Capstone Project – 1
Airbnb Booking Analysis

Team Members

Puneet Suthar
Ankush Yadav
Anvesh Agarwal
Sayed Faraz Ashraf

What are we talking about ?

Content :

- Introduction to Airbnb
- General overview of the dataset
- Data cleaning
- Exploratory Data Analysis
- Conclusion

Introduction

Airbnb is an open online platform where people list their own housing for rent. Since 2008, it has grown in popularity and specially for those community which frequently use to travel. It is becoming a strong competitor to the hotel industry. It has millions of listing, which generate lots of data. We are analyzing these data for making business decision, for looking best room type etc.

How our dataset Look like ?

General overview of the dataset

16 Columns

&

48895 Rows

General overview of the dataset

16 Column

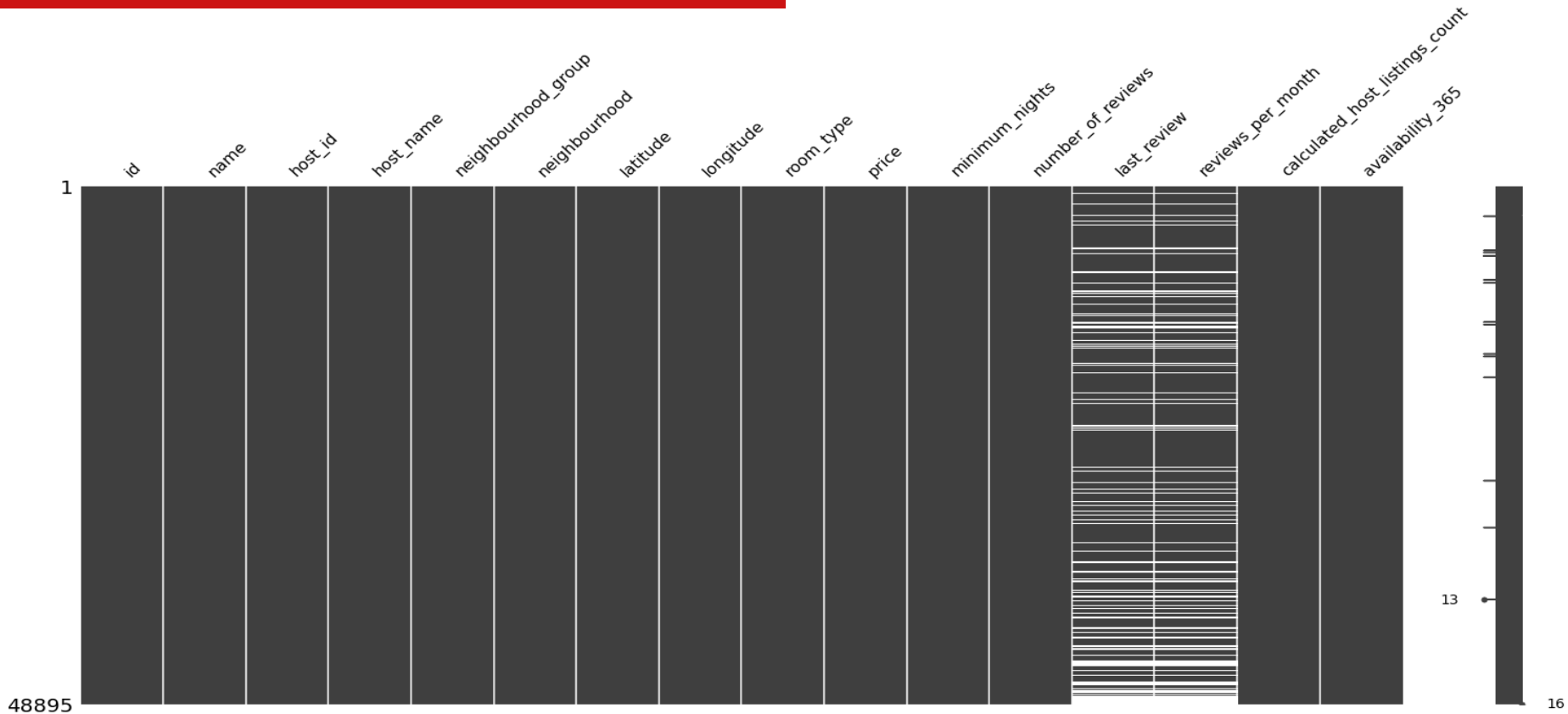
1. id : Unique listing id.
2. name : Name of the property .
3. host_id : unique id for each listed host.
4. host_name : Name of the host.
5. neighbourhood_group : Location
6. neighborhood : Area
7. latitude : Latitude coordinates
8. longitude : Longitude coordinates
9. room_type : Listing space types

General overview of the dataset

16 Column

10. price : Price in dollars
11. minimum_nights : minimum nights required to stay
12. number_of_reviews : No. of reviews written for the listing
13. last_review : Last reviewed date for the listing
14. reviews_per_month : Total review per month for the listing
15. calculated_host_listings_count : Total no of listing against the host id
16. availability_365 : Number of days when listing is available for booking.

