Capstone Project

EDA

Airbnb Bookings Analysis

Team Members:

Akshada Dani

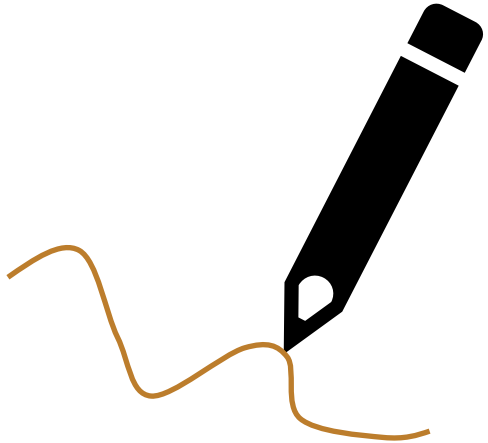
Pritesh Ashok Lonkar

Somnath Patnaik

Subham Choudhary

CONTENT

- About EDA
- AirBnB
- Assumptions
- Map of New York
- Data Overview
- Data Cleaning –
  Outliers, Missing & duplicates
- Data Analysis and Visualisations
- Conclusion

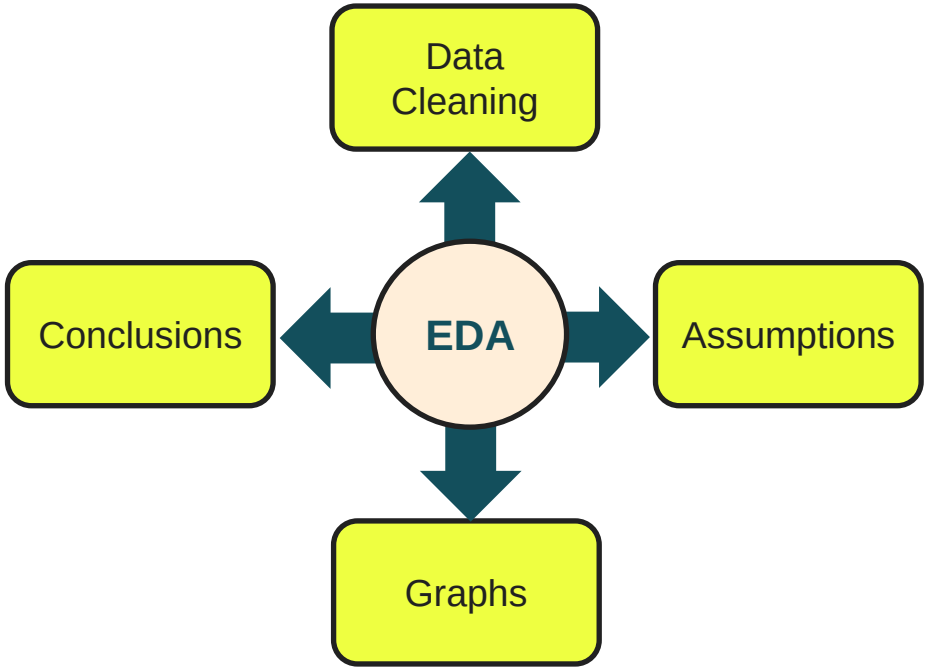


Exploratory Data Analysis (EDA)

EDA is used to analyze and investigate data sets and summarize the data.

It uses both analytical and visualization tools like plots to analyze.

It helps determine patterns, spot anomalies, test a hypothesis, or check assumptions.



AirBnB

Airbnb is digital marketplace where we book lodging facility

The company was founded in 2008 by Brian Chesky, Nathan Blecharczyk, Joe Gebbia

The Platform can be accessed through mobile & web

Airbnb has listings in **more than 220 countries** and regions

Airbnb's headquarters at 888 Brannan Street, in San Francisco, California

Airbnb

Note: The Dataset we are using is from USA

Assumptions

- The Data is true and from reliable source.
- The Conditions and expected to stay as it is.
- Airbnb is impartial in nature and does not show any bias towards any particular neighbourhood.

MAP


NEW YORK
Yonkers
New Rochelle
Huntington
Long Island Sound
Paterson
Clifton
NEW JERSEY
The Bronx
Manhattan
Hudson River
Long Island
Newark
Jersey City
LaGuardia Airport
Queens
Newark Liberty Int. Airport
New York
Hempstead
Statue of Liberty
John F. Kennedy Int. Airport
Elizabeth
Upper N.Y. Bay
Brooklyn
Great South Bay
Staten Island
Woodbridge Township
Lower New York Bay
Edison
NORTH ATLANTIC OCEAN
Raritan Bay
Sandy Hook Bay
Sandy Hook
0 5 km
0 5 mi

Importing and Mounting

Pandas is a high-level data manipulation

NumPy can be used to perform mathematical operations on arrays

```
# Importing neccesary libraries

import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

Matplotlib uses for data visualization and graphical plotting

Seaborn used for making graphics

```
from google.colab import drive
drive.mount('/content/drive')
```

Columns

a) Id: Unique serial number.
b) Name: Description given to each accommodation.
c) Host id: Unique serial number given to each host.
d) Host Name: Name of every host.
e) Neighbourhood Group: Various boroughs(town/district) within New York city.
f) Neighbourhood: Various divisions within each neighbourhood group.
g) latitude and longitude: It is geographic coordinates that specify the position of a particular location.
h) Room Type: Variety of rooms depending on the size.
i) Price: Cost of the rooms.
j) Minimum Nights: Number of nights, hosts stay in that accommodation.
k) Number of reviews: Number of times hosts give reviews.
l) Last Review: Date of last review.
m) Reviews per month: Ratio of number of reviews to number of days in each month.
n) Calculated host listings count: Number of times a host visited that particular room.
o)availability_365: Number of days, rooms are available in a year.

