Oklahoma State University

Spears School of Business

MSIS-5193-71999

Programming for Data Science and Analytics I

Prof. Xiao Luo, PhD

Airbnb Analysis

By

Srivarshika Gadde – A20450847

Kishore Madanagopal – A20450297

Somunath Reddy Sirasanambeti – A20446153

Saidabi Karumanchi – A20443812

12-01-2023

Table of Contents

ABSTRACT

This project presents a thorough analysis of Airbnb listings, utilizing Python for data extraction, preprocessing, exploratory analysis, and modeling. The primary objective is to uncover trends, pricing patterns, and determinants of customer satisfaction within Airbnb listings. Utilizing a comprehensive dataset from Airbnb, we employed Python tools such as Pandas for data manipulation, Matplotlib and Seaborn for visualization.

The analysis commenced with an extensive data cleaning phase, addressing missing values and data inconsistencies, followed by an exploratory data analysis (EDA) to gain initial insights. Key statistical measures were computed, and visualizations were generated to understand the distribution and relationships between various features.

Subsequent to the EDA, we focused on feature engineering to derive new insightful attributes and prepare the data for modeling. Various machine learning models were experimented with, aiming to predict listing prices and customer satisfaction levels. These models were rigorously evaluated using standard performance metrics.

The findings from this project provide valuable insights into the factors influencing Airbnb pricing and customer satisfaction, such as location, amenities, and host reputation. These insights can aid both hosts in optimizing their listings and customers in making informed choices. Additionally, the project demonstrates the efficacy of Python in handling and analyzing large-scale datasets, making it a valuable resource for data analysts and researchers in the field of hospitality and accommodation services.

This report not only details the methodologies and findings but also discusses the limitations and potential future extensions of the analysis, providing a comprehensive view of the project and its implications in the domain of data-driven decision-making in the sharing economy.

1. Introduction

In the era of dynamic and ever-changing travel accommodations, Airbnb has undeniably emerged as a revolutionary disruptor, fundamentally altering the way people engage with and explore new destinations. This transformative platform has reshaped the hospitality landscape, offering unique and personalized experiences that extend beyond the traditional confines of hotels.

Our research, which explores the dynamic "New York City Airbnb Open Data," is a monument to the analytical power of Python and its powerful libraries, which include pandas, numpy, matplotlib, and Seaborn. Our goal was very clear: we had to break down the dataset's complexity, understand its core features, and extract valuable information that would help us understand the spirit of Airbnb in New York City.

A thorough exploratory data analysis (EDA) set the groundwork for our investigation. Equipped with descriptive statistics and a range of visual aids, we set out to unravel the connections and patterns present in the dataset. Our investigations included the intricate link between price and the minimum number of nights, the dance between availability and pricing, and the interaction between price and the number of reviews.

Each facet of our analysis unfolded with precision and clarity. Our code, a manifestation of robustness and adaptability, was crafted to ensure the repeatability and scalability of our endeavors. The journey was not merely a technical exercise; it was a narrative, with detailed interpretations accompanying each visualization, providing a rich context for our discoveries.

This document extends an invitation—a portal into our exploration. We encourage you to traverse the intricacies of our approach, engage with our findings, and emerge with a profound understanding of the Airbnb market in the vibrant tapestry of New York City. The insights contained within these digital pages are more than just observations; they represent a deeper comprehension of the dynamics that shape the Airbnb experience in one of the world's most iconic cities. Come explore with us. The heartbeat of New York City's Airbnb market awaits your discovery.

2. Literature Review

2.1 "The evolution of Airbnb research: A systematic literature review using structural topic modeling" (Heliyon. 2023 Jun):

This paper employs advanced text-mining techniques, specifically structural topic modeling, to comprehensively assess the evolution of research on Airbnb up to the year 2022. By leveraging these techniques, the study aims to uncover emerging research concerns and trends within the Airbnb literature. The application of structural topic modeling allows for a nuanced understanding of the evolving landscape, identifying key themes, and shedding light on the dynamics of scholarly discussions. The findings of this paper could provide valuable insights for researchers, policymakers, and industry practitioners seeking to understand the trajectory of Airbnb research and its implications.

2.2 "Airbnb research: an analysis in tourism and hospitality journals" (Luisa Andreu, Enrique Bigne, Suzanne Amaro, Jesús Palomo. 11 February 2020):

This research employs bibliometric methodologies to analyze Airbnb-related publications within the context of tourism and hospitality journals. By applying research performance analysis, the study not only identifies prominent journals but also pinpoints influential authors and countries contributing to the body of knowledge on Airbnb. This approach offers a systematic examination of the academic landscape surrounding Airbnb, providing a quantitative understanding of research patterns and highlighting areas of concentrated scholarly activity. Researchers, journal editors, and institutions can benefit from the insights provided by this paper to navigate the academic landscape of Airbnb research effectively.

2.3 "Airbnb: The future of networked hospitality businesses" (Jeroen Oskam, Albert Boswijk, March 2015):

This article, dated March 2015, takes a forward-looking approach by analyzing the nature of Airbnb as a phenomenon and predicting its potential growth over the subsequent five years. By exploring the anticipated effects on city destinations, hotels, and the broader tourism sector, the study aims to contextualize the history of networked hospitality. The authors synthesize cumulative implications,

providing businesses and local governments with strategic insights to navigate the evolving landscape. This foresight-oriented research offers a foundation for anticipating challenges and opportunities in the networked hospitality domain and guiding stakeholders in formulating adaptive strategies for the future.

3. Dataset

3.1 Source:

The dataset utilized in this report is sourced from Kaggle, a prominent data science competition platform and online community under Google LLC. Kaggle serves as a hub for data scientists and machine learning practitioners, providing access to a diverse range of datasets, collaborative environments, and a platform for sharing knowledge and insights.

3.2 Dataset Overview:

The specific dataset under consideration focuses on Airbnb listings in New York City (NYC). This comprehensive dataset encapsulates a wealth of information related to individual Airbnb listings, offering valuable insights for analyses in the realms of tourism, hospitality, and urban data.

3.3 Key Variables:

The dataset encompasses various key variables, each contributing to a multifaceted understanding of Airbnb operations in NYC. Notable variables include:

- **ID:** Unique identifier for each Airbnb listing.
- **Name:** Descriptive name associated with the listing.
- **Host ID and Host Name:** Identifiers and names of the hosts managing the listings.
- **Neighborhood:** Information about the neighborhood where the listing is situated.
- **Geographical Coordinates (Latitude and Longitude):** Location coordinates of the listings.
- **Room Type:** Classification of the room (e.g., entire home, private room, shared room).
- **Price:** Cost associated with renting the listing.
- **Minimum Nights:** The minimum number of nights required for booking.
- **Number of Reviews:** Total number of reviews received for the listing.
- **Last Review and Reviews Per Month:** Details about the last review received and the average reviews per month.
- **Calculated Host Listings Count:** Count of listings associated with the host.
- **Availability 365:** Number of days the listing is available for booking in a year.

3.4 Dataset Characteristics:

- **Number of Rows:** 48895
- **Number of Columns:** 16

3.5 Data Provenance:

Given that the dataset is from Kaggle, it is crucial to acknowledge the platform's commitment to data quality, collaborative learning, and reproducibility. Kaggle often provides detailed documentation, discussion forums, and kernels (code notebooks) associated with datasets, facilitating transparency and shared understanding within the data science community.

3.6 Utilization:

This dataset serves as the foundation for our analysis, allowing us to explore trends, patterns, and relationships within the context of Airbnb listings in NYC. Through rigorous examination and interpretation of these variables, we aim to derive meaningful insights that contribute to a deeper understanding of the dynamics of the Airbnb market in this vibrant city.

