

**NYC AIRBNB DATA ANALYSIS AND VISUALISATION**

A project report submitted to the Bharathiar University in the partial fulfillment of the requirements for the award of the degree of

## B.Sc., COMPUTER SCIENCE WITH DATA ANALYTICS

Submitted by

### ASHARAF ALI S (Reg.NO:2228B0317)

Under the guidance of

### Mr. U. SATCHITHANANTHAM.,

(Head of The Department, Department of Computer Science with Data Analytics)

## DEPARTMENT OF COMPUTER SCIENCE WITH DATA ANALYTICS AJK COLLEGE OF ARTS AND SCIENCE(AUTONOMOUS)

(Affiliated to Bharathiar University, Re-Accredited with A+ grade by NAAC)

## NAVAKKARAI, COIMBATORE – 641 105.

**OCTOBER 2024**



**NYC AIRBNB DATA ANALYSIS AND VISUALISATION**

A project report submitted to the Bharathiar University in the partial fulfillment of the requirements for the award of the degree of

## B.Sc., COMPUTER SCIENCE WITH DATA ANALYTICS

Submitted by

### ASHARAF ALI S (Reg.NO:2228B0317)

Under the guidance of

### Mr. U. SATCHITHANANTHAM.,

(Head of The Department, Department of Computer Science with Data Analytics)

## DEPARTMENT OF COMPUTER SCIENCE WITH DATA ANALYTICS

**AJK COLLEGE OF ARTS AND SCIENCE(AUTONOMOUS)**

(Affiliated to Bharathiar University, Re-Accredited with A+ grade by NAAC)

## NAVAKKARAI, COIMBATORE – 641 105.

**OCTOBER 2024**



## PROJECT WORK

**NYC AIRBNB DATA ANALYSIS AND VISUALISATION**

**Bonafide Work Done by ASHARAF ALI S (Reg.No.2228B0317)**

The project submitted in partial fulfillment of the requirements for the award of BSc.,Computer Science with Data Analytics of Bharathiar University, Coimbatore – 641 046

### GUIDE HEAD OF THE DEPARTMENT

**Submitted for the Viva-Voce Examination held on**

**INTERNAL EXAMINER EXTERNAL EXAMINER**

**OCTOBER 2024**

# DECLARATION

## DECLARATION

I, **ASHARAF ALI S** (**2228B0317**) herby declare that the project entitled **“NYC AIRBNB DATA ANALYSIS AND VISUALISATION** ” is my project work carried out during the Third year B.Sc. Computer Science with Data Analytics at AJK College of Arts and Science, Coimbatore, under the guidance of **Mr.U.SATCHITHANANTHAM ,** Head of The Department, Department of Computer Science with Data Analytics, AJK College of Arts and Science, and has not submitted previously for the award of any other degree or diploma by me to any institution or university according to the best of my knowledge.

### SIGNATURE OF THE CANDIDATE

**Place:** Coimbatore

**Date:**

# ACKNOWLEDGEMENT

## ACKNOWLEDGEMENT

I express my sincere thanks to all those who have provided a valuable guidance towards the completion of this system as part of the syllabus of the Third Year BSc Computer Science with Data Analytics Course.

I am deeply indebted to my guide **Mr.U. SATCHITHANANTHAM,** Head of the Department, Department of Computer Science with Data Analytics, for making available his intimate knowledge and experience in making “**NYC AIRBNB DATA ANALYSIS AND VISUALISATION”.**

I am deeply thank to **Mr. U. SATCHITHANANTHAM,** Head of the Department, Department of Computer Science with Data Analytics for his effective guidance and constant encouragement, which let this information work to its successful completion.

I extend my gratitude to **Dr. S. RAJU, MBA., Ph.D.,** Principal, AJK College of Arts and Science, whose primary aim is to establish and promote educational institutions of excellence and eminence in all fields so as to initiate and equip the youth with the originality of thinking, self-reliance and technological expertise.

I strive to the almost of my sincerity to repay a millionth of my indebtedness with profound gratitude by acknowledging the inestimable support and extensive guidance by our management members, **AJK College of Arts and Science**, whose dedicated care has come down a long way not only in competing support and venture but also in making our dreams into reality.

I would like to express my heartfelt thanks to my parents for their blessings, my dear friends, classmates and all faculties for their help and wishes for the successful completion of this project.

# CONTENTS

**CONTENTS**

|  |  |  |
| --- | --- | --- |
| **S.NO** | **TITLE** | **PAGE No** |
|  | **SYNOPSIS** |  |
| **1** | **INTRODUCTION** |  |
|  | 1.1 About the Project |  |
| **2** | **SYSTEM STUDY AND ANALYSIS** |  |
|  | 2.1 Existing System |  |
|  | 2.2.1 Disadvantages of the Existing System |  |
|  | 2.2 Proposed System |  |
|  | 2.2.2 Advantages of the Proposed System |  |
| **3** | **SYSTEM CONFIGURATION** |  |
|  | 3.1 Hardware Configuration |  |
|  | 3.2 Software Configuration |  |
|  | 3.3 Modules |  |
| **4** | **SYSTEM TESTING AND IMPLEMENTATION** |  |
|  | 4.1 System Implementation |  |
| **5** | **CONCLUSION** |  |
| **6** | **FURTHER SCOPE OF THE PROJECT** |  |
| **7** | **BIBILIOGRAPHY** |  |
| **8** | **APPENDIX** |  |
|  | 8.1 Project Design |  |

# SYNOPSIS

## SYNOPSIS

The project titled "NYC Airbnb Data Analysis and Visualization" delves into the intricate realm of New York City's Airbnb ecosystem, with a primary focus on conducting data analysis and visualization to unveil insightful patterns and trends. By harnessing a comprehensive dataset encompassing diverse facets of Airbnb listings throughout the city, the study endeavors to investigate factors including pricing dynamics, spatial distribution, and neighborhood popularity. Through the application of robust analytical techniques and compelling visual representations, this research aims to provide a nuanced comprehension of Airbnb's operations within the dynamic urban landscape of New York City. Such insights are intended to offer valuable perspectives on both the market itself and its implications for residents and visitors alike.

The frontend of the project has been developed utilizing Python incorporating libraries such as NumPy, Pandas, Seaborn, Matplotlib, and Folium. Visualaising the data using power BI. Concurrently, the backend is powered by MySQL. The selected operating system for the project is Microsoft Windows.

In conclusion, "NYC Airbnb Data Analysis and Visualization" has provided a comprehensive exploration of New York City's dynamic Airbnb market. Through rigorous data analysis and insightful visualizations, the project has illuminated key patterns and trends in pricing dynamics, spatial distribution, and neighborhood popularity among Airbnb listings. project consists of five modules, **Data Collection and Preprocessing** handles acquiring and cleaning the data**. Exploratory Data Analysis** uncovers key patterns in neighborhoods, room types, and prices. **Geospatial Analysis** maps listings to examine spatial trends. **Predictive Analytics and Pricing**  use machine learning to forecast prices, and **Visualization in Power BI** presents the insights through interactive dashboards.

# INTRODUCTION

## INTRODUCTION

The analysis and visualization of Airbnb data in New York City offers a compelling glimpse into the dynamics of the city’s rental market, revealing patterns in pricing, availability, and geographic distribution. By leveraging comprehensive datasets from Airbnb listings, researchers and analysts can uncover insights into how factors such as neighborhood, property type, and seasonal trends influence rental rates and occupancy. This exploration not only aids potential renters and investors in making informed decisions but also assists city planners and policymakers in understanding the impact of short-term rentals on local housing markets and communities. Through sophisticated data visualization techniques, these insights become more accessible and actionable, highlighting trends and anomalies that drive the evolving landscape of NYC’s accommodation sector.

### 1.1 ABOUT THE PROJECT

The NYC Airbnb Data Analysis and Visualization project aims to provide a comprehensive examination of short-term rental trends across New York City by analyzing a rich dataset of Airbnb listings. This project focuses on uncovering patterns related to rental pricing, occupancy rates, and geographic distribution, offering insights into how various factors such as neighborhood characteristics, property types, and seasonal fluctuations affect the rental market. By employing advanced data visualization techniques, the project not only highlights key trends and anomalies but also facilitates a deeper understanding of the interplay between Airbnb rentals and local housing dynamics.

### PROJECT OVERVIEW

The NYC Airbnb Data Analysis and Visualization project seeks to explore and illustrate key trends and patterns in the city's short-term rental market, focusing on pricing, availability, and geographic distribution to provide actionable insights for stakeholders.

### PURPOSE

The purpose of analyzing and visualizing Airbnb data for New York City is to gain deeper insights into the dynamics of the local short-term rental market. By examining factors such as pricing trends, geographic distribution, occupancy rates, and host behaviors, we can identify patterns and correlations that inform both potential guests and hosts. This analysis not only helps prospective renters make informed decisions but also assists policymakers in understanding the impact of short-term rentals on housing availability and neighborhood dynamics.

### FEATURES

* **Listing Attributes**:

Price, location, room type, and amenities.

### Host Information:

Experience level, response rates, and host activity.

### Occupancy Patterns:

Booking frequency and seasonal trends.

### Geographic Distribution:

Heat maps and cluster analysis of listing density and neighborhood trends.

### Temporal Analysis:

Trends over time, including price fluctuations and booking trends.

### TECHNOLOGY STACK

NYC Airbnb data analysis and visualizations typically includes a range of tools and platforms. Data extraction and cleaning are often managed with Python libraries such as Pandas and NumPy, while MYSQL databases. For data visualization, tools such as Power BI provide interactive dashboards, while Python libraries like Matplotlib, Seaborn offer customizable visualizations. Geographic data analysis leverages GIS tools like Folium for mapping and spatial analysis.

### BUSINESS MODEL

The business model for NYC Airbnb data analysis and visualizations focuses on generating revenue through multiple channels. It includes offering subscription-based access to comprehensive dashboards and real-time analytics for real estate professionals, investors, and property managers. Custom analytical reports and actionable insights are provided on a project basis, catering to specific client needs. Additionally, licensing detailed datasets to market researchers and firms seeking in-depth market intelligence creates another revenue stream. Partnering with hospitality and travel companies for targeted advertising and strategic collaborations further enhances business opportunities, leveraging data to drive informed decision-making and optimize rental strategies.

### IMPACT

The impact of NYC Airbnb data analysis and visualizations is significant, as it provides critical insights into the short-term rental market's dynamics. By revealing trends in pricing, occupancy, and geographic distribution, this analysis helps stakeholders make data-driven decisions, optimizing rental strategies and improving profitability. For policymakers, the data offers a clearer understanding of the effects of short-term rentals on local housing markets and neighborhoods, informing regulations and housing policies. Additionally, these insights enhance transparency, enabling consumers to make informed choices and fostering a more balanced and equitable rental environment in one of the world's most competitive markets.

# SYSTEM STUDY AND

**ANALYSIS**

## SYSTEM STUDY AND ANALYSIS

### EXISTING SYSTEM

NYC Airbnb data analysis and visualizations typically involve a combination of commercial and open-source tools. Platforms like Airbnb’s own data insights, along with third-party services such as Inside Airbnb, provide baseline data on listings, prices, and occupancy. Analytical frameworks often utilize SQL databases for data storage and processing, while tools like Tableau and Power BI offer visualization capabilities. Geographic analysis is supported by GIS platforms such as ArcGIS and QGIS. However, these systems can vary in their depth of analysis and ease of access, with many relying on public datasets or limited proprietary data, potentially lacking the granularity needed for comprehensive market insights.

### DISADVANTAGES OF EXISTING SYSTEM

* + - * **Limited Granularity**:

Publicly available data may lack detailed insights, affecting the depth of analysis.

### Outdated Information:

Many systems rely on static or infrequent updates, leading to outdated or inaccurate data.

### Data Integration Challenges:

Difficulty in integrating diverse data sources can result in incomplete analyses.

### High Costs:

Proprietary tools and platforms may be expensive, limiting access for smaller stakeholders.

### Lack of Customization:

Limited flexibility in existing tools can restrict users’ ability to perform specific, advanced analyses.

### Specialized Knowledge Required:

Complex tools may require significant expertise, making them less accessible to non- specialists.

### PROPOSED SYSTEM

The proposed system for NYC Airbnb data analysis and visualizations aims to address current limitations by integrating real-time data feeds with advanced analytical and visualization tools. This system would utilize a unified platform that combines comprehensive datasets from multiple sources, including real-time listings, bookings, and reviews, to provide up-to-date and granular insights. It would employ machine learning algorithms for predictive analytics and trend forecasting, while interactive dashboards and customizable reports enable users to tailor analyses to specific needs.

### ADVANTAGES OF PROPOSED SYSTEM

* + - * **Real-Time Data Integration**:

Provides up-to-date insights by incorporating real-time data feeds from multiple sources.

### Enhanced Granularity:

Offers detailed and comprehensive analysis with access to granular data on listings, bookings, and reviews.

### Advanced Analytics:

Utilizes machine learning and predictive algorithms to forecast trends and identify patterns.

### Customizable Dashboards:

Allows users to tailor visualizations and reports to specific needs and preferences.

### User-Friendly Interface:

Features intuitive design and accessibility, making advanced analytics easier for both experts and non-specialists.

### Robust Data Integration:

Seamlessly combines diverse data sources for more accurate and holistic insights.