```
air_df.columns

Index(['id', 'name', 'host_id', 'host_name', 'neighbourhood_group',
       'neighbourhood', 'latitude', 'longitude', 'room_type', 'price',
       'minimum_nights', 'number_of_reviews', 'last_review',
       'reviews_per_month', 'calculated_host_listings_count',
       'availability_365'],
      dtype='object')
```

Dataset Overview

	price	minimum_nights	number_of_reviews	reviews_per_month	calculated_host_listings_count	availability_365
count	48895.000000	48895.000000	48895.000000	38843.000000	48895.000000	48895.000000
mean	152.720687	7.029962	23.274466	1.373221	7.143982	112.781327
std	240.154170	20.510550	44.550582	1.680442	32.952519	131.622289
min	0.000000	1.000000	0.000000	0.010000	1.000000	0.000000
25%	69.000000	1.000000	1.000000	0.190000	1.000000	0.000000
50%	106.000000	3.000000	5.000000	0.720000	1.000000	45.000000
75%	175.000000	5.000000	24.000000	2.020000	2.000000	227.000000
max	10000.000000	1250.000000	629.000000	58.500000	327.000000	365.000000

```
Rows: 48895
Columns: 16
```

```
Data Types (Columns):
float64(3), int64(7), object(6)
```

25% of Listings have 69$, 50% bookings have 106$ & 75% have cost 175$

Mean Cost : 153$ Max Cost is 10,000 $ Avg Nights Booked: 7

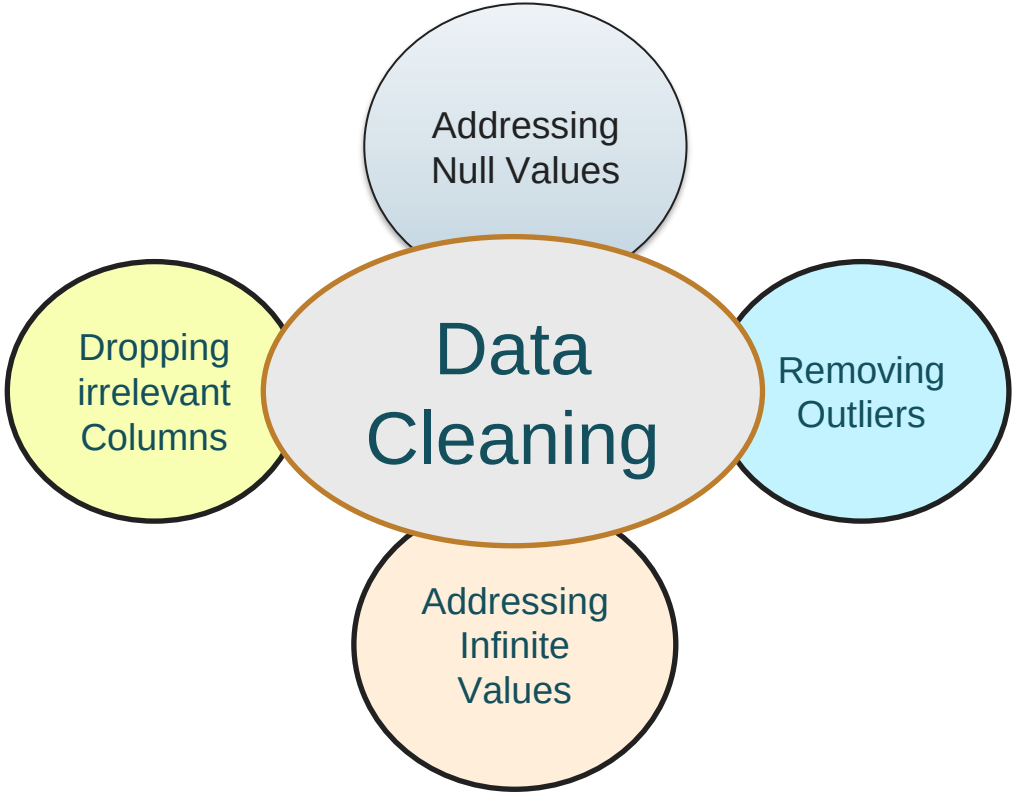
Data Frame

```
] air_df.head()
```

	id	name	host_id	host_name	neighbourhood_group	neighbourhood	latitude	longitude	room_type	price	minimum_nights	number_of_reviews	last_review	reviews_per_month	calculated_host_listings_count	availability_3
0	2539	Clean & quiet apt home by the park	2787	John	Brooklyn	Kensington	40.64749	-73.97237	Private room	149	1	9	2018-10-19	0.21	6	3
1	2595	Skylit Midtown Castle	2845	Jennifer	Manhattan	Midtown	40.75362	-73.98377	Entire home/apt	225	1	45	2019-05-21	0.38	2	3
2	3647	THE VILLAGE OF HARLEM....NEW YORK !	4632	Elisabeth	Manhattan	Harlem	40.80902	-73.94190	Private room	150	3	0	NaN	NaN	1	3
3	3831	Cozy Entire Floor of Brownstone	4869	LisaRoxanne	Brooklyn	Clinton Hill	40.68514	-73.95976	Entire home/apt	89	1	270	2019-07-05	4.64	1	1
4	5022	Entire Apt: Spacious Studio/Loft	7192	Laura	Manhattan	East Harlem	40.79851	-73.94399	Entire	80	10	9	2018-11-19	0.10	1	

“ Head Command top 5 rows ”

Data Cleaning



Data Cleaning

Check for null values in each column and removing Null Values

