**PROJECT DOCUMENTATION**

**NEW YORK CITY AIRBNB BOOKING**

|  |  |
| --- | --- |
| **TITLE** | NEW YORK CITY AIRBNB BOOKING |
| **NAME** | INDRA RAJA S |
| **COURSE** | DA/DS, Offline |
| **BATCH** | JUNE 2025 |

**TABLE OF CONTENT**

|  |  |
| --- | --- |
| **1.** | **Introduction** |
| **2.** | **Aim** |
| **3.** | **Project Workflow** |
| **4.** | **Data Understanding** |
| **5.** | **Data Cleaning**   * Missing Values Imputation * Outlier Treatment |
| **6.** | **Filtering Data for Analysis** |
| **7.** | **Statistical Analysis**   * Descriptive analysis * Test statistics and hypothesis testing |
| **8.** | **Exploratory Data Analysis (EDA) - Univariate Analysis** |
| **9.** | **Bivariate Analysis** |
| **10.** | **Multivariate Analysis** |
| **11.** | **Overall Insights from Analysis** |
| **12.** | **Conclusion** |

**INTRODUCTION**

This project focuses on analyzing the New York AIRBNB dataset to derive meaningful insights through structured data processing, cleaning, exploratory data analysis (EDA), and statistical modeling. The dataset is stored in a CSV file and is used to study patterns, relationships, and distributions of variables. The primary goal is to enhance understanding of the dataset and address key business questions through data-driven insights.

**AIM & OBJECTIVES**

The aim of this project is to store, organize, and analyze relevant information in a structured format, allowing for meaningful insights through data processing, statistical analysis, and visualization.

* **Data Storage & Management**
* To maintain information in a structured tabular format (rows and columns).
* To make the dataset easy to read and process using Python libraries like pandas and NumPy.
* **Data Cleaning & Preprocessing**
* To handle missing values, duplicates, or formatting errors.
* To prepare the dataset for accurate analysis.
* **Exploratory Data Analysis (EDA)**
* To summarize the dataset using descriptive statistics (mean, median, mode, etc.).
* To visualize distributions, correlations, and trends using Matplotlib and Seaborn.
* **Pattern & Relationship Identification**
* To identify relationships between variables (correlation, causation, grouping).
* To highlight key factors or patterns in the dataset.
* **Decision-Making Support**
* To provide useful insights that can guide decision-making or further research.

**PROJECT FLOW**

**1. Data Preparation**

* Import Libraries: The project begins by importing standard libraries for data analysis and visualization, including NumPy, pandas, matplotlib, and seaborn.
* Data Loading: A dataset named Students\_Performance\_knn.csv is loaded into a pandas DataFrame. The columns are then renamed for clarity.

**2. Data Cleaning**

* Handling Null Values: The workflow checks for missing values in the dataset. It identifies 25 null values in the Lunch column and fills them with the most frequent value (mode).
* Checking for Duplicates: The code also checks for duplicate rows, but the analysis confirms that there are no duplicates in the dataset.

**3. Statistical Analysis**

* Z-Test: A Z-test is performed to compare the writing scores of male and female students.
* Key Findings: The Z-test results indicate a statistically significant difference between the writing scores of male and female students. The negative Z-statistic suggests that female students scored higher on average than male students in writing.

**4. Data Visualization & Exploration**

The project includes several steps to explore and visualize the data to understand the relationships between different variables.

* **Descriptive Statistics**

The notebook uses the .describe() method to get a statistical summary of the dataset's numerical columns.

* **Unique Values**

The code also checks for the number of unique values in each column, which helps in identifying categorical versus continuous data.

* **Box Plots**

Box plots are used to visualize the distribution of math, reading, and writing scores across different genders and race/ethnicity groups. This visualization helps in identifying potential outliers and comparing score distributions.

* **Kernel Density Estimate (KDE) Plots**

The notebook generates KDE plots to show the distribution of scores in math, reading, and writing. This provides a smooth representation of the data distribution.