# SYSTEM CONFIGURATION

## 3.SYSTEM CONFIGURATION

### HARDWARE CONFIGURATION

* + - **Model:** Lenovo ideapad 3 15ITL6
    - **Processor:** 11th Gen Intel(R) Core(TM) i3-1115G4 @ 3.00GHz
    - **Chipset:** Intel Q570
    - **Memory:** 16GBDDR4
    - **Network:** Integrated realtek Ethernet LAN 10/100/1000
    - **Intel WI-FI** 6 AX201+Bluetooth 5.1
    - **Hard Drive:** 512GB NVMe class 35 SSD
    - **Chassis:** AIO – 13.54”\*21.26”\*2.07”
    - **Display:** “23.8” FHD Non-Touch Anti-Glare
    - **Keyboard:** Lenovo KB216 Wired Multimedia USB Keyboard
    - **Mouse:** Lenovo MS116 optical Mouse

### SOFTWARE CONFIGURATION

* + - **OPERATING SYSTEM:**

### MICROSOFT WINDOWS

* + - **TOOL:**

### MYSQL DATABASE

* + - * **PYTHON (PANDAS, NUMPY, MATPLOTLIB, SEABORN, FOLIUM)**

### MICROSOFT POWER BI

* + - * **GOOGLE COLAB**

### MYSQL DATABASE:

In the NYC Airbnb data analysis and visualizations project, MySQL serves as a robust and reliable relational database management system for storing and managing structured data. MySQL’s efficiency in handling large volumes of data makes it ideal for storing detailed information on Airbnb listings, bookings, and user reviews. Its SQL query capabilities enable complex data retrieval and manipulation, which are essential for performing detailed analyses and generating insights. By leveraging MySQL, the project benefits from its scalability, ease of integration with various data processing tools, and support for data normalization, ensuring data integrity and consistency across

the analysis workflow. Additionally, MySQL’s compatibility with popular data visualization and analysis tools enhances the overall efficiency of data extraction and reporting processes.

### USES OF MYSQL DATABASE:

* **Efficient Data Storage**:

Stores large volumes of structured data on Airbnb listings, bookings, and user reviews.

### Data Retrieval:

Allows for precise querying to extract specific datasets needed for analysis.

### Data Management:

Ensures data consistency and integrity through normalization and relational database design.

### Complex Query Handling:

Supports complex SQL queries for aggregations, filtering, and joining multiple data tables.

### Performance Optimization:

Utilizes indexing and query optimization features to improve performance on large datasets.

### Integration with Tools:

Integrates easily with data processing and visualization tools for seamless analysis and reporting.

### Historical Data Analysis:

Facilitates the management and querying of historical data for trend analysis and time- series insights.

### PYTHON:

In the NYC Airbnb data analysis and visualizations project, Python plays a crucial role in data processing, analysis, and visualization. Leveraging libraries such as Pandas and NumPy, Python facilitates efficient data cleaning, manipulation, and analysis, allowing for detailed exploration of Airbnb datasets. For data visualization, libraries like Matplotlib, Seaborn enable the creation of insightful charts and interactive plots that help reveal trends and patterns. making Python an invaluable tool for generating actionable insights and comprehensive reports.

### PANDAS:

* **Data Manipulation**:

Provides powerful data structures for cleaning, transforming, and analyzing Airbnb datasets.

### Data Aggregation:

Facilitates grouping and aggregating data to uncover trends and insights.

### Data Integration:

Simplifies merging and joining multiple data sources for comprehensive analysis.

### NUMPY:

* **Numerical Operations**:

Supports efficient handling of large arrays and matrices, essential for performing complex mathematical computations.

### Statistical Analysis:

Provides functions for statistical operations and data transformation, aiding in detailed numerical analysis.

### MATPLOTLIB:

* **Basic Plotting**:

Enables the creation of static, high-quality plots such as histograms, scatter plots, and line charts to visualize key metrics.

### Customization:

Offers extensive options for customizing plot aesthetics, including labels, colors, and annotations.

### SEABORN:

* **Advanced Visualization**:

Enhances Matplotlib with additional functionalities for creating attractive and informative statistical plots, such as heatmaps, pair plots, and violin plots.

### Data Insights:

Simplifies the process of exploring relationships between variables and visualizing distributions with improved visual appeal and clarity.

### FOLIUM:

* **Geospatial Mapping**:

Creates interactive maps to visualize geographic data, such as listing locations and heatmaps of rental density.

### Interactive Visualization:

Allows for dynamic map features like zooming and panning, making it easier to explore spatial patterns in the data.

### MICROSOFT POWER BI:

In the NYC Airbnb data analysis and visualizations project, Microsoft Power BI serves as a powerful tool for creating interactive and visually compelling dashboards. Power BI enables users to integrate data from various sources, including MySQL databases and CSV files, providing a unified platform for analysis. With its drag-and-drop interface, users can easily build and customize reports, visualizations, and interactive elements such as slicers and filters. Power BI's advanced visualization capabilities include charts, maps, and KPI indicators that help in presenting complex data insights in an accessible format. Additionally, its real-time data refresh feature ensures that stakeholders always have access to the most current information, enhancing decision-making and strategic planning. Power BI's ability to share and collaborate on reports within teams and across organizations further supports effective communication and data-driven insights.

### Interactive Dashboards:

Allows the creation of dynamic dashboards with interactive elements such as filters, slicers, and drill-through capabilities to explore data in-depth.

### Data Integration:

Supports integration with various data sources, including MySQL databases, CSV files, and cloud services, for a comprehensive data analysis.

### Customizable Visualizations:

Provides a range of visualization options like bar charts, line graphs, maps, and KPI indicators, enabling the presentation of data in various formats.

### Real-Time Data Refresh:

Facilitates automatic data updates to ensure that reports reflect the latest information, enhancing accuracy and timeliness.

### Advanced Analytics:

Offers built-in analytical tools and functions, such as DAX (Data Analysis Expressions), for creating complex calculations and metrics.

### User-Friendly Interface:

Features a drag-and-drop interface that simplifies the process of building and customizing reports, making it accessible for users without advanced technical skills.

### Collaboration and Sharing:

Enables easy sharing of reports and dashboards with stakeholders, promoting collaboration and ensuring that insights are communicated effectively across teams.

### GOOGLE COLAB:

Google Colab is an excellent tool for data analysis projects, offering a collaborative environment where users can write and execute Python code effortlessly. With built-in support for libraries like Pandas, NumPy, and Matplotlib, Colab simplifies data manipulation, visualization, and exploration. Users can easily import datasets from various sources, including Google Drive and online repositories. Additionally, the platform provides access to powerful computing resources, such as GPUs, allowing for efficient processing of large datasets. Its user-friendly interface and sharing capabilities make it an ideal choice for teams working on data-driven projects.

### Collaborative Environment:

Enables multiple users to work on the same notebook in real-time.

### Integration with Libraries:

Supports essential data analysis libraries like Pandas, NumPy, and Matplotlib for data manipulation and visualization.

### Easy Data Access:

Allows importing datasets directly from Google Drive, GitHub, and other online sources.

### Powerful Hardware:

Provides access to GPUs and TPUs, facilitating faster processing of large datasets.

### User-Friendly Interface:

Features an intuitive interface for both beginners and experienced users, enhancing productivity.

### Sharing and Publishing:

Easily share notebooks with team members or publish them for broader audiences.

**3.3 Modules:**

**1.Data Collection and Preprocessing:**

Objective: Import the NYC Airbnb dataset (from CSV, API, or database). Tasks: Use Python libraries like pandas to load the dataset, handle missing data, convert data types, and filter unnecessary columns.

**2. Exploratory Data Analysis:**

Objective: Perform an initial analysis to understand the data’s structure and key metrics. Tasks: Generate descriptive statistics (mean, median, mode, etc.) and visualizations such as histograms, bar charts, and scatter plots to understand distributions and correlations.

**3. Geospatial Analysis:**

Objective: Analyze geographic patterns in Airbnb listings across NYC neighborhoods. Tasks: Use libraries like Folium or Geopandas to visualize the distribution of Airbnb listings on a map, examining hotspots and clustering of properties.

**4. Predictive Analytics and Pricing Models:**

Objective: Build a model to predict Airbnb listing prices based on factors like location, room type, and number of reviews. Tasks: Use machine learning models such as linear regression or decision trees to predict prices.

**5. Visualization and Reporting in Power BI:**

Objective: Create interactive dashboards and reports to visualize Airbnb trends. Tasks: Import the processed data into Power BI and build visualizations such as geographic heatmaps, price vs. rating charts, and availability trends by neighborhood**.**

# SYSTEM IMPLEMENTATION

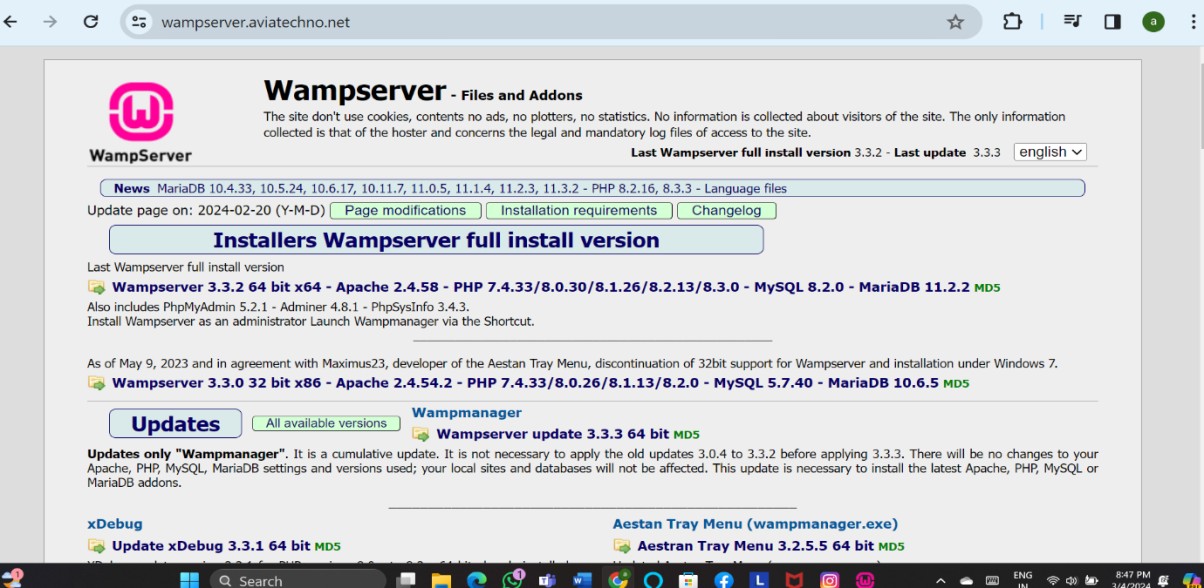
## SYSTEM IMPLEMENTATION

### 4.1 SYSTEM IMPLEMENTATION

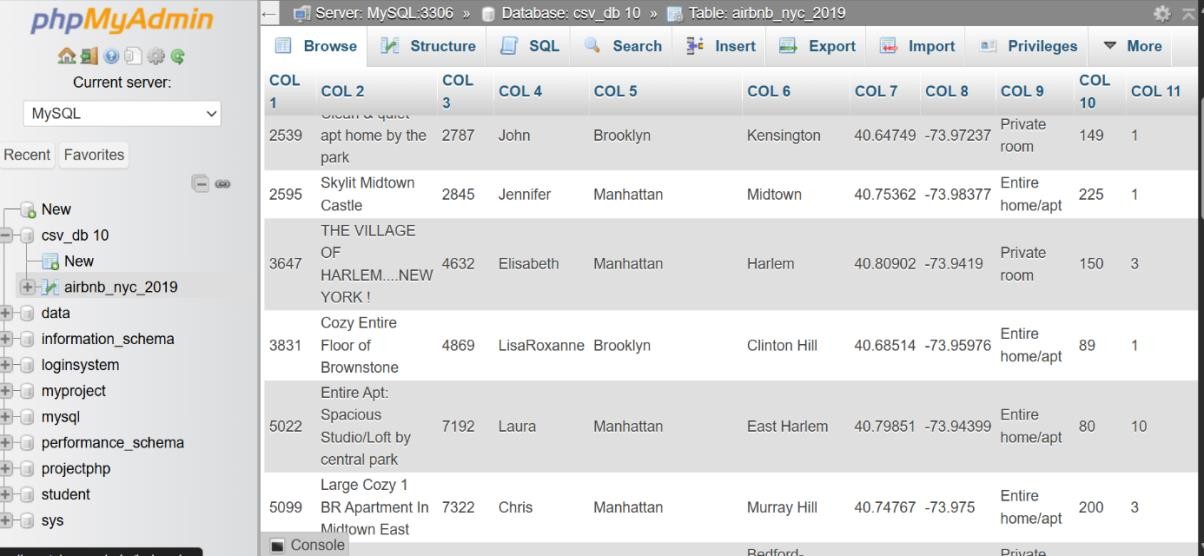
The first step in the project is system implementation, which requires setting up a database. To accomplish this, we download the WampServer.

Go to the browser and type wampserver download

***“https://wampserver.aviatechno.net/”***



After that we want to insert the data’s into the database and save into the future .