```
air_df.isna().sum()
```

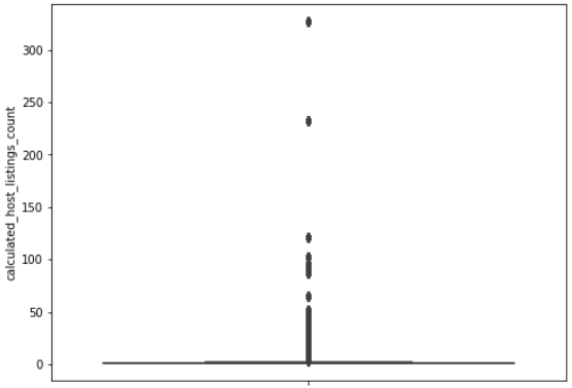
```
id                                  0
name                               16
host_id                             0
host_name                          21
neighbourhood_group                 0
neighbourhood                       0
latitude                            0
longitude                           0
room_type                           0
price                               0
minimum_nights                      0
number_of_reviews                   0
last_review                     10052
reviews_per_month               10052
calculated_host_listings_count      0
availability_365                    0
dtype: int64
```

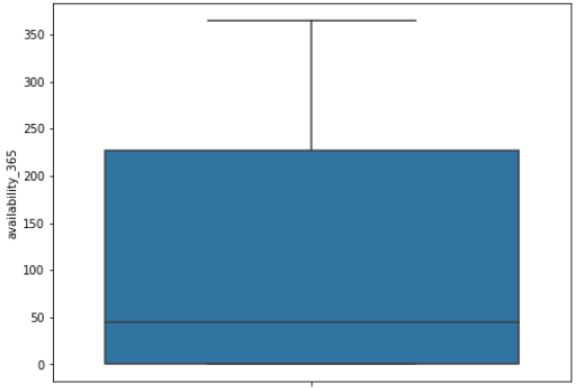
```
new_air_df.isna().sum()
```

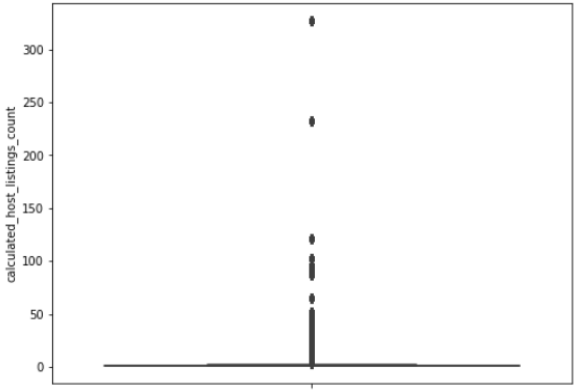
```
host_name                       0
neighbourhood_group             0
neighbourhood                   0
latitude                        0
longitude                       0
room_type                       0
price                           0
