Capstone Project
Airbnb Bookings Analysis - EDA

Team
Kaimur Data stimulators

Members
Nethinti Ramakrishna (Team Leader)
Sahil Sukhdeve

Let’s Explore and Analyse Airbnb NYC 2019

Introduction

Problem Statement

Key objectives we are focusing on

Data Preparation & Data Wrangling

Exploratory Data analysis

Key findings

Tableau & Excel Dashboarding

Challenges faced

Conclusion

Introduction

What is Airbnb?

Airbnb, Inc. is an American company that operates an online marketplace for lodging, primarily homestays for vacation rentals, and tourism activities. Based in San Francisco, California, the platform is accessible via website and mobile app. Airbnb does not own any of the listed properties; instead, it profits by receiving commission from each booking. The company was founded in 2008 by Brian Chesky, Nathan Blecharczyk and Joe Gebbia. Airbnb is a shortened version of its original name, AirBedandBreakfast.com.

Problem Statement

- Data analysis on millions of listings provided through Airbnb is a crucial factor for the company. These millions of listings generate a lot of data - data that can be analyzed and used for security, business decisions, understanding of customers' and providers' (hosts) behavior and performance on the platform, guiding marketing initiatives, implementation of innovative additional services and much more.

- This dataset has around 49,000 observations in it with 16 columns and it is a mix between categorical and numeric values.

Key objectives

- ? What can we learn about different hosts and areas?
- ? What can we learn from predictions? (ex: locations, prices, reviews, etc)
- ? Which hosts are the busiest and why?
- ? Is there any noticeable difference of traffic among different areas and what could be the reason for it?
- ? Finding price difference at different localities and different room types.
- ? Finding average prices.
- ? Understanding customer behaviour.

Data Preparation & Data Wrangling

- **Dropping unnecessary data:**
- **as "last_review" and "reviews_per_month" have more than 10,000 null values, it affects the outcomes of Data analysis; So, we are removing these columns and also as we are not doing any analysis specifically on latitude and longitude, we're also removing these variables as well**

- **Verifying Data quality:**
- **We have gone through whole data and checked null values and reviewed any missing data or wrong data. And prepared whole data ready for exploratory data analysis**

- **Basic data exploration:**
- **Using describe() and info () and size functions of pandas. Gone through a basic exploration of data before entering into EDA.**

Data Description

- **id**: unique reference number for each different hotel.
- **name**: name of different hotels of various neighborhood groups.
- **host_id**: unique reference id of each individual host.
- **host_name**: name of host hosting different hotels.
- **neighbourhood_group:** aggregate group of neighborhood cities of some particular regions.
- **neighbourhood:** cities present in NYC.
- **latitude:** latitude is a geographic coordinate that specifies the north–south position of a point on the Earth's surface. Latitude is an angle which ranges from 0° at the Equator to 90° at the poles.

Data Description continued...

- **longitude**: Longitude is a geographic coordinate that specifies the east–west position of a point on the Earth's surface, or the surface of a celestial body.
- **room_type**: Different room types available for booking, which contains Private room, Entire home/apt, Shared room.
- **price:** price per each night stay of different room types at various hotels.
- **minimum_nights**: minimum nights booked in particular hotel.
- **number_of_reviews**: count of reviews got for each hotel.
- **last_review:** date of last review got by a customer to a particular hotel.
- **reviews_per_month:** count of reviews getting per month of a particular hotel.
- **calculated_host_listings_count**: It represents total number of listings made by a specific host. In some cases, the properties are same but some of the other features differ like(room_type).

