**AIRBNB CASE STUDY: DATA METHODLOGY**

**📄Purpose:** This document outlines the methodology for data analysis on Airbnb in New York City, with the aim of gaining insights into the company's operations, revenue growth trends, and customer behavior.

**👤 Team:** The analysis will be conducted by Sudheer, Jyothi, and Hemant, who will collaborate to ensure comprehensive and accurate results.

**🗓️ Date:** The analysis will take place from 2nd to 9th May 2023.

**A. Methodology Approach**

**1. Research Problem:**

Over the past few months, Airbnb has experienced a significant reduction in revenue as a result of pandemic-related lockdowns. With restrictions now being lifted and people beginning to travel more frequently, Airbnb is seeking to ensure its readiness for this shift.

**2. Business Understanding:**

Airbnb is a San Francisco-based online marketplace that facilitates lodging and tourism activities primarily through homestays and vacation rentals. The platform, accessible via website or app, has two types of customers: hosts and renters. Airbnb earns commissions from both ends and strives to provide the best services at reasonable prices while using advanced technology to make the booking process easy for renters.

**3. Type of Data required to Analyze**

To analyze the decline in revenue for Airbnb during the pandemic, we need to gather the following data:

* Determine whether the hosted sites on the platform offer a satisfactory user experience
* Identify if competitors are capturing market share
* Focus on the boroughs of New York City (the Bronx, Brooklyn, Manhattan, Queens, and Staten Island) for analysis
* Collect data on the location and details of the hosted sites
* Gather information on the hosts, including their profiles and contact details
* Look at the prices of the hosted sites and their reviews received by end consumers

By examining this data, we can identify opportunities for improving the user experience & revenue for Airbnb.

**4. How was the Data acquired? (Assumption)**

* The data used in the analysis was captured from Airbnb's CRM tool for managing customers who host sites on their platform.
* The reviews in the data set are positive as there is no mention of their sentiment, but this assumption needs to be validated by examining the data.

**5. Whom are we presenting?**

* **Data Analysis Managers:** They manage the data analysts for processes, and their technical expertise is basic.
* **Lead Data Analyst:** The lead data analyst oversees the entire team of data and business analysts and is technically sound.
* **Head of Acquisitions and Operations, NYC:** This role is responsible for property and host acquisitions and operations, including negotiating prices and services offered by properties.
* **Head of User Experience, NYC:** This role optimizes the order of property listings in neighborhoods and cities to improve customer preferences and traction for properties on the website and Airbnb app.

**6. Recommendations**

Based on our analysis, we recommend the following:

1. Conduct one-on-one interactions with property owners in Staten Island, Queens, and the Bronx to identify their challenges in being fully functional for a maximum number of days in a year. Allow a booking of more than 10 days of minimum night stays to address these challenges.
2. Facilitate interactions between the top 5 hosts to share their experiences and ideas for better improvement and value generation. This will help the wider community of hosts to improve their offerings and generate more value.
3. Provide discounted commission rates to property owners who keep the minimum night stay booking window for more than 10 days and make their property functional for a maximum number of days in a year. This will incentivize hosts to provide better services and improve their property listings.

**B. Method of Analysis along with code:**

**1. Data Understanding and Preparation**

To prepare for data analysis, we imported relevant Python libraries to assist in the process. The specific libraries used may vary depending on the specific needs of the analysis.

**# Importing Libraries**

import numpy as np

import pandas as pd

import seaborn as sns

import matplotlib.pyplot as plt

%matplotlib inline

plt.style.use('seaborn-dark-palette')

from scipy import stats

import datetime as dt

import plotly

import plotly.express as px

**# Ignore warnings**

import warnings

warnings.filterwarnings("ignore")

pd.set\_option('display.max\_columns', None)