Next, we want to obtain the Airbnb dataset, which we are getting from Kaggle. After that, we will analyze the data using Python. Google Colab serves as the interface to run the code. Here, we have already uploaded the data into the database.

**The Data**

We'll begin by importing:

* **pandas** for data manipulation;
* **seaborn** and **matplotlib** for data visualization;
* **folium** to deal with geographical data.

### STEP 1:

Open the google colab and give the inputs

### CODE:

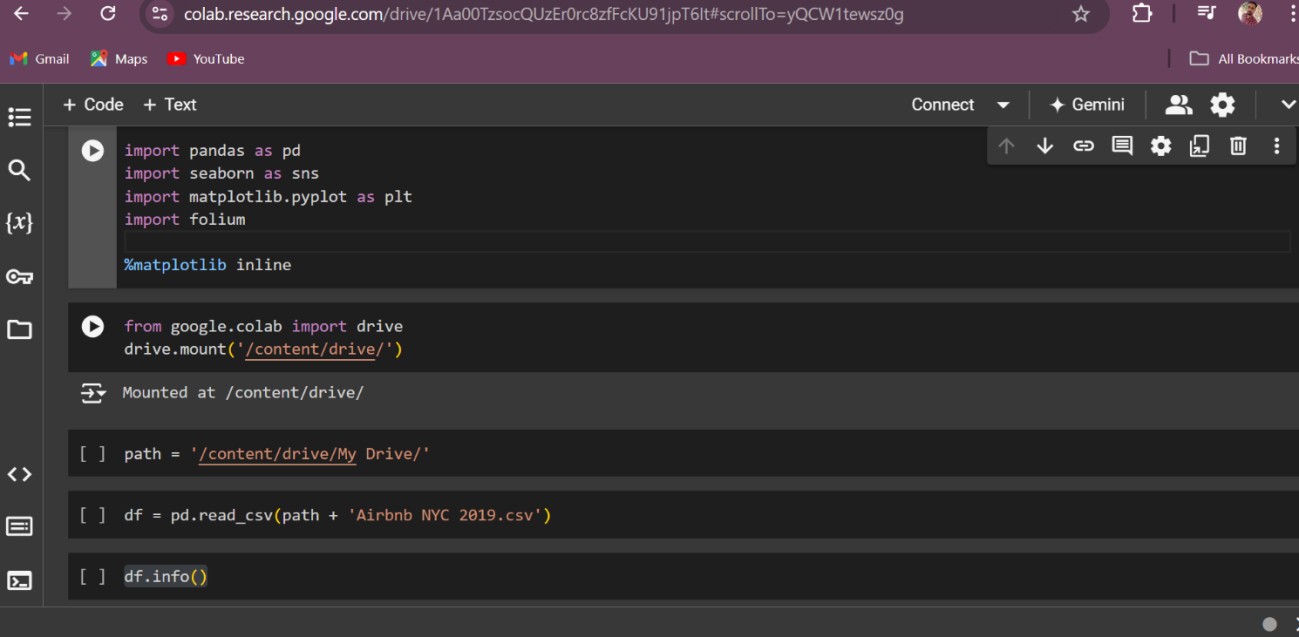
import pandas as pd import seaborn as sns

import matplotlib.pyplot as plt import folium

%matplotlib inline

from google.colab import drive drive.mount('/content/drive/') path = '/content/drive/My Drive/'

df = pd.read\_csv(path + 'Airbnb NYC 2019.csv') df.info()



### STEP 2:

df.info()is a method used to get a concise summary of a Data Frame. It displays:

* the number of columns and rows
* column labels,
* column data types,
* memory usage,
* the number of non-null values in each column.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 0 | id | 48895 | non-null | int64 |
| 1 | name | 48879 | non-null | object |
| 2 | host id | 48895 | non-null | int64 |
| 3 | host name | 48874 | non-null | object |
| 4 | neighbourhood\_group | 48895 | non-null | object |
| 5 | neighbourhood | 48895 | non-null | object |
| 6 | latitude | 48895 | non-null | float64 |
| 7 | longitude | 48895 | non-null | float64 |
| 8 | room\_type | 48895 | non-null | object |
| 9 | price | 48895 | non-null | int64 |
| 10 | minimum nights | 48895 | non-null | int64 |
| 11 | number of reviews | 48895 | non-null | int64 |

### STEP 3:

df.head()is used to display the first 5 rows of the dataframe df. You can specify the number of rows to display as an argument. For example, to display the first 10 rows you would use df.head(10)

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | id | name | host id | host name | neighbourhood\_group | neighbourhood | latitude | longitude | room\_type | price |
| 0 | 2539 | apt home by the park | 2787 | John | Brooklyn | Kensington | 40.64749 | -73.97237 | Private room | 149 |
| 1 | 2595 | Skylit Midtown Castle | 2845 | Jennifer | Manhattan | Midtown | 40.75362 | -73.98377 | Entire home/apt | 225 |
| 2 | 3647 | THE VILLAGE OF HARLEM....NEW YORK ! | 4632 | Elisabeth | Manhattan | Harlem | 40.80902 | -73.9419 | Private room | 150 |
| 3 | 3831 | Floor of Brownstone | 4869 | LisaRoxanne | Brooklyn | Clinton Hill | 40.68514 | -73.95976 | home/apt |  |
| 4 | 5022 | Entire Apt: Spacious Studio/Loft by | 7192 | Laura | Manhattan | East Harlem | 40.79851 | -73.94399 | Entire home/apt | 80 |

### STEP 3:

df.info()

Is a method in pandas that prints information about a DataFrame. This includes:

* The number of columns and rows.
* Column labels, data types and non-null values.
* Memory usage.

### (round(df.isnull().sum() / df.shape[0] \* 100, 2)).sort\_values(ascending=False)

This code calculates the percentage of missing values (NaN) in each column of the DataFrame named df. First, df.isnull().sum() calculates the number of missing values in each column. Then, it divides the result by df.shape[0] (the total number of rows in the DataFrame) and multiplies by 100 to get the percentage. Finally, round(..., 2) rounds the percentages to two decimal places, and .sort\_values(ascending=False) sorts the result in descending order, showing the columns with the highest percentage of missing values first.

### STEP 4:

df.hist(bins=15, figsize=(15,10));

This line of code generates a histogram for each numerical column in the DataFrame df.

* df.hist() creates the histograms.
* bins=15 specifies that each histogram should have 15 bins.
* figsize=(15, 10) sets the size of the figure to 15 inches in width and 10 inches in height.
* The semicolon (;) suppresses the output of the last command in the cell.

### STEP 5:

df.hist(bins=15, figsize=(15,10));

This line of code generates a histogram for each numerical column in the DataFrame df.

* df.hist() creates the histograms.
* bins=15 sets the number of bins for each histogram to 15.
* figsize=(15,10) sets the size of the figure to 15 inches in width and 10 inches in height.

The semicolon at the end suppresses the output of the function.

### STEP 6:

df\_clean.hist(bins=15, figsize=(15,10)); meaning

This code snippet cleans the dataframe df and stores it in df\_clean.

1. It filters the dataframe to only include rows where price is less than or equal to 500 and greater than 0, and minimum\_nights is less than or equal to 30.
2. The columns reviews\_per\_month and last\_review are dropped from the dataframe.
3. The index is reset.
4. Histograms are plotted for the numerical columns of the cleaned dataframe.

This code aims to remove outliers and unnecessary columns from the dataframe and then visualize the distribution of the remaining data.

### STEP 7:

area = areas\_reviews['neighbourhood\_group']

review = areas\_reviews['number\_of\_reviews']

fig = plt.figure(figsize = (10, 5))

# creating the bar plot

plt.bar(area, review, color ='maroon',

width = 0.4)

plt.xlabel("area")

plt.ylabel("review")

plt.title("Area vs Number of reviews")

plt.show()

This code displays the relationship between areas and the number of reviews in a bar chart. It starts by extracting the 'neighbourhood\_group' as area and 'number\_of\_reviews' as review from the 'areas\_reviews' DataFrame. Next, it sets up the plot figure with a size of 10 by 5. Then it creates a bar plot with 'area' on the x-axis and 'review' on the y-axis, using maroon bars with a width of 0.4. Finally, it adds labels for the x-axis ('area'), y-axis ('review'), title ('Area vs Number of reviews') and displays the plot.

**STEP 8:**

room\_type = traffic\_areas['room\_type']

stayed = traffic\_areas['minimum\_nights']

fig = plt.figure(figsize = (10, 5))

# creating the bar plot

plt.bar(room\_type, stayed, color ='maroon', width = 0.4)

plt.xlabel("Room Type")

plt.ylabel("Minimum number of nights stayed")

plt.title("Traffic Areas")

plt.show()

This code visualizes the relationship between room type and the minimum number of nights stayed using a bar chart.

Here's a breakdown:

* room\_type = traffic\_areas['room\_type']: Extracts the 'room\_type' column from the 'traffic\_areas' DataFrame and assigns it to the variable room\_type. This will be used for the x-axis of the bar chart.
* stayed = traffic\_areas['minimum\_nights']: Extracts the 'minimum\_nights' column from the 'traffic\_areas' DataFrame and assigns it to the variable stayed. This will be used for the y-axis of the bar chart.
* fig = plt.figure(figsize = (10, 5)): Creates a new figure with a size of 10 by 5 inches.
* plt.bar(room\_type, stayed, color ='maroon', width = 0.4): Creates a bar chart with 'room\_type' on the x-axis and 'stayed' on the y-axis. The bars are colored maroon and have a width of 0.4.
* plt.xlabel("Room Type"): Sets the label for the x-axis.
* plt.ylabel("Minimum number of nights stayed"): Sets the label for the y-axis.
* plt.title("Traffic Areas"): Sets the title of the chart.
* plt.show(): Displays the chart.

Execute the code yourself to see the output. This will help you visualize which room types have longer minimum stay requirements.

**STEP 9:**

traffic\_areas = new\_df.groupby(['neighbourhood\_group','room\_type'])['minimum\_nights'].count().reset\_index()

traffic\_areas = traffic\_areas.sort\_values(by='minimum\_nights', ascending=False)

traffic\_areas

|  |  |  |  |
| --- | --- | --- | --- |
| Column1 | neighbourhood\_group | room\_type | minimum\_nights |
| 6 | Manhattan | Entire home/apt | 13199 |
| 4 | Brooklyn | Private room | 10132 |
| 3 | Brooklyn | Entire home/apt | 9559 |
| 7 | Manhattan | Private room | 7982 |
| 10 | Queens | Private room | 3372 |
| 9 | Queens | Entire home/apt | 2096 |
| 7 | Bronx | Private room | 652 |
| 8 | Manhattan | Shared room | 480 |
| 5 | Brooklyn | Shared room | 413 |
| 0 | Bronx | Entire home/apt | 379 |
| 11 | Queens | Shared room | 198 |

**FLOWCHART:**

DECISION MAKING

CREATE POWER BI VISUALIZATIONS

SHRING AND COLLABORATION

DATA CLEANING AND PREPARATION

DATA COLLECTION

DATA VISUALIZATIONS

DATA ANALYSIS

# CONCLUSION

## CONCLUSION

NYC AIRBNB DATA ANALYSIS AND VISUALIZATION project provided valuable insights into the Airbnb market in New York City through the use of Python for data processing and Power BI for visualization. By leveraging Python libraries such as pandas, matplotlib, and seaborn, we were able to clean, analyze, and uncover trends in the dataset, such as price distributions, neighborhood performance, host activity, and availability patterns. Using Power BI, these insights were further enhanced through interactive dashboards, allowing stakeholders to explore data visually, compare listings, and identify potential areas for growth or optimization. This project not only demonstrates the power of combining data science with modern visualization tools but also provides actionable insights for Airbnb hosts, real estate investors, and city planners aiming to optimize Airbnb services in New York City.

### KEY FEATURES:

### Data Collection and Preprocessing:

### Utilized the publicly available NYC Airbnb dataset. Cleaned the dataset by handling missing values, outliers, and ensuring consistency in data types. Conducted feature engineering to derive new insights, such as calculating price per neighborhood, room type analysis, etc.

* **Exploratory Data Analysis (EDA) Using Python:**

Performed descriptive statistics to understand distributions of features like price, location, room types, and availability. Analyzed correlations between different variables using Python libraries such as Pandas, NumPy, and Matplotlib/Seaborn.

* **Geospatial Analysis:**

Mapped Airbnb listings using Geopandas and Folium to visualize geographical distribution. Created heatmaps to represent popular areas for Airbnb listings and pricing patterns across the city.

* **Price vs. Location Analysis**:

Compared prices of listings by neighborhoods and proximity to landmarks (e.g., Central Park, Times Square). Identified price variations by boroughs, highlighting trends like higher prices in Manhattan and affordable listings in Brooklyn.

* **Host and Review Insights**:

Analyzed host activities, including number of listings per host and identifying potential super hosts. Reviewed guest ratings and reviews to understand factors affecting guest satisfaction and host performance.

* **Time-Series Analysis**:

Conducted time-series analysis to identify seasonal trends in Airbnb bookings and pricing using Python. Examined listing availability and pricing fluctuations throughout the year.

* **Power BI Visualization**:

Created interactive dashboards and reports using Power BI to visualize the key findings. Neighborhood-level pricing maps. Dynamic filtering for room types, prices, and availability. Interactive charts for host listings, guest reviews, and booking patterns. Provided insights into business performance, supply-demand analysis, and market trends in NYC's Airbnb market.

