but also for its role in enhancing market efficiency and empowering consumers. By promoting fair and transparent pricing practices, this analysis contributes to the overall sustainability and vibrancy of the Airbnb marketplace.

II. Potential of The Project:

The project holds the potential to significantly enhance market efficiency, transparency, and consumer empowerment within the Airbnb ecosystem, ultimately fostering economic growth and development.

- Market Efficiency & Transparency:

  Contribution: By identifying key factors influencing Airbnb listing prices, the project enhances market efficiency by facilitating better resource allocation and pricing decisions among hosts. Through data-driven insights, it promotes transparency, enabling hosts to adjust rates according to demand fluctuations, thus fostering a more balanced supply and demand equilibrium.

  Significance: This contributes to a more dynamic and responsive marketplace, encouraging innovation and competition while ensuring fair and transparent pricing practices. Ultimately, it leads to a more efficient and transparent Airbnb ecosystem.

- Consumer Empowerment:

  Contribution: The project empowers consumers by providing them with insights into Airbnb listing prices, enabling them to make informed decisions based on their preferences and budget constraints. Transparent pricing practices foster trust and satisfaction among guests, driving repeat bookings and positive recommendations.

  Significance: This empowerment enhances consumer welfare and promotes trust in the Airbnb platform, ultimately leading to improved guest experiences and sustained growth in user engagement. It also encourages hosts to offer competitive rates and high-quality accommodations, further benefiting consumers.

- Economic Impact:

Contribution: By optimizing Airbnb listing prices, the project stimulates economic activity by enabling hosts to maximize their revenue potential and generate income from underutilized assets. This contributes to economic growth, job creation, and entrepreneurship in local communities where Airbnb operates.

Significance: The project's economic impact extends beyond individual hosts to benefit broader economic stakeholders, including businesses, local governments, and service providers. It fosters a more vibrant and sustainable economy while promoting the efficient use of resources within the Airbnb marketplace.

Data Sources:

The raw input data file for the above problem statement is taken from the site :

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

This dataset is a CSV file which has 48895 data points and 16 Features.

- ➢ Features include:
  - **id**: Unique identifier for each Airbnb listing (numeric).
  - **name**: Name of the listing (text).
  - **host_id**: Unique identifier for each host (numeric).
  - **host_name**: Name of the host (text).
  - **neighbourhood_group**: Grouping of neighborhoods (text).
  - **neighborhood**: Specific neighborhood of the listing (text).
  - **latitude**: Latitude coordinate of the listing (numeric).
  - **longitude**: Longitude coordinate of the listing (numeric).
  - **room_type**: Type of room available (text).
  - **price**: Price per night for the listing (numeric).
  - **minimum_nights**: Minimum number of nights required for booking (numeric).
  - **number_of_reviews**: Total number of reviews for the listing (numeric).
  - **last_review**: Date of the last review (text).
  - **reviews_per_month**: Average number of reviews per month (numeric).
  - **calculated_host_listings_count**: Total number of listings managed by the host (numeric).
  - **availability_365**: Number of days the listing is available within the next 365 days (numeric).

- Data types:
  - **Numeric features**:id,host_id,latitude,longitude,price, `minimum_nights`, number_of_reviews,reviews_per_month,calculated_host_listings_count ,availability_365.
  - **Text features:**name, host_name, neighbourhood_group, neighborhood, room_type, last_review.
  - **Memory usage:** Approximately 6.0 MB.

Data Cleaning/ Processing:

- Data Cleaning Steps:

Convert Host Name to Uppercase:
To maintain consistency and facilitate easier data manipulation, the 'host_name' column is converted to uppercase. This ensures that all host names are uniformly represented in uppercase format, eliminating variations in capitalization that could potentially cause discrepancies in data analysis. By standardizing the case of host names, tasks such as sorting, filtering, and comparing data become more straightforward and accurate. Additionally, converting host names to uppercase enhances data cleanliness and readability, contributing to the overall quality and usability of the dataset for subsequent analysis and modeling purposes.

Drop Rows with Missing Values in 'Name' and 'Hostname' Columns:

Removing rows with missing values in the 'name' and 'host_name' columns is essential to maintain the integrity and completeness of the dataset. Missing values can introduce inaccuracies and biases into the analysis, potentially skewing results and conclusions drawn from the data. By eliminating these incomplete rows, we ensure that each listing in the dataset has essential information such as the name of the listing and the name of the host associated with it. This promotes data reliability and ensures that subsequent analyses are based on a comprehensive and representative sample of Airbnb listings. Moreover, having complete data allows for more robust modeling and more accurate insights into factors influencing listing prices and market dynamics. Overall, removing rows with missing values in these key columns enhances the overall quality and usability of the dataset for analysis and decision-making purposes.

Remove Unnecessary Columns:

Removing the 'id' and 'host_id' columns from the dataset was done to streamline the data and reduce unnecessary complexity in subsequent analyses. These columns contain unique identifiers assigned to each Airbnb listing and host, respectively, which do not provide meaningful insights into the factors influencing listing prices or market dynamics. By eliminating these columns, we simplify the dataset and focus on variables that are more relevant to the analysis, such as the name of the listing, the host's name, location, room type, price, and availability. This reduction in complexity improves the interpretability of the dataset and facilitates more efficient data processing and modeling. Ultimately, removing extraneous columns like 'id' and 'host_id' allows us to concentrate on the essential features that drive pricing decisions and guest preferences in the Airbnb marketplace.

Rename 'Name' Column to 'Hotel Name':

Changing the column name from 'name' to 'hotel_name' was done to enhance clarity and consistency in data representation. The term 'hotel_name' more accurately reflects the nature of the information contained within the column, which pertains to the names of accommodations listed on Airbnb. This change not only improves the descriptive accuracy of the dataset but also ensures consistency with industry terminology, making it easier for users to understand and interpret the data. By adopting a standardized naming convention, we facilitate smoother data handling processes and reduce the likelihood of confusion or misinterpretation during analysis. Additionally, this change aligns with best practices for data organization and documentation, contributing to the overall usability and effectiveness of the dataset for analytical purposes.

Handle Missing Values in 'Last Review' and 'Reviews per Month' Columns:

Filling missing values in the 'last_review' column using the forward fill method and in the 'reviews_per_month' column with the median value was essential to ensure data completeness and integrity. In the case of the 'last_review' column, using the forward fill method allows us to propagate the last observed non-null value forward to fill missing entries. This approach is appropriate for temporal data, such as review dates, where adjacent values are likely to be similar. By doing so, we maintain the temporal sequence of reviews and avoid introducing biases into the analysis. Similarly, filling missing values in the 'reviews_per_month' column with the median value ensures that the distribution of review frequencies remains representative of the dataset as a whole. Using the median, a robust measure of central tendency, helps mitigate the impact of outliers and preserves the overall distribution of review frequencies. Together, these data imputation techniques enhance the completeness and reliability of the dataset, enabling more accurate and comprehensive analyses of Airbnb listing data.

Convert 'Last Review' to DateTime Format:

Converting the 'last_review' column to datetime format was necessary for proper date handling and analysis. By representing dates as datetime objects, we enable more accurate and intuitive manipulation, calculation, and comparison of temporal data. This conversion allows us to perform various date-related operations, such as calculating time intervals, extracting specific components (e.g., year, month, day), and filtering data based on temporal criteria. Moreover, datetime format standardized the representation of dates, ensuring consistency and compatibility with other date-related functions and libraries. This enhances the reliability and interpretability of the dataset, enabling more sophisticated analyses and insights into temporal patterns, trends, and dynamics within the Airbnb listings data. Overall, converting the 'last_review' column to datetime format facilitates more robust and meaningful analyses of temporal aspects of the dataset.