availability_365: number of available days for booking in a year

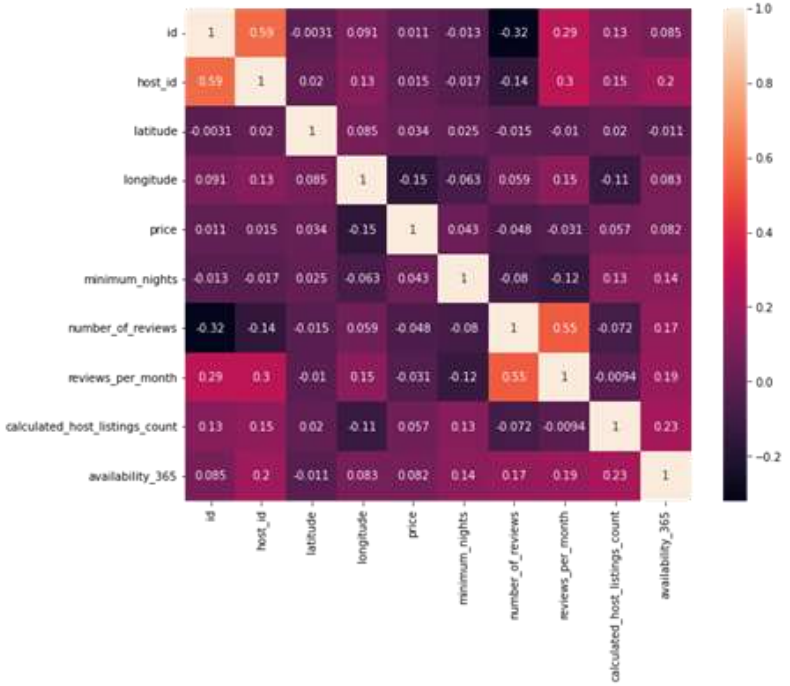
Exploratory Data Analysis

- **Tools used for EDA:**
- **Programming Language: Python**
- **Libraries: Pandas, Matplotlib, Seaborn**
- **Tableau Desktop**
- **Microsoft Excel**

Correlation between data

Drawn a correlation between all the numerical variables in the dataset by Plotting a heatmap using seaborn.

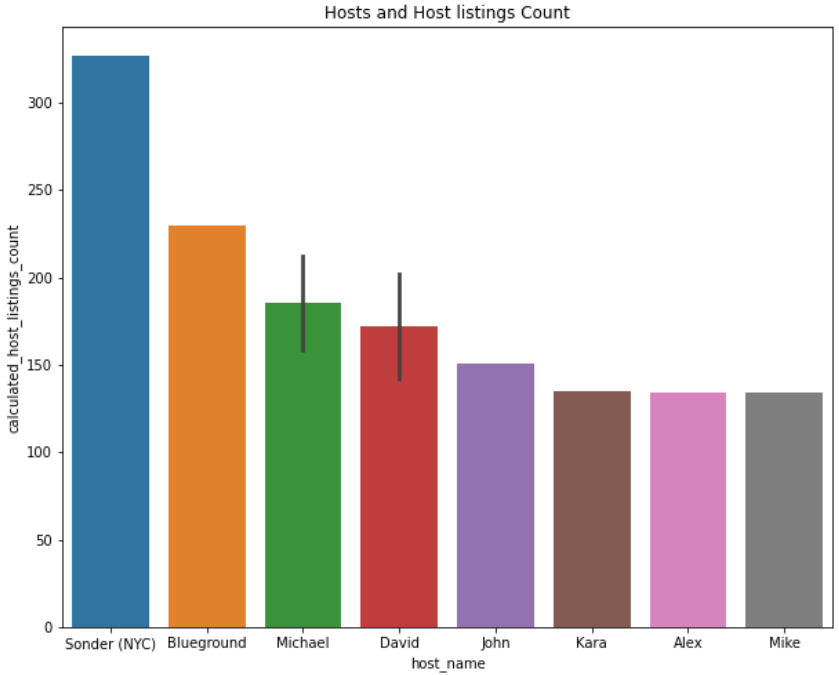
```
# Correlation in the data
corr = airbnb.corr()
plt.figure(figsize=(10,8))
sns.heatmap(corr, annot=True)
```



1) Top Hosts and their listings count:

For analyzing this we used group by function and took 'host_name', 'neighbourhood_group', and "calculated_host_listings_count" and calculated the top 10 hosts.

	host_name	neighbourhood_group	calculated_host_listings_count
13217	Sonder (NYC)	Manhattan	327
1834	Blueground	Manhattan	230
9742	Michael	Manhattan	212
3250	David	Manhattan	202
9741	Michael	Brooklyn	159
6808	John	Manhattan	151
3249	David	Brooklyn	142
7275	Kara	Manhattan	135
432	Alex	Manhattan	134
9856	Mike	Manhattan	134



