Methodology

Tools Used for Data Wrangling: Python - Jupyter notebook

Step 1: Data Understanding

Loaded the data properly and understood the meaning of variables and their importance; How each variable would be useful for this particular analysis; Statistically understanding data and checked the datatypes of each variable

Number of rows :48895

Number of columns: 16

Step 2: Data Wrangling

Datatype correction:

Changed the data type of last_review column from object to date

```
# to view the datatypes
df.dtypes
```

```
id                                 int64
name                              object
host_id                            int64
host_name                         object
neighbourhood_group               object
neighbourhood                     object
latitude                         float64
longitude                        float64
room_type                         object
price                              int64
minimum_nights                     int64
number_of_reviews                  int64
last_review                       object
reviews_per_month                float64
calculated_host_listings_count     int64
availability_365                   int64
dtype: object
```

```
#Converting last_review to date type
df['last_review'] = pd.to_datetime(df['last_review'])
```

Handling Null Values:

```
# To view percentage of null values
df.isnull().mean()*100
```

```
id                                 0.000000
name                               0.032723
host_id                            0.000000
host_name                          0.042949
neighbourhood_group                0.000000
neighbourhood                      0.000000
latitude                           0.000000
longitude                          0.000000
room_type                          0.000000
price                              0.000000
minimum_nights                     0.000000
number_of_reviews                  0.000000
last_review                       20.558339
reviews_per_month                 20.558339
calculated_host_listings_count     0.000000
availability_365                   0.000000
dtype: float64
```

- The **last_review** and **reviews_per_month** columns have about 20 percent missing values
- For the null values in the **reviews_per_month** column, we assume that customers have not given reviews for those listings, indicating that these listings are less preferred by customers. Therefore, we will fill the null values with 0
- For the **last_review** column, we will not impute the null values and leave them as blanks throughout the analysis. We assume that these null values indicate that customers have not given any reviews yet. Since it is a date column, we will not impute it with any values.
- The few null values in the **name** and **host_name** columns suggest that these values are missing by chance, so this information should be collected by the relevant team. For now, we will leave these fields blank

Column Segmentation:

Segmenting fields into categorical , numerical, location and date columns

```
Categorical Variables:
     - room_type
     - neighbourhood_group
     - neighbourhood

Continous Variables(Numerical):
     - Price
     - minimum_nights
     - number_of_reviews
     - reviews_per_month
     - calculated_host_listings_count
     - availability_365
- Continous Variables could be binned in to groups too

Location Varibles:
     - latitude
     - longitude

Time Varibale:
     - last_review
```

Dropping off unwanted fields for analysis :

Id and host_id has been deleted

```
# Dropping few columns which will not be used for analysis
df.drop("id",axis=1,inplace=True)
df.drop("host_id",axis=1,inplace=True)
```

Extracting the useful data:

Created two new columns by extracting year and month from last_review

```
# Extracting month,year from last_review
df['last_reviews_month'] = df['last_review'].dt.month
df['last_reviews_year'] = df['last_review'].dt.year
```

Data misspelling:

Found a misspelling in neighbourhood and corrected it

```
# Replacing misspelt neighbourhood
df["neighbourhood"]=df["neighbourhood"].replace("Bay Terrace, Staten Island","Bay Terrace")
```