Remove Special Characters from 'Hotel Name': Removing special characters, digits, and symbols from the 'hotel_name' column was performed to standardize the data and enhance readability for analysis. Special characters, digits, and symbols in the 'hotel_name' column can introduce noise and inconsistency, making it challenging to accurately interpret and analyze the data. By eliminating these extraneous elements, we ensure that the 'hotel_name' column contains only relevant textual information, improving its clarity and coherence. Standardizing the data in this manner also facilitates easier comparison and manipulation of 'hotel_name' values, as well as enhances the overall quality and reliability of the dataset. Moreover, clean and standardized data promotes more accurate analysis and modeling outcomes, enabling stakeholders to derive meaningful insights and make informed decisions based on the Airbnb listings data. Therefore, removing special characters, digits, and symbols from the 'hotel_name' column plays a crucial role in preparing the data for comprehensive and reliable analysis.

Combine Latitude and Longitude into 'Location' Column:

Concatenating latitude and longitude values into a single 'location' column was carried out to simplify geographic representation and analysis of the dataset. By combining these two geographical coordinates into a single column, we create a more compact and easily interpretable representation of the spatial information associated with each Airbnb listing. This 'location' column provides a concise yet comprehensive description of the geographical position of each listing, facilitating spatial analysis and visualisation tasks. Moreover, having latitude and longitude values stored together in a single column streamlines data handling processes and reduces the complexity of working with geographical data. This simplification enhances the efficiency and effectiveness of geographic analyses, enabling stakeholders to gain valuable insights into spatial patterns, trends, and relationships within the Airbnb listings dataset. Therefore, concatenating latitude and longitude values into a 'location' column serves as a practical and convenient approach to representing geographic information for subsequent analysis and interpretation.

Normalize 'Availability_365' Column:

Normalizing the values in the 'availability_365' column to a range between 0 and 1 was undertaken to ensure uniformity and comparability across the dataset. By scaling the values within a standardized range, we mitigate the influence of magnitude differences and ensure that all data points are on the same scale. This normalization process facilitates fair comparisons and analyses by removing biases that may arise from disparate value ranges. Additionally, scaling the availability values between 0 and 1 makes them easier to interpret and incorporate into predictive models, as they now represent proportions or percentages of availability throughout the year. Overall, normalizing the 'availability_365' column enhances the consistency and reliability of analyses involving availability data, thereby improving the accuracy and effectiveness of subsequent modeling and decision-making processes.

Filter Outliers in 'Minimum Nights' Column:

Removing outliers in the 'minimum_nights' column by retaining only values below the 99th percentile was conducted to enhance data quality and modeling accuracy. Outliers, which are extreme values that deviate significantly from the majority of the data, can introduce noise and bias into statistical analyses and predictive models. By excluding these outliers, we ensure that the dataset is more representative of the underlying distribution of 'minimum_nights' values, thereby improving the robustness of subsequent analyses and modeling efforts. Moreover, focusing on the majority of the data while disregarding extreme values helps prevent skewed results and erroneous conclusions. This process of outlier removal promotes a more accurate understanding of the relationship between 'minimum_nights' and other variables, ultimately leading to more reliable insights and decision-making in the context of Airbnb listing data analysis.

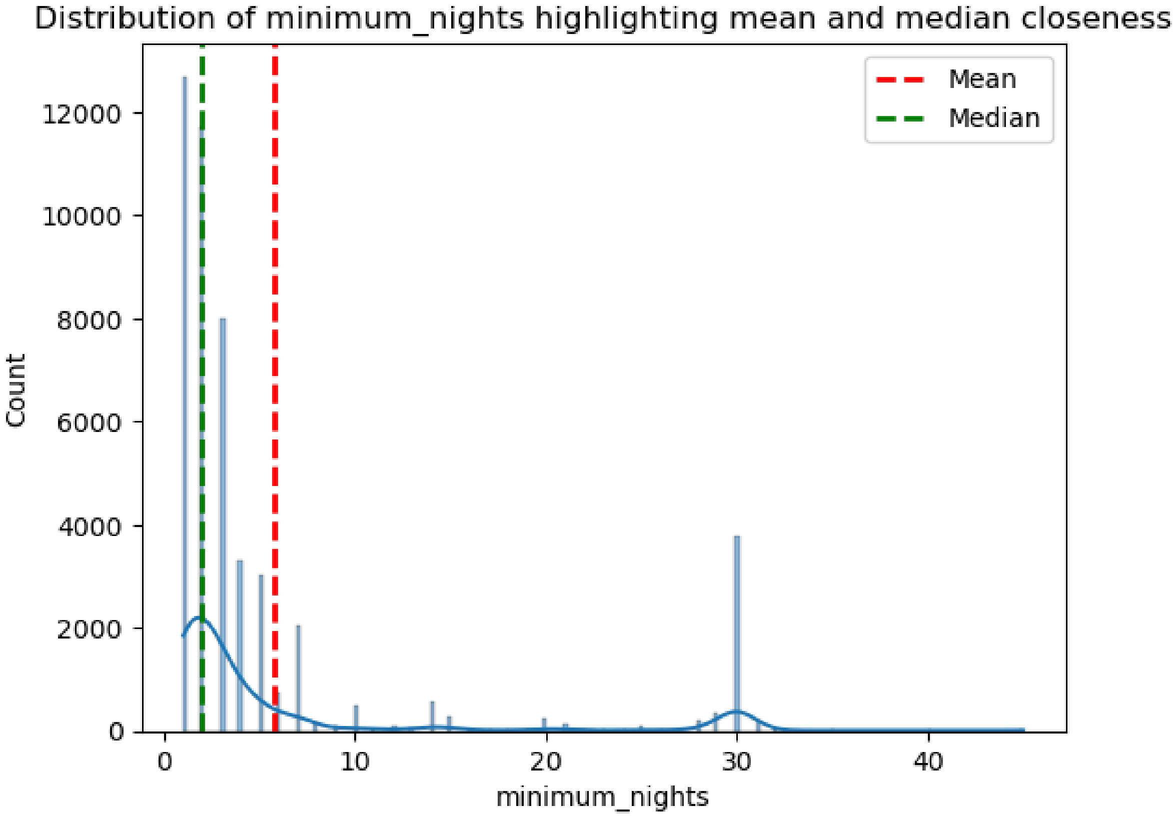
Encode Categorical Variables:

Encoding categorical variables such as 'neighborhood', 'neighbourhood_group', and 'room_type' using label encoding was performed to convert them into a numerical format suitable for analysis while retaining their categorical nature.

Label encoding assigns a unique numerical label to each category within a categorical variable, thereby allowing mathematical operations and analyzes to be performed on the data. This conversion enables machine learning algorithms to effectively process and interpret categorical features, which are essential for many analytical tasks. By preserving the categorical nature of the variables, label encoding maintains the interpretability of the data, ensuring that the encoded values still represent distinct categories rather than arbitrary numerical values. This approach facilitates seamless integration of categorical variables into analytical pipelines and enhances the accuracy and effectiveness of predictive modeling and data-driven decision-making processes.