* **Insights and Recommendations**:

**For Hosts**: Pricing strategies based on location and competition to maximize revenue.

**For Tourists**: Identified best-value neighborhoods based on budget and room type preferences.

**For Investors**: Suggested high-demand areas for new Airbnb investments based on geospatial

data and occupancy rates.

# FURTHER SCOPE OF THE PROJECT

## FURTHER SCOPE OF THE PROJECT

NYC Airbnb Data Analysis and Visualization project includes expanding the analysis to cover multiple cities globally. By building an app, the platform can integrate Airbnb data from various cities, allowing users to explore comparative analytics between locations. This will enable insights into pricing trends, demand patterns, and guest preferences across different regions, offering more comprehensive data to hosts, travelers, and investors. The app can also incorporate real-time data updates, machine learning models for price prediction and demand forecasting, and personalized recommendations based on user preferences. Additionally, the platform could feature dynamic visualizations, allowing users to interact with the data for any selected city, facilitating better decision-making and market understanding.

### Key Features:

### Incorporating Additional Cities:

### Expand the analysis to include Airbnb data from other major cities (e.g., Los Angeles, London, Paris). This will allow for comparative studies of different cities' markets and trends.

### Automated Data Ingestion:

### Develop a system that automatically ingests and updates Airbnb data periodically using APIs, ensuring real-time analysis for multiple cities without manual intervention.

### Enhanced Geospatial Analysis:

### Implement advanced geospatial techniques to analyze urban zones, proximity to landmarks, and travel hubs across multiple cities, providing more in-depth location-based insights.

### Advanced Machine Learning Models:

### Integrate machine learning models (e.g., price prediction, demand forecasting) to provide hosts with recommendations for pricing strategies based on factors like seasonality and competition in various cities.

* **User Personalization:**

Add user-centric features in the app, such as custom filters or dashboards, allowing hosts, tourists, or real estate investors to view tailored insights relevant to their preferences and cities of interest.

* **Scalable and Cross-Platform App Development:**

Develop a scalable app with cloud integration to handle large datasets across different cities, ensuring performance and accessibility on multiple platforms (e.g., mobile and web).

* **Data Visualization Enhancement:**

Expand Power BI visualizations to support multi-city comparisons, providing users with city-level insights on pricing, room types, host activity, and guest reviews through interactive and comparative dashboards.

* **City-Specific Trend Reports:**

Generate detailed trend reports for each city, highlighting factors like seasonal demand, tourism trends, and popular neighborhoods, helping stakeholders make data-driven decisions tailored to each city.

# 

# BIBILIOGRAPHY

## BIBILIOGRAPHY

**BOOKS REFERRED:**

* McKinney, W. (2017). Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython. O'Reilly Media.
* Wesley, D. (2020). Mastering Geospatial Analysis with Python. Packt Publishing.
* Microsoft. (2021). Power BI Documentation. Microsoft.
* Airbnb. (2022). Airbnb Open Data Portal.
* VanderPlas, J. (2016). Python Data Science Handbook: Essential Tools for Working with Data. O'Reilly Media.
* Chen, C., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.
* OpenStreetMap Contributors. (2022). OpenStreetMap: Geospatial Data.
* Ritchie, H., & Roser, M. (2021). Tourism Data. Our World in Data.

### WEBSITES REFERRED:

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

### <https://github.com/praveendecode/Airbnb_Analysis>

### <https://github.com/roopak08/New-York-Airbnb-Analysis-using-Power-BI>

### <https://github.com/ExplainData/Airbnb_Data_Analysis>

### [Analysing New York City Airbnb Data.ipynb - Colab (google.com)](https://colab.research.google.com/github/otavio-s-s/data_science/blob/master/Analysing%20New%20York%20City%20Airbnb%20Data.ipynb)

# APPENDIX

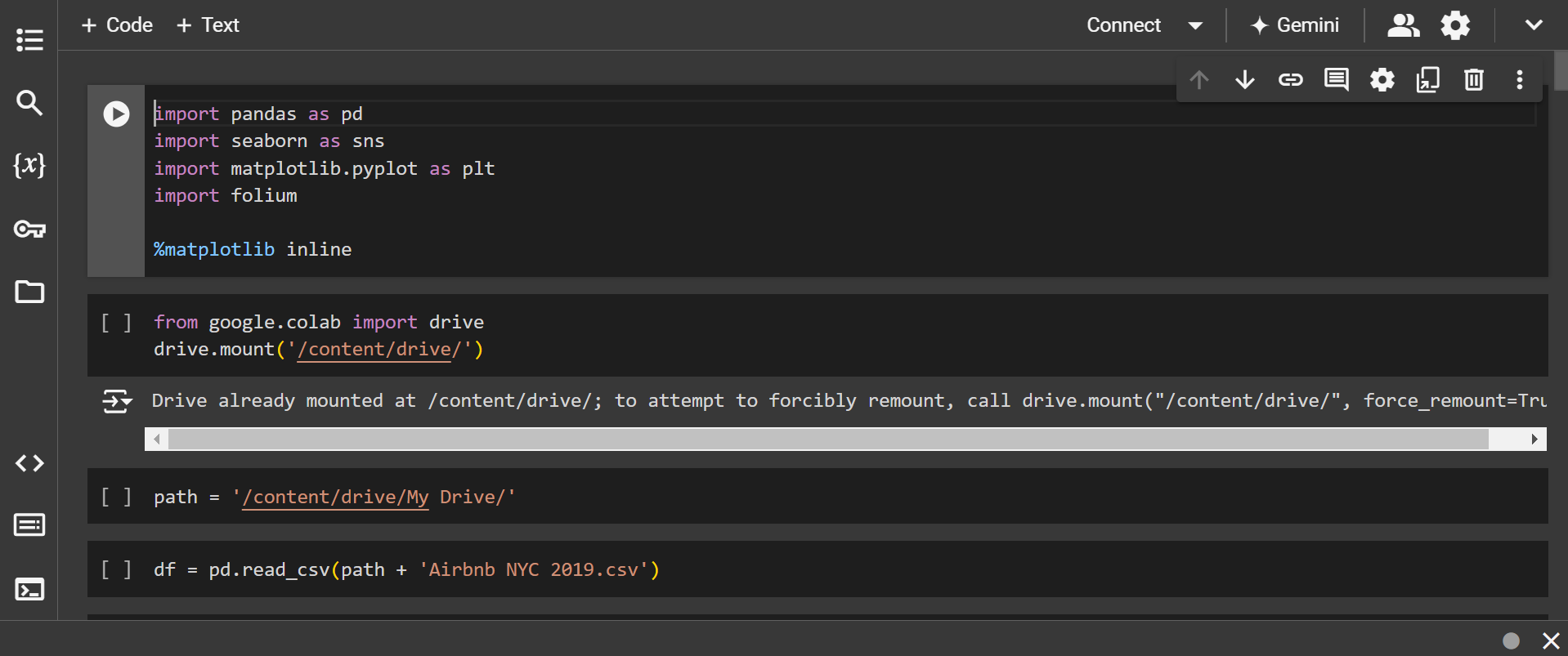
## 8. APPENDIX

**8.1 Project Design**

First, we need to collect our data, so download our dataset from Kaggle. After that, we will analyze the data using Python. Here is the code to run in Google Colab: open Google Colab and input the necessary parameters to analyze the data. We have already uploaded the dataset to our drive, so we can run our code.

**OPEN THE GOOGLE COLAB:**

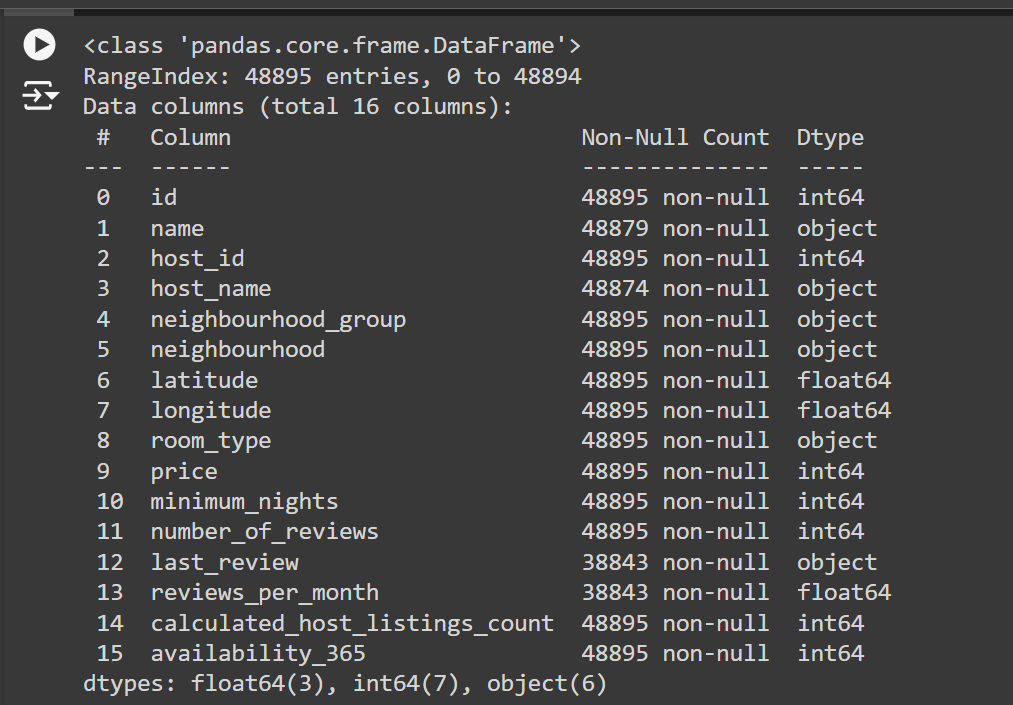
Here we are importing the essential python libraries,



After that we are analyzing the dataset

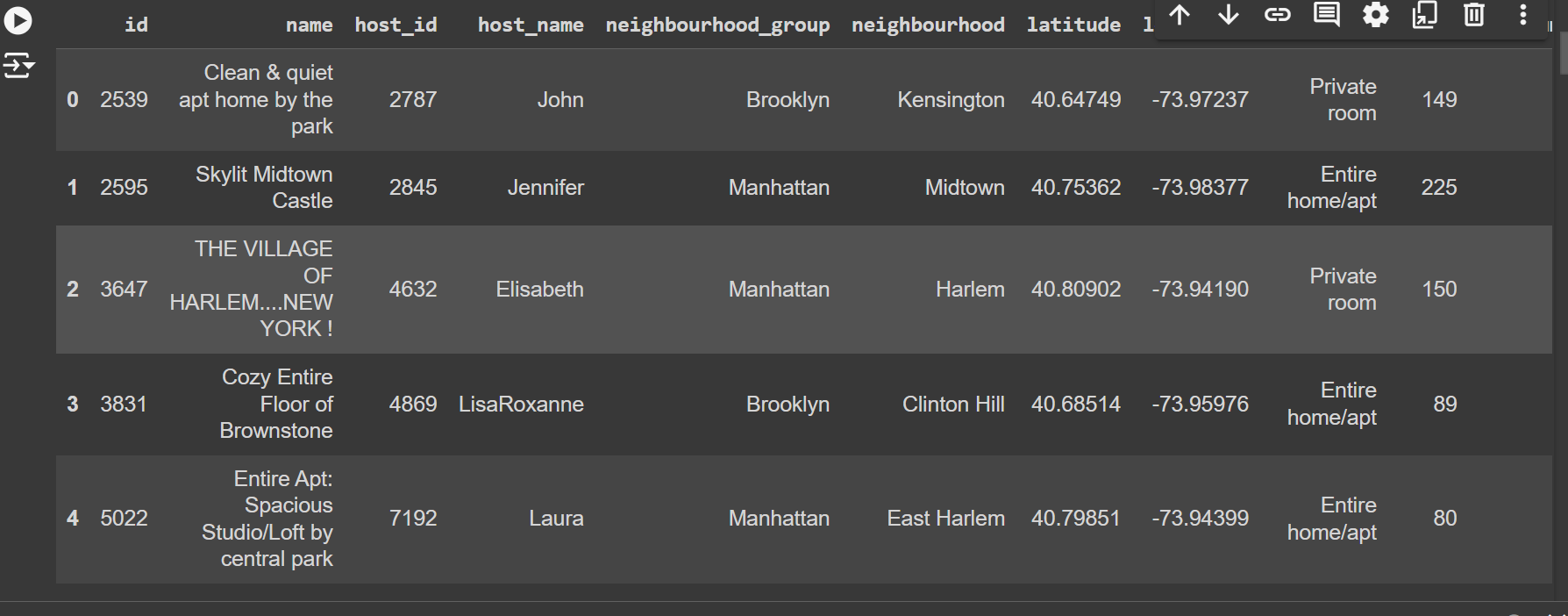
df.info()

It shows you information about a DataFrame including the index, column names, column data types, non-null values and memory usage.