minimum_nights                  0
number_of_reviews               0
reviews_per_month               0
calculated_host_listings_count  0
availability_365                0
dtype: int64
```

“ Replaced Null Values with 0 and dropped few least relevant columns ”

Outliers





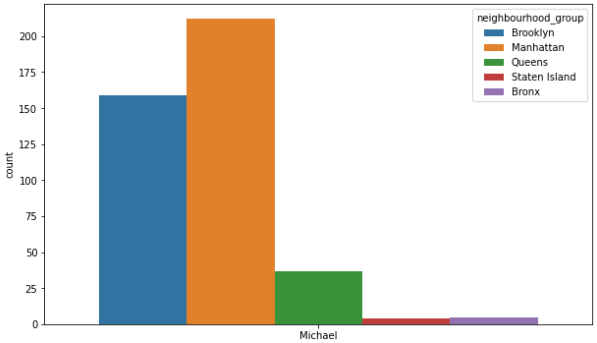


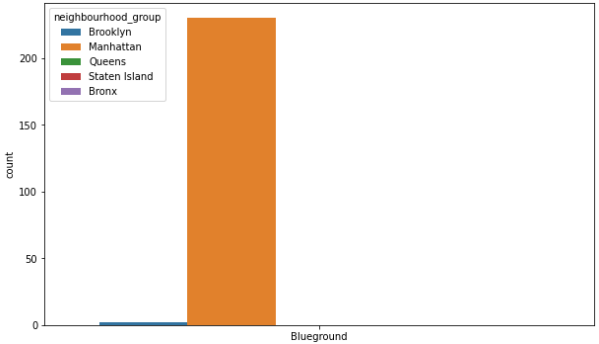
“ In this dataset, values are genuine customer behavior. So, we do not remove any data value which is seen ae outlier. ”

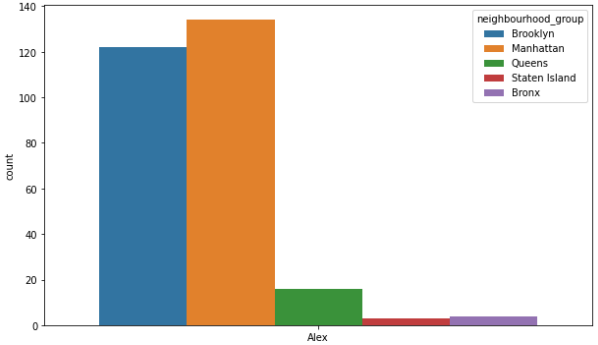
EDA

Host Name & Preferred Room type

a) Host Name

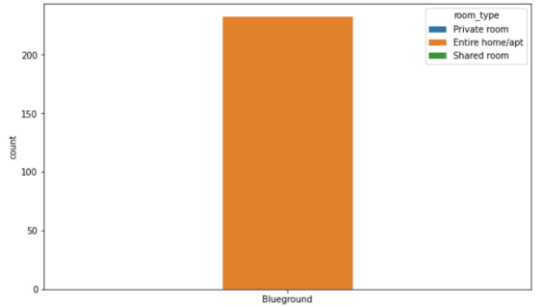


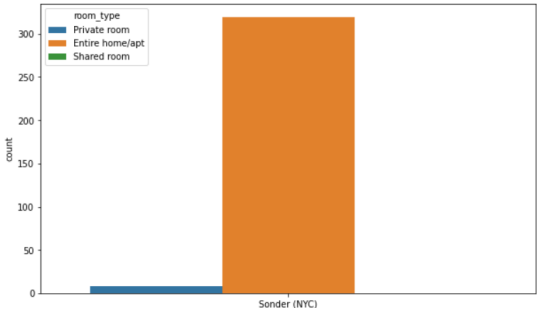


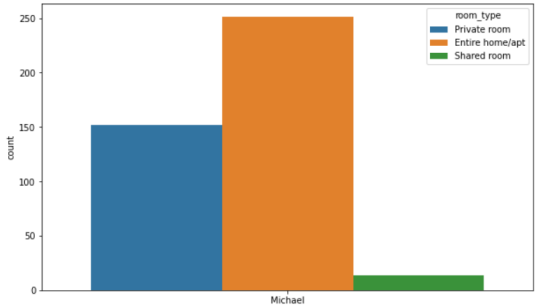


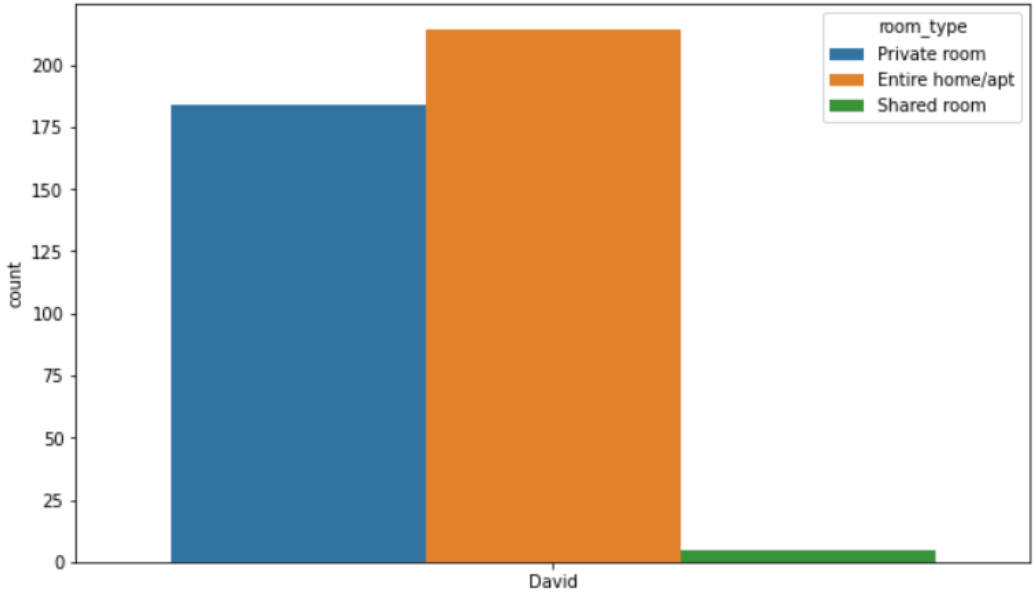
" After analysing top 10 Spenders, it was found that Manhattan and Brooklyn were booked most. "

b) Room Type





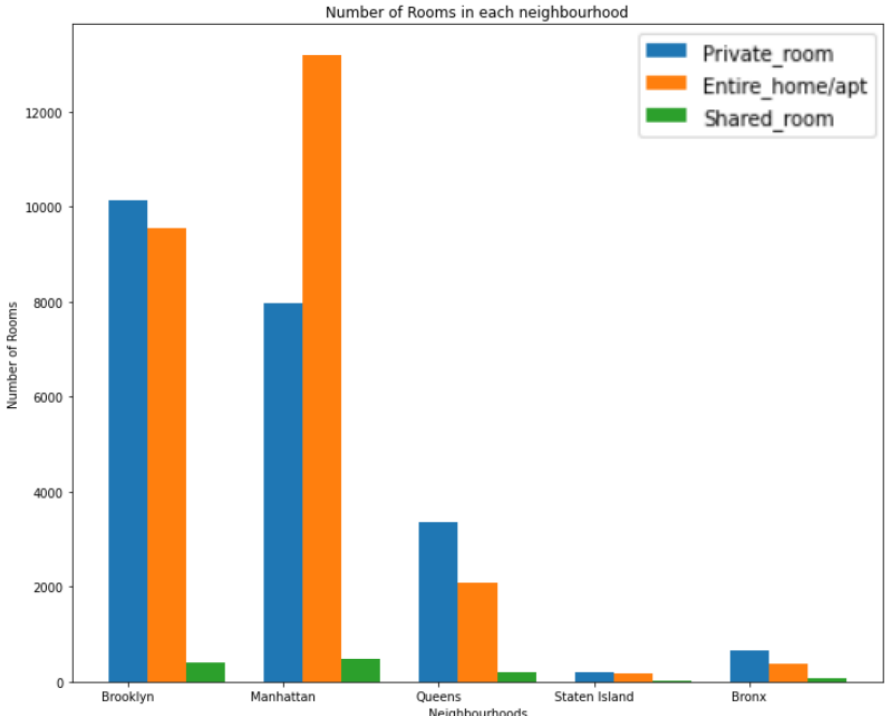




“ Top spenders preferred mostly Entire home/apt and Private room ”

Grouping count of different rooms in different neighbourhood group