Exploratory Data Analysis (EDA)

1.The distribution of 'minimum_nights' with overlaid mean and median lines:



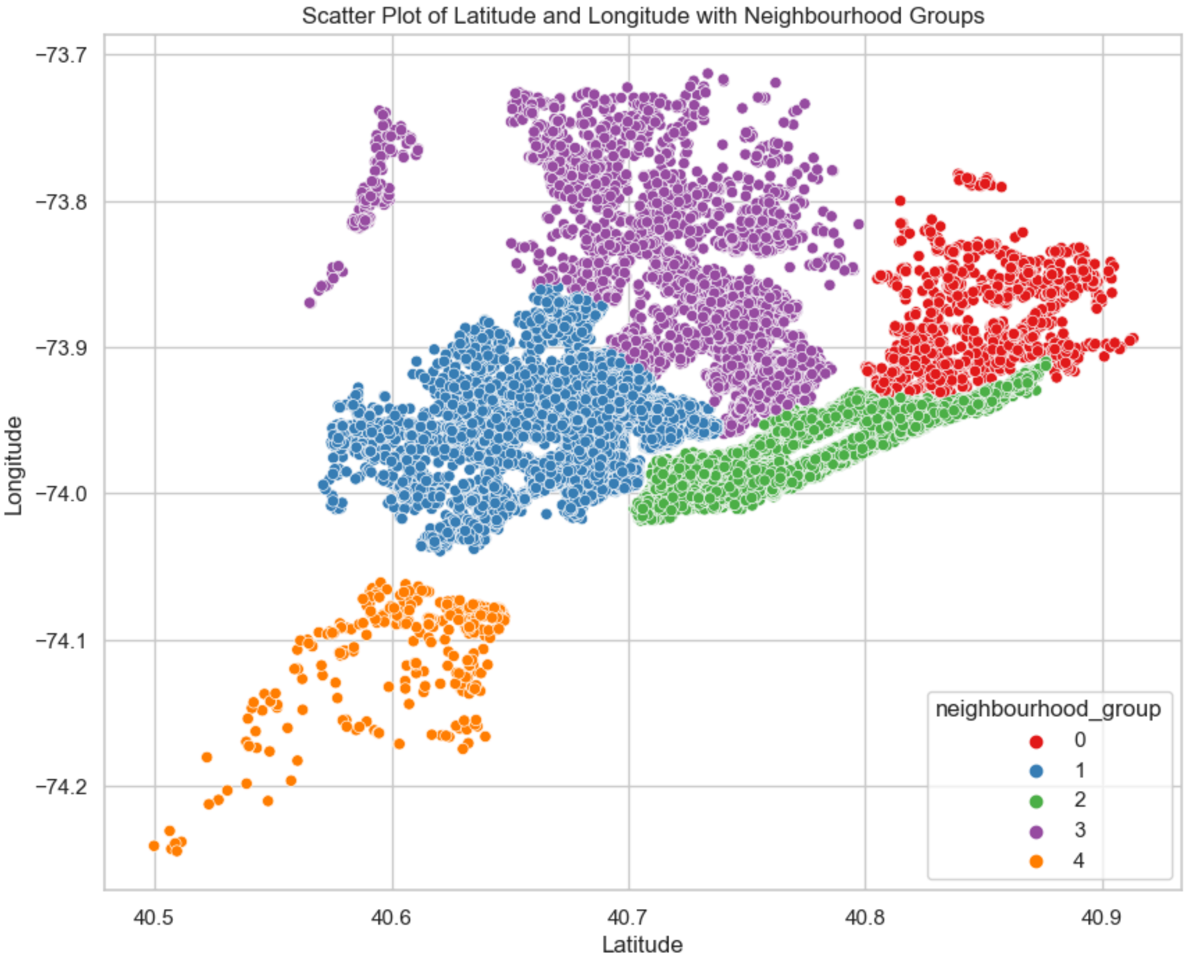
The plot shows a histogram of the "minimum_nights" variable, which represents the minimum number of nights a guest can book a particular listing. The x-axis represents the different values of "minimum_nights," while the y-axis shows the count of listings with those specific values.

The majority of listings have a minimum_nights value of 2 or 3, with a slight skew towards a minimum_nights value of 2. This indicates that most hosts prefer shorter stays for their listings.

The mean (average) of the minimum_nights distribution is approximately 6, while the median (middle value) is around 2. There is a small peak around a minimum_nights value of 30, which could be due to a specific type of listing or a particular host with a high minimum_nights requirement.

In summary, the distribution of minimum_nights shows that most listings have a minimum requirement of 2 or 3 nights, while there is a small portion of listings with significantly higher minimum_nights requirements. This could impact the accessibility and popularity of these listings, as guests might prefer shorter minimum stays.

2. Geographical distribution of listings with respect to latitude and longitude, coloured by neighborhood groups:

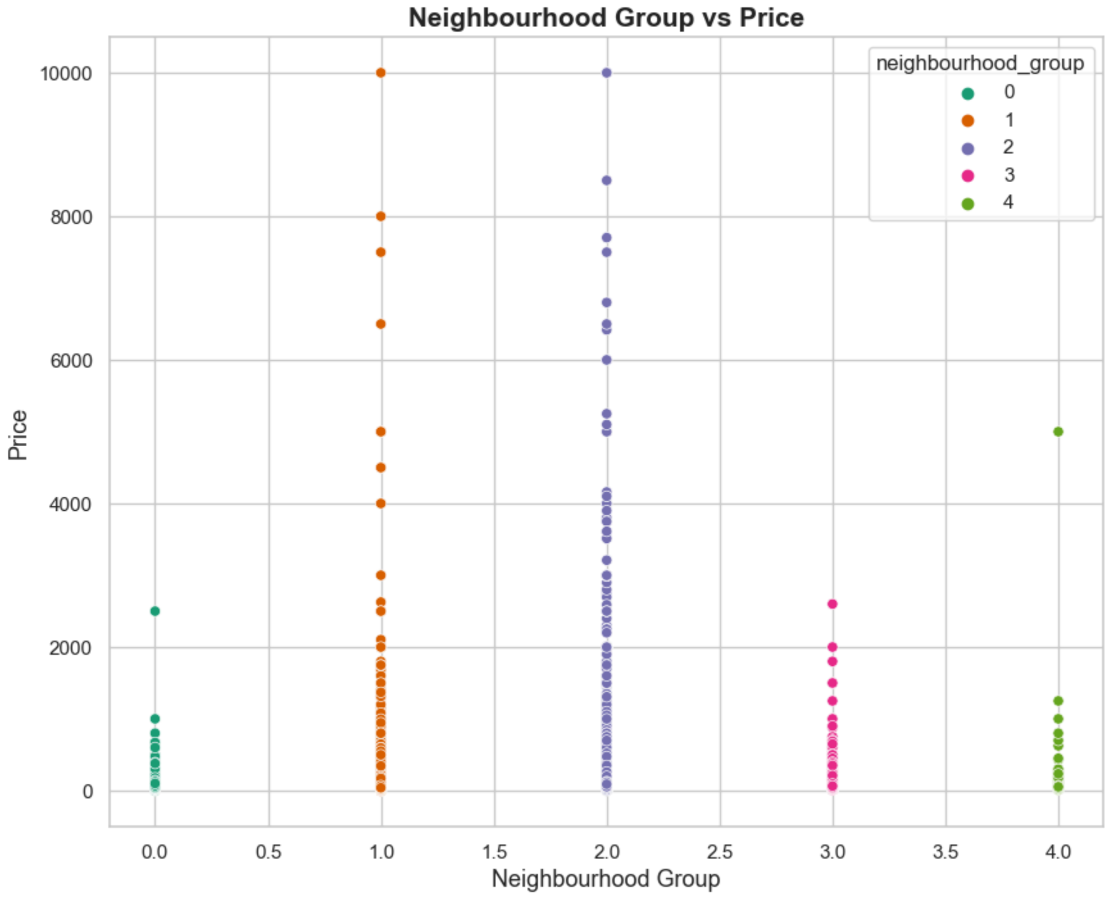


0: Bronx, 1: Brooklyn, 2: Manhattan, 3: Queens, 4: Staten Island

The provided scatter plot displays latitude and longitude coordinates, with different neighborhood groups in New York City (Bronx, Brooklyn, Manhattan, Queens, and

Staten Island) color-coded for easy identification. The Bronx neighborhood group occupies the northwestern part of the plot, characterized by longitude values ranging from -73.7 to -73.9 and latitude values ranging from 40.5 to 40.7. Meanwhile, the Brooklyn neighborhood group is situated in the southwestern part, with longitude values spanning from -73.7 to -74.2 and latitude values ranging from 40.5 to 40.8. Manhattan, located in the southern part of the plot, features longitude values between -73.9 to -74.0 and latitude values from 40.7 to 40.8. Queens, situated in the eastern region, has longitude values ranging from -73.7 to -74.0 and latitude values between 40.5 to 40.8. Lastly, the Staten Island neighborhood group appears in the southern part of the plot, with longitude values ranging from -74.0 to -74.2 and latitude values from 40.5 to 40.7. The plot also indicates that neighborhood groups are predominantly separated by the East River and the Hudson River, with varying densities of listings across different areas. This geographical insight aids in understanding the distribution of listings across New York City and can assist in making informed decisions when selecting accommodations.