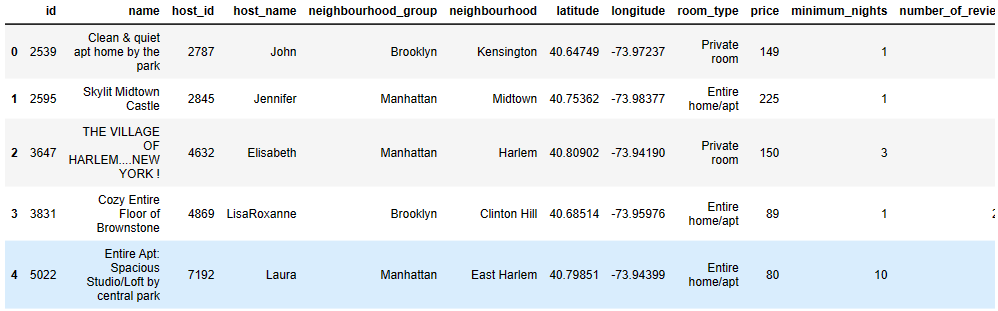
pd.set\_option('display.max\_rows', None)

To gain a better understanding of the data, we executed basic functions to load and interpret variables, data types, dimensions, and size of the dataframe in Python. The code used for this step may vary depending on the specific data and analysis needs.

**# load the dataset**

airbnb = pd.read\_csv("AB\_NYC\_2019.csv")

airbnb.head()



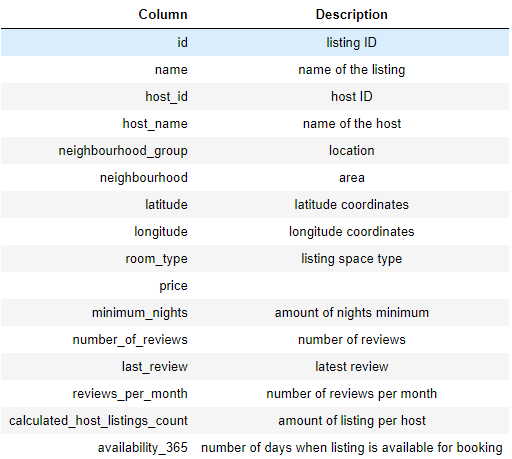
**# Dimensions**

Airbnb.shape

**# Data-Types**

airbnb.info()

**2. Variables in the dataframe:**



The above observations helped us to perform a more detailed analysis of both numerical and categorical variables by utilizing appropriate functions and techniques. We also conducted some basic data cleaning steps, such as handling missing values and outliers, before proceeding with the analysis.

**# Numeric Analysis**

airbnb.describe()

**# Analyzing categorical values**

airbnb.select\_dtypes(include=['object']).describe()

**3. Handling Missing values and outliers:**

* Two columns (last\_review and reviews\_per\_month) had a missing value rate of 20.56%, while two other columns (host\_name and name of the place) had minimal missing values.
* The missing values in the last\_review and reviews\_per\_month columns were not missing at random and indicated that the corresponding places might not be preferred by future customers, leading to a negative impact on business.
* Some host and place names were missing by chance, and we left those rows blank.
* We imputed the missing values in the reviews\_per\_month column with 0.
* We used the code provided below and the missingno library to identify missing values.

Below is the code we used to identify missing values. We also imported missingno library to do the same.

**# Checking missing values columns**

import missingno as msno

msno.bar(airbnb)

**# Checking missing values percentages**

def null\_values(airbnb):

return round((airbnb.isnull().sum()/len(airbnb)\*100).sort\_values(ascending = False),2)

null\_values(airbnb)

After addressing the missing values, we proceeded to handle the spread of outliers in the data frame. We used the code below to identify the spread of the outliers.

**# Extracting Numeric columns:**

int\_cols = airbnb.select\_dtypes(include=['int64', 'float64']).columns

**# Tagging them:**

list(enumerate(int\_cols))

**# Plotting the spread of outliers:**

plt.figure(figsize=([20,22]))

for n,col in enumerate(int\_cols):

plt.subplot(5,2,n+1)

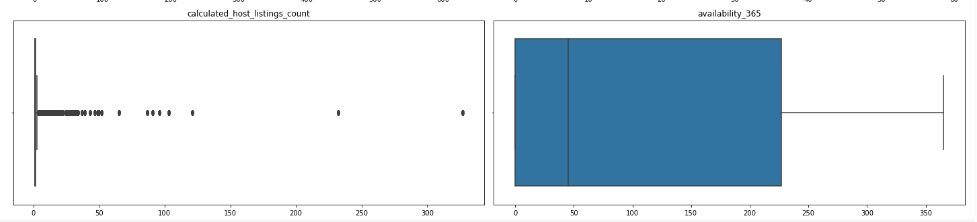
sns.boxplot(airbnb[col], orient = "h")

plt.xlabel("")