```
Brooklyn and Private room 10128
Brooklyn and Entire home/apt 9554
Brooklyn and Shared room 413
Manhattan and Private room 7976
Manhattan and Entire home/apt 13196
Manhattan and Shared room 480
Queens and Private room 3370
Queens and Entire home/apt 2096
Queens and Shared room 198
Staten Island and Private room 188
Staten Island and Entire home/apt 176
Staten Island and Shared room 9
Bronx and Private room 652
Bronx and Entire home/apt 378
Bronx and Shared room 60
```



" Every Neighbourhood had almost equal numbers of Private Rooms or Entire Home/Apts and Brooklyn had MAXIMUM Rooms "

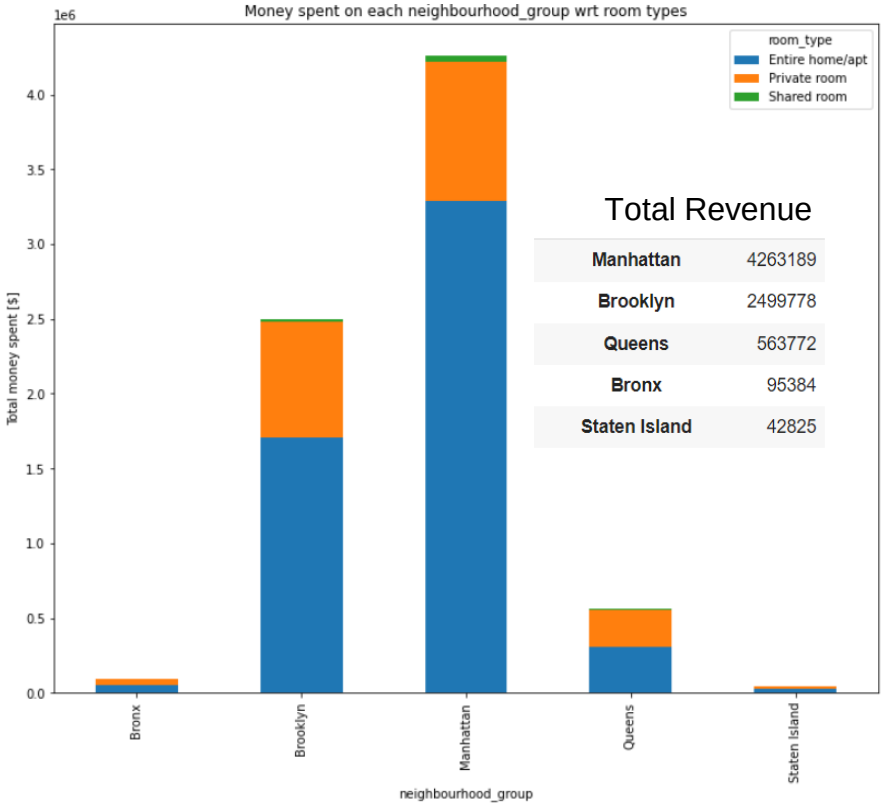
Revenue Generated by Various Groups

Neighbourhood Group	Entire Home/apt	Private Room	Shared Room
Bronx	48250	43546	3588
Brooklyn	1704008	774902	20868
Manhattan	3288982	931498	42709
Queens	308218	241888	13666