****

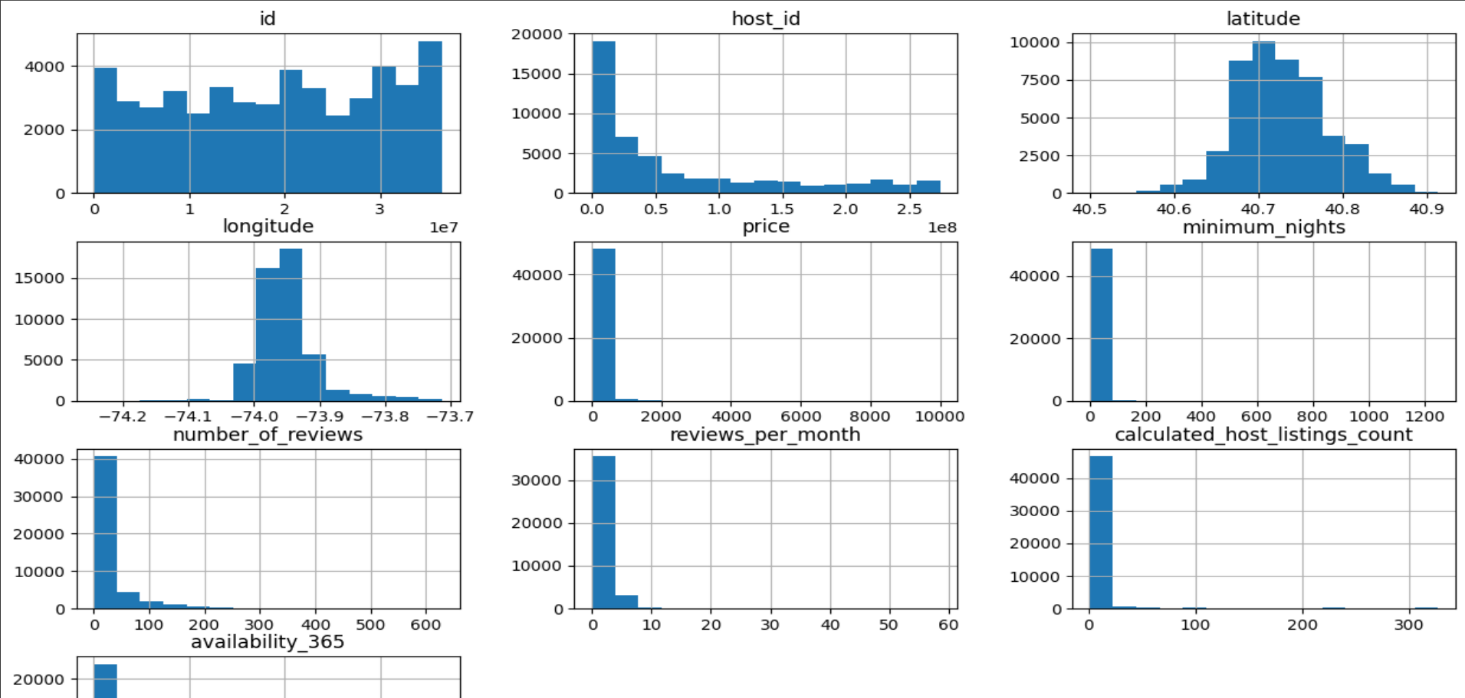
**Exploring the Dataset:**

df.head()

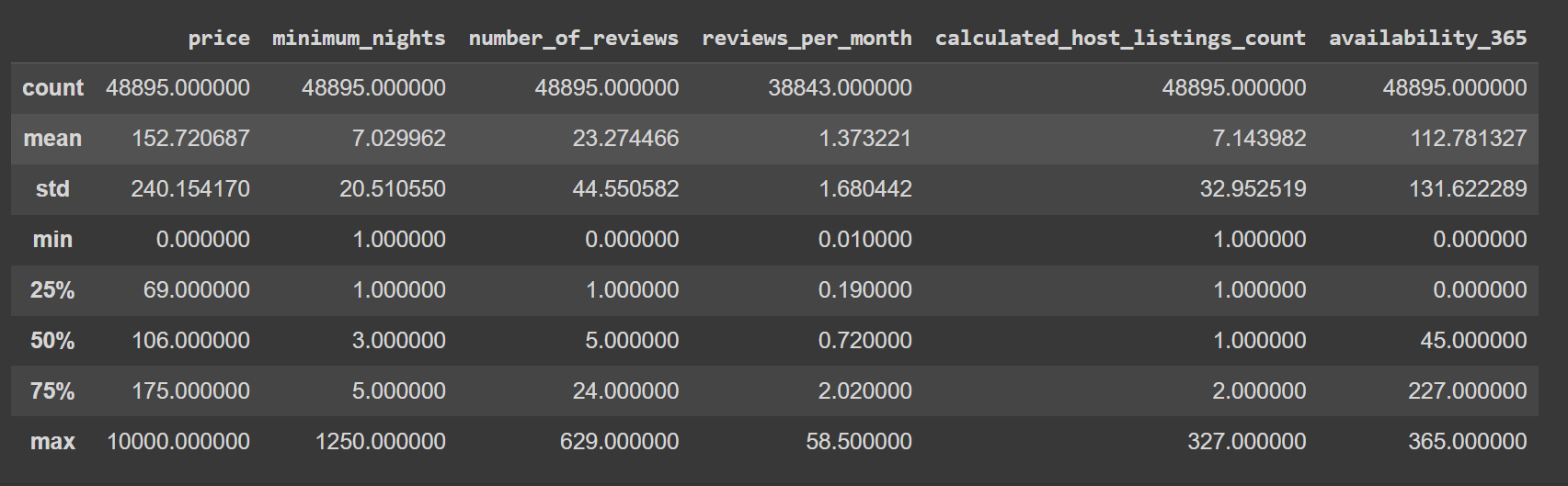


**Variable Distribution:**

df.hist(bins=15, figsize=(15,10));



df[['price', 'minimum\_nights', 'number\_of\_reviews', 'reviews\_per\_month', 'calculated\_host\_listings\_count', 'availability\_365']].describe()



**Removing Outliers:**

df['price'].plot(kind='box', vert=False, figsize=(15,3))

plt.show()

# Calculating the number and the percentage of prices that are equal to $0 and over $500.00

print('Values over $500.00: ')

print(len(df[df['price'] > 500]))

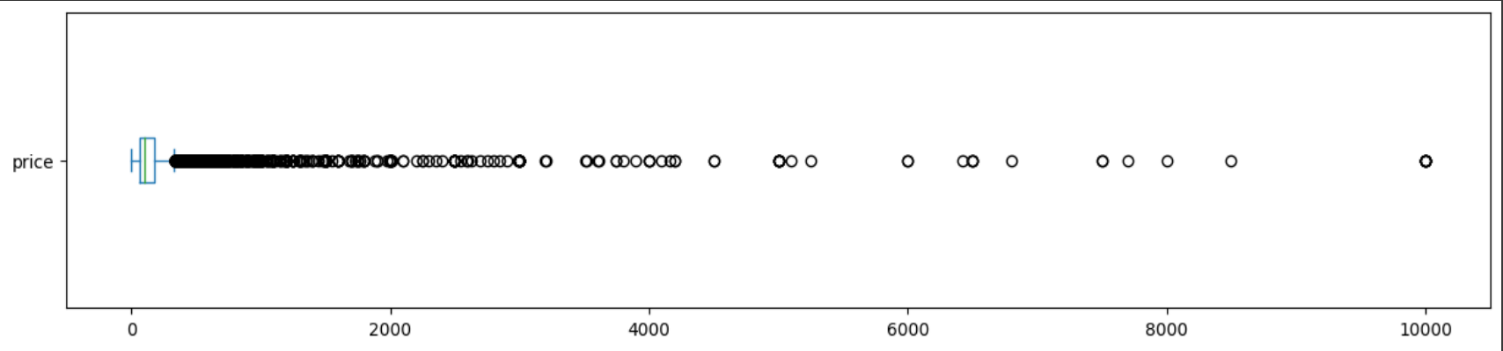
print('{:.4f}%'.format((len(df[df['price'] > 500]) / df.shape[0]) \* 100))

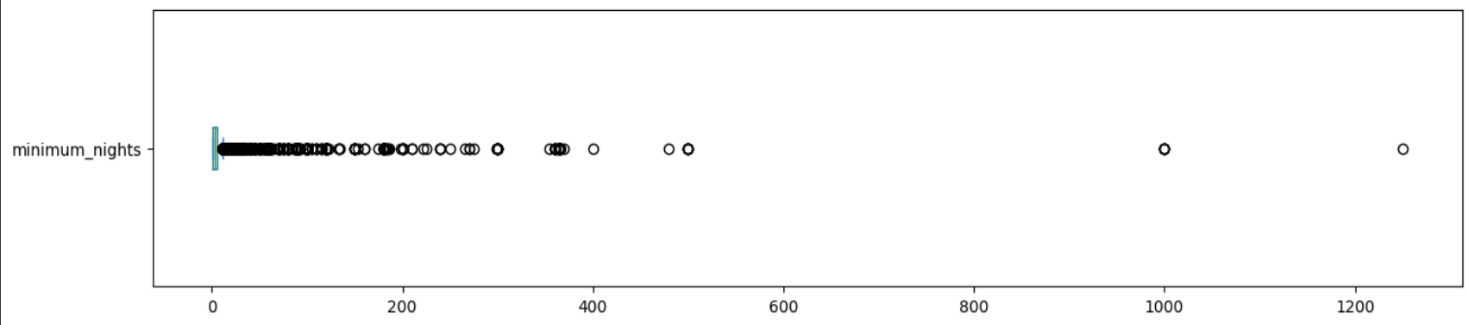
print('\n')

print('Values equal to $0: ')

print(len(df[df['price'] == 0]))

print('{:.4f}%'.format((len(df[df['price'] == 0]) / df.shape[0]) \* 100))





df['minimum\_nights'].plot(kind='box', vert=False, figsize=(15,3))

plt.show()

# Calculating the number and the percentage of rooms with the minimum nights value over 30

print('Values over 30 nights: ')

print(len(df[df['minimum\_nights'] > 30]))

print('{:.4f}%'.format((len(df[df['minimum\_nights'] > 30]) / df.shape[0]) \* 100))

df\_clean = df[(df['price'] <= 500) & (df['price'] > 0) & (df['minimum\_nights'] <= 30)].copy()

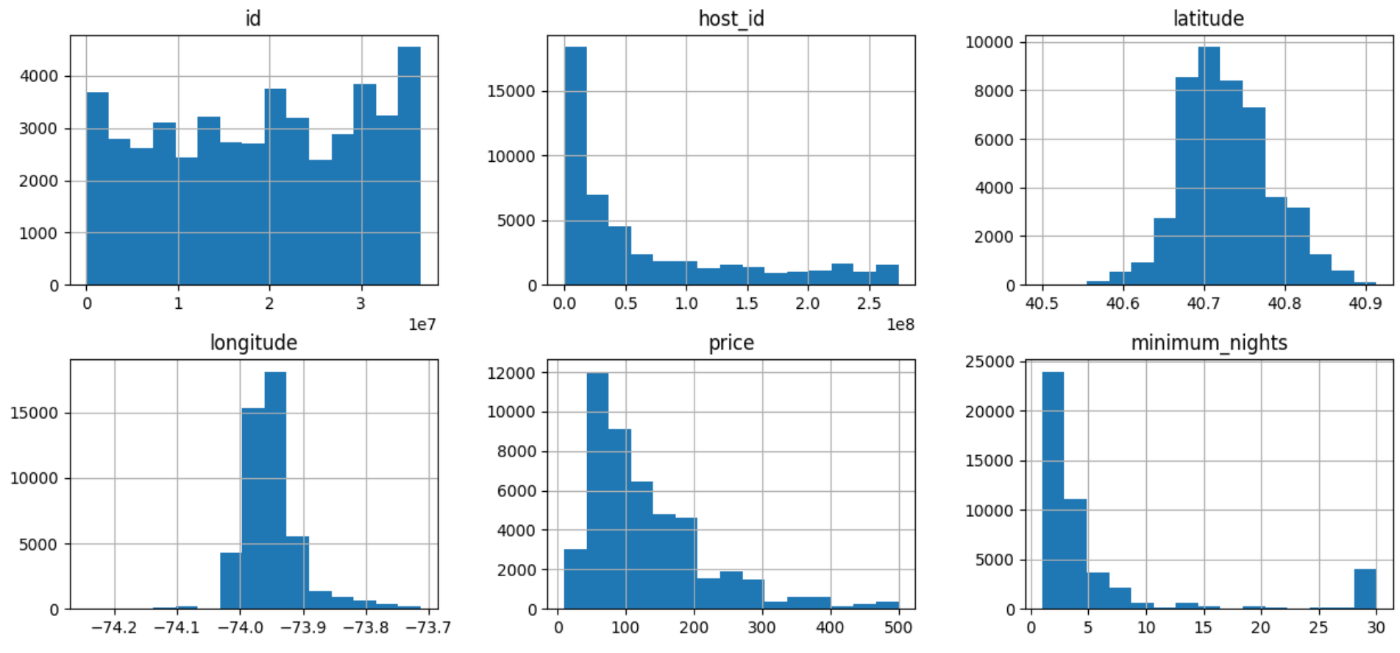
# Droping the columns

df\_clean.drop(['reviews\_per\_month', 'last\_review'], axis=1, inplace=True)

df\_clean.reset\_index(drop=True, inplace=True)