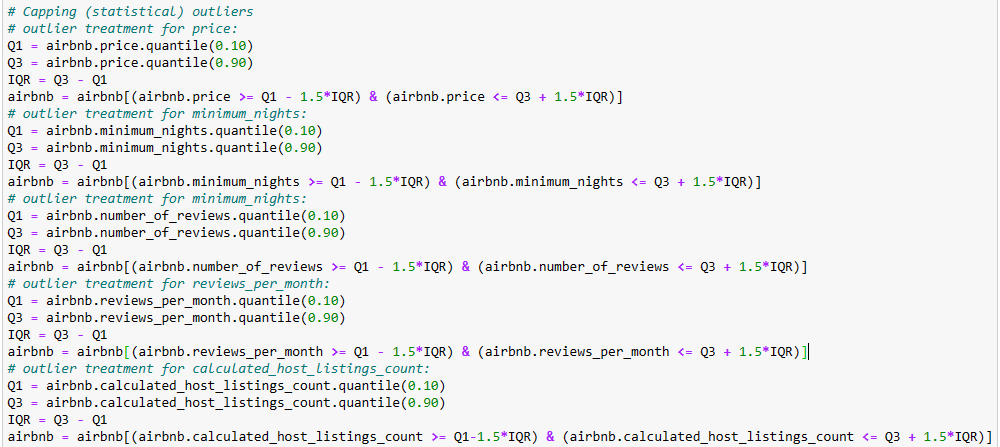
plt.ylabel("")

plt.title(col)

plt.tight\_layout()

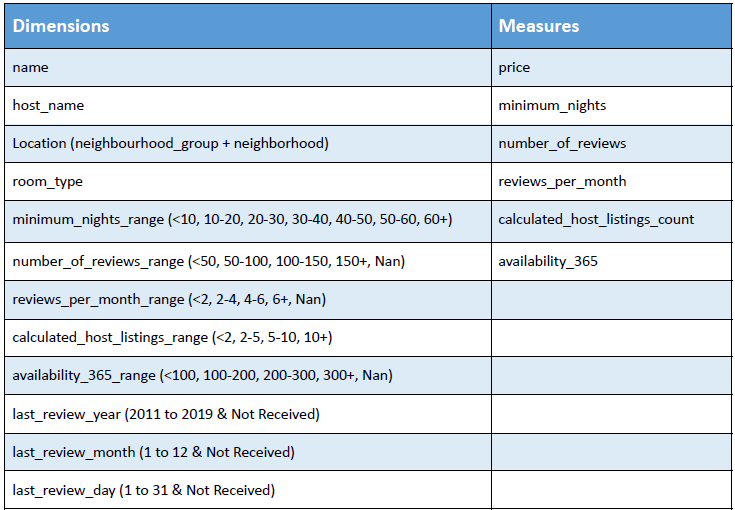


We treated the outliers in the dataframe by capping them at 10%. The code for capping the outliers is provided below.



**4. Feature Selection / Engineering**

We transformed some numeric features into categorical variables through binning during our Data Preprocessing, but still kept the original numerical features for analysis. The resulting table displays all the engineered variables for further analysis.



**5. Analyzing Data:**

* For data analysis, we conducted both Univariate and Bi-Multivariate Analysis.
* In the Univariate Analysis, we used Seaborn library's Distribution plot for Numeric columns and Countplot for Categorical columns.
* We plotted the count of Neighborhood Groups using the Countplot.
* For Numeric columns, we also created a pairplot using Seaborn library in Bi-Multivariate Analysis.
* All the plots were created in Python and the corresponding codes were written for each plot.