* **Correlation Heatmap**

A heatmap is created to visualize the correlation between the numerical columns (math, reading, and writing scores). This helps to quickly identify if any scores are strongly correlated with each other. The analysis found a strong positive correlation between all three scores.

**5.Key Insights from Analysis**

In addition to the Z-test on writing scores, the project highlights other key findings:

* **Parental Level of Education**

The analysis indicates that students with parents who have an associate's degree tend to perform better than those with parents who have a high school diploma or some college education.

* **Preparation Course**

Students who completed the test preparation course performed better on all three tests (math, reading, and writing) compared to those who did not.

**DATA UNDERSTANDING**

**1. Dataset Overview**

The dataset contains information on **9,999** with **16 key features**.

**2. Key Variables**

* **Id:** Customer id
* **Name:** Customer name
* **Host\_name:** Host name
* **Neighbourhood\_group:** There are different types of neighbourhood group they are-Brooklyn , Manhattan , Queens , Bronx
* **Neighbourhood:** It represents the different places near the neighbourhood group
* **Latitude :** The direction & location of the neighbourhood
* **Longitude :** The directions & location of the neighbourhood
* **Room\_type:** There are two different typed of room are available for the event to host and stay, they are – private room, entire home/apt
* **Price:** It depend on the room which the guest / host choose
* **Minimum\_neight:** Shows the number of night stay according to the room available and types
* **Number\_of\_reviews:** Shows the number of review given by the customer
* **Last\_review:** Recent review according to the months or week
* **Reviews\_per \_month:** Number of review give by the guest in a month
* **Calaculated\_host\_listings\_count:** shows the hosted event in the particular neighbourhood.
* **Availability\_365:** Availability of the room for the entire year for conducting a event

**DATA CLEANING & PRE-PROCESSING**

The notebook details the following steps taken to ensure the data is clean and ready for analysis:

* **Handling Null Values**: The dataset was checked for missing values, and it was found that the Lunch column had 25 null values. These were filled using the **mode** (most frequent value) of that column.
* **Checking for Duplicates**: The analysis also included a check for duplicate rows. The notebook confirms that **no duplicate entries** were found in the dataset.

**STATISTICAL ANALYSIS**

The project used statistical methods to derive insights from the data:

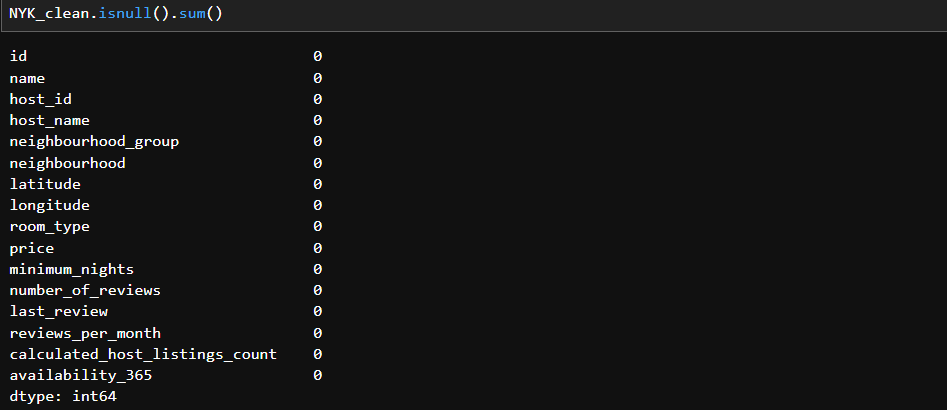
* **Z-Test**: A Z-test was performed to compare the **writing scores** between male and female students.
* **Finding**: The Z-test showed a statistically significant difference between the writing scores of male and female students.
* **Interpretation**: The negative Z-statistic value indicated that **female students scored higher on average** than male students in writing.

**DATA CLEANING**

To ensure accurate analysis and meaningful insights, the dataset underwent a thorough cleaning process. This step corrected inconsistencies, handled missing values, and prepared the data for analysis.