# Plotting the histograms againd

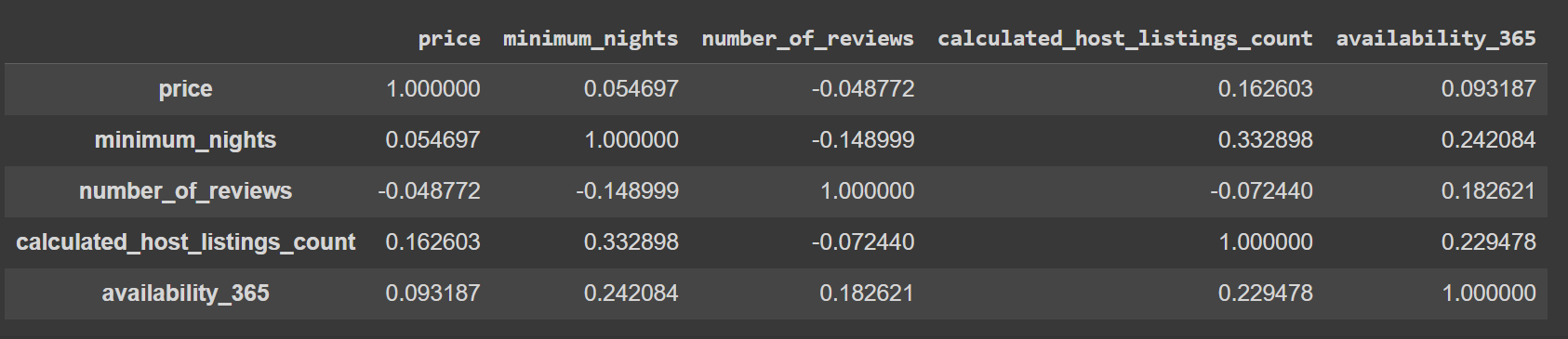
df\_clean.hist(bins=15, figsize=(15,10));



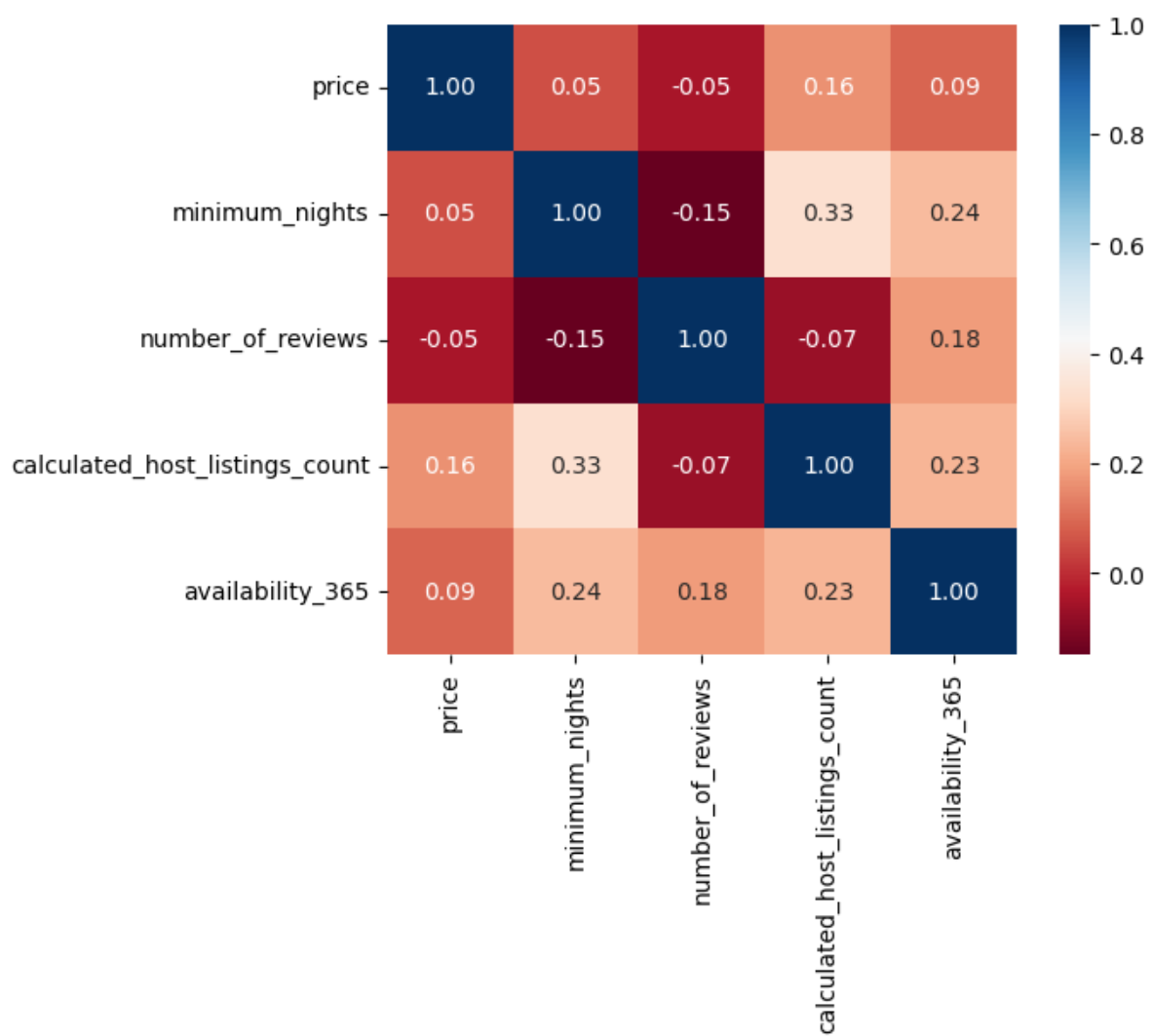
**Correlations:**

corr = df\_clean[['price', 'minimum\_nights', 'number\_of\_reviews', 'calculated\_host\_listings\_count', 'availability\_365']].corr()

corr



sns.heatmap(corr, cmap='RdBu', fmt='.2f', square=True, linecolor='white', annot=True);



**Predictions:**

area = areas\_reviews['neighbourhood\_group']

review = areas\_reviews['number\_of\_reviews']

fig = plt.figure(figsize = (10, 5))

# creating the bar plot

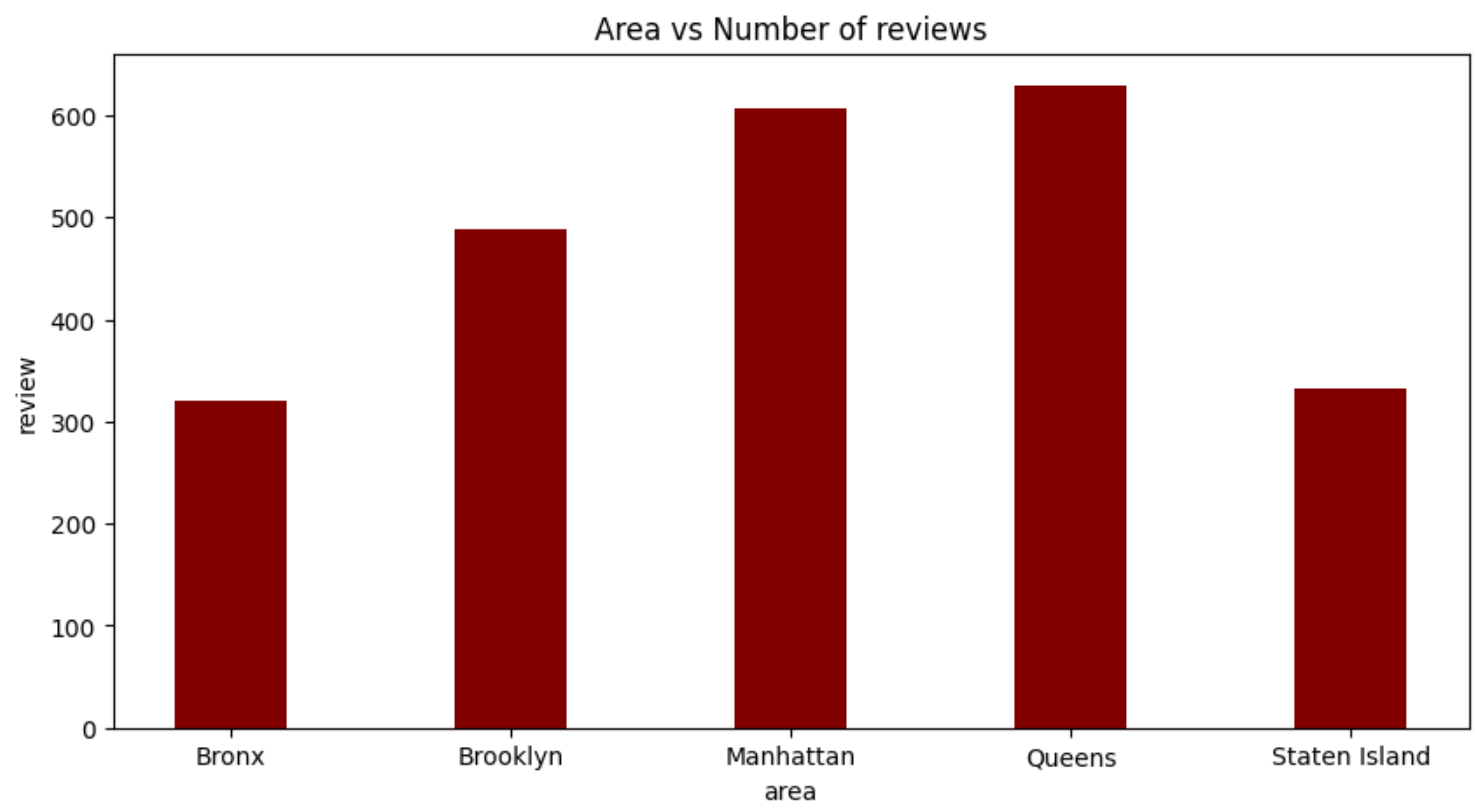
plt.bar(area, review, color ='maroon', width = 0.4)

plt.xlabel("area")

plt.ylabel("review")

plt.title("Area vs Number of reviews")

plt.show()



area = price\_area['price']

price = price\_area['number\_of\_reviews']

fig = plt.figure(figsize = (10, 5))

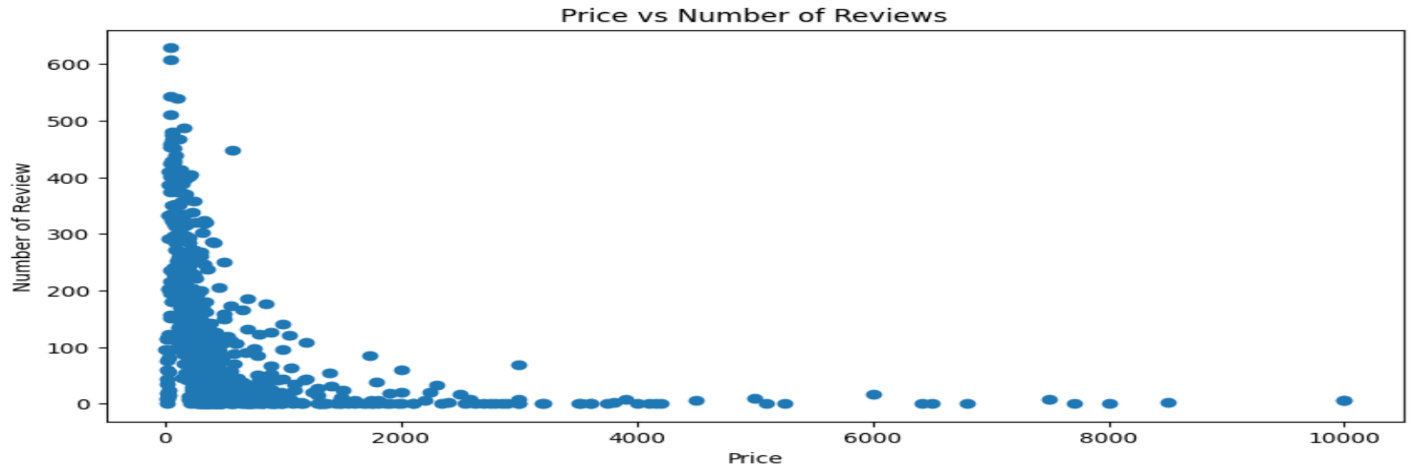
plt.scatter(area, price)

plt.xlabel("Price")

plt.ylabel("Number of Review")

plt.title("Price vs Number of Reviews")

plt.show()



**Which hosts are the busiest:**

name = busiest\_hosts['host\_name']

reviews = busiest\_hosts['number\_of\_reviews']

fig = plt.figure(figsize = (10, 5))