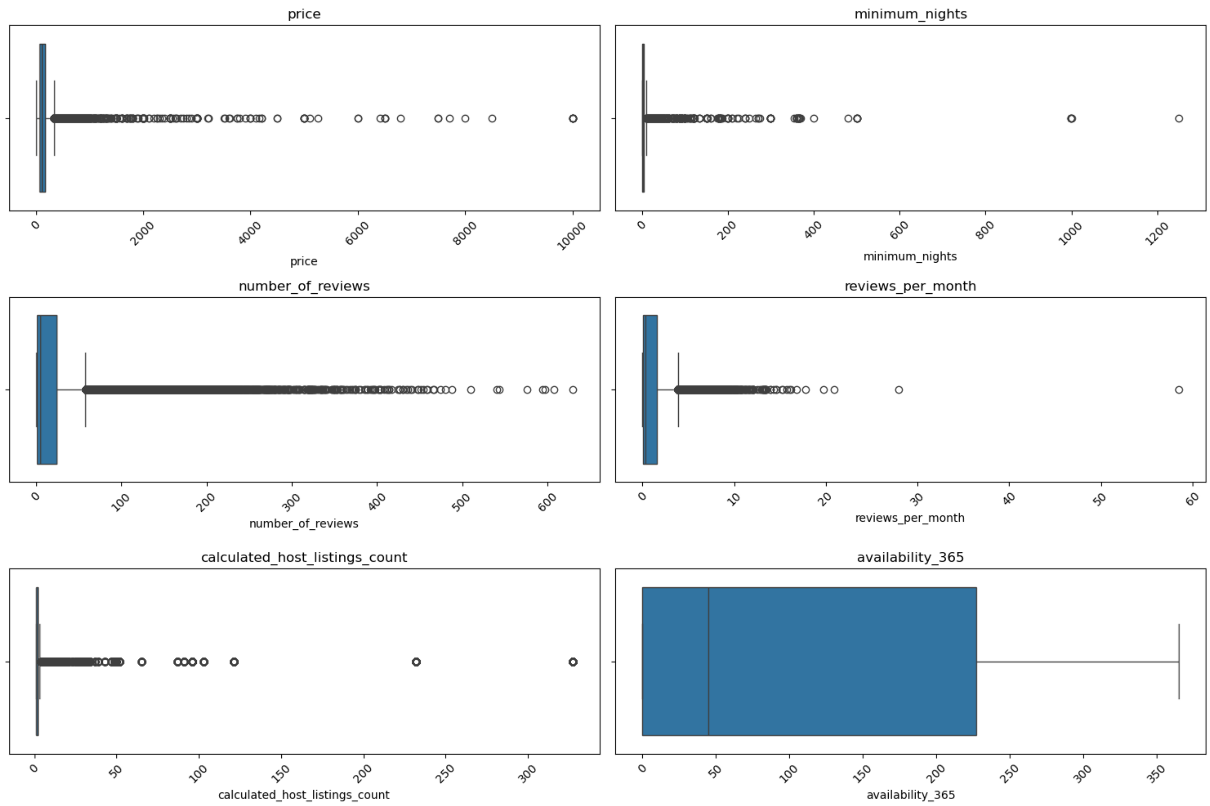
Type_Of_Host:

Created Type_Of_Host as new column based on below logic

```
# Categorizing Host as Individual and Professional based on number of listings they possess
df["Type_Of_Host"]= df['calculated_host_listings_count'].apply(lambda x: 'Individual_host' if x < 2 else 'Professional_Host')
```

Outlier Handling:

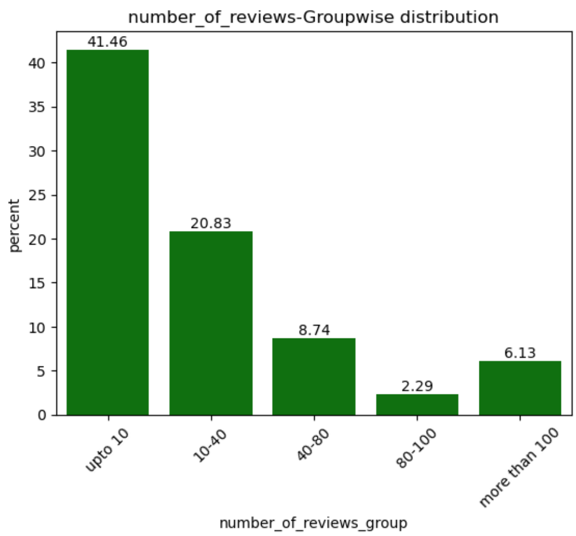
Found outliers in **price, minimum_nights, number_of_reviews, reviews_per_month, calculated_host_listings_count** columns