**6. Matrix used for Analysis**

We developed a 2x2 Matrix to guide our graph creation process while analyzing the data. The Matrix incorporated values for various combinations of Dimensions and Measures, including:

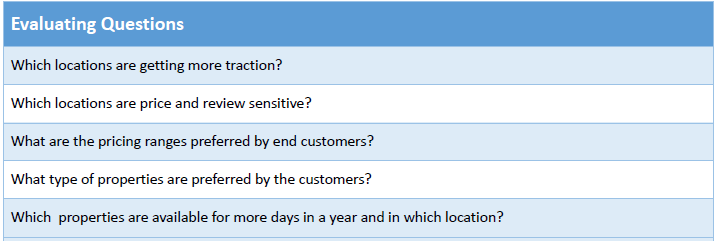
* Categorical & Numerical
* Categorical & Categorical
* Numerical & Numerical
* Numerical & Categorical

This Matrix served as a roadmap to identify which dimensions and measures were used to obtain insights from the data. The Matrix is displayed below.



**7. Evaluation of Methods**

We evaluated the Matrix at every step by asking relevant questions to extract insights from the data that were useful to our target audience. Below are some examples of the questions we curated for creating graphs.



**C. Findings & Insights**

**1. Basic Data Interpretation:**

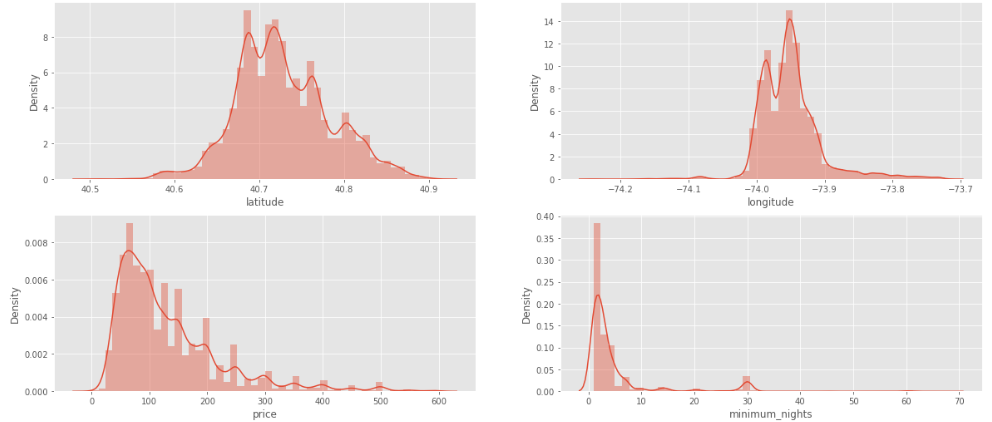
Here's a brief summary of the basic data interpretation:

* The dataframe has 16 columns and 48,895 rows.
* The dataframe contains 3 floats, 7 integers, and 6 objects data types.
* There are many columns with missing values in the dataframe.
* Further investigation is required to determine the reason behind the missing values and feature engineering may be needed.

**2. Numeric and Categorical Analysis:**

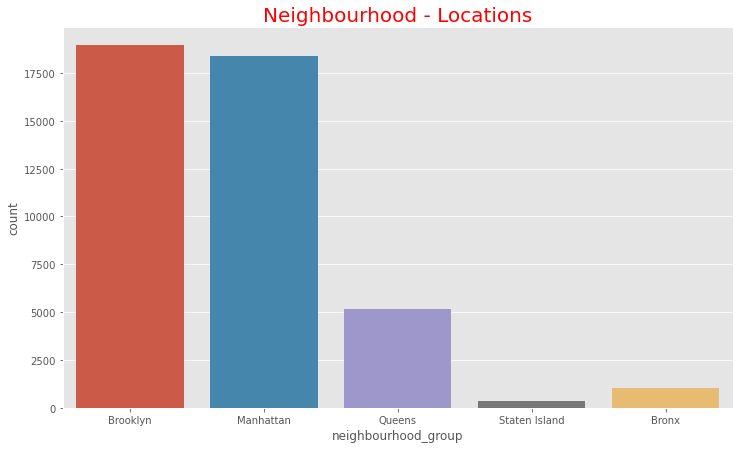
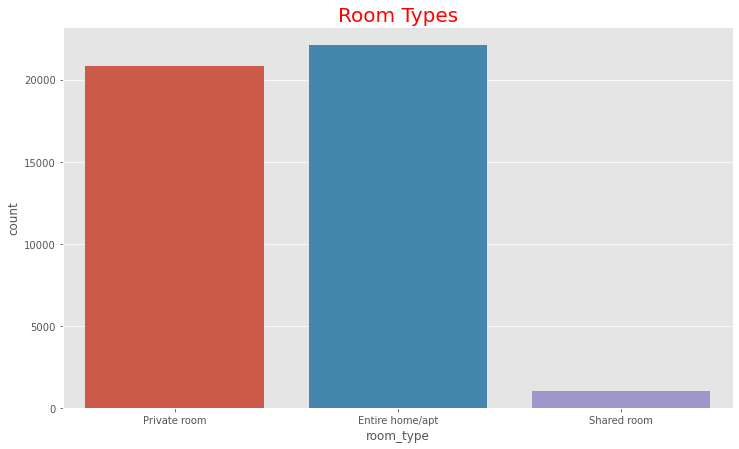
* The data set contains 16 columns and 48,895 rows, with 3 floats, 7 integers, and 6 objects.
* Numeric and categorical analysis reveal that the data has 0 dollar prices, significant variance in certain columns, and that Manhattan is the most welcoming neighborhood while the Hillside Hotel is the most frequently hosted place by Michael.
* Further investigation and feature engineering are needed to address missing values and explore the reasons for the 0 dollar prices and variance in certain columns.