3. Relationship between neighborhood groups and prices of listings:



0: Bronx, 1: Brooklyn, 2: Manhattan, 3: Queens, 4: Staten Island

The plot is a bar graph that shows the average price for different neighborhood groups in New York City.

The graph shows the average price for each neighborhood group, with the Bronx having the lowest average price and Manhattan having the highest average price.

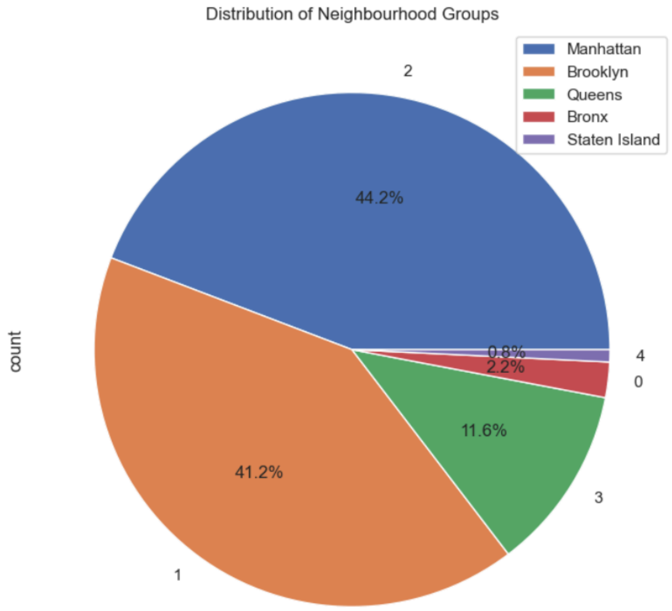
It shows that the difference in average price between the Bronx and Manhattan is significant, with Manhattan having a much higher average price compared to the Bronx.It shows the difference in average price between Queens and Brooklyn is small, with Queens having a slightly lower average price compared to Brooklyn.

The graph shows that Staten Island has a higher average price compared to the Bronx and Queens, but a lower average price compared to Brooklyn and Manhattan.

The graph is useful for understanding the price differences between different neighborhood groups in New York City, and can help travelers make informed decisions when booking accommodations.

The above plot is a bar graph that shows the average price for different neighborhood groups in New York City. The graph shows that the Bronx has the lowest average price, while Manhattan has the highest average price. The difference in average price between the Bronx and Manhattan is significant, with Manhattan having a much higher average price compared to the Bronx. The difference in average price between Queens and Brooklyn is small, with Queens having a slightly lower average price compared to Brooklyn. Staten Island has a higher average price compared to the Bronx and Queens, but a lower average price compared to Brooklyn and Manhattan. This information is useful for understanding the price differences between different neighborhood groups in New York City and for making informed decisions when booking accommodations.

4. Distribution of Neighborhood Groups:



The pie chart illustrates the distribution of neighborhood groups in a dataset, presenting the number of listings in each borough of New York City.Manhattan comprises the highest proportion of listings, accounting for 41.2% of the total. This aligns with Manhattan's status as a prominent tourist destination and a central hub for business and entertainment. Following Manhattan, Brooklyn ranks as the second most popular borough, representing 44.2% of total listings. This reflects Brooklyn's increasing popularity as a tourist destination, characterized by its diverse neighborhoods.

Staten Island exhibits a relatively small percentage of listings, comprising only 0.8% of the total. However, considering Queens' expansive land area and diverse population, its low percentage may not fully indicate its potential as a destination.The Bronx presents an even smaller percentage of listings, contributing only 2.2% to the total. The Bronx's location on the mainland and perception as less tourist-friendly compared to other boroughs may contribute to its lower listing percentage.Queens records the smallest percentage of listings, representing only 11.6% of the total. Staten Island, being the least populous borough and often overlooked as a tourist destination, may explain its limited listing percentage.

In summary, the data underscores that Manhattan and Brooklyn hold the highest number of listings, while Queens, the Bronx, and Staten Island exhibit considerably fewer listings. This distribution may be influenced by factors such as popularity, accessibility, and tourism infrastructure.

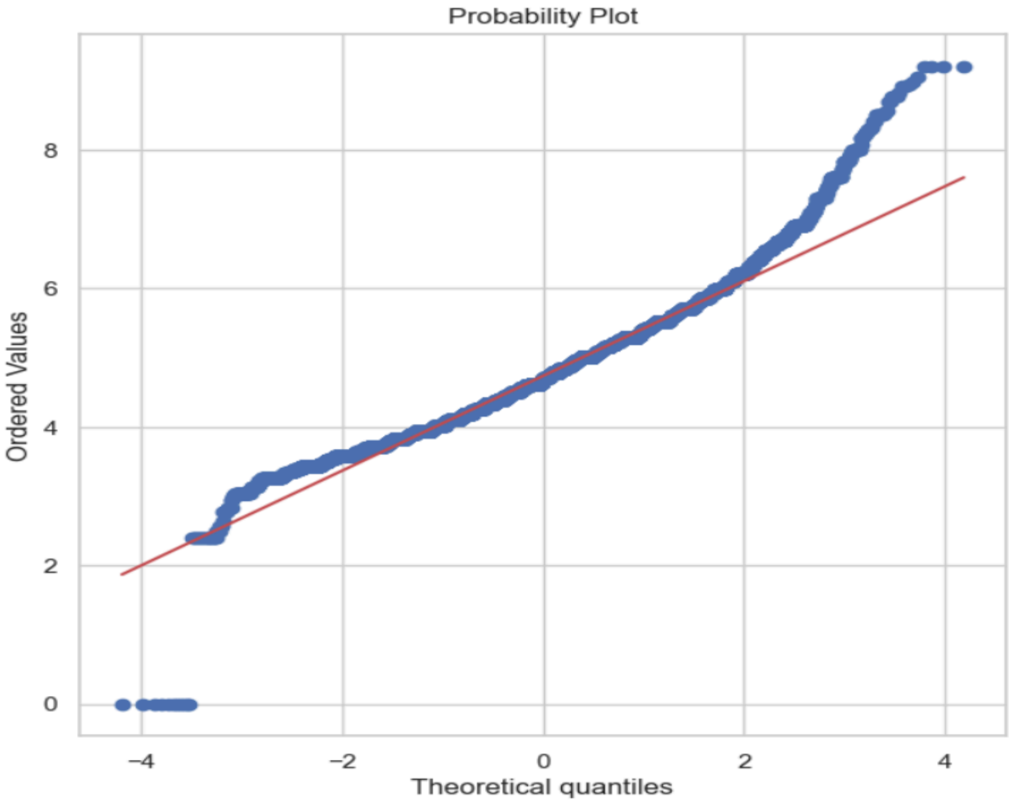
5. Probability Plot:

The below plot is a probability plot, specifically a Q-Q plot (Quantile-Quantile plot), which compares the distribution of a dataset to a theoretical distribution. In this case, the theoretical distribution is likely a normal distribution.