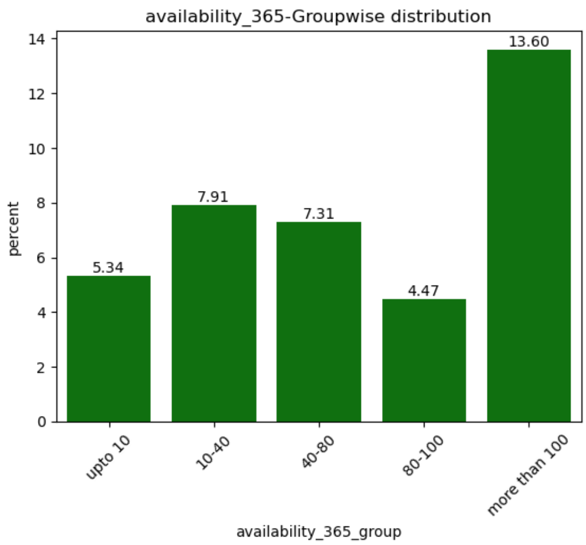


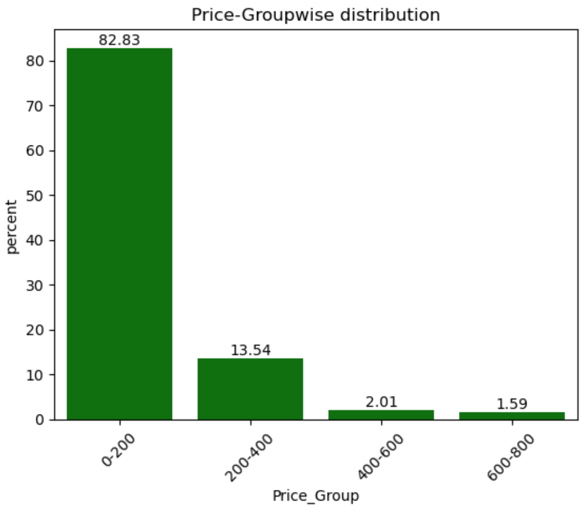
Handled the outliers by capping values above the 99th percentile at the 99th percentile value, as there was a significant difference between the 99th percentile and the maximum values. This method was applied to all columns where outliers were present to ensure consistency and prevent extreme values from skewing the analysis

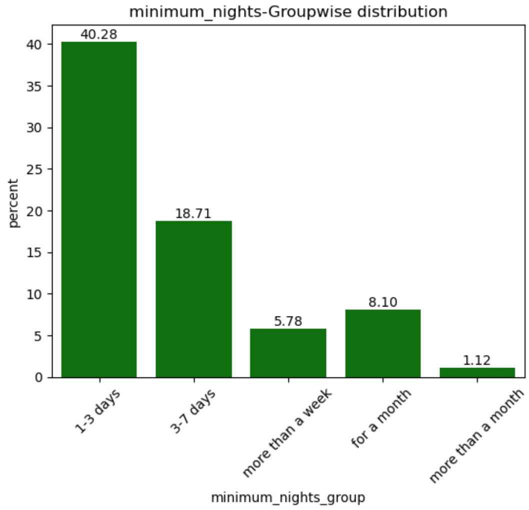
Binning the values in numerical columns for analysis