4. Platform:

4.1 PyCharm Community Edition:

PyCharm is a highly popular Integrated Development Environment (IDE) designed specifically for Python development. Developed by JetBrains, it offers a wide range of features that make it suitable for various types of Python development, including web development, scientific computing, and data analysis. The community edition is a free and open-source version of PyCharm, which includes essential tools like an intelligent Python editor, code inspections, a graphical debugger, and version control integration.

Here are some additional details about PyCharm:

- **Intelligent Code Editor**: PyCharm's editor offers advanced features like code completion, error detection, and quick fixes, which significantly enhance coding efficiency.

- **Integrated Debugger and Test Runner**: The built-in debugger and test runner make it easier to identify and fix issues in your code. It supports visual debugging, which allows you to see variable values without printing them.

- **Remote Development Capabilities**: You can run, debug, and test your Python code on remote machines or virtual environments, which is particularly useful for working with large datasets or in a team environment.

- **Database Tools**: PyCharm Community Edition was utilized for the project's development due to its robust Python support, including intelligent code completion, on-the-fly error checking, quick fixes, and easy project navigation. This free, open-source IDE provided all the necessary tools for efficient Python coding and debugging, streamlining the development process without additional cost.

- **Version Control Integration**: It has seamless integration with version control systems like Git, Subversion, and Mercurial, making it easier to manage your codebase and collaborate with others.

- **Web Development Support**: With support for web development frameworks, PyCharm can be used for developing web applications, including front-end technologies like HTML, CSS, and JavaScript.

5. Methods

5.1 Data Loading and Inspection

5.1.1 Pandas for Data Loading :

- **pd.read_csv()**: This function is used to read the Airbnb dataset from a CSV file. It's a versatile function that handles various delimiters and types of data.
- **df.head()**: Displays the first few rows of the Data Frame. It's crucial for getting a quick overview of the data structure, column names, and types of data that you will be working with.
- **df.info()**: Provides concise information about a Data Frame. This includes the number of entries, columns, the data type of each column, and the number of non-null values. It's invaluable for identifying columns with missing values and understanding data types.

5.2 Data Cleaning and Preprocessing

5.2.1 Handling Missing Values :

- **df.isnull().sum()**: This calculates the sum of missing values in each column. Identifying missing data is essential for cleaning and preprocessing.
- **df.dropna()**: This method is used to drop rows or columns with missing values. Deciding how to handle missing data (removing it, imputing values, etc.) is critical for the integrity of your analysis.

5.2.2 Duplicate Detection and Removal:

- **df.duplicated()**: Identifies duplicate rows in the Data Frame, which is essential for ensuring the uniqueness of data entries.
- **df.drop_duplicates()**: Removes duplicate rows from the Data Frame, preventing skewed analysis results due to repeated entries.

5.3 Descriptive Statistical Analysis

5.3.1 NumPy for Statistics :

- Functions like **np.mean()**, **np.max()**, **np.min()**, **np.std()**, and **np.median()** are used to compute basic statistical metrics. These metrics provide insights into the central tendency, spread, and overall distribution of rental prices in the dataset.

5.4 Data Exploration

5.4.1 Unique Value Identification:

- **pd.unique()**: This function identifies unique values in specific columns (like 'neighbourhood_group' and 'room type'), aiding in understanding the diversity and categories present in the Airbnb listings.

5.5 Advanced Data Visualization

5.5.1 Seaborn and Matplotlib for Graphs:

- Creating bar plots using **sns.barplot()** and customizing them with labels and titles using Matplotlib functions. These visualizations are key for making comparisons and understanding distributions in a visually intuitive manner.
- **Plotly for Interactive Plots:**
- **px.bar()**: Generates interactive bar charts, which enhance the interactivity of data presentations, especially useful for web-based dashboards or interactive reports.

5.6 Aggregation and Grouped Analysis

5.6.1 Pandas GroupBy:

- **df.groupby()**: This method groups the data by specified criteria, followed by aggregation functions like **.max()** and **.mean()**. It's a powerful method for segmenting the data into subgroups and calculating statistics for each group, which is vital for comparative analysis.

5.7 Combining and Comparing Data

5.7.1 Dual-Axis Plotting:

- Creating dual-axis plots allows the comparison of two different data types or scales on the same graph, like room type count and average reviews, providing a more nuanced understanding of the relationships between different data elements.

5.7.2 Pivot Tables for Heatmaps:

- **df.pivot_table()**: Used to reorganize and reshape data, which is then visualized using **sns.heatmap()**. Heatmaps are excellent for displaying complex data matrices and uncovering patterns and correlations.

5.8. Distribution Analysis

5.8.1 Histograms for Frequency Distribution:

- **sns.histplot()**: This function plots histograms, which are fundamental for analyzing the frequency distribution of a variable (like the number of reviews). It helps in understanding the spread and skewness of the data.

6. Libraries Used

6.1 PANDAS:

Pandas is a popular open-source data analysis and manipulation library for Python. It provides a powerful set of tools for working with structured data, including the ability to read and write data from various file formats, perform data cleaning and preprocessing, handle missing data, merge and join data sets, and perform various types of data aggregation and transformation. The main data structures in Pandas are the Series and DataFrame. A Series is a one-dimensional array-like object that can hold any data type, while a Data Frame is a two-dimensional table-like data structure with rows and columns. Data Frames can be thought of as a collection of Series, with each Series representing a column of data. Pandas also provides a wide range of functions for data analysis and visualization, including statistical functions, time series analysis tools, and plotting tools. Overall, Pandas is a powerful tool for data analysis and is widely used in data science, machine learning, and other data-driven fields.

6.2 NUMPY:

NumPy is a Python library for numerical computing. It provides a high-performance multidimensional array object, along with tools for working with these arrays.

Some of the key features of NumPy include:

- **Array:** A fast, flexible container for large arrays of homogeneous data, including numerical data.
- **Broadcasting**: A powerful mechanism that allows NumPy to perform operations on arrays of different shapes and sizes.
- **Vectorized operations**: NumPy provides a suite of element-wise functions that operate on whole arrays, making it easy to perform complex numerical computations.
- **Linear algebra**: NumPy provides a suite of functions for linear algebra operations, including matrix multiplication, eigenvalue decomposition, and singular value decomposition.
- **Fourier transforms:** NumPy provides functions for computing fast Fourier transforms and other spectral analysis techniques.

NumPy is widely used in scientific computing, data analysis, and machine learning. It is a fundamental library in the Python scientific computing ecosystem, along with libraries like Pandas, Matplotlib, and Scikit-Learn.

6.3 SEABORN:

Seaborn is a Python data visualization library based on Matplotlib. It provides a high-level interface for creating informative and attractive statistical graphics. Seaborn builds on top of Matplotlib and integrates well with Pandas data structures, making it a popular choice for data scientists and analysts. Some of the key features of Seaborn include:

- **Easy data exploration:** Seaborn provides a few built-in datasets that can be easily loaded and visualized to help explore the data.
- **Statistical visualization:** Seaborn provides a wide range of statistical visualization tools, including heatmaps, bar plots, line plots, scatter plots, and more.
- **Customizable aesthetics:** Seaborn provides a variety of customizable options for controlling the appearance of the visualizations, such as color palettes, styles, and themes.
- **Integration with Pandas:** Seaborn integrates well with Pandas data structures, making it easy to visualize data stored in data frames.
- **Advanced visualization techniques:** Seaborn also provides advanced visualization techniques, such as faceting, which allows you to create multiple plots based on subsets of the data. Overall, Seaborn is a powerful and flexible data visualization library that can help you to quickly create informative and attractive visualizations for your data.