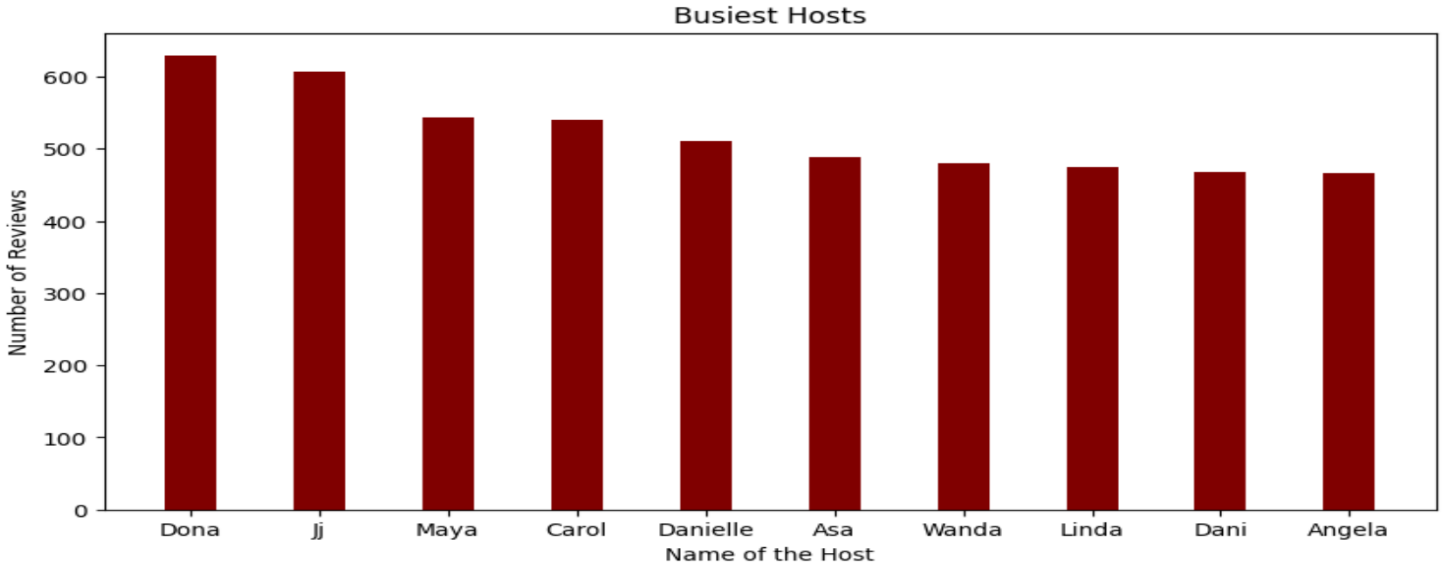
plt.bar(name, reviews, color ='maroon', width = 0.4)

plt.xlabel("Name of the Host")

plt.ylabel("Number of Reviews")

plt.title("Busiest Hosts")

plt.show()

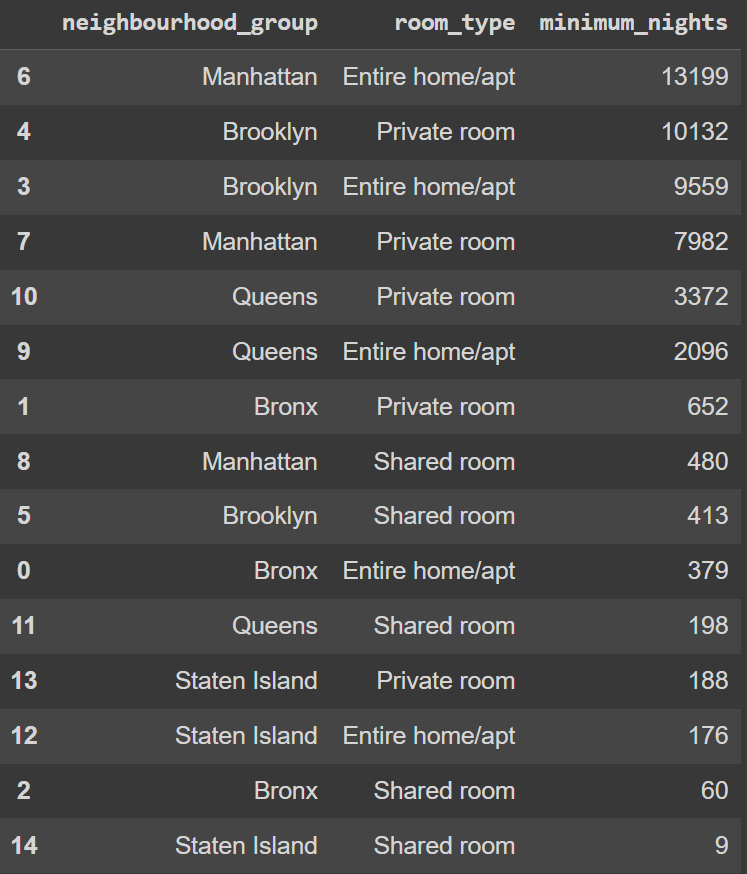


**Is There a Noticeable Difference in Traffic Among Different Areas?**

traffic\_areas = new\_df.groupby(['neighbourhood\_group','room\_type'])['minimum\_nights'].count().reset\_index()

traffic\_areas = traffic\_areas.sort\_values(by='minimum\_nights', ascending=False)

traffic\_areas



room\_type = traffic\_areas['room\_type']

stayed = traffic\_areas['minimum\_nights']

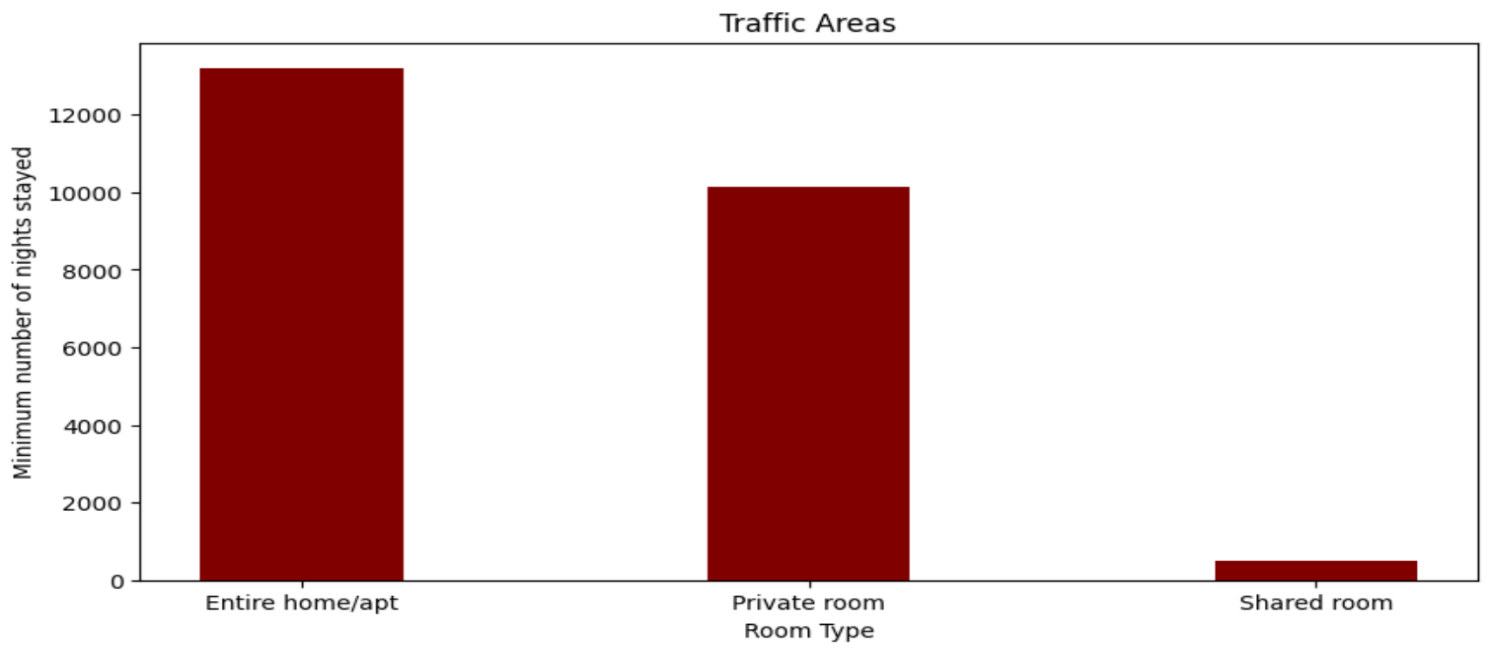
fig = plt.figure(figsize = (10, 5))plt.bar(room\_type, stayed, color ='maroon', width = 0.4)

plt.xlabel("Room Type")

plt.ylabel("Minimum number of nights stayed")

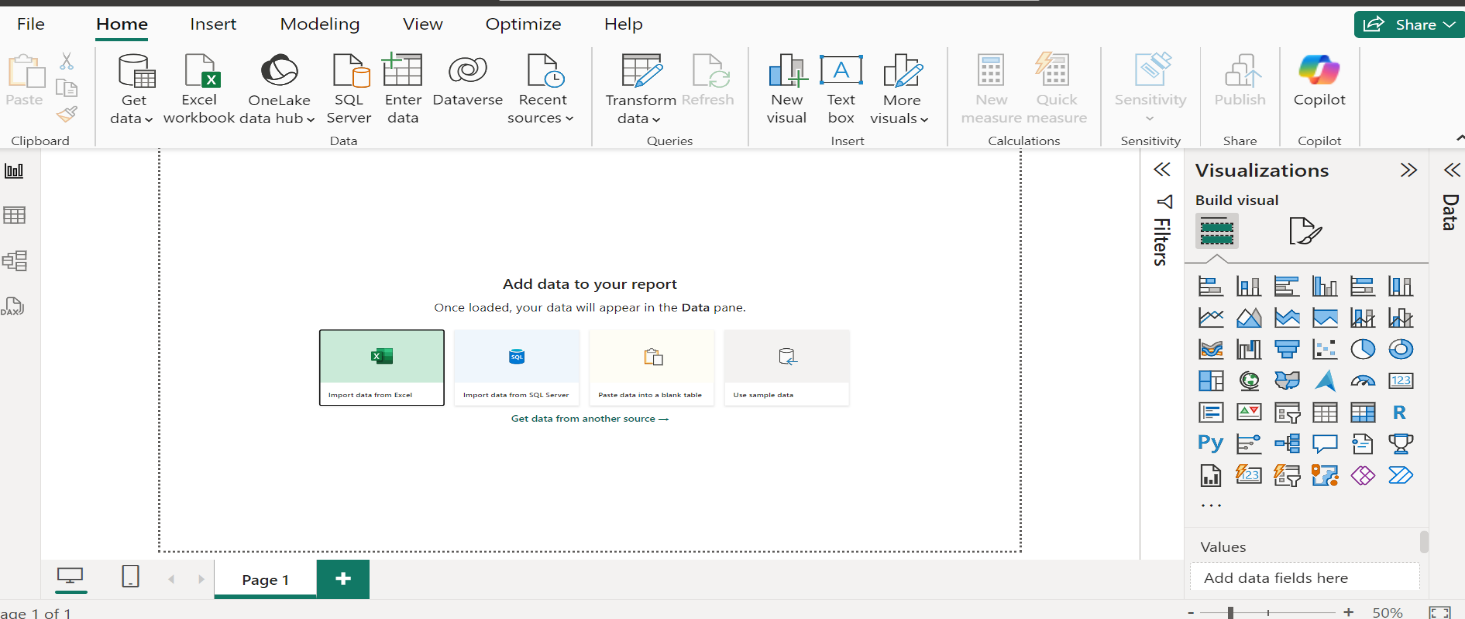
plt.title("Traffic Areas")

plt.show()