The Q-Q plot shows a linear pattern, which suggests that the dataset being plotted follows a normal distribution.The majority of the data points fall along the diagonal line, indicating that the data is well-approximated by a normal distribution.

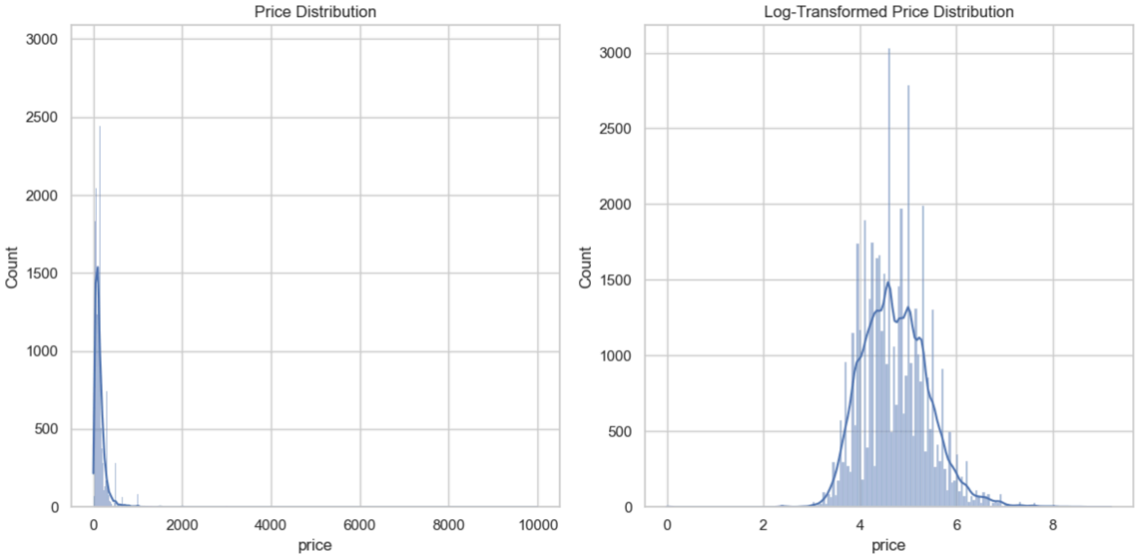
There are a few data points that deviate from the diagonal line, particularly in the lower left and upper right corners of the plot. These points may represent outliers or extreme values in the dataset.

The plot shows a concentration of data points in the center, which is consistent with a normal distribution.The plot suggests that the dataset has a mean and standard deviation that are similar to a normal distribution.



In summary, the Q-Q plot suggests that the dataset being plotted follows a normal distribution, with some deviations in the form of outliers or extreme values. This information could be useful for understanding the distribution of a dataset and identifying any unusual or unexpected values.

6. Price Distribution & Log-Transformed Price Distribution:



Upon applying a logarithmic transformation to the prices of listings in various neighborhood groups in New York City, a notable change in distribution pattern emerges. Initially, the bar graph displayed diverse price distributions among neighborhood groups, with some groups featuring substantially higher prices than others. However, post-transformation, prices appear more uniformly distributed across neighborhood groups. This is evident from the bar graph, where the differences in price distributions among groups appear less pronounced, and price variations across groups seem more consistent. The transformation effectively mitigates price distribution disparities, resulting in a more balanced representation of prices across neighborhood groups.

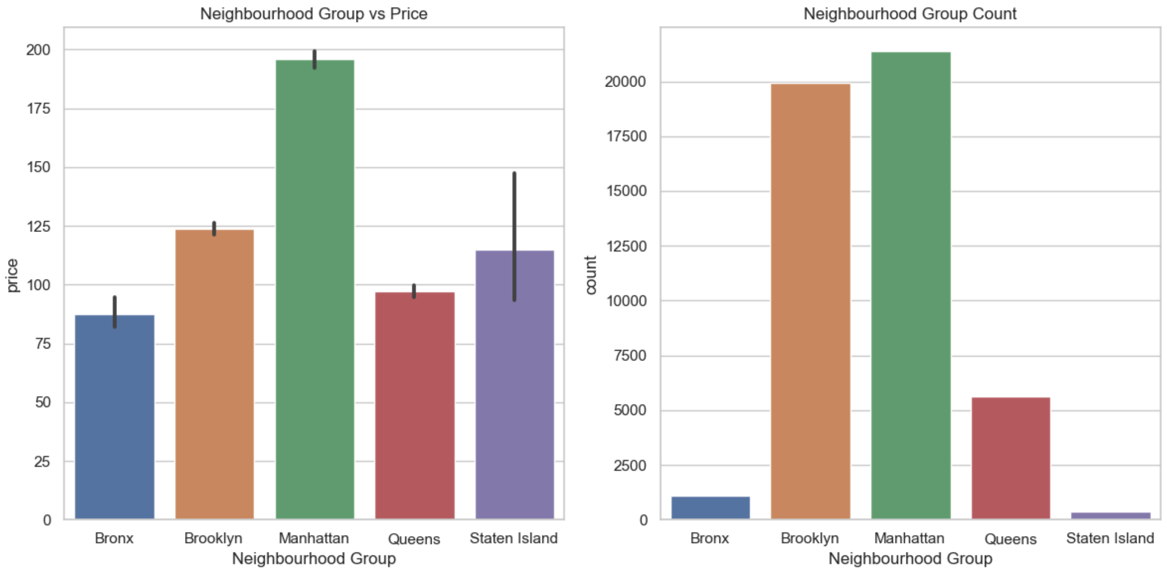
The uniform price distribution post-transformation indicates successful normalization of the price data. By employing a logarithmic scale, the transformation compresses price ranges, particularly for neighborhoods with initially higher prices, while simultaneously expanding ranges for those with lower prices. This compression and expansion yield a more equitable price distribution across all groups, aligning better with assumptions of statistical analyses such as linearity and homoscedasticity, thereby enhancing subsequent analytical robustness and reliability.

Moreover, the uniform price distribution suggests the logarithmic transformation reduces the influence of outliers on overall price distribution. Outliers, representing extreme values or anomalies, can skew distributions and affect statistical accuracy. Logarithmic transformation attenuates outlier impact by compressing their values, bringing them closer to data central tendencies. Consequently, the transformed data exhibits a more consistent and stable price distribution across neighborhood groups, offering a more accurate representation of New York City's real estate market trends.

In summary, the provided image illustrates price distribution across different New York City neighborhood groups before and after logarithmic transformation. The transformation diminishes data skewness, enhancing visualization and analysis. The Bronx has the highest number of free or unpriced listings, while Brooklyn and Manhattan feature the highest overall number of listings.