6.4 MATPLOTLIB

Matplotlib is a data visualization library in Python that is used to create static, interactive, and animated visualization. It provides a wide variety of tools for creating visualizations such as line charts, scatter plots, bar charts, histograms, heatmaps, and more. Matplotlib provides a wide range of customization options for visualizations, such as changing the colors, labels, markers, adding legends, and more.

6.5 Plotly Express (px)

Plotly.express is a module from Plotly, a versatile graphing library in Python used for creating interactive and visually appealing plots.

Here's more about these libraries:

- **Simplified Interface for Basic Plots:** Plotly Express provides a simple, declarative syntax for creating common types of plots. It's designed to be a high-level interface for rapid plotting.
- **Wide Range of Plots:** It supports a variety of plots, including scatter plots, line charts, bar charts, box plots, histograms, scatter matrix (pair plots), 3D charts, and geographical maps.
- **Interactivity:** One of the key features of Plotly Express is its built-in support for interactivity, such as hover tools, zooming, and panning.
- **Customization and Styling:** While being simpler to use, it still offers extensive customization options for plot aesthetics like colors, legends, and axis titles.
- **Ease of Use:** It’s particularly favored for its ease of use and the ability to create complex visualizations with minimal code.

7. Project framework

7.1 Data Preparation:

We began by downloading the dataset from Kaggle. The first step is to collect a large dataset of numbered values. The dataset should be diverse and include a variety of features, including room type, neighborhood, availability, and reviews.

7.2 Data Preprocessing:

The collected data is preprocessed to remove null and irrelevant information, such as special duplicate values and errors. Addressing missing values, inconsistent formatting, outliers, and errors in the dataset to ensure accuracy in analysis. The resulting text is normalized to ensure consistency across the dataset. This ultimately leads to an improvement in data quality.

7.3 Data visualization:

After data processing and cleaning up the data we made sure we have visualized them in a way that was more understandable and precise. We have used graphs and plots (like histograms, bar graphs and heatmaps) to visualize data for better understanding and to identify patterns, trends, and anomalies. Data visualization is an important skill in applied statistics and machine learning. Statistics does indeed focus on quantitative descriptions and estimations of data. Data visualization provides an important suite of tools for gaining a qualitative understanding.

7.4 Exploratory Data analysis:

Exploratory Data Analysis (EDA) in the context of Airbnb involves a set of procedures aimed at understanding the characteristics, patterns, and potential insights within Airbnb. This process is crucial for making informed decisions, whether for business strategy, market analysis, or enhancing user experiences.

8. Code explanation

Section 1: Importing the Required Libraries

- **Libraries**: Pandas, NumPy, Matplotlib, Seaborn, Plotly Express, and Plotly Graph Objects are imported. These libraries are essential for data manipulation, analysis, and visualization.
- **Data Loading**: The Airbnb dataset for New York City in 2019 is loaded into a DataFrame df.

Section 2: Initial Data inspection

- **Initial Exploration**: Displays the first few rows with df.head() to get an overview of the dataset. The df.info() function is used to summarize the DataFrame, providing information about column names, non-null counts, and data types.

Section 3: Handling Missing Values

- **Missing Values**: Calculates the number of missing values in each column and prints the result. Rows with missing values in the 'reviews_per_month' column are dropped, as this data might be crucial for the analysis.

Section 4: Identifying and Removing Duplicate Rows

- **Duplicates**: Identifies and prints duplicate rows, then removes them from the DataFrame to ensure data accuracy.

Section 5: Descriptive Statistics

- **Statistical Analysis**: Calculates basic statistics like average, maximum, minimum, standard deviation, and median for rental prices. These metrics provide an overview of the pricing landscape of Airbnb listings.

Section 6: Unique Values Analysis

- **Unique Values**: Extracts and prints unique values for the 'neighbourhood_group' and 'room_type' columns, providing insights into the variety of listings available.

Section 7: Data Visualization

- **Graph 1:** Visualizes the most expensive rentals by region.
- **Graph 2:** A histogram showing the distribution of properties per host.
- **Graph 3**: Two heatmaps representing average availability and price by neighborhood group.
- **Graph 4**: Visualizes a bar chart that represents average rental prices by region in New York.
- **Graph 5**: Visualizes a horizontal stacked bar with maximum rentals price by region and room type in New York.
- **Graph 6:** A histogram displaying the distribution of the number of reviews.
- **Graph 7**: Visualizes a combo chart that represents the room type count and average reviews by neighborhood group.

Purpose: These visualizations are used to explore different aspects of the Airbnb dataset, such as pricing, room types, and reviews across different neighborhoods. They use various plotting techniques and styles for effective data presentation.

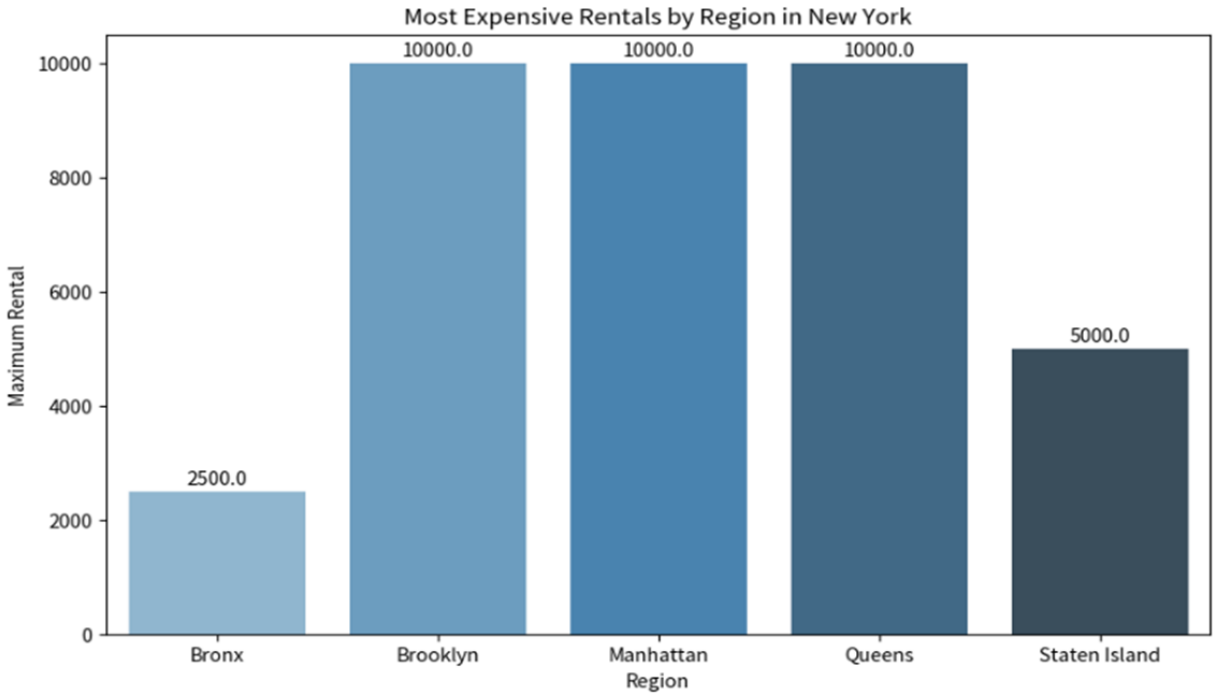
General Overview

The code provides a detailed analysis of Airbnb listings, including data cleaning, statistical analysis, and comprehensive visualizations.

- It uses a combination of descriptive statistics and graphical representations to uncover insights about the Airbnb market in New York City.
- The code is structured to first understand and clean the data, followed by analyzing and visualizing it in various ways to draw meaningful conclusions.