Staten Island	30597	11711	517

Avg Room Cost

Manhattan – 196.89, Bronx – 87.5, Queens – 99.5

Brooklyn– 124.39, Staten Island – 114.8



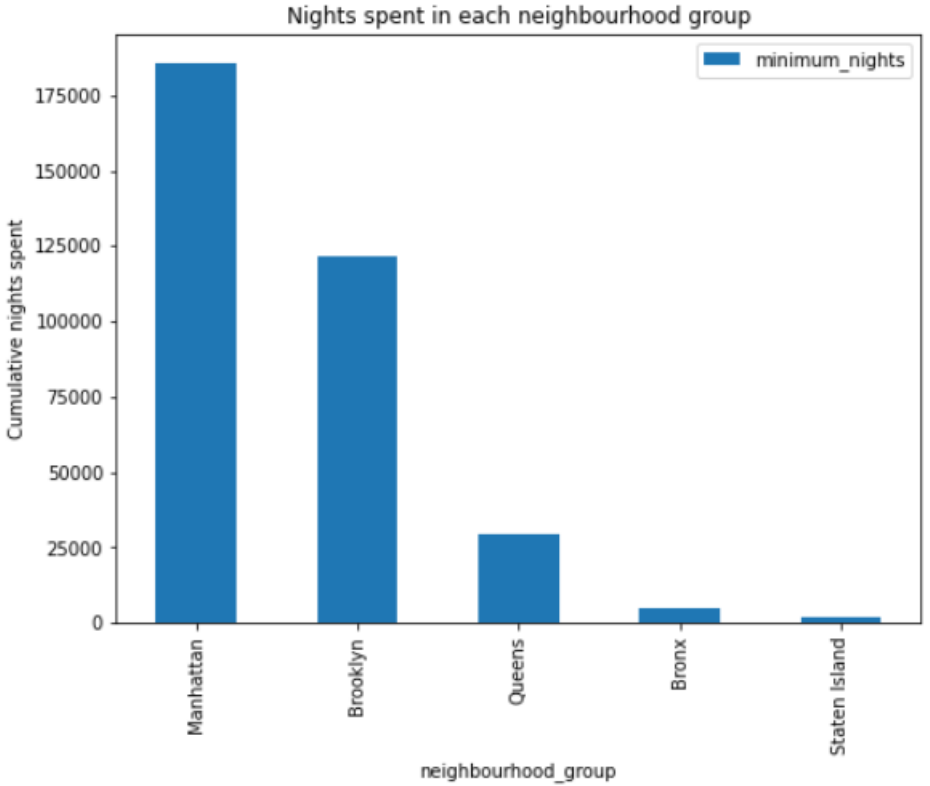
Nights Spent

Interim Conclusion

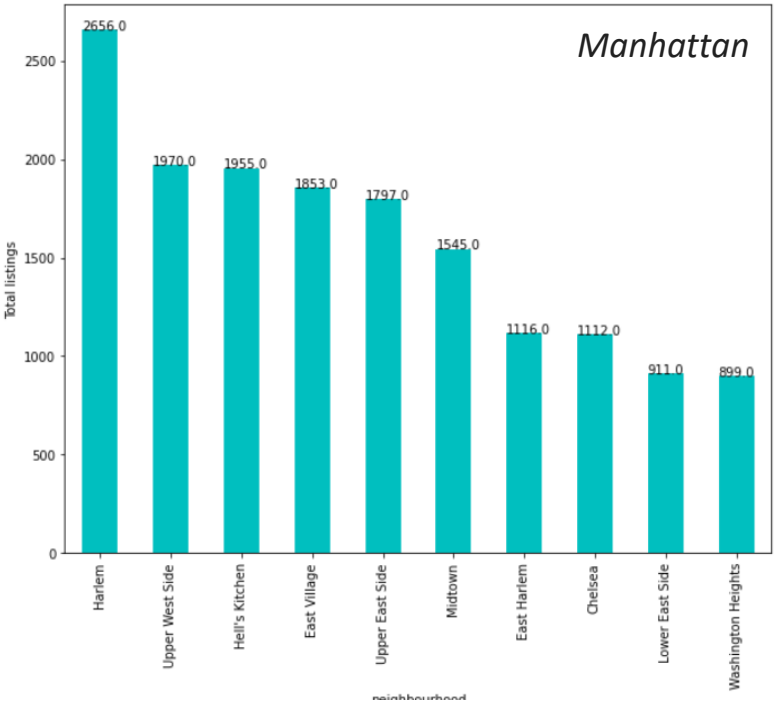
Manhattan has highest average cost, maximum rooms & maximum revenue followed by Brooklyn.

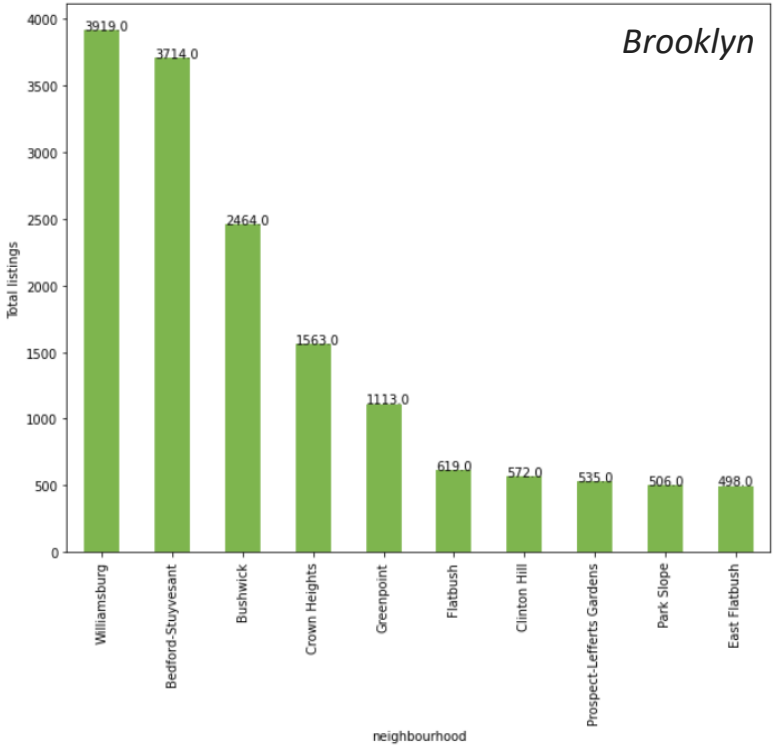
These places are occupied on weekdays and holidays.

So, these places may be business hub cum tourist place



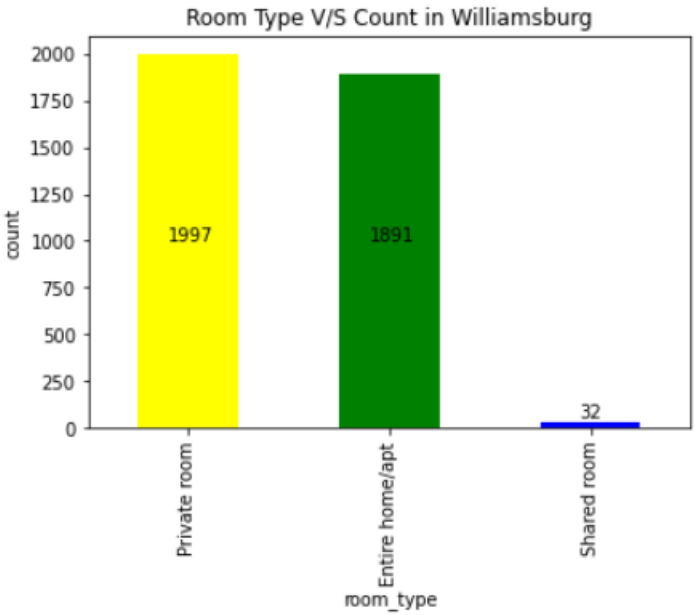
Top 10 Bookings from Manhattan & Brooklyn

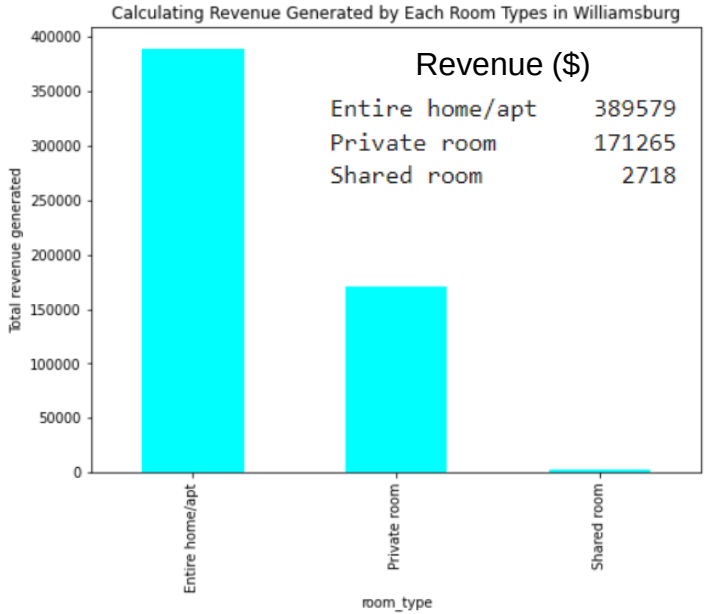




" So, One destination (Williamsburg) had 8.02 % bookings "

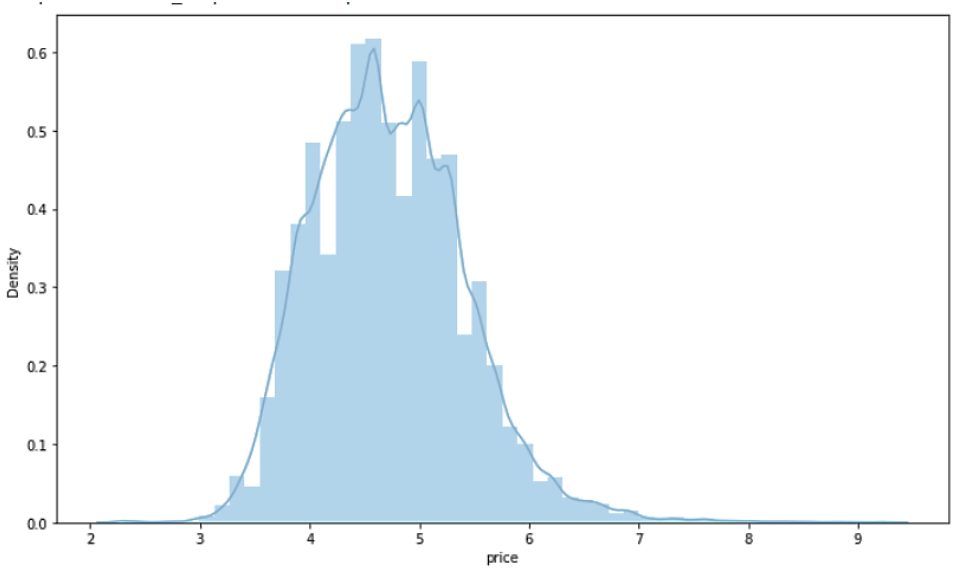
EDA on Most Frequently Booked Neighbourhood

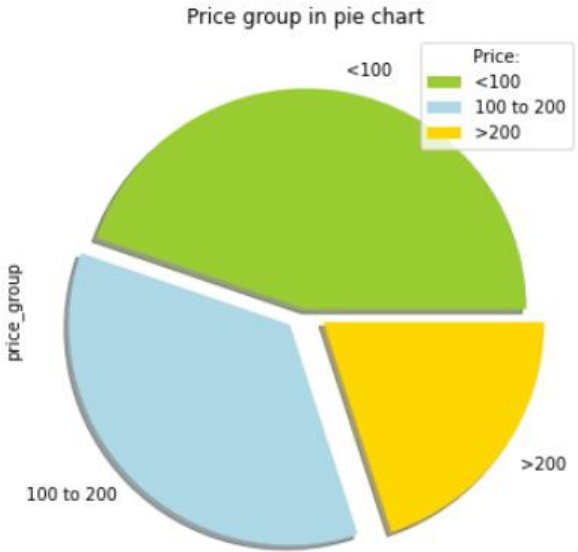