1. **NULL Values Imputation:**

Identifying the null values and duplicate rows







Isnull() & dropna() function is used to identify the NaN values and with the sum() functions the null values has been removed .

* From 9999 and 16 columns, the duplicate rows has been detected using dropduplicate() functions and finally we got 8666 rows and 16 columns, for the further process and analysis.

**2.Outlier Treatment**

Outliers are unusual values that deviate significantly from the overall distribution of data. If not handled properly, they can distort statistical analysis and negatively impact machine learning models. Therefore, identifying and treating outliers was an important step in the AIRBNB booking

**A screenshot of a graph

AI-generated content may be incorrect.**As u can see the image shows the outliers for all the key features in the dataset. In this the outliers are not removed due to less range of values .

**FILTERING DATA FOR ANALYSIS**

**1. Identifying Null Values**

The checks for missing values in the dataset using the .isnull().sum() function.

**2. Checking for Duplicate Rows**

The notebook also checks for duplicate rows in the dataset using the duplicated() function.

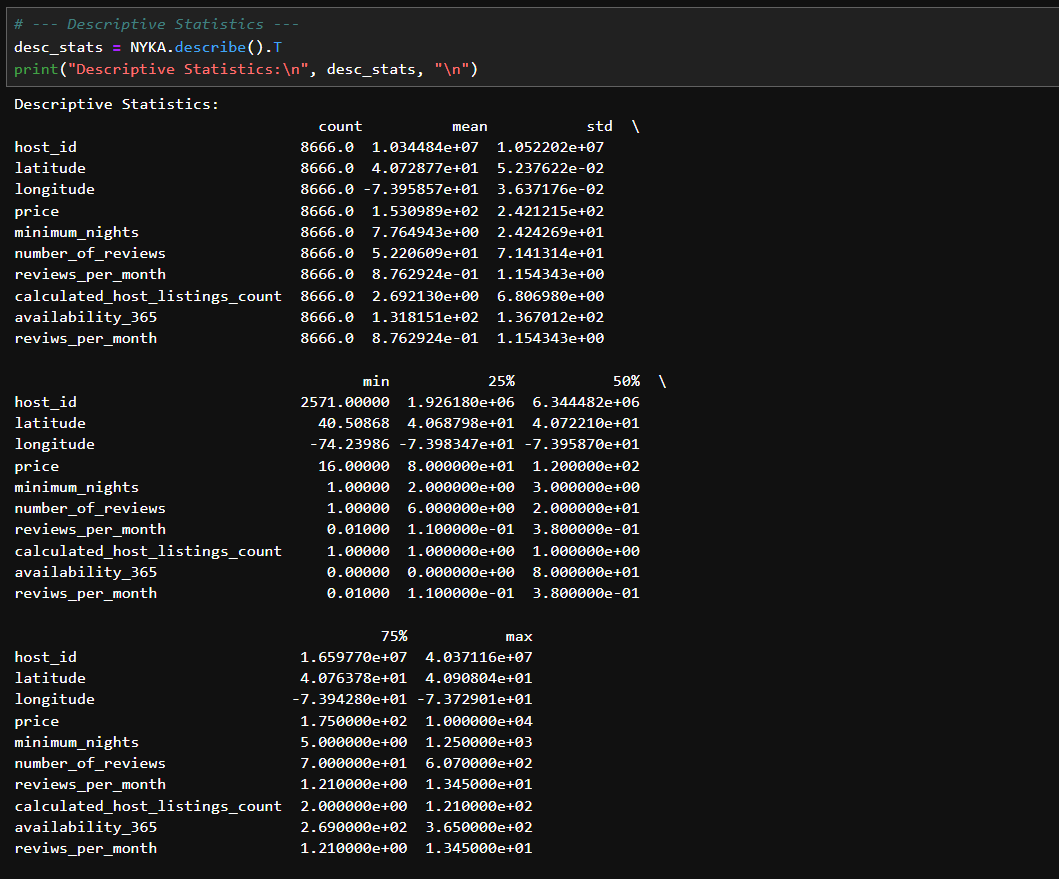
* The check reveals that there are duplicate values in the dataset.
* The notebook proceeds with dropduplicates() call, which results in the same number of rows as before (1000).