9. Results

Graph 1: Most Expensive Rentals by Region in New York.



"Most Expensive Rentals by Region in New York" is the title of the bar chart in this graph. It contrasts the highest rental costs in the Bronx, Brooklyn, Manhattan, Queens, and Staten Island regions of New York.

Here are the key points:

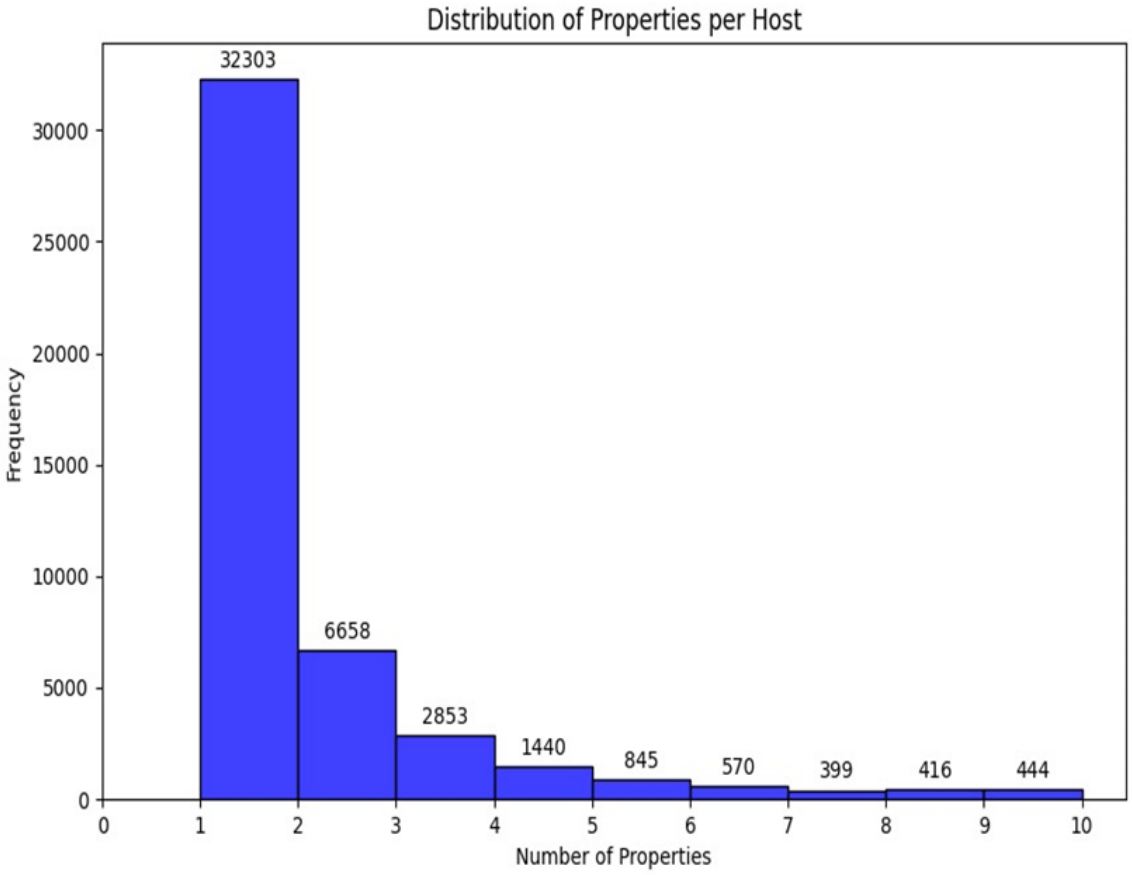
- The vertical axis (y-axis) represents the maximum rental price in dollars.
- The horizontal axis (x-axis) lists the five regions.
- Each bar represents a region and its height correlates to the maximum rental price for that area.

From the chart, we can observe the following:

- Manhattan and Queens have the highest maximum rental prices, both at $10,000, indicating that they are the most expensive regions for rentals in this dataset.
- Brooklyn follows with a maximum rental price of $10,000 as well, equal to Manhattan and Queens.
- Staten Island has a notably lower maximum rental price at $5,000.
- The Bronx has the lowest maximum rental price among the regions shown, at $2,500.

This bar graph is helpful for rapidly comparing the highest and lowest possible rental costs as well as determining which parts of New York are most and least expensive to rent. It is evident that rental prices are greater in Manhattan, Brooklyn, and Queens, and lower in the Bronx and Staten Island.

Graph 2: Distribution of Properties per Host



The displayed graph is called "Distribution of Properties per Host" and is a histogram. It shows how frequently hosts occur based on how many properties they have listed. The features are broken down as follows:

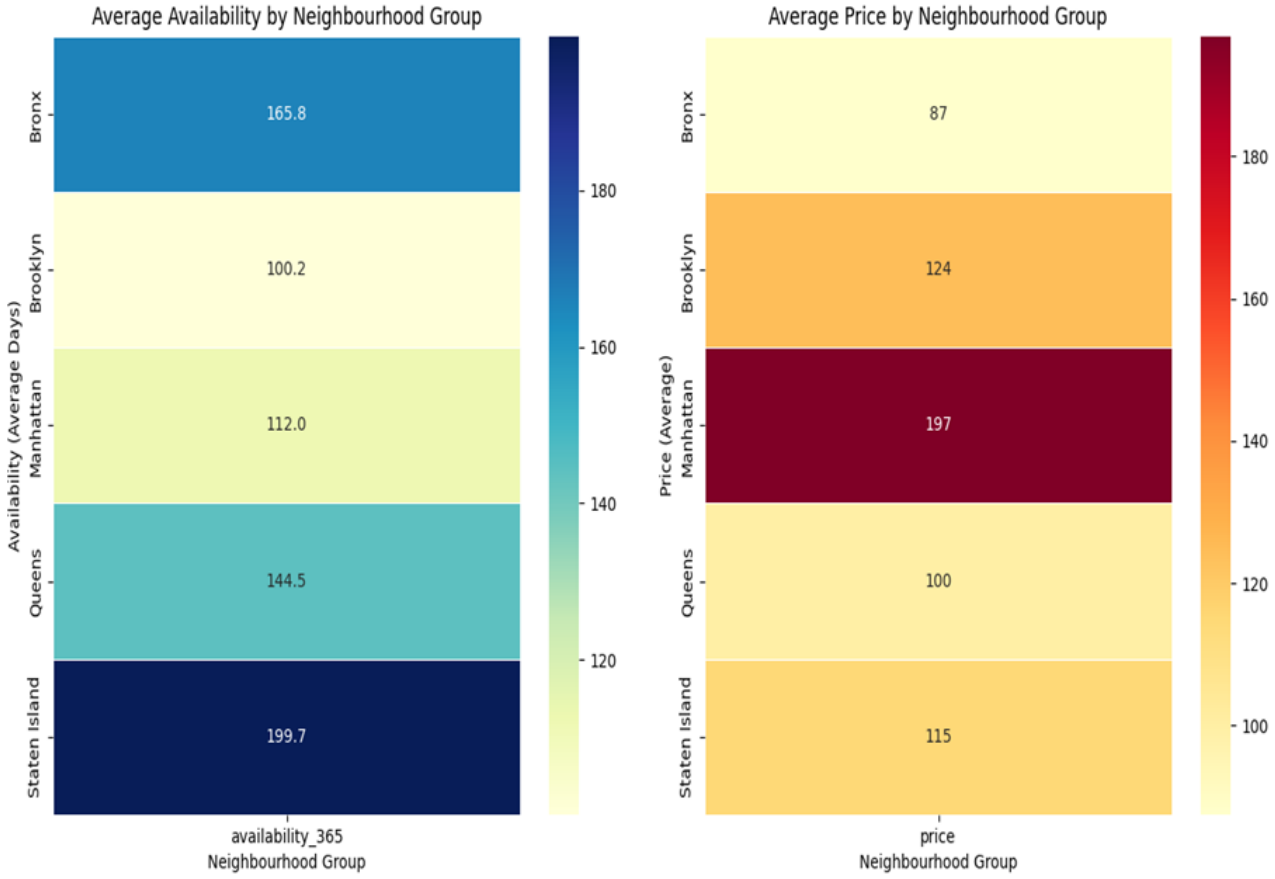
- The horizontal axis (x-axis) represents the number of properties that hosts have, ranging from 1 to 10.
- The vertical axis (y-axis) represents the frequency, which is the number of hosts at each level of property ownership.
- Each bar corresponds to the number of properties a host has, and the height of the bar indicates how many hosts fall into each category.