Grouped the numerical columns for easy visualization



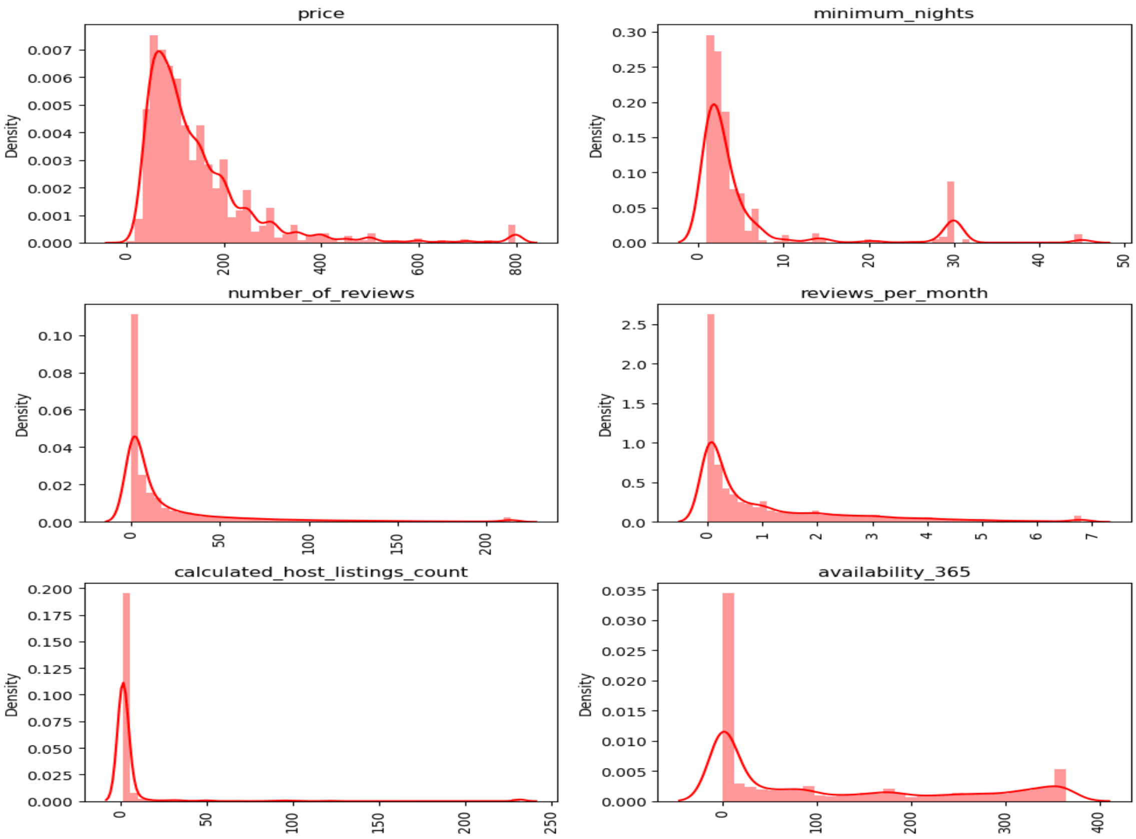




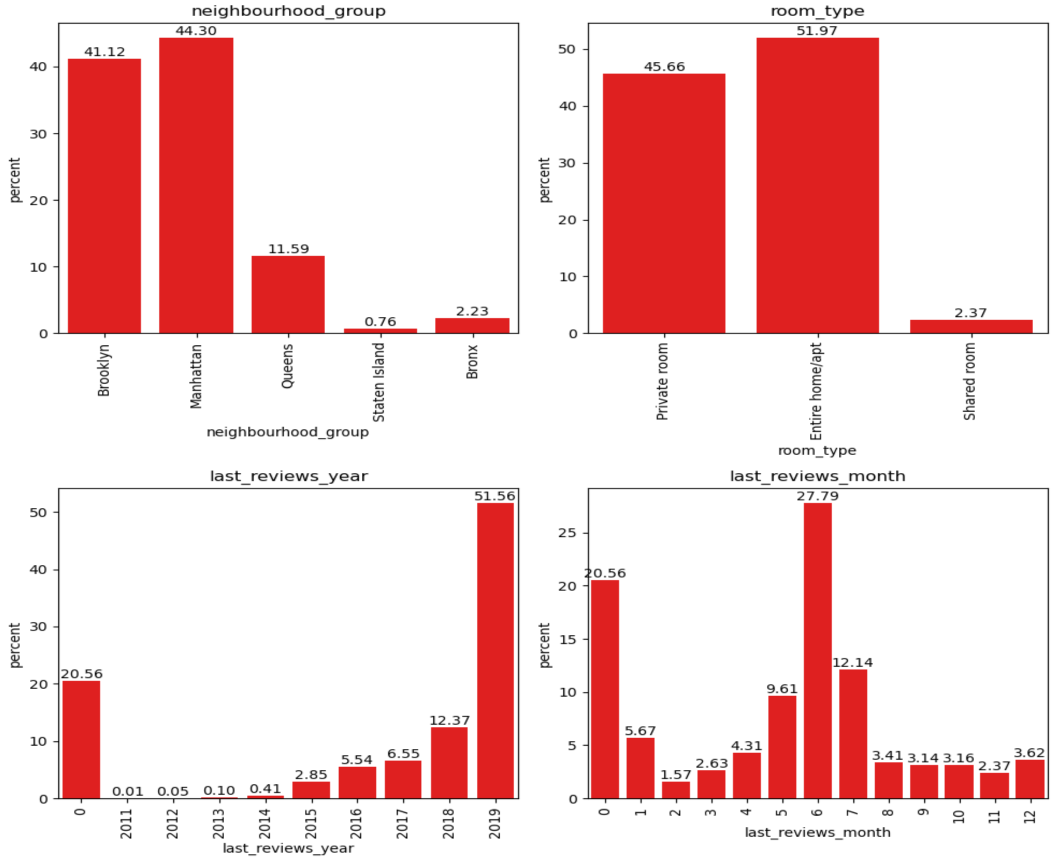


Step 3: Univariate Analysis

Univariate analysis on numerical columns:

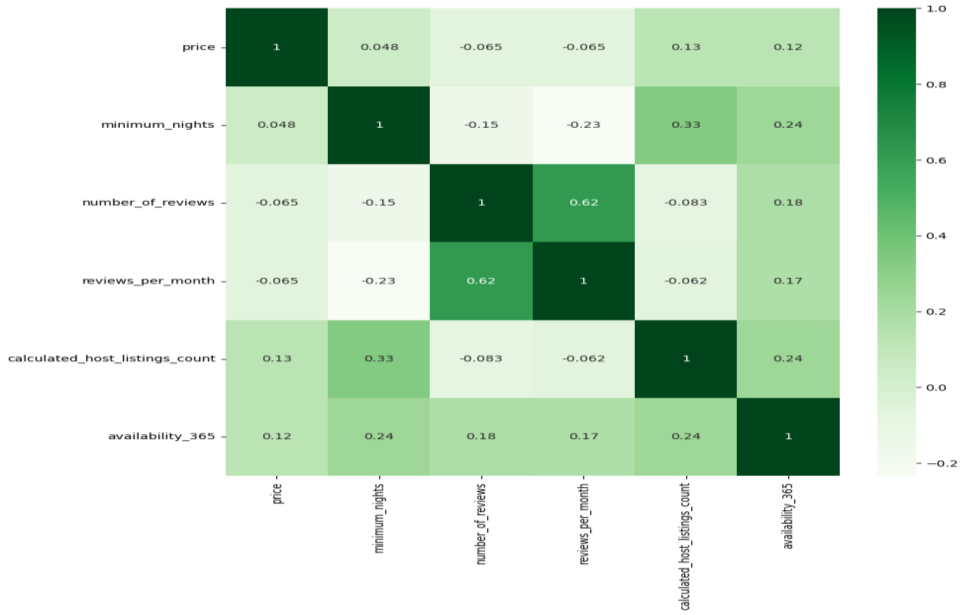


Univariate Analysis on categorical columns:



Step 4: Multivariate Analysis

Multivariate analysis doesn't show any meaningful correlation between variables; Reviews_per_month and number_of_reviews showed a positive correlation but they should be obviously related to each other ; Apart from this no other variables shown significant correlation.



	price	minimum_nights	number_of_reviews	reviews_per_month	calculated_host_listings_count	availability_365
price	1	0.048	-0.065	-0.065	0.13	0.12
minimum_nights	0.048	1	-0.15	-0.23	0.33	0.24
number_of_reviews	-0.065	-0.15	1	0.62	-0.083	0.18
reviews_per_month	-0.065	-0.23	0.62	1	-0.062	0.17
calculated_host_listings_count	0.13	0.33	-0.083	-0.062	1	0.24
availability_365	0.12	0.24	0.18	0.17	0.24	1

After completing the data wrangling and analysis steps, I exported the cleaned and processed data to a new file, which was then used for further visualization and analysis in Tableau

After creating the visualizations in Tableau, I used them to develop a PowerPoint presentation according to the project's needs. The presentation highlighted key insights and findings, incorporating the visualizations to effectively communicate the results. This ensured that the data-driven insights were presented clearly and aligned with the project's objectives