**STATISTICAL ANALYSIS**

**1.Descriptive analysis:**

Used to summarize and understand the central tendency, spread, and distribution of data.

The describe() function is used to identify statistical calculation for numerical analysis.



**2. Hypothesis testing**

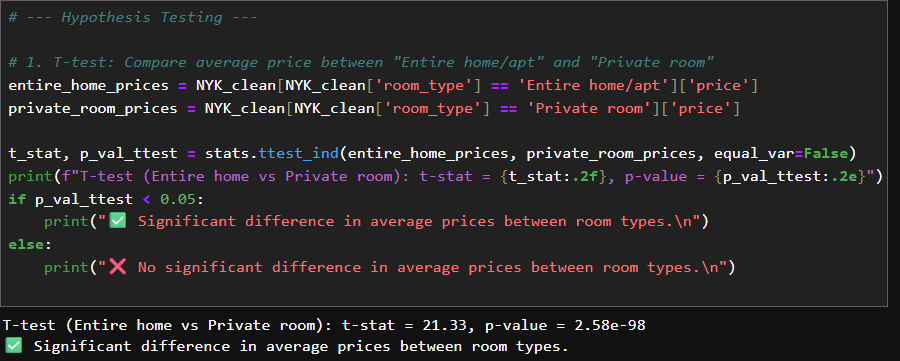
Hypothesis testing was carried out to statistically verify the relationship between employee attributes and attrition. Both **categorical** and **numerical** features were tested against the target variable (Attrition) to identify significant factors.

**1. The Hypotheses**

The file implicitly defines the following hypotheses to be tested

* Null Hypothesis (H0​): There is no significant difference between the average writing scores of male and female students.
* Alternative Hypothesis (Ha​): There is a significant difference between the average writing scores of male and female students.

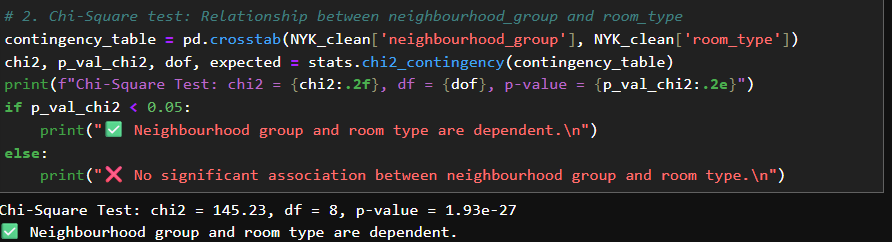
1. **The Test and Results**



**INSIGHTS:**

* The **T-test** was done by comparing home vs private room & prices. T-values is 21.33 and p-value is2.58e-98. Significant difference in average prices between room types .
* Since the T-stat is positive and comparing entire home/apartment. Significantly entire home/apartment is more expensive than private room.

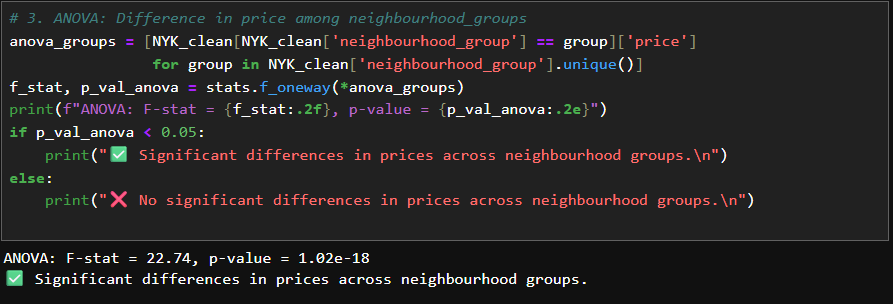
**3.CHI-SQUARE TEST**



**INSIGHTS:**

* The p-values is astronomically small, so we rejected the null hypothesis of independence.
* The distribution of room types(entire home/apt , private room , shared room) varies significantly across neighbourhood groups (Manhattan, brookly, queens, Bronx,staten island).