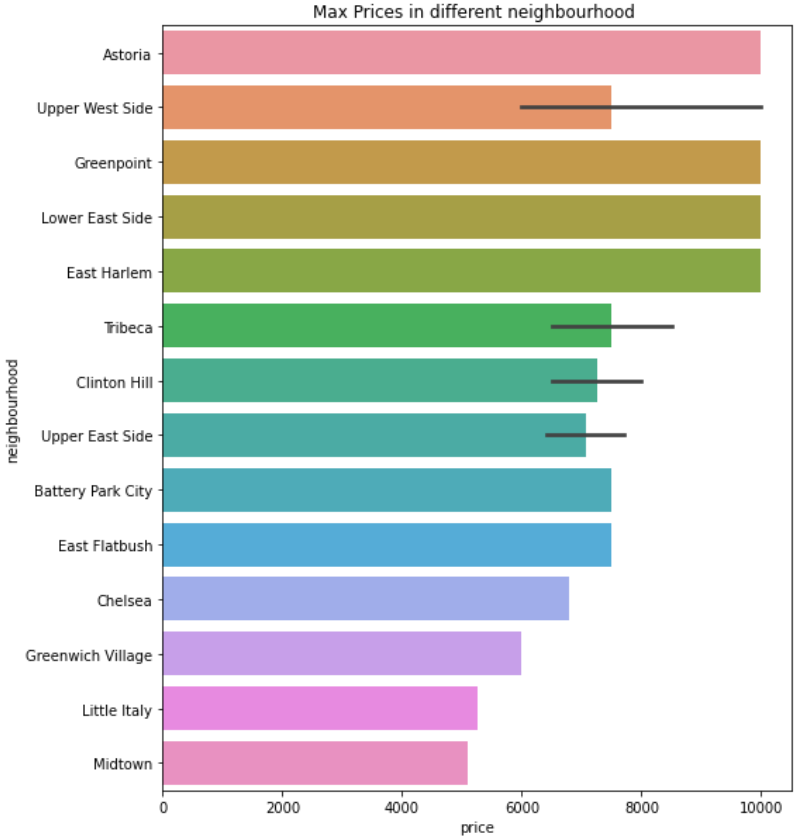
Key findings:

- 8 out of top 10 hosts are from the ''Manhattan'' neighbourhood group.
- 2 out of top 10 hosts are from the ''Brooklyn'' neighbourhood group.
- [Sonder (NYC), Blue ground, Michael, David, John, Kara, Alex, Mike] are the top hosts of Manhattan neighbourhood group.
- [Michael, David] are the top hosts in Brooklyn neighbourhood group.
- Manhattan neighbourhood group hosts are out-performing in listings.

2) Max prices in different neighborhood:

For this question, we approached with, 'name', 'neighbourhood_group', 'neighbourhood', 'price','minimum_nights', 'number_of_reviews'

	name	neighbourhood_group	neighbourhood	price	minimum_nights	number_of_reviews
20222	Furnished room in Astoria apartment	Queens	Astoria	10000	100	1
1112	1-BR Lincoln Center	Manhattan	Upper West Side	10000	30	1
27228	Luxury 1 bedroom apt. -stunning Manhattan views	Brooklyn	Greenpoint	10000	5	1
36159	Quiet, Clean, Lit @ LES & Chinatown	Manhattan	Lower East Side	9999	99	1
41053	Spanish Harlem Apt	Manhattan	East Harlem	9999	5	1
2225	2br - The Heart of NYC: Manhattans Lower East ...	Manhattan	Lower East Side	9999	30	1
7092	Beautiful/Spacious 1 bed luxury flat-TriBeCa/Soho	Manhattan	Tribeca	8500	30	1
19735	Film Location	Brooklyn	Clinton Hill	8000	1	1
18366	East 72nd Townhouse by (Hidden by Airbnb)	Manhattan	Upper East Side	7703	1	1
2749	70' Luxury MotorYacht on the Hudson	Manhattan	Battery Park City	7500	1	1
20464	Gem of east Flatbush	Brooklyn	East Flatbush	7500	1	1
2451	3000 sq ft daylight photo studio	Manhattan	Chelsea	6800	1	1
4205	Apartment New York \nHell's Kitchens	Manhattan	Upper West Side	6500	30	1
37639	SUPER BOWL Brooklyn Duplex Apt!!	Brooklyn	Clinton Hill	6500	1	1
27697	Luxury TriBeCa Apartment at an amazing price	Manhattan	Tribeca	6500	180	1
32010	Park Avenue Mansion by (Hidden by Airbnb)	Manhattan	Upper East Side	6419	1	1