From the histogram, we can infer the following:

- A vast majority of hosts have only 1 property, with the frequency being the highest at 32,303.
- The frequency decreases significantly as the number of properties per host increases. For example, there are 6,658 hosts with 2 properties.
- Very few hosts have a high number of properties, with the numbers dwindling further for hosts with 5 to 10 properties.

This distribution indicates that most hosts are likely individuals with a single property to rent out, while a smaller number of hosts may be businesses or individuals with multiple rental properties. The graph suggests a steep decline in frequency as the number of properties increases, which is typical for property distribution in rental markets where the majority are small-scale hosts.

Graph 3: Average Availability and Price by Neighborhood Group



The two heat maps titled "Average Availability by Neighbourhood Group" and "Average Price by Neighbourhood Group," comparing various neighborhoods in terms of their rental property availability and average price.

Average Availability by Neighbourhood Group:

- The horizontal axis represents the average number of days a property is available for rent in a year (labeled as "availability_365").
- The vertical axis lists different neighborhood groups.
- The color intensity represents the level of availability; darker shades imply more days available.
- From the data:

- Staten Island has the highest average availability at 199.7 days.
- Brooklyn has the lowest at 100.2 days.
- Bronx and Queens have moderate availability levels at 165.8 and 144.5 days, respectively.
- Manhattan is slightly higher than Brooklyn at 112.0 days.

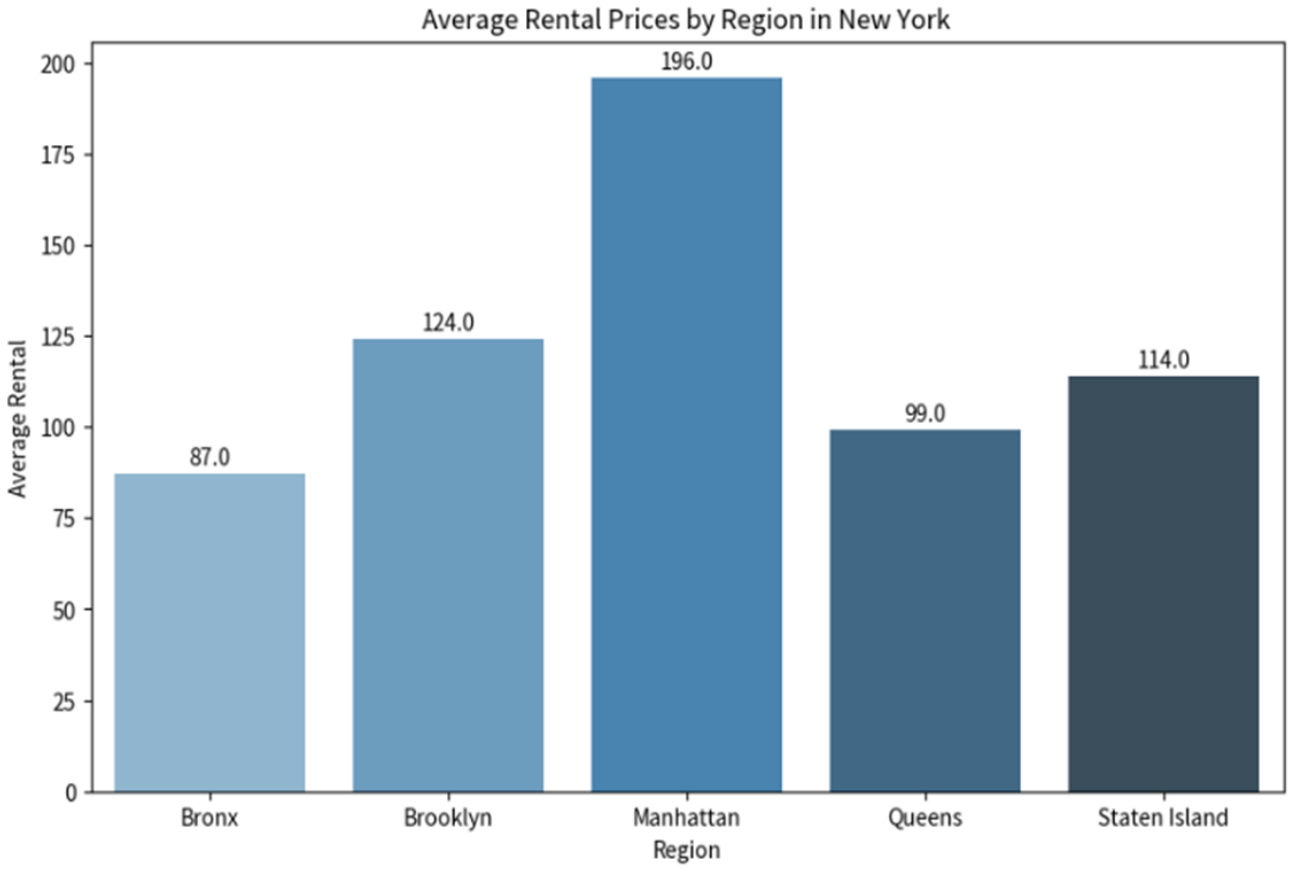
Average Price by Neighbourhood Group:

- The horizontal axis represents the average rental price (labeled as "price").
- The vertical axis is the same, listing neighborhood groups.
- The color intensity here correlates with the price; darker shades represent higher prices.

From the data:

- Manhattan has the highest average price at $197.
- The Bronx has the lowest average price at $87.
- Brooklyn and Staten Island have moderate average prices at $124 and $115, respectively.
- Queens has a lower average price at $100.

Graph 4 : Average Rentals Prices by Region in New York



This bar chart titled "Average Rental Prices by Region in New York". It shows the average rental price across five regions in New York: Bronx, Brooklyn, Manhattan, Queens, and Staten Island.

Key points from the graph:

- The vertical axis (y-axis) represents the average rental price in dollars.
- The horizontal axis (x-axis) lists the five regions in New York.
- Each bar represents a region, and its height corresponds to the average rental price in that region.