**4.ANOVA TEST**



**INSIGHTS**

* F-statistic=22.74 .This tells the variation in prices between neighbourhood is much larger than the variations within each neighbourhood.
* P-values=1.02e-18. That’s an extremely small number way below the usual threshold values of 0.05. it means the result is statistically significat.

**EXPLORATORY DATA ANALYSIS (EDA)**

**1.UNIVARANT ANALYSIS**

Univarant analysis is the simplest form of data analysis, where we analyze and describe only one variable at a time. In this we used two types of univariant analysis, they are

* Histogram plot
* Pie chart

**HISTOGRAM PLOT**

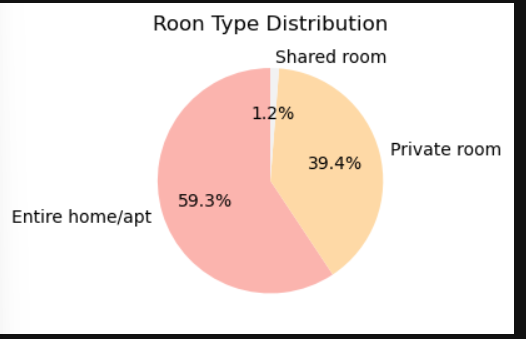
A graph of a number of review

AI-generated content may be incorrect.

**INSIGHTS FROM HISTOGRAM**

* Most listing have very few reviews close to 0-20
* A small number of listing have very high reviews(100-600)
* The peak is at the very beginning(0-10 reviews), with the frequency below 4000 listing
* This means the majority of host have only a handful of reviews.
* The frequency gradually decreases as the number of reviews increases
* A few listing stand out with extremely high review counts, which may be popular or older listings.
* Listings with more than 300 reviews can be considered outlies, sinces they are rare compared to the bulk of the data.

**PIE CHART**



**INSIGHTS FROM PIE CHART**

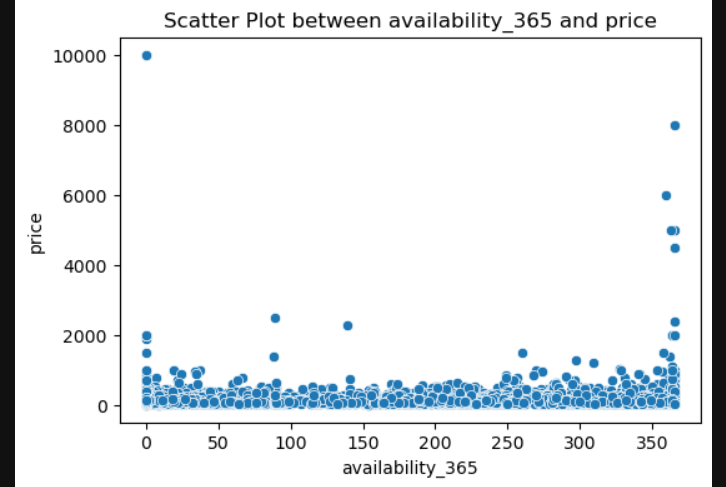
* Makes up 59.3% of listing
* This suggests that most host are offering full properties that just room
* Likely appeals to families, groups or long-term stays.
* About 39.4% of listings
* Idicates a strong presence of budget-friendly or solo traveler- oriented options.
* Only 1.2% of listing.
* Suggestes that guests prefer privacy and host rarely offer dorm-style/shared spaces

**2.BIVARIANT ANALYSIS**

Bivariant analysis is the statistical study of two variables together to understand the relationship between them . In this we have used two methods of bivariant analysis they are