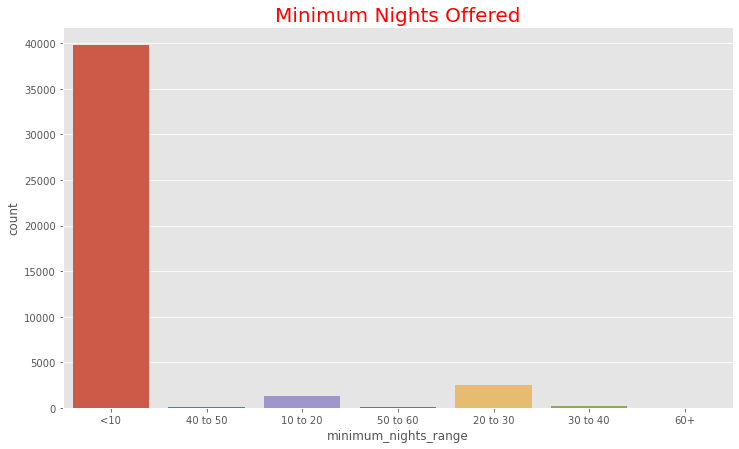
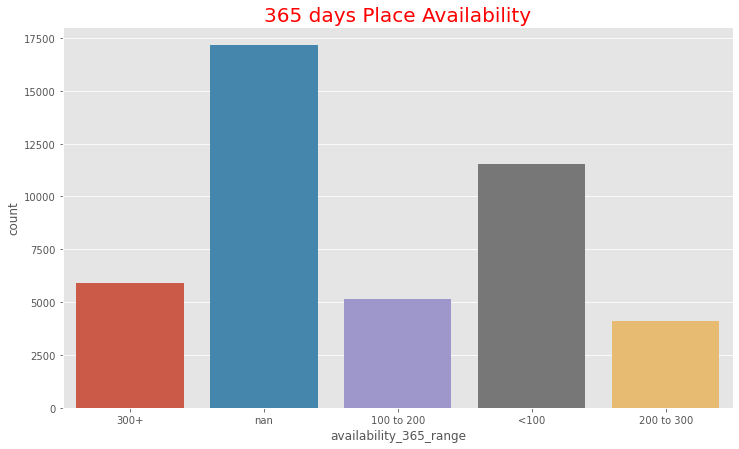
**3. Numerical Univariate Analysis:**



**Insights:** The majority of the listed sites on Airbnb are priced between 80 and 160 dollars per day, but there are still some that cost over 500 dollars per day.