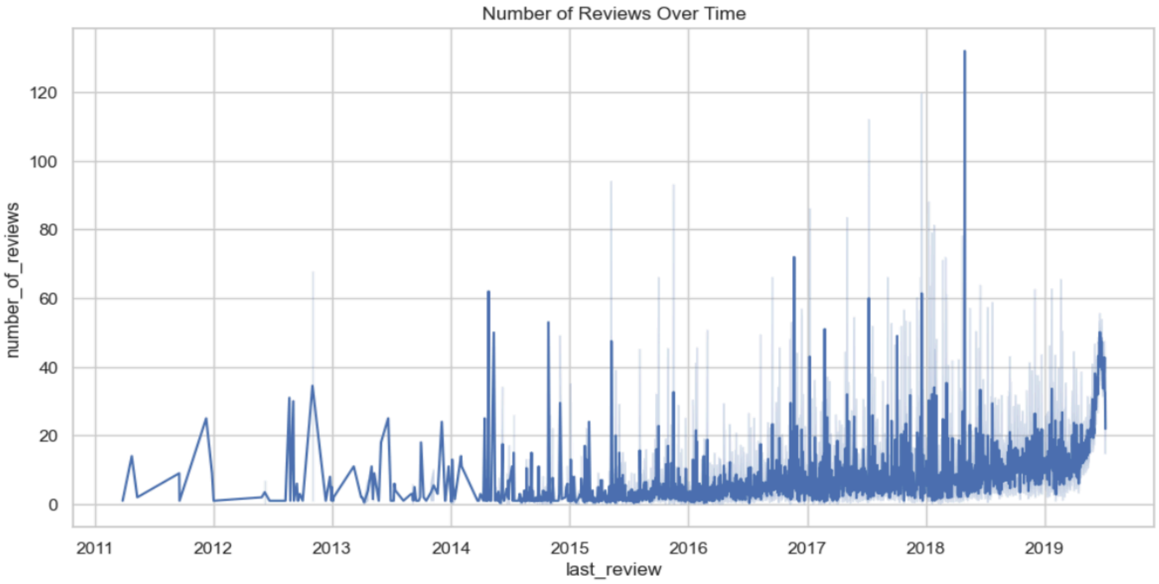
7. The relationship between neighborhood groups and price, as well as the count of listings per neighborhood group.

In our analysis, we observe two distinct visualizations: a plot depicting the mean price across different neighborhoods and a bar plot illustrating the count of neighborhoods within each borough. Interestingly, Staten Island exhibits the lowest count of neighborhoods, while Manhattan boasts the highest count, indicating a significant variation in the distribution of neighborhoods across boroughs.



Despite Brooklyn having a greater number of neighborhoods compared to Staten Island, their mean prices are nearly similar. This suggests that while Brooklyn may have a larger residential presence, the average pricing within its neighborhoods aligns closely with those of Staten Island, possibly indicating comparable socioeconomic factors or market dynamics influencing housing prices. This insight underscores the complexity of neighborhood dynamics within New York City, where boroughs with differing neighborhood counts may still exhibit similar average pricing trends, reflecting the nuanced interplay of various socio-economic and geographical factors within the real estate market.

8. Number of Reviews Over Time

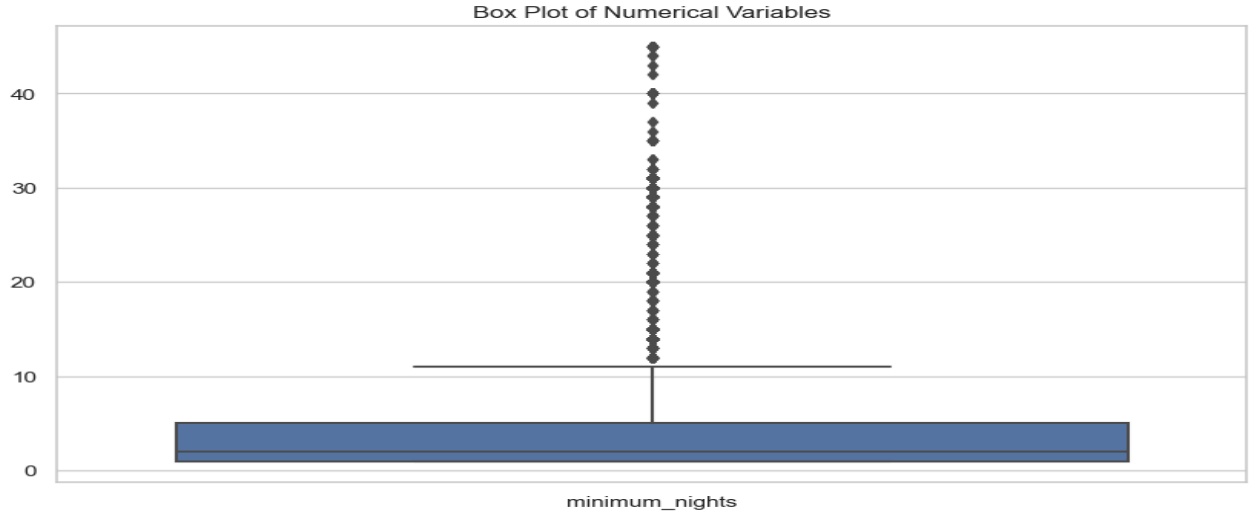


The time series analysis of Airbnb reviews relative to the last review date indicates a gradual increase in review numbers over the observed period, reflecting a rising level of engagement with Airbnb listings. Starting from an initial baseline, the reviews show a consistent upward trend, suggesting growing popularity and utilization of Airbnb accommodations. Factors such as expanded market reach, improved user experience, and enhanced listing visibility may be contributing to this trend.

Notably, a peak in activity is identified in the mid-2018 to 2019 timeframe, characterized by the highest number of reviews recorded. This surge may be influenced by seasonal travel demand, special events in New York City, promotional campaigns, or external factors like economic conditions and tourism trends. Overall, the analysis highlights the dynamic nature of Airbnb usage, with an upward trajectory in review numbers and a significant peak in activity during the specified period.

Understanding these temporal patterns can offer valuable insights for hosts, property managers, and policymakers aiming to optimize offerings, improve guest experiences, and capitalize on periods of increased demand in the New York City Airbnb market.

9. Box Plot of Numerical Variables

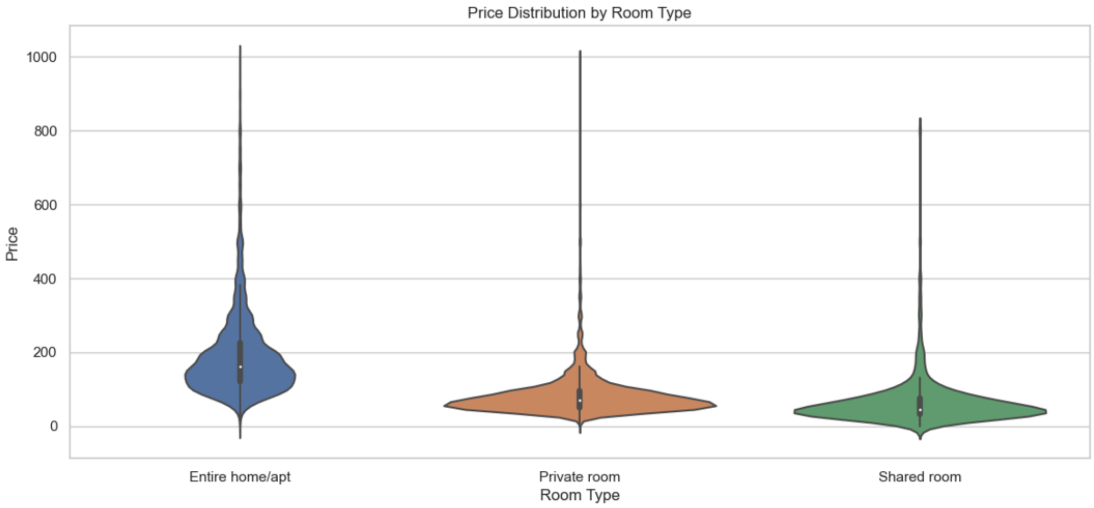


The box plot of minimum nights with an interquartile range (IQR) of 4 nights and a median of 2 reveals several key insights. The IQR represents the range within which the middle 50% of the data falls, while the median indicates the central tendency of the data.