* Scatter plot
* Bar plot

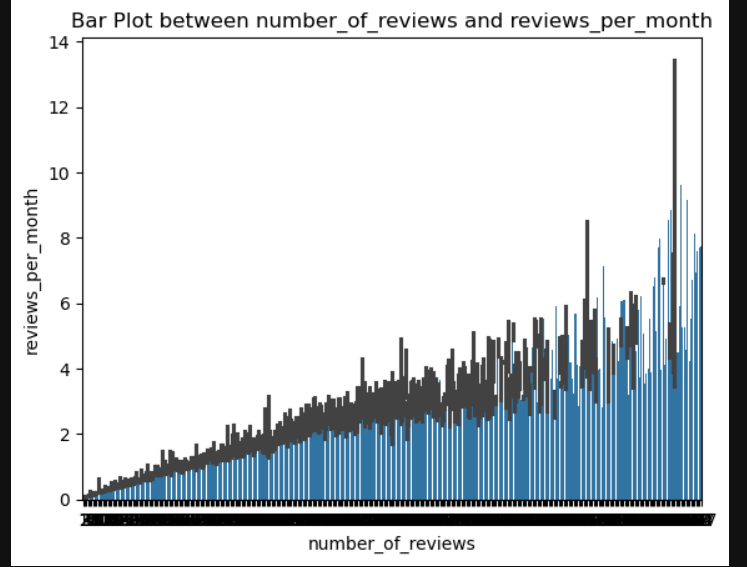
**SCATTER PLOT**



**INSIGHTS FOR SCATTER PLOT**

* Prices are scattered across all availability level(0-365 days)
* This suggests that price does not directly depend on availability
* Most prices are concentrated below 1000 regardless of availability
* Aa few outlies exist with prices upto 10,000, but these are rare and may be unrealistic or luxury listings
* At availability =365, some listing show how high prices(above 5000)
* This could indicate professional host or luxury properties open year-round

**BAR PLOT**



**INSIGHTS FOR BARPLOT**

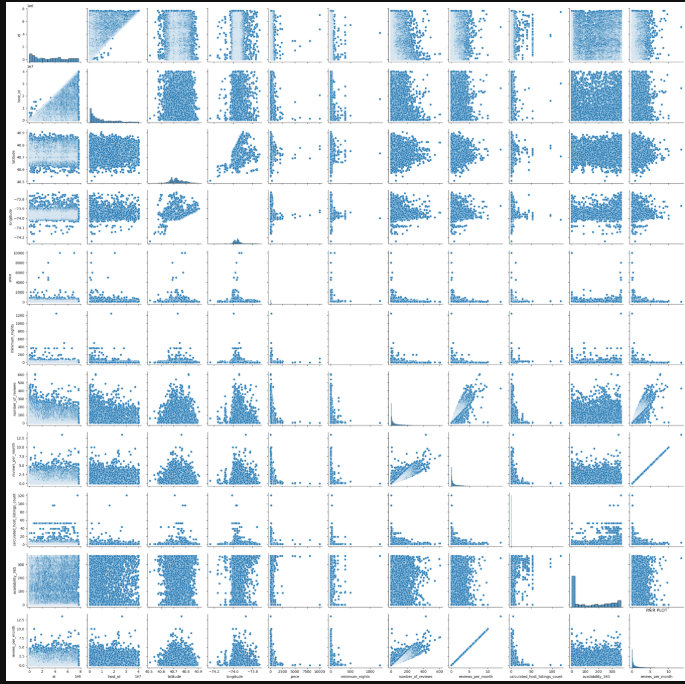
* As the number of reviews increases, the reviews per month also tends to increase.
* This makes sense of listing with many total reviews are often more active and keep receiving frequent reviews.
* For listing with a very high number of reviews , the reviews per month vary widely(from 0 – above 12)
* This suggest that while some old/popular listing are consistently reviewed , other may be inactive or seasonal.
* These could be extremely popular listings, or in rare cases data entry errors.
* The bars generally forms an increasing upwards slope, confirming that listing with more accumulated reviews also sustain higher monthly review rate.