**4. Categorical Univariate Analysis:**

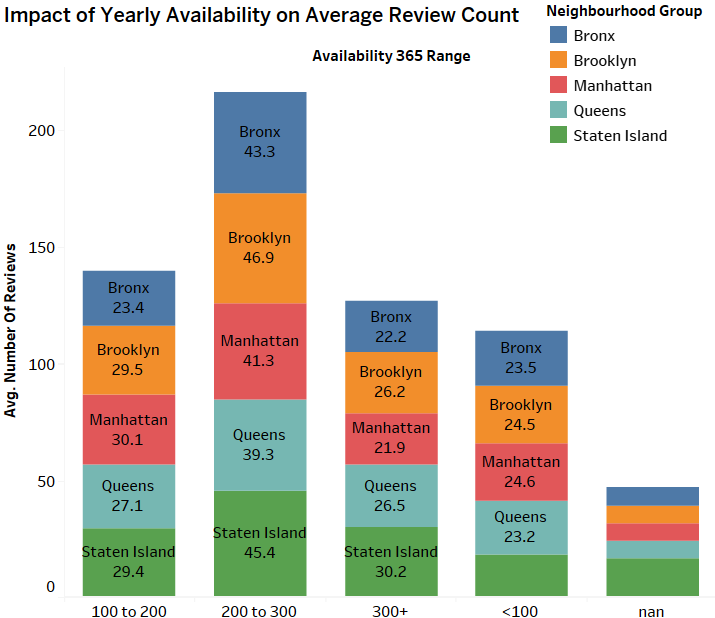
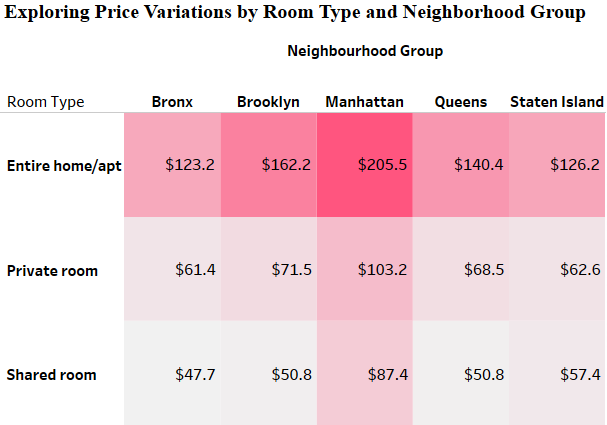
 

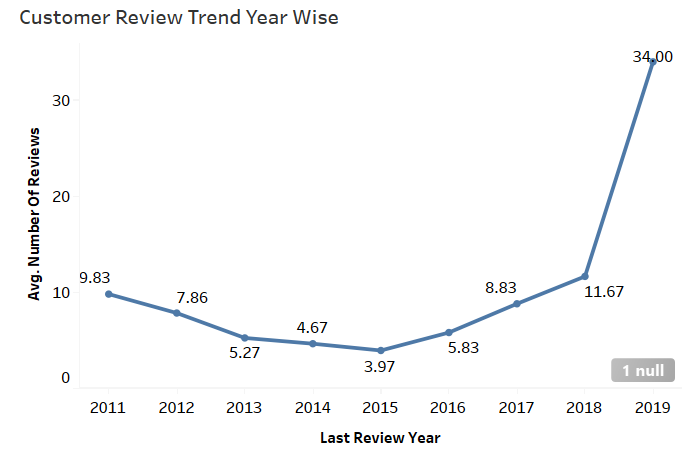
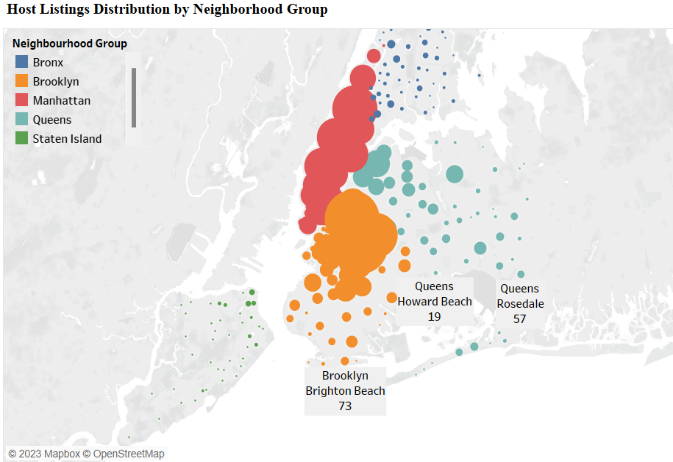
 