In this box plot, the majority of listings have minimum nights requirements that fall within the range of approximately 0 to 6 nights, as indicated by the length of the box (IQR). The median value of 2 nights suggests that half of the listings require guests to stay for 2 nights or fewer, while the other half have minimum night requirements above this threshold.

However, there are notable outliers in the dataset, particularly at the higher end of the spectrum. Listings with minimum night requirements exceeding 40 nights are evident as individual data points beyond the upper whisker of the box plot. These outliers represent a minority of listings but may have important implications for certain types of travelers or specific rental scenarios.

10. Price Distribution by Room Type

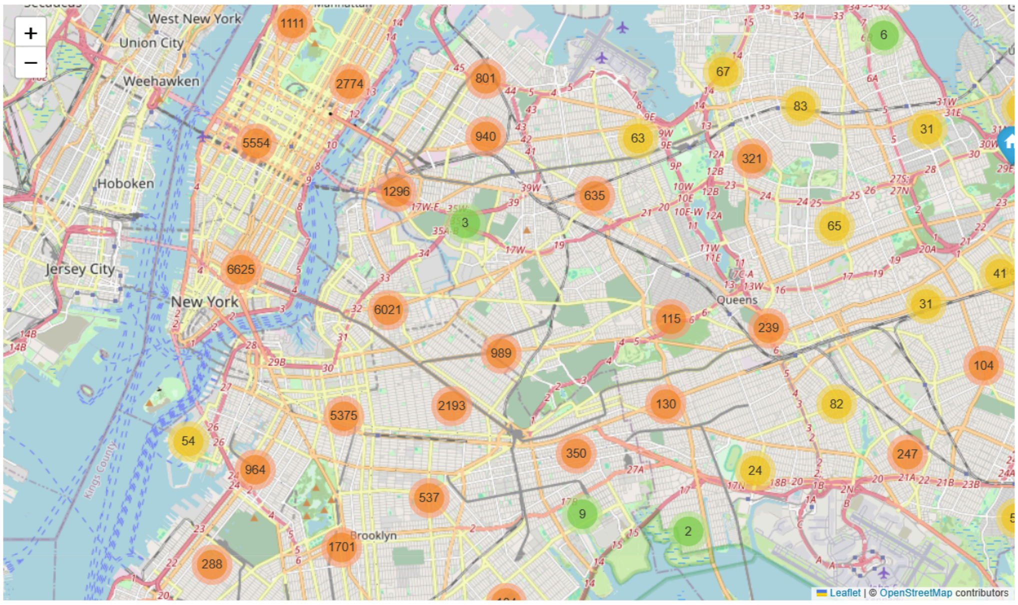


The provided plot shows a violin plot that compares the price distribution for three different room types: "Entire home/apt", "Private room" and “Shared room".

The "Entire home/apt" room type has a higher average price compared to the others.The "Entire home/apt" room type has a wider range of prices,The "Private room" and “Shared room"room type has a narrower range of pricesThe "Entire home/apt" room type has a higher number of high-priced listings and "Shared room" room type has a higher number of low-priced listings.

11. A map displaying the distribution of listings based on latitude and longitude.

This interactive plot features a dynamic geo-map displaying the geographical distribution of hotels. Users can easily zoom in and out of the map to explore various areas and observe the concentration of hotels within each region. The map provides a comprehensive overview of hotel locations, allowing users to visually assess the density of hotels across different areas and gain insights into the spatial distribution of accommodation options.This interactive feature enhances the user experience by enabling seamless navigation and exploration of hotel clusters and distribution patterns.



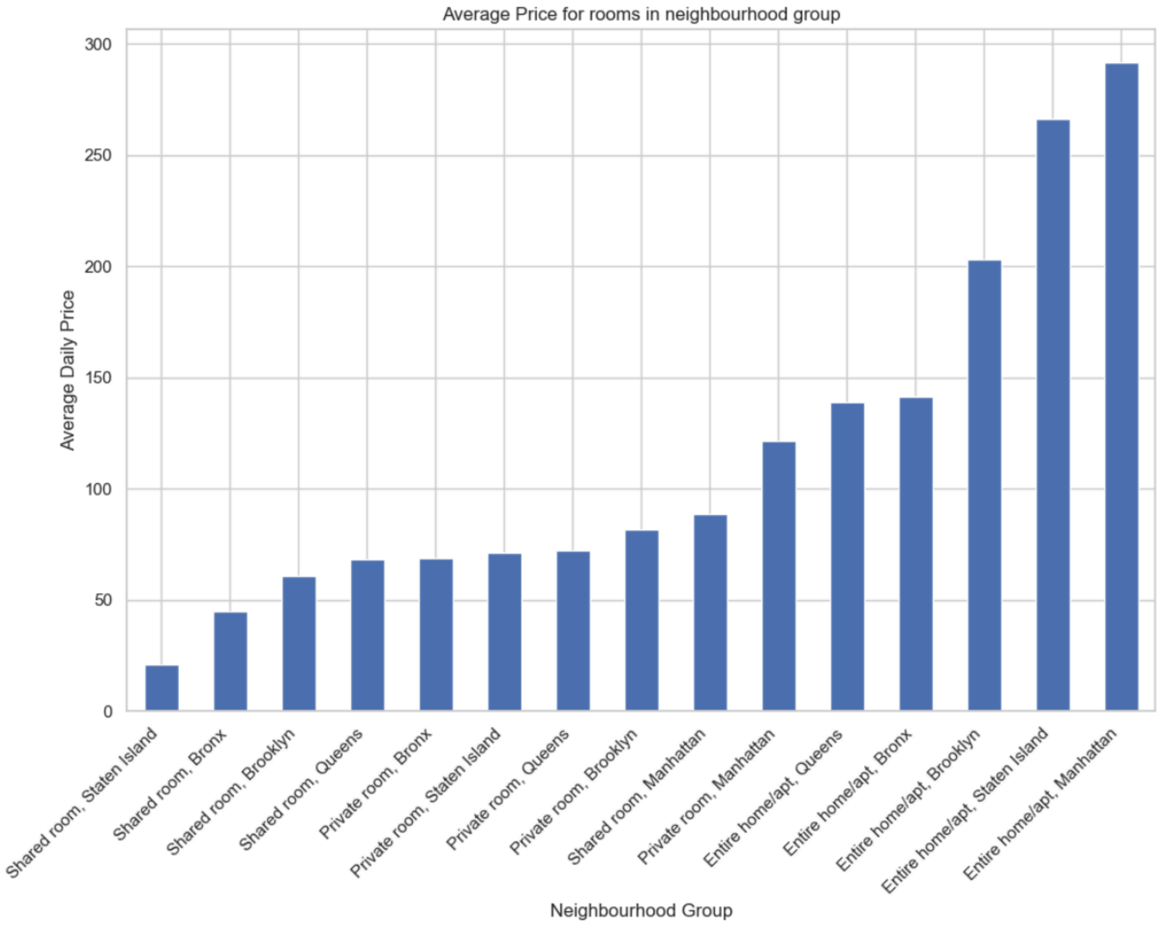
Whether zooming in to analyze specific neighborhoods or zooming out for a broader perspective, users can efficiently visualize and evaluate the abundance of hotels in different geographical areas. Overall, this interactive geo-map serves as a valuable tool for understanding the spatial dynamics of the hotel industry and facilitating informed decision-making for travelers, researchers, and industry professionals alike.