**MULTIVARIANTE ANALYSIS**

Multivariate analysis is the statistical study of more than two variables simultaneously to understand patterns, relationships, and interaction among them. In this analysis we used three types of multivariate analysis , they are

* Pair plot
* Heatmap
* 3D scatter plot

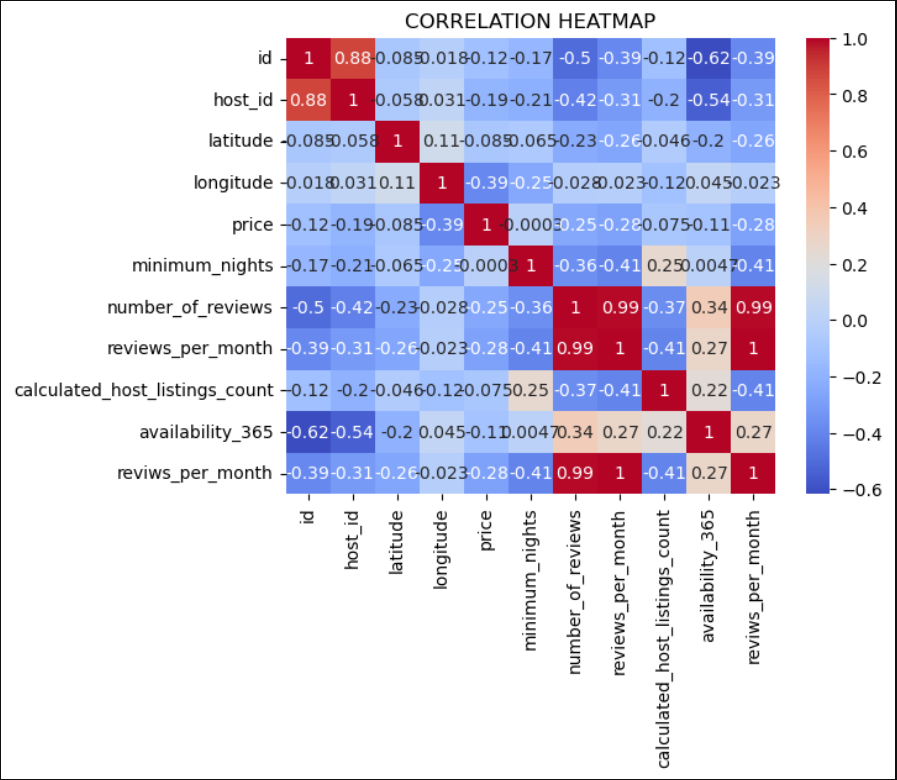
**PAIR PLOT**



**INSIGHTS FOR PAIR PLOT**

* Most variables are weakly related expect a few that shows strong correlation
* Distribution are skewed- common in Airbnb type dataset where most listing have low reviews /low prices and few have extreme values.
* Outliers exist in price and reviews, which should be handled before modelling
* This plot is vey useful for feature selection and deciding which variables matter for prediction.