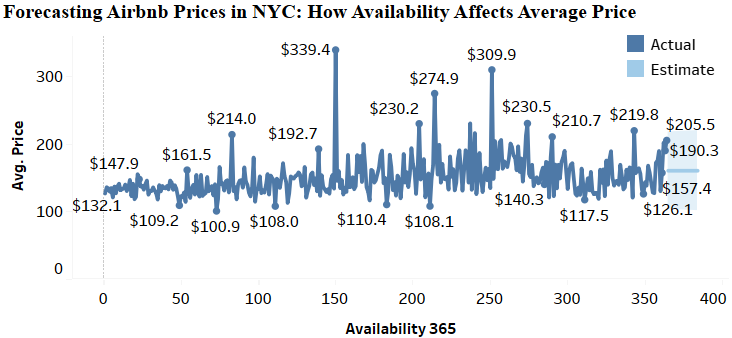
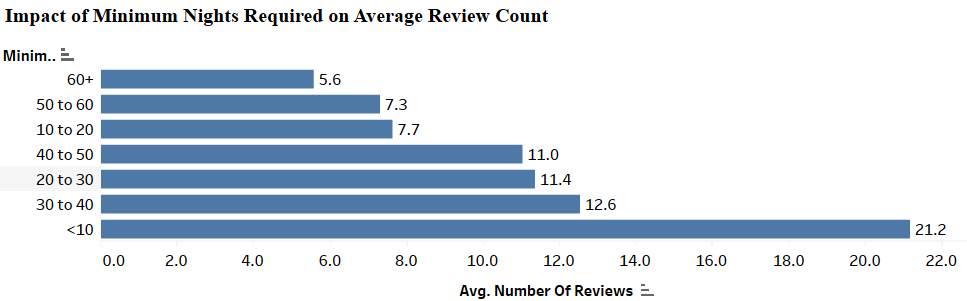
**Insights:**

* Last reviews were not provided most of the time on a daily basis, with most reviews being provided on the 6th and 7th day of the month.
* June receives the most last reviews, followed by May.
* Majority of the sites have less than 100 days availability, and most of them have provided 0 days availability.
* Majority of the hosts have less than 2 sites hosted by them on the platform.
* Majority of the sites have received less than 2 reviews per month & less than 50 reviews till date.
* Majority of the sites provide less than 10 nights stay at a time, and most of them are either Private rooms or Entire apartments in Brooklyn and Manhattan.

**5.Multivariate Analysis:**

**Insights:**

* There is no correlation between the numerical variables.
* Manhattan has the highest-priced properties on average, while other neighborhoods offer medium-priced properties.
* High-priced entire home/apt properties are available for fewer days than private or shared rooms.
* Most hosts have less than 2 properties listed, but some in Manhattan have over 10.
* Brooklyn receives the most reviews for properties open over 200 days, while some Staten Island sites receive low reviews due to being closed all year.
* Most customers prefer to pay $81-160 per night, based on their reviews.
* Michael, David, Alex, John, and Daniel are the top hosts with high-priced listings and many reviews.
* Silver Lake, Richmondtown, Eltingville, Huguenot, and Manhattan Beach are the top low-priced locations with high reviews, while Neposit, NoHo, Tribeca, Willowbrook, and Flatiron District have low reviews despite high prices.
* The top 6 properties with the highest reviews are in Brooklyn.
* Entire home/apt properties are preferred, followed by private and shared rooms due to their longer minimum stay window.
* The top 5 properties with the highest prices are located in Manhattan and Brooklyn.
* Manhattan and Brooklyn offer the most extended minimum night stays.
* The number of reviews has increased significantly since 2015-2016, with most reviews occurring in May-July or at the start of the year.
* Brooklyn and Manhattan have the most properties open for over 300 days, while properties open less than 50 days are in Queens or Staten Island.
* The most critical factors that influence customers to provide reviews are the property's minimum or maximum stay window and how many days it's open in a year.

**D.Toolse Used**

* Python is utilized for data understanding, pre-processing, and general univariate and multivariate analysis.
* Tableau used for in-depth bi-multivariate analysis.
* The combination of these tools allows for comprehensive exploration and understanding of the data.

Thank You !