" Though Private rooms & Entire/apt Home have similar bookings, revenue generated by Entire Home/apt is more. So, if someone is willing to start new business, Private Rooms in Williamsburg is most profitable and safest option "

Price Density Distribution





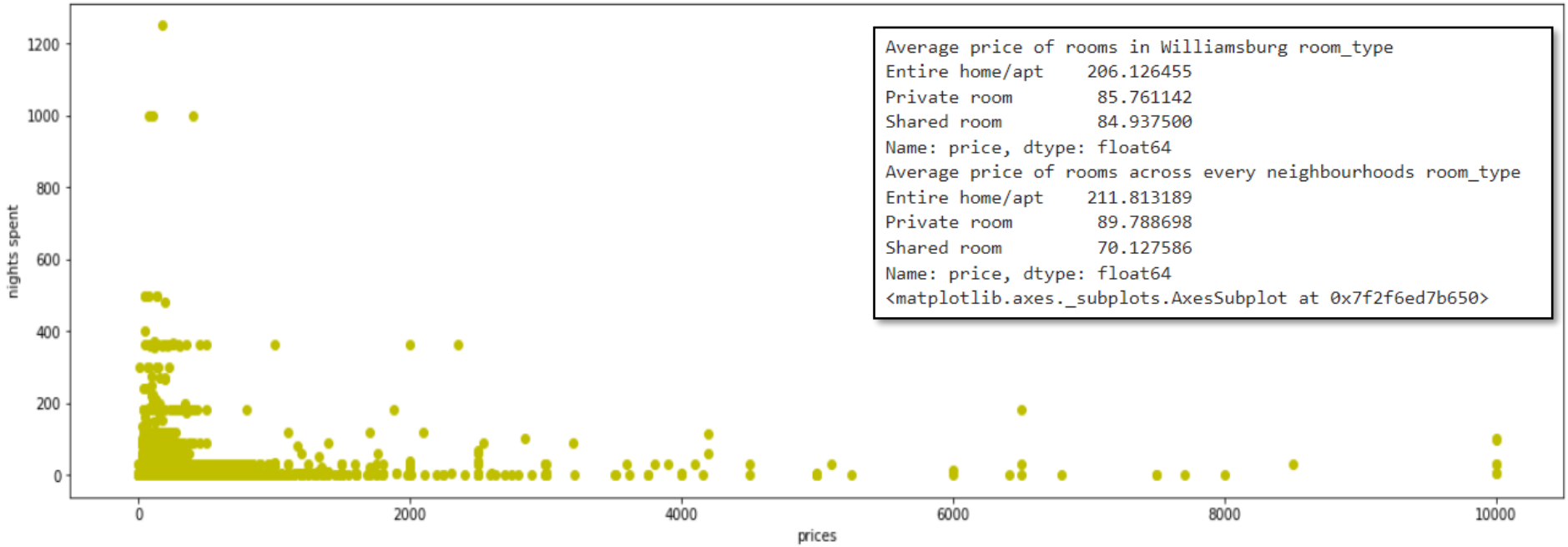
21877 rooms having price less than 100

17233 rooms havind price between 100 to 200

9785 rooms having price greater than 200

Most rooms are below 200 price range with almost 80% and very few are more than 200.

Price V/S Nights and Average Price




```
Average price of rooms in Williamsburg room_type
Entire home/apt    206.126455
Private room        85.761142
Shared room         84.937500
Name: price, dtype: float64
Average price of rooms across every neighbourhoods room_type
Entire home/apt    211.813189
Private room        89.788698
Shared room         70.127586
Name: price, dtype: float64
<matplotlib.axes._subplots.AxesSubplot at 0x7f2f6ed7b650>
```

" The Costs of National Average and Most Booked Neighbourhood (Williamsburg) are similar and booking were maximum when price is up to 200 $!! "

CONCLUSION

- Sonder(NYC), Blueground, Michael and David are top 4 most spending customers. Their popular destinations is mainly Manhattan and Brooklyn.
- Top spenders prefer Private rooms or Entire apartments to stay. So, it's may be a tourist place cum business hub.
- Manhattan had generated maximum revenue i.e., 4,263,189 $ and Williamsburg in Brooklyn is the most preferred destination with 3,919 bookings.
- Entire Homes/Apts had generated maximum revenue and had average price of 200$.

So, If anyone wants to start a new business he/she can start with Entire Home/Apts located at Williamsburg and having price range of around 200 $ are mostly booked.

THANK YOU