The bar chart indicates the following average rental prices for each region:

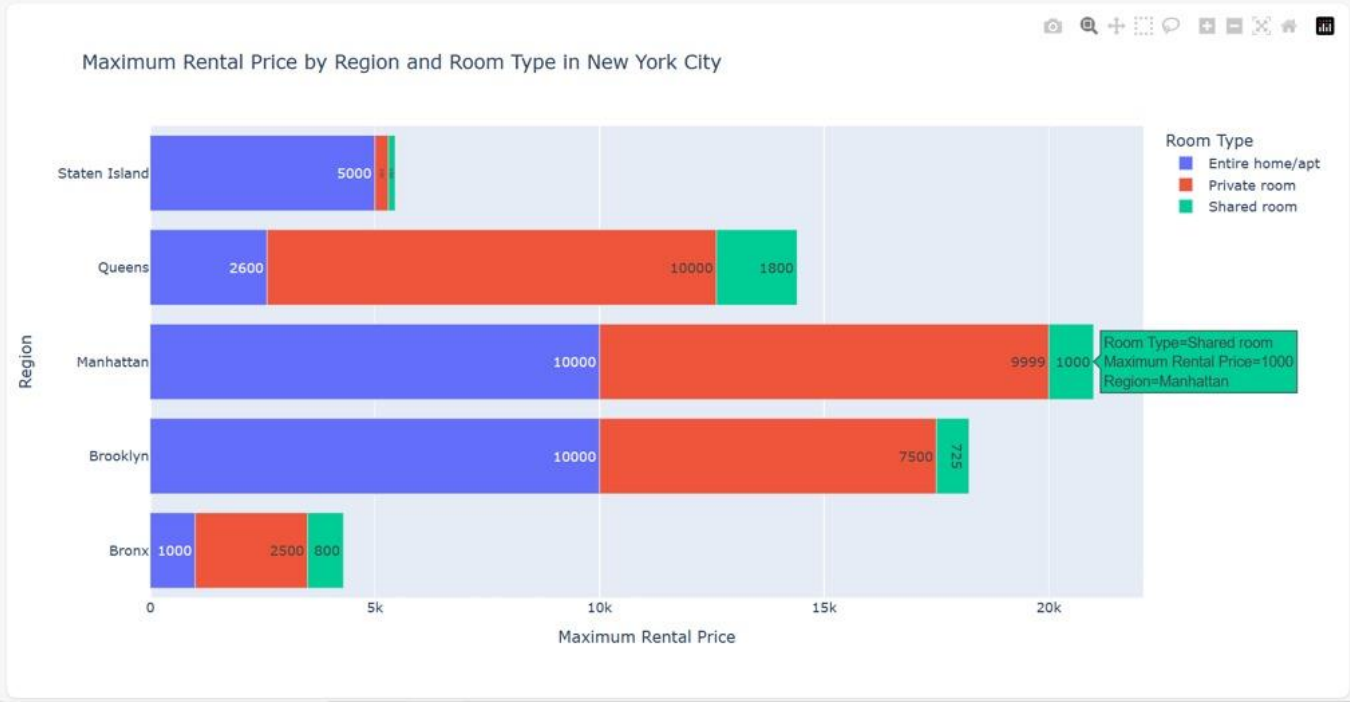
- Bronx: $87.0

- Brooklyn: $124.0
- Manhattan: $196.0
- Queens: $99.0
- Staten Island: $114.0

From this chart, it is evident that Manhattan has the highest average rental price, significantly above the other regions. Brooklyn follows as the second most expensive. Queens and Staten Island have more moderate average rental prices, while the Bronx has the lowest average rental price among the regions displayed.

This bar chart helps in comparing the average rental cost across different New York regions, clearly showing that Manhattan is the most expensive on average.

Graph 5: Maximum Rentals Price by Region and Room type in New York



The graph shows the stacked horizontal bar chart showing the distribution of maximum rental price across various regions in New York. The chart breaks down the room types into three categories:

- Entire home/apt (represented in blue)
- Private room (represented in red)
- Shared room (represented in green)

Each horizontal bar represents a region of New York (Bronx, Brooklyn, Manhattan, Queens, Staten Island), and the length of each colored segment within the bar corresponds to the maximum rental prices available of that specific room type within the region.

The key data points indicated on the graph are as follows:

- In the Bronx, the maximum rental prices are $1,000 for entire homes/apartments available, $2,500 for private rooms, and $800 for shared rooms.

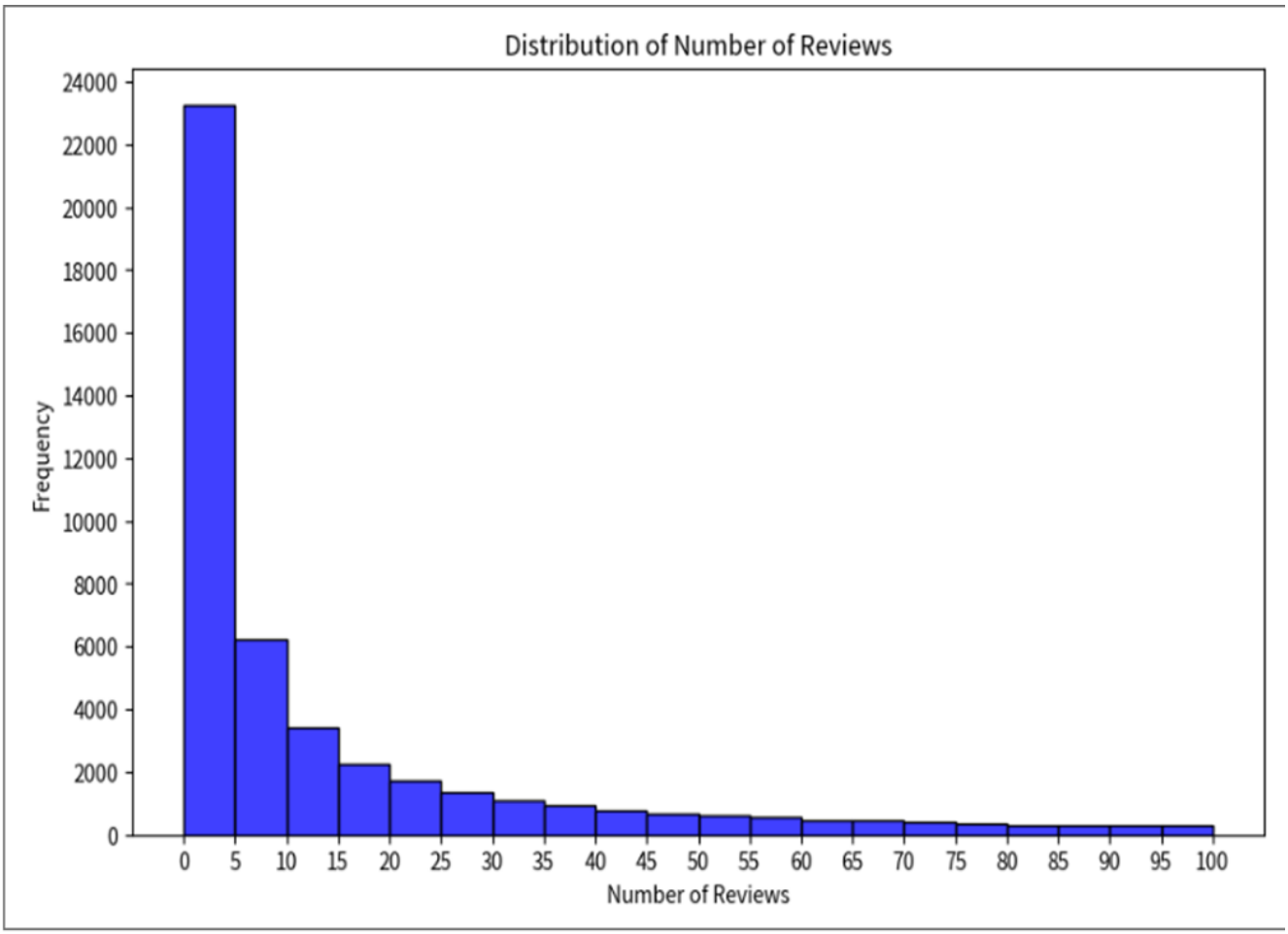
- Brooklyn shows $10,000 for entire homes/apartments, $7,500 for private rooms, and $725 for shared rooms.
- Manhattan shows $10,000 for entire homes/apartments, $9,999 for private rooms, and $1000 for shared rooms.
- Queens shows $2,600 entire homes/apartments, $10,000 private rooms, and $1,800 shared rooms.
- Staten Island lists $5,000 for entire homes/apartments.