12. Average Price for rooms in neighborhood group:

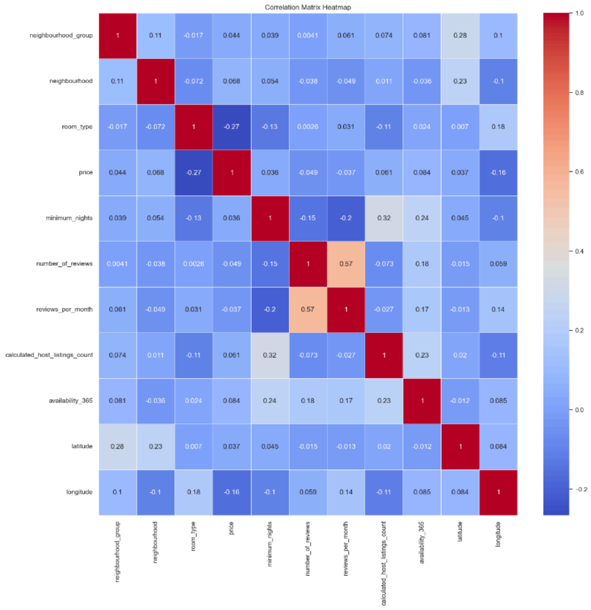
The bar graph illustrates the comparison of average daily prices for various room types and neighborhood groups in New York City. Notably, the "Entire home/apt" room type exhibits the highest average daily price and widest price range, with values ranging from approximately $140 to $290. In contrast, the "Private room" type demonstrates a narrower price range, spanning approximately $70 to $120. The "Shared room" type features the lowest prices, with values ranging around $70.

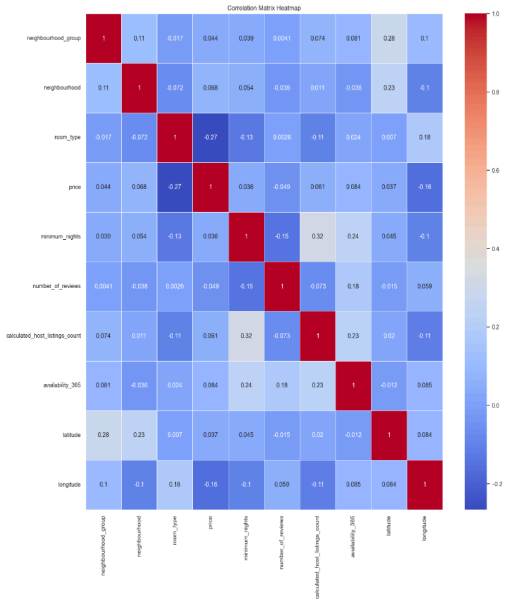
Regarding neighborhood groups, Manhattan commands the highest average daily price, trailed by Brooklyn, Queens, Bronx, and Staten Island. Conversely, Staten Island records the lowest average daily price, followed by the Bronx, Queens, Brooklyn, and Manhattan. The discrepancy in average daily prices between the highest and lowest-priced neighborhood groups is substantial, underscoring Manhattan's premium prices and Staten Island's affordability.

The bar graph serves as a valuable tool for comprehending the pricing differentials across room types and neighborhood groups in New York City, aiding travelers in making informed accommodation choices.



13. Correlation Matrix Heatmap:





The strong positive correlation of 0.57 between the variables "number_of_reviews" and "reviews_per_month" suggests that as the number of reviews increases, there is a corresponding increase in the reviews per month. This indicates a consistent pattern where accommodations receiving more reviews tend to have a higher monthly review rate, and vice versa. However, in statistical analysis, high correlation between two variables may lead to multicollinearity issues, which can affect the accuracy and reliability of predictive models. Therefore, to mitigate this, it might be prudent to drop one of the correlated variables, in this case, "reviews_per_month," to enhance the robustness of the analysis and avoid redundancy in the information captured by both variables.

After dropping "reviews_per_month," it is advisable to recompute the correlation matrix to reassess the relationships between the remaining variables. This step ensures that the analysis is based on independent and non-redundant information, providing a clearer understanding of the factors influencing the number of reviews. The updated correlation matrix will help in identifying any new patterns or relationships between variables, contributing to a more accurate interpretation of the data and facilitating more reliable statistical modeling if necessary.

14. Pairwise relationships between selected numerical variables.

The provided pairplot of table with several variables related to a dataset. The variables include price, minimum nights, number_of_reviews, availability_365, and reviews_per_month.

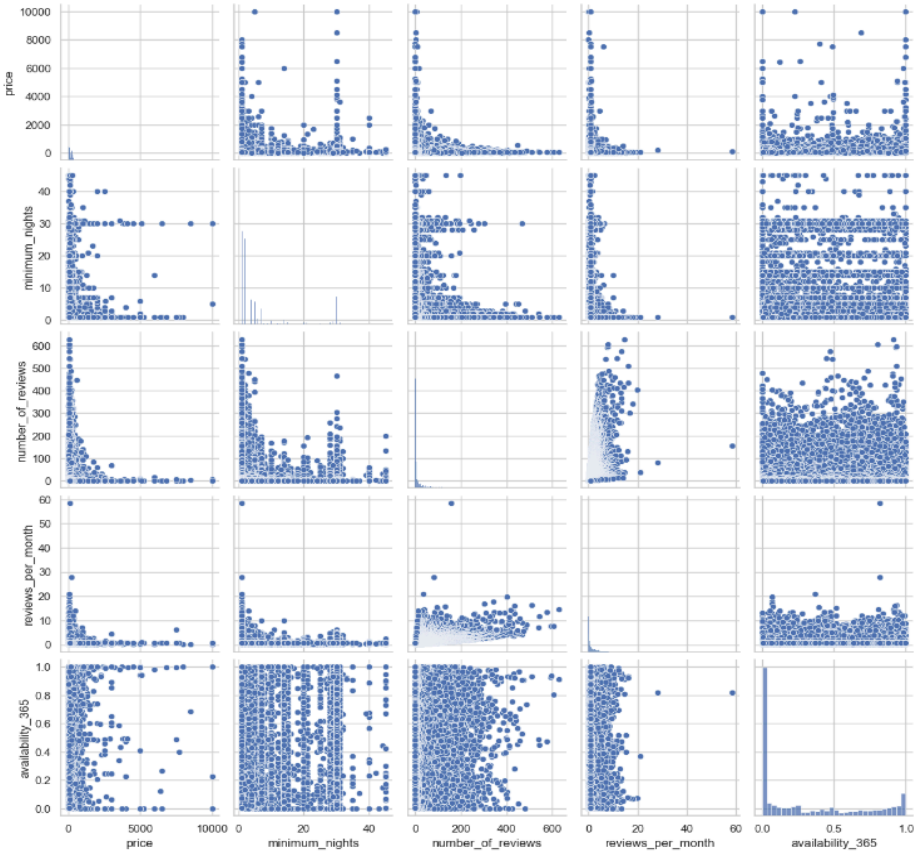
The price variable ranges from 2000 to 10000, indicating a wide range of prices in the dataset.

The minimum nights variable ranges from 0 to 40, suggesting that some listings have no minimum night requirement while others require up to 40 nights.

The number_of_reviews variable ranges from 0 to 600, indicating that some listings have very few reviews while others have many.

The availability_365 variable ranges from 0 to 1.0, suggesting that some listings have no availability for the entire year while others are available for the entire year.

The reviews_per_month variable ranges from 0 to 600, indicating that some listings receive very few reviews per month while others receive many.



In summary, the table provides information on several variables related to a dataset of listings. The variables show a wide range of values, indicating that there is a lot of variability in the dataset. There is no clear pattern or correlation between the variables, suggesting that they may not be strongly related. The information in the table could be useful for understanding the characteristics of the dataset and for identifying potential trends or patterns in the data.