Key findings:

- **The highest price is 10,000 $ and can be seen in Astoria, Upper west side, lower east side, Greenpoint and East Harlem.**

- **Most minimum nights spent in top 20 price list are,**

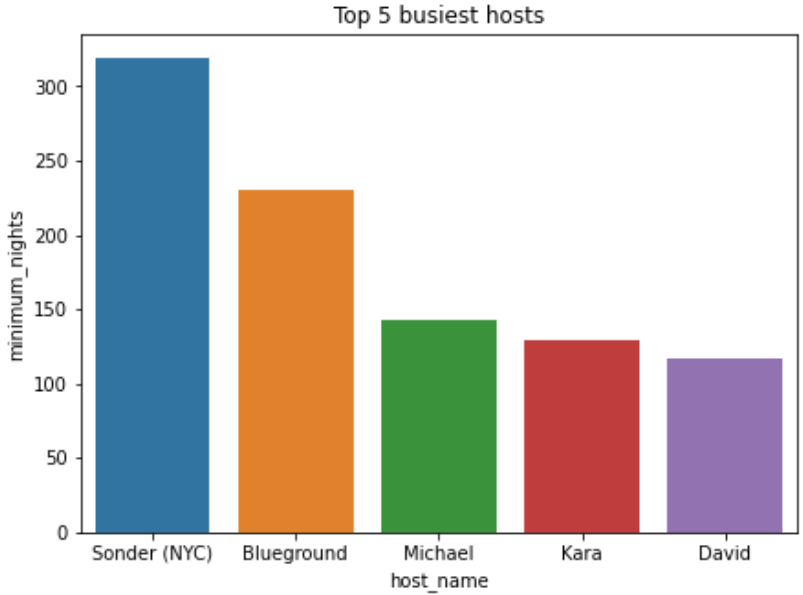
  - **Luxury TriBeCa Apartment at an amazing price - 180 nights**
  - **Furnished room in Astoria apartment - 100 nights**
  - **Quiet, Clean, Lit @ LES & Chinatown - 99 nights**

- **Highest priced rooms (i.e., 10,000$) are all present in Manhattan, Brooklyn and Queens neighbourhood group.**

3) Finding Busiest Host

Next, we were interested in finding the busiest hosts by considering “minimum nights” bookings in their hotels. We took, ‘host_name','neighbourhood_group','room_type', "minimum_nights" and got the following result.

	host_name	neighbourhood_group	room_type	minimum_nights
16549	Sonder (NYC)	Manhattan	Entire home/apt	319
2295	Blueground	Manhattan	Entire home/apt	230
12299	Michael	Manhattan	Entire home/apt	143
9190	Kara	Manhattan	Entire home/apt	129
4128	David	Manhattan	Entire home/apt	117



Key findings:

- 5 out of top 5 are all from " Manhattan" neighbourhood group.

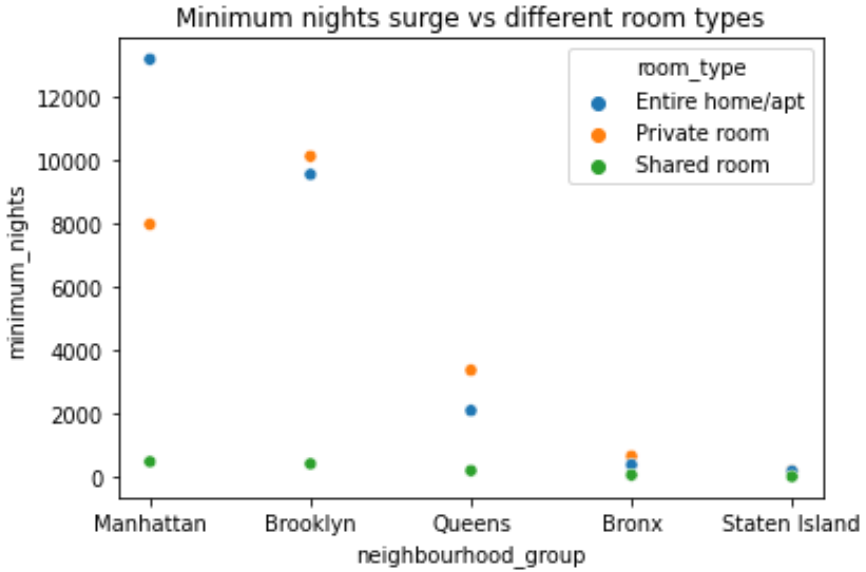
- Sonder (NYC), Blue ground, Michael, Kara, David are the top 5 most busiest hosts.