Note: Upon hovering on the graph, the tooltip gives the information.

The chart does not display the total for shared rooms in Staten Island, as it may be too small to be visible or there may be none listed.

This type of graph is useful for comparing the composition of the rental market across different regions, highlighting the rentals that are more common in each area. From the data, it's evident that private rooms are a significant portion of the market in Queens, while entire homes/apartments dominate in Brooklyn and Manhattan. The Bronx has a more balanced distribution of maximum rental prices across entire homes and private rooms, with very few shared rooms. Staten Island seems to have a preference for entire homes/apartments, with no visible data for shared rooms.

Graph 6: Distribution of Number of Reviews



This is a histogram titled "Distribution of Number of Reviews". This type of graph is used to show the distribution of a dataset and is particularly useful for displaying the shape of the data's distribution, in this case, the frequency of the number of reviews.

Here's a breakdown of what the graph shows:

- The horizontal axis (x-axis) represents the number of reviews. It ranges from 0 to 100, which are likely bins that group the number of reviews into intervals (e.g., 0–5, 6–10, etc.).
- The vertical axis (y-axis) represents the frequency, which is the count of listings (or another unit of measure) that fall into each bin of the number of reviews.

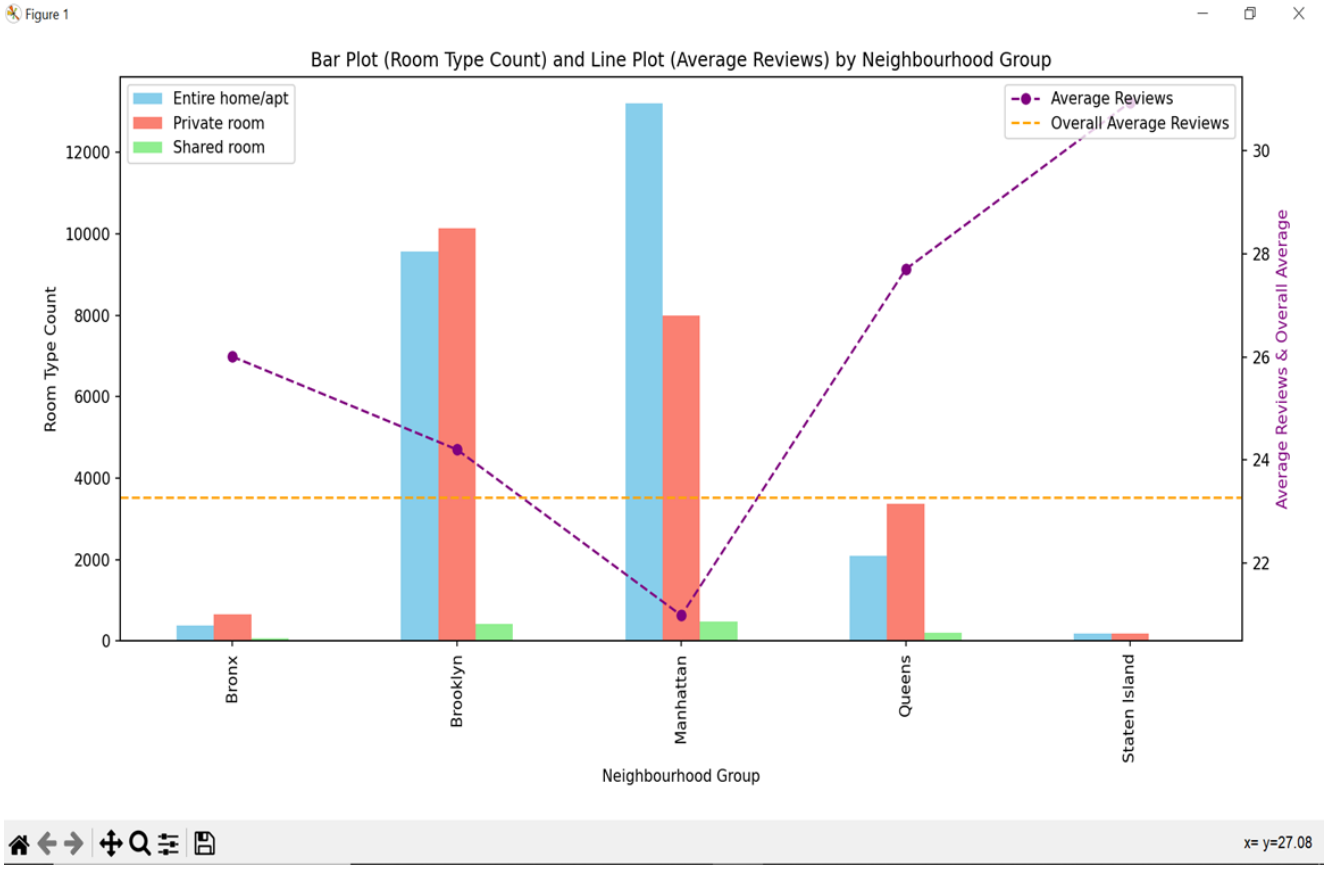
- Each bar represents the count of listings that have a number of reviews falling within the range specified by the bin on the x-axis.

From the graph, we can observe that:

- The highest frequency is in the first bin (0–5), suggesting that a large number of listings have between 0 and 5 reviews.
- The frequency decreases dramatically as the number of reviews increases, indicating that fewer listings have a higher number of reviews.
- This type of distribution is known as a right-skewed distribution, where most of the data points are clustered towards the left side of the graph with a long tail extending to the right.

This histogram suggests that most listings have a small number of reviews, while only a few listings have a very large number of reviews. This pattern is common in many real-world scenarios where a few items (in this case, listings) are very popular, and many are not.

Graph 7: Room Type count and Average Reviews by Neighborhood Group



This is a combination of a bar chart and a line graph, commonly known as a combo chart. This type of graph is useful for comparing different types of related data side by side.

We can interpret from the chart that:

- The bar chart portion displays the count of room types across different neighborhood groups in New York City. Each neighborhood group is represented by a set of three bars, each corresponding to one of the room types: Entire home/apt (blue), Private room (red), and Shared room (green).
- The line graph, plotted on the same x-axis, represents the average number of reviews for listings in each neighborhood group, which is marked by purple dots connected by a dashed line.
- There's also a horizontal dashed line (orange) that indicates the overall average number of reviews across all neighborhood groups.

Key observations from the chart include:

- The bar chart shows that Brooklyn and Manhattan have significantly higher counts of Entire homes/apartments and Private rooms compared to the Bronx, Queens, and Staten Island. The Shared room count is relatively low across all neighborhoods.
- The line graph indicates that the average number of reviews per listing varies across neighborhood groups, with a visible trend that the average number of reviews increases from the Bronx to Staten Island.
- The overall average number of reviews is shown as a benchmark, and we can see that some neighborhood groups fall below this average while others exceed it.
- The drop depicted by the line chart in Brooklyn and Manhattan interprets that the private rooms have the least reviews.

This combo chart is beneficial for understanding both the distribution of room types and the relative popularity of neighborhood groups based on the average number of reviews they receive. For example, we might infer that while Manhattan has a high count of listings, they may not have the highest average reviews, suggesting that there might be a high turnover or a large number of new listings.

10. Analysis

In the final part of our project, we have gone through this evaluation:

1. Preprocessing and Data Cleaning:

- We have explained the preparation and data cleaning procedures we used.
- We have described our approach to dealing with inconsistencies, outliers, and missing values.