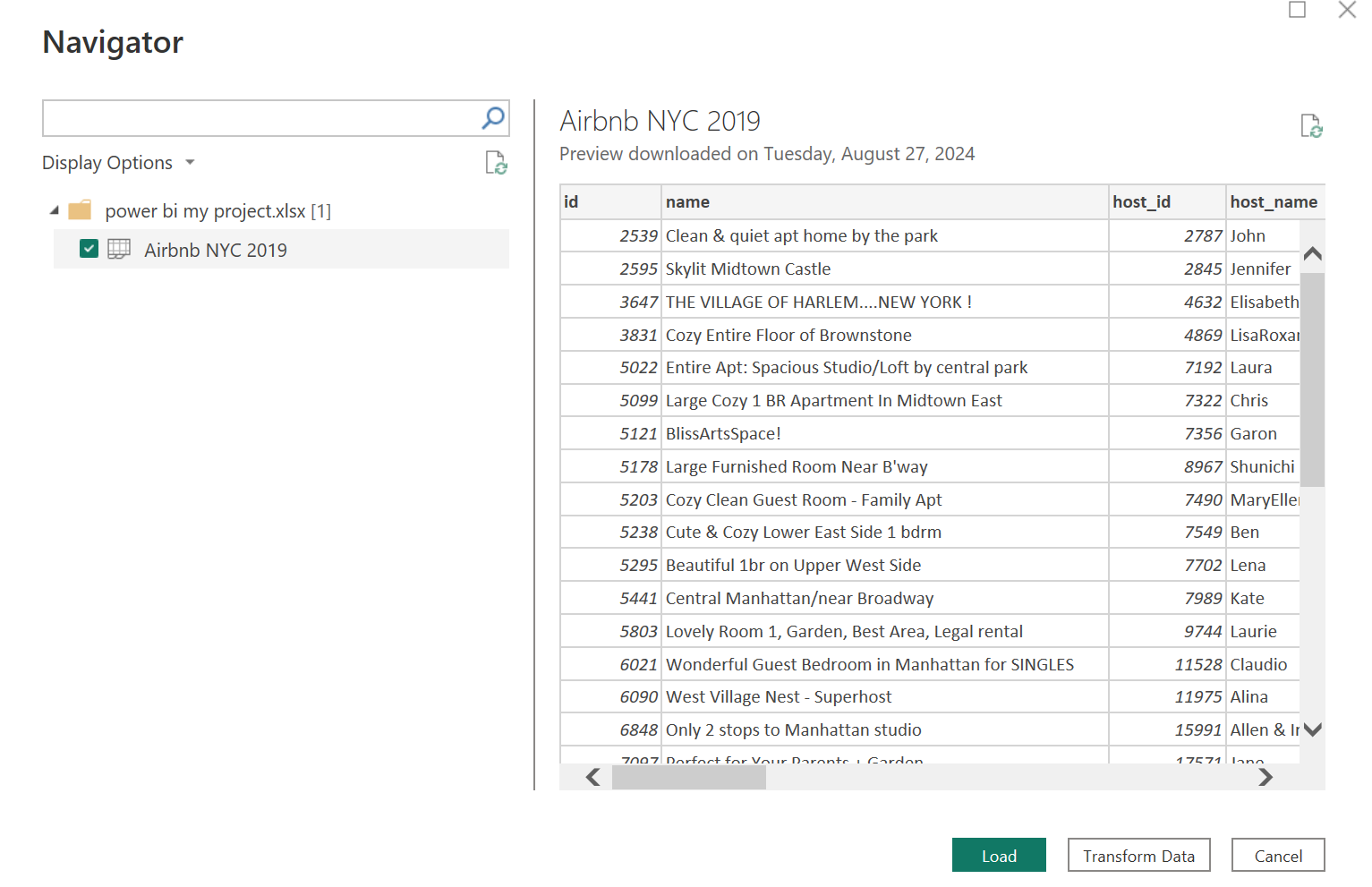


**MICROSOFT POWER BI VISULAIZATION**

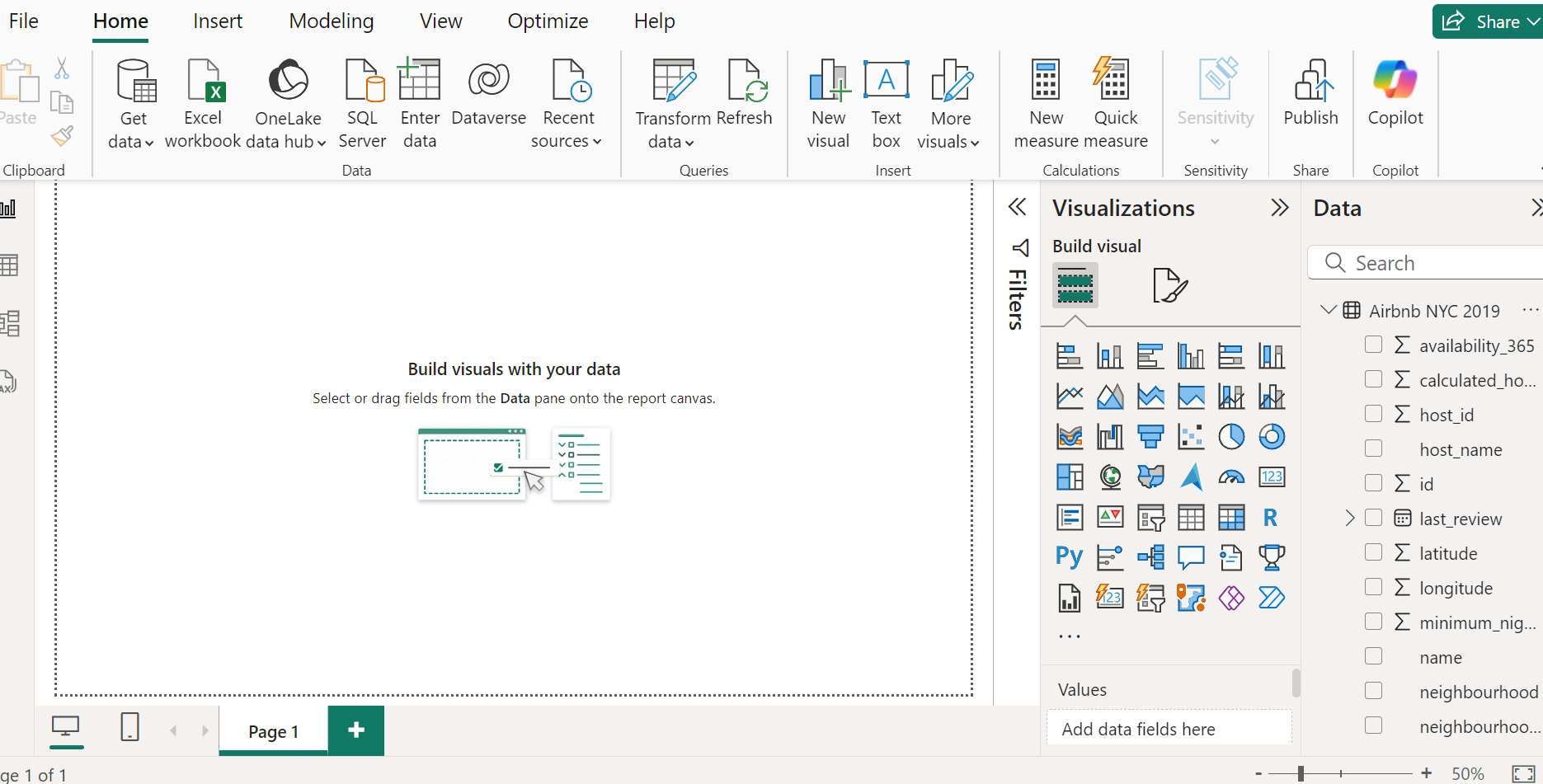
Open Power BI Desktop, insert the dataset into Power BI, and then visualize our data.



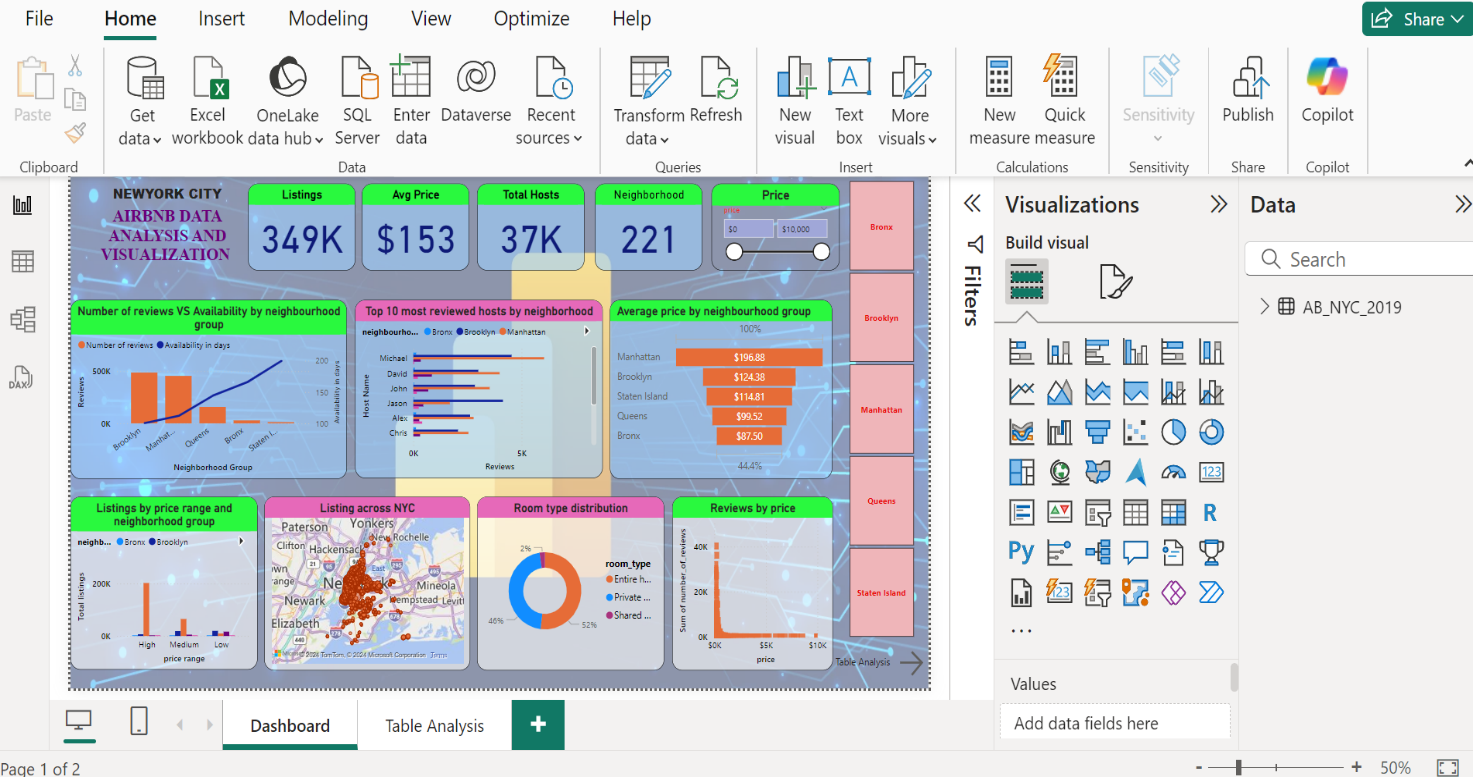
Click excel to import the data and load.



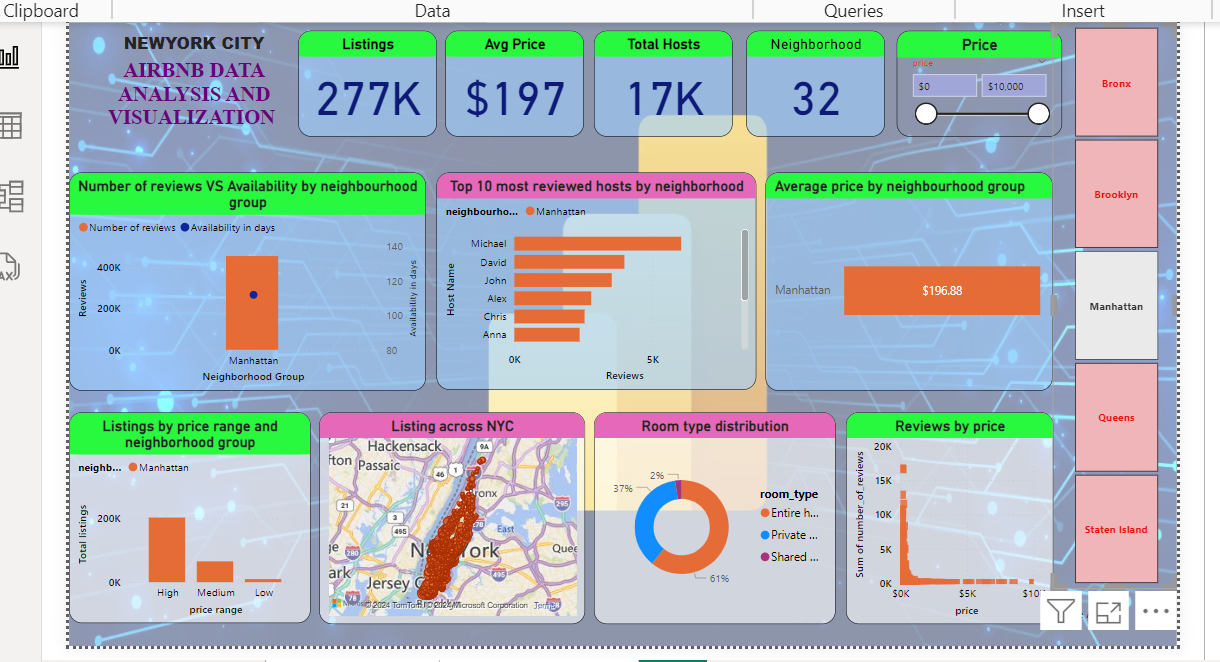
Now we want to visualize the massive data to meaningful insight.



Select the data and visualizing the data . selecting charts,card etc to visualize the data. Here is the final output.



When we want to analyze the entire dataset quickly, we can simply click a button. For instance, if we’re interested in Manhattan information only, clicking that button will transform the entire dashboard to display only the selected city.



So, in this way to we can perform the datasets.

**Conclusion**

During the project, we performed some interesting analysis of the New York City Airbnb data and managed to answer some questions, such as:

* What kind of room is more common in New York City Airbnb?
* What is the price difference between different types of rooms?
* What are the most expensive regions to stay in New York?

To accomplish such goals, we went through major data manipulation steps, such as exploring, cleaning, analyzing, and visualizing data.

With all that said, the conclusions are:

* Private rooms and entire apartments are the most common room types;
* Hotel rooms and entire apartments are usually more expensive than private and shared rooms;
* Over 80% of the rooms are located in Manhattan and Brooklyn, which are also the most expensive regions;
* Yes, if you want to stay close to the major attractions of the city you'll probably expend more money.