4) Traffic among different room types and different neighborhood

Our main motive through this step was to find traffic among different types of rooms at different neighborhood groups.

For this we took, 'neighbourhood_group', 'room_type', 'minimum_nights' to analyze this question

	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
1	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
13	Staten Island	Private room	188
12	Staten Island	Entire home/apt	176
2	Bronx	Shared room	60
14	Staten Island	Shared room	9



Key findings:

- In Manhattan, people are preferring "Entire Home/apt".
- But, in Brooklyn, Queens and Bronx people are preferring private rooms.
- In Staten Island, people are having equal preference over all three types of rooms.

Tableau Dashboard - 1

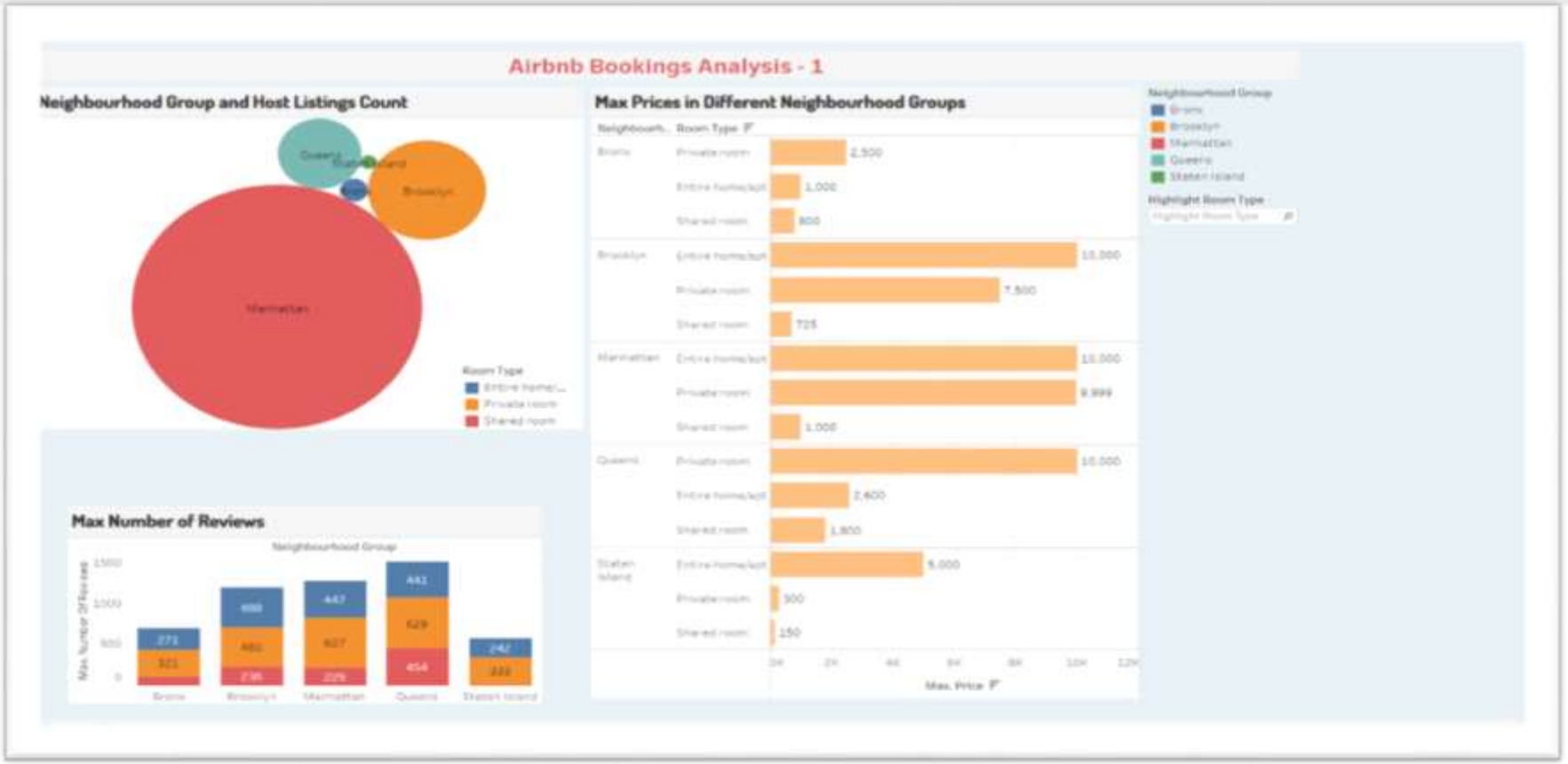
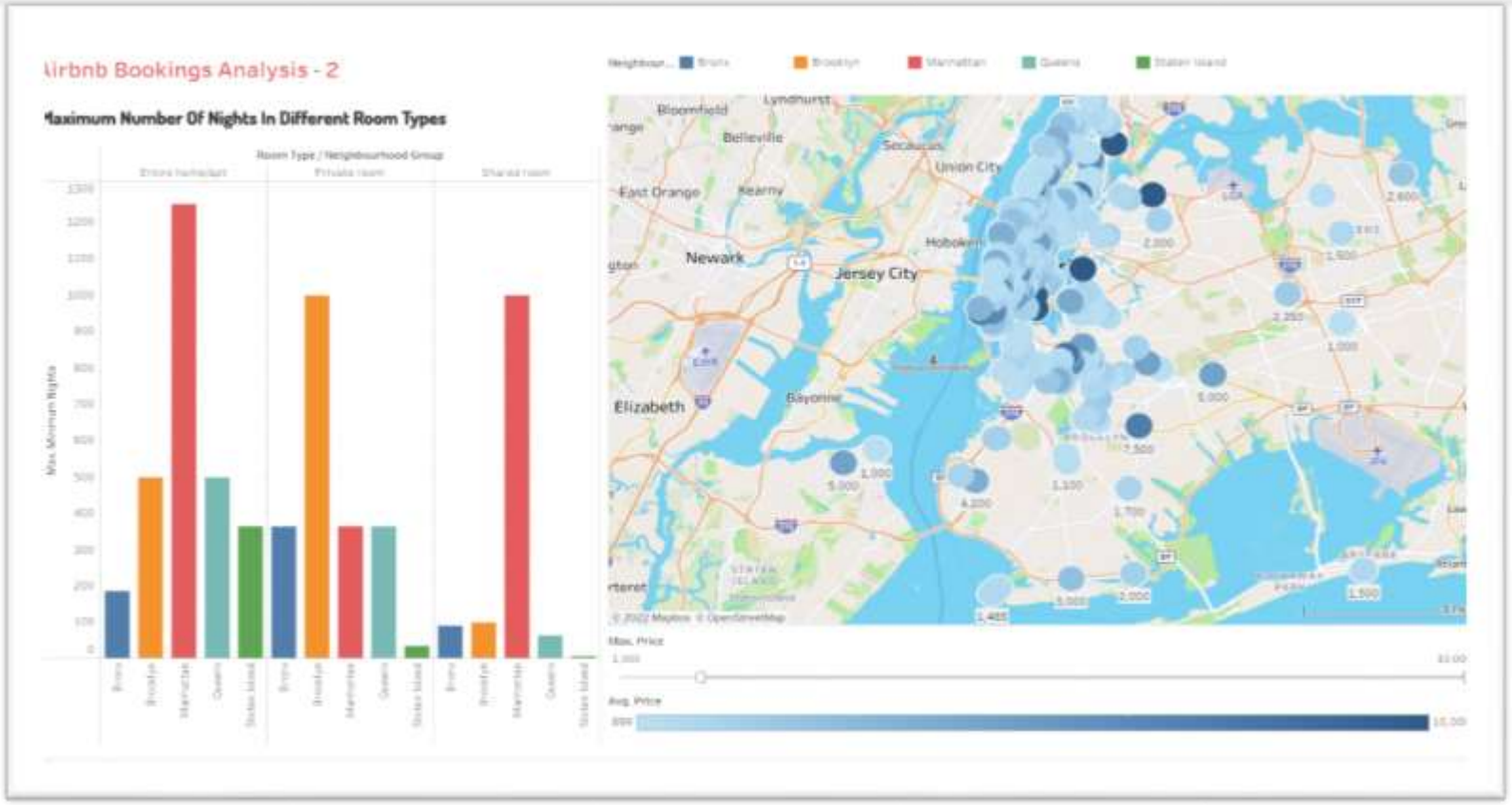
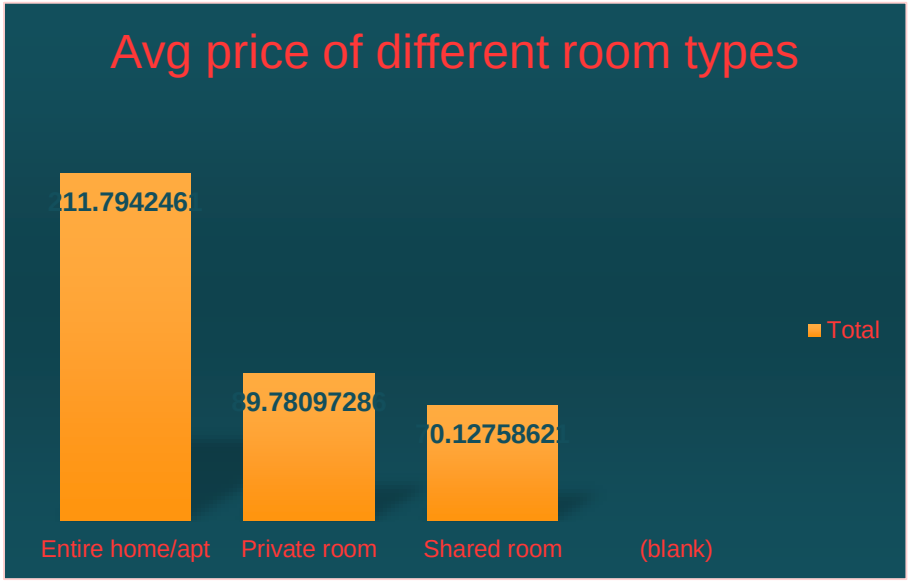