**HEATMAP**



**INSIGHTS FOR HEATMAP**

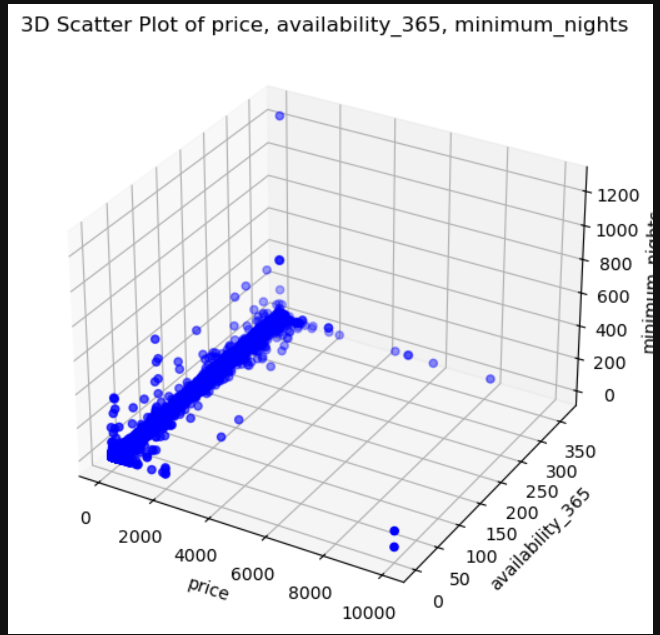
**STRONG RELATIONSHIP**

* Reviews per month & number of reviews correlation~0.99. these two are rarely identical in behaviour-unsurprising, since review per month is likely derived from total review over time

**NEGATIVE RELATIONSHIP**

* Price & reviews per month :correlation~-0.41 higher-priced listing tend to receive fewer monthly reviews , hinting at lower booking frequency or niche appeal
* Host id & availability: correllation~-0.54 hosts with more listing nights have less availability per listing, possible due to calendar management or seasonal operations.

**3D SCATTER PLOT**



**INSIGHTS FOR 3D SCATTER PLOT**

**PRICE DISTRIBUTION**

* Most listing are clustered at lower price(below-2000)

**MINIMUM NIGHTS**

* A few extreme cases have unusually high= minimum nights which might indicate data error or very restrictive hosts

**AVAILABILITY\_365:**

* Many properties seems available for either very few days or almost the entire years.

**RELATIONSHIP BETWEEN VARIABLES**

* No strong linear correlation is clearly visible amoung prices , minimum nights and availability\_365
* Most of the dense cluster lies in the low price, low minimum nights and availability.
* High priced listing do not corresponding to high availability or high minimum nights

**CONCLUSION**

**Key Findings and Insights**

The exploratory data analysis (EDA) revealed several key insights about the dataset, which appears to be focused on Airbnb listings.

**Data Cleaning**

The initial dataset contained missing values, which were successfully handled by dropping the corresponding rows, resulting in a cleaned dataset of 8666 rows.

**Price and Minimum Nights**

While there is no strong linear correlation between price, minimum nights, and availability, the data shows that most dense clusters of listings have low prices and low minimum nights.

**Availability\_365**

Many properties are either available for very few days or for almost the entire year.

High-priced Listings:

High-priced listings do not necessarily correspond to high availability or a high number of minimum nights.

**Data Anomalies**

A few extreme cases with unusually high minimum nights (over 1000) were identified, which may be data errors.

**Recommendations**

Based on these findings, here are some recommendations for further analysis and potential business strategies:

**Investigate Price Anomalies**

The presence of listings with unusually high prices (above $2000) should be further investigated. It would be beneficial to determine if these are data entry errors or if they represent a specific niche of high-end properties.

**Analyze High-availability Properties**

Since many properties are available for almost the entire year, a deeper dive into these listings could be useful. This could involve looking at factors like location, host reputation, and room type to understand what makes them continuously available.

**Segment the Market**

The analysis shows that different market segments behave differently. It is recommended to segment the market based on factors like room\_type, neighbourhood, and price to tailor recommendations and strategies for different groups of hosts and guests.

**Future Work and Next Steps:**

To build on this initial analysis, the following steps are recommended:

**Advanced Data Cleaning:** A more nuanced approach to handling missing values and data anomalies, rather than simply dropping them, could be explored. This might involve using imputation techniques for missing values or consulting with domain experts to understand and handle the extreme cases.

**Geospatial Analysis:** Given that the dataset contains latitude and longitude, future work could include a geospatial analysis to visualize the distribution of listings and prices across different neighborhoods.

**Predictive Modeling:** The cleaned and prepared dataset could be used to build predictive models. For example, a model could be developed to predict the optimal price for a new listing or to forecast booking rates based on various features.

**Time-Series Analysis:** With the last\_review column, a time-series analysis could be conducted to understand the seasonality of bookings and how review frequency changes over time.