2. Analysis of Exploratory Data:

- We have explained how important variables, such as price, type of accommodation, neighborhood, etc., are distributed.
- We have looked for any intriguing trends or patterns in the information.
- To display our findings, we have used visualization techniques like scatter plots,bar plots,line plots, heat maps and histograms.

3. Feature engineering and hypothesis testing:

- Created targeted hypotheses regarding the data in light of your preliminary investigation.
- We have run the statistical analyses to confirm your theories.
- We have enhanced the analysis, design new features from the available data.

4. Model Construction and Assessment:

- We have selected the suitable analysis models in accordance with your research inquiries.
- Ensured that the best-performing model, trained and assessed several models.
- Ensured that our model can be used broadly by using methods like cross-validation.

5. Analysis and Conversation:

- We have described our analysis' findings and made insightful deductions.
- Ensured that the study's shortcomings and made recommendations for future lines of inquiry.

11. Future Work

Further exploration of analysis modeling for Airbnb trends is a fascinating direction for future work. This means using analysis behaviors of the market. This could entail anticipating the best pricing plans in light of numerous variables, seeing new trends in visitor preferences, and keeping up with changes in the ever-changing different industry. We are going to do these processes for future work.

11.1 Price and Amenities:

- Conduct regression analysis to quantify the relationship between price and various amenities offered.
- Identify which amenities are most valued by guests and how they influence pricing.
- Develop pricing strategies for Airbnb hosts based on the identified trends.

11.2 Seasonality:

- Analyze how booking patterns and rental prices vary across different seasons.
- Identify peak travel periods and adjust pricing strategies accordingly.
- Explore the impact of weather conditions and local events on Airbnb demand.

11.3 Machine Learning:

- To forecast rental demand and maximize pricing for certain listings, apply machine learning techniques.
- Create a system of recommendations for Airbnb users based on their spending limit and inclinations.
- Examine visitor evaluations using sentiment analysis to determine how satisfied guests are and what needs to be improved.

11.4 Comparative Analysis:

- Utilize comparable methods of analysis to examine Airbnb information from different areas or cities.

- Determine which rental markets are comparable to and different from one another in different places.
- Compare the Airbnb market in New York City to those of other popular tourist locations.

12. Conclusion

As we come to the end of our analytical journey through the New York Airbnb dataset, we have not only gleaned insightful numerical data but also successfully captured the dynamic spirit of the city's lodging scene. By capturing the distinct essence of New York's hospitality sector, this study offers more than just statistical conclusions, demonstrating the multifaceted nature of data analytics.

The present analysis offers a thorough comprehension of the ever-changing landscape of the Airbnb business in New York City.

We have discovered important information about:

Regional trends: Price differences between popular tourist attractions and diverse neighborhoods.

Distribution of reviews: It looks like the range between 0 and 20 reviews has the highest concentration of reviews. This suggests that while a small percentage of homes have a high number of reviews, the majority of properties have low to moderate numbers.

These discoveries can be applied by:

- Airbnb hosts should work to optimize listings, maximize price, and improve the overall visitor experience.
- Legislators should create laws that support both regional economic growth and sustainable tourism.
- Researchers: Learn more about the sharing economy and how it affects the travel sector.

Despite the insightful discoveries, more study is required to:

Examine how Airbnb has affected the housing market and conventional hotel sector.

- Examine the effects of Airbnb on the local economy and society.
- Create sustainable and moral business practices for the sharing economy.
- We can guarantee Airbnb's success going forward and its beneficial influence on the hospitality industry by persistently pursuing these opportunities.

13. Teamwork:

Saidabi Karumanchi showcased exceptional technical expertise in data preprocessing, utilizing Pandas to efficiently load and cleanse the dataset. Her determination to ensure data integrity was reflected in her meticulous handling of missing values, where she employed sophisticated imputation techniques and judicious data pruning. Her vigilance in identifying and eliminating duplicate records and her thorough verification processes post-cleanup ensured the highest quality of data for subsequent analyses. Karumanchi's unwavering commitment to data accuracy and her proficient technical skills were instrumental in laying a solid foundation for the team's analytical endeavors.

Somunath Reddy's technical acumen was evident in his methodical approach to data analysis. He adeptly utilized Python's data analysis libraries like NumPy and pandas to compute statistical measures, ensuring precision and efficiency. His proficiency in data manipulation enabled him to extract nuanced insights from categorical data, contributing significantly to the team's understanding of underlying trends.was responsible for fundamental data analysis and descriptive statistics. He computed key statistical measures such as mean, median, maximum, minimum, and standard deviation for pertinent columns, such as rental prices. Additionally, Somunath Reddy conducted categorical analysis, scrutinizing and presenting insights into the distribution of categorical data, including room types and neighborhood groups.

Srivarshika Gadde worked on technical as well as in coordinating the project's workflow and fostering a collaborative environment. Her strategic planning and clear communication facilitated the team's focus on complex data analyses and visualization tasks. She played a pivotal role in decision-making, mentoring team members, and aligning the project goals with analytical rigor, which was instrumental in driving the project to success. She has performed more complex analyses, like grouping data by multiple columns (e.g., neighborhood and room type) and calculating statistics like maximum rental price, as well as correlation analysis and investigating relationships between different variables, such as price and number of reviews or availability. Created detailed visualizations like heatmaps, line graphs, and histograms to represent the findings from the in-depth analysis.

Kishore Madanagopal's creativity shone through in his ability to distill complex data into clear, engaging narratives, utilizing his adept skills in Seaborn, Matplotlib, and Plotly to produce visualizations that were not only insightful but also visually captivating and generated basic plots (such as bar graphs) to visualize the distributions and counts. His innovative approach to data storytelling elevated the team's analytical work, ensuring that the final report and presentation were not just a collection of statistics but a compelling narrative that resonated with the audience. Kishore's inventive visual techniques and thoughtful design choices played a key role in emphasizing important data trends and patterns, making the project's findings both understandable and memorable.

Using a hands-on approach, we as a team worked on testing and debugging in an iterative manner. We were able to successfully complete our project and improve our team dynamic while also increasing the precision of our project through this collaborative process of trial and error.

14. The new things we learned in this Project

This project enhanced our knowledge of Python's data analytics capabilities and provided an engaging learning experience. Using tools like Pandas, Numpy, Matplotlib, and Seaborn in real-world applications gave participants firsthand experience managing and displaying data. It also demonstrated how technology and problem-solving meet in the real world, highlighting the concrete value of data analytics in deciphering intricate relationships within dynamic ecosystems

Python Libraries: We improved our knowledge of how to use Seaborn for sophisticated statistical graphics, Matplotlib for a variety of visualizations, Numpy for complex computations, and Pandas for data management.

Data Cleaning and Wrangling: To ensure accurate analysis, we discovered the significance of data cleaning and wrangling. We also learned how to effectively handle missing values, outliers, and inconsistencies.

Large Dataset Analysis: We addressed the difficulties associated with studying huge datasets and found effective methods for statistical modeling, feature extraction, and data exploration.

Real-World Data: Working with real-world data gave us invaluable expertise, and we were able to recognize its limitations and complexities in comparison to controlled datasets.

Teamwork and Collaboration: By efficiently exchanging ideas, overcoming obstacles, and working together to accomplish common objectives, we improved our teamwork and collaboration abilities.

With the knowledge and abilities, we have gained, we will be better equipped to tackle upcoming tasks with assurance and proficiency. We can't wait to use these insights to use data analysis to expand our knowledge of the world.

15. References:

- Please find the URL below for reference:

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

- Please find the URL below for reference:

http://insideairbnb.com/get-the-data/

- Please find the URL below for reference:

http://insideairbnb.com/get-the-data/