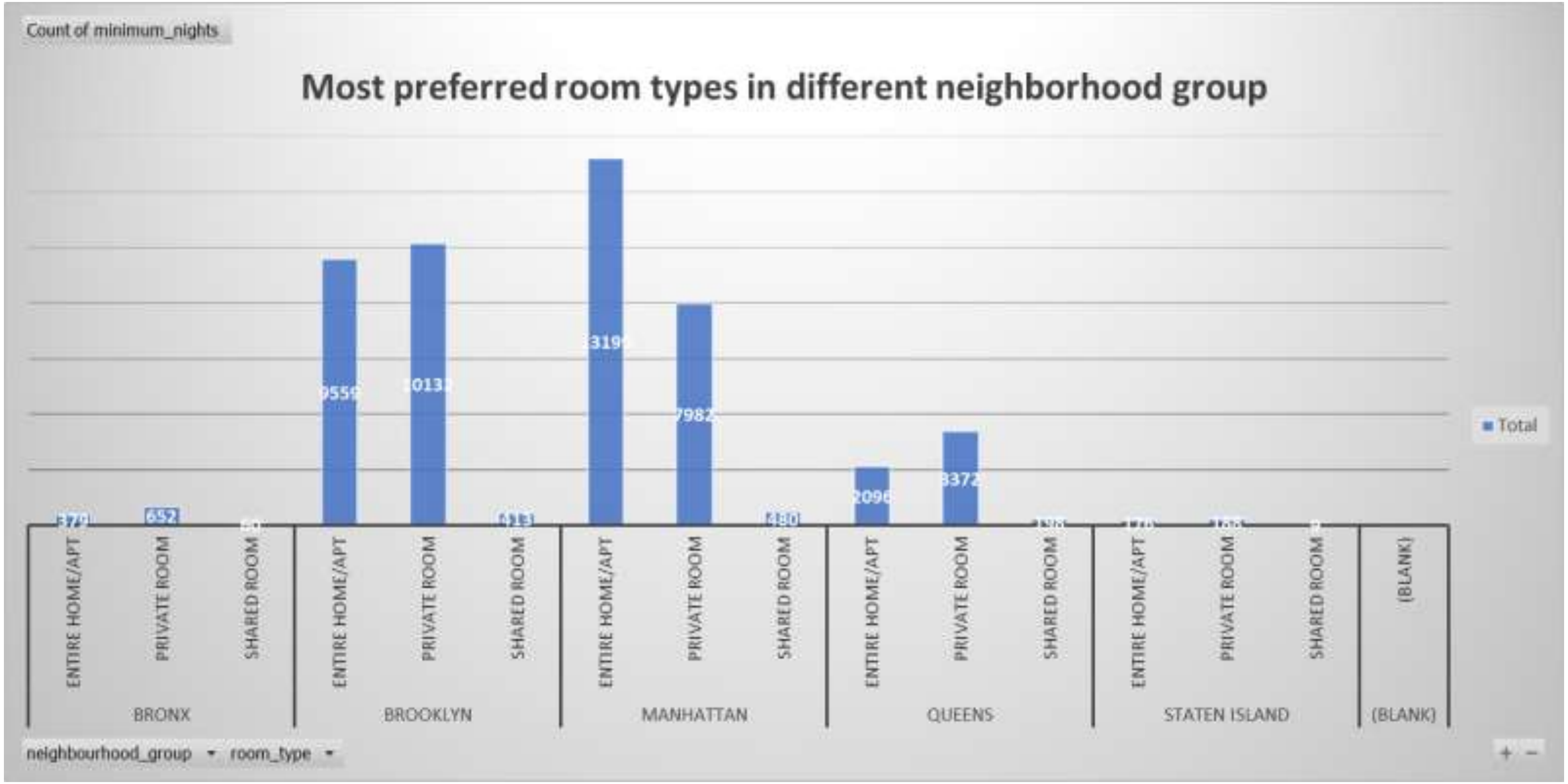
Tableau Dashboard - 2



Microsoft Excel Pivot Charts



Continued..


Count of minimum_nights
Most preferred room types in different neighborhood group
379
652
9559
10132
413
13199
7982
480
2096
3372
Total
ENTIRE HOME/APT
PRIVATE ROOM
SHARED ROOM
ENTIRE HOME/APT
PRIVATE ROOM
SHARED ROOM
ENTIRE HOME/APT
PRIVATE ROOM
SHARED ROOM
ENTIRE HOME/APT
PRIVATE ROOM
SHARED ROOM
ENTIRE HOME/APT
PRIVATE ROOM
SHARED ROOM
(BLANK)
BRONX
BROOKLYN
MANHATTAN
QUEENS
STATEN ISLAND
(BLANK)
neighbourhood_group
room_type

Challenges faced

- Verifying quality of such huge data and looking for error values.
- Dropping down irrelevant data and making the whole data getting ready for full pledged data analysis.
- Understanding and visualizing complex numerical data, and communicating business solutions.
- Analysing and solving various queries and presenting clear cut outputs.

Conclusion

Airbnb dataset-2019 appeared to be a very rich dataset with a variety of columns that allowed us to do deep data exploration on each significant column presented.

First, we have found hosts that take good advantage of the Airbnb platform and provide the most listings; we found that our top host has 327 listings. After that, we proceeded with analysing boroughs and neighbourhood listing densities and what areas were more popular than another.

From the entire analysis on Airbnb bookings analysis, our assumptions before analysis went totally different after getting results from the analysis. The whole EDA process gave very fascinating results and insights that will be helpful for business development and expansion, budget allocations and focusing on things people prefer.

Project & Dashboard links

- **Colab Notebook:**
  https://colab.research.google.com/drive/1Ha0oVQV4PBwc3HBGAk7bTrMs4QbycJMG#scrollTo=9tbqx-j3nmW7

- **Github Repository :**
- https://github.com/rknethinti/EDA-Airbnb-bookings-analysis/blob/main/Airbnb_Bookings_Analysis_Capstone_Project.ipynb

- **Tableau Dashboard :**
- https://public.tableau.com/app/profile/ramakrishna.nethinti/viz/AirbnbBookingsAnalysis/Dashboard1

Thank You