Null values in the dataset



Null values in the dataset


```
# Finding null values in the dataset
df.isnull().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
```

- Last_reviews – **10052** null values
- reviews_per_month - **10052** null values
- host_name – **21** null values
- name – **16** null values

Have a look on **unique** values

```
# Insight about unique values
df.nunique()
```

Have a look on **unique** values

```
# Insight about unique values
df.nunique()
```

```
id                                48895
name                              47905
host_id                           37457
host_name                         11452
neighbourhood_group                   5
neighbourhood                       221
latitude                          19048
longitude                         14718
room_type                             3
price                               674
minimum_nights                      109
number_of_reviews                   394
last_review                        1764
reviews_per_month                   937
calculated_host_listings_count       47
availability_365                    366
dtype: int64
```

neighbourhood_ groups

&

room_types

```
# Finding types of neighbourhood Groups
pd.DataFrame(df.neighbourhood_group.unique(),columns = ['neighbourhood_group'])
```

```
# Finding types of Room type
pd.DataFrame(df.room_type.unique(),columns = ['room_type'])
```

neighbourhood_ groups

&

room_types

```
# Finding types of neighbourhood Groups
pd.DataFrame(df.neighbourhood_group.unique(),columns = ['neighbourhood_group'])
```

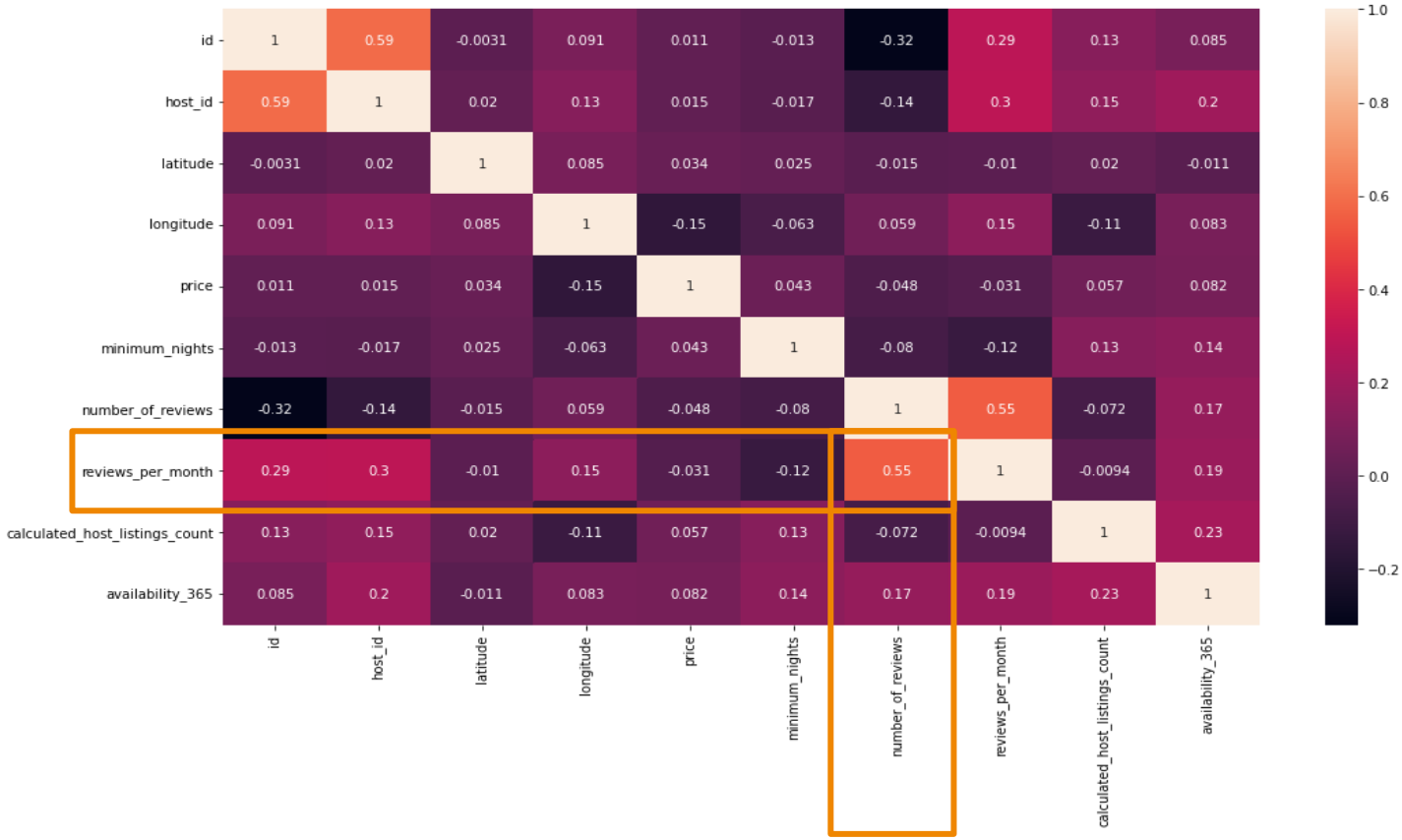
	neighbourhood_group
0	Brooklyn
1	Manhattan
2	Queens
3	Staten Island
4	Bronx

```
# Finding types of Room type
pd.DataFrame(df.room_type.unique(),columns = ['room_type'])
```

	room_type
0	Private room
1	Entire home/apt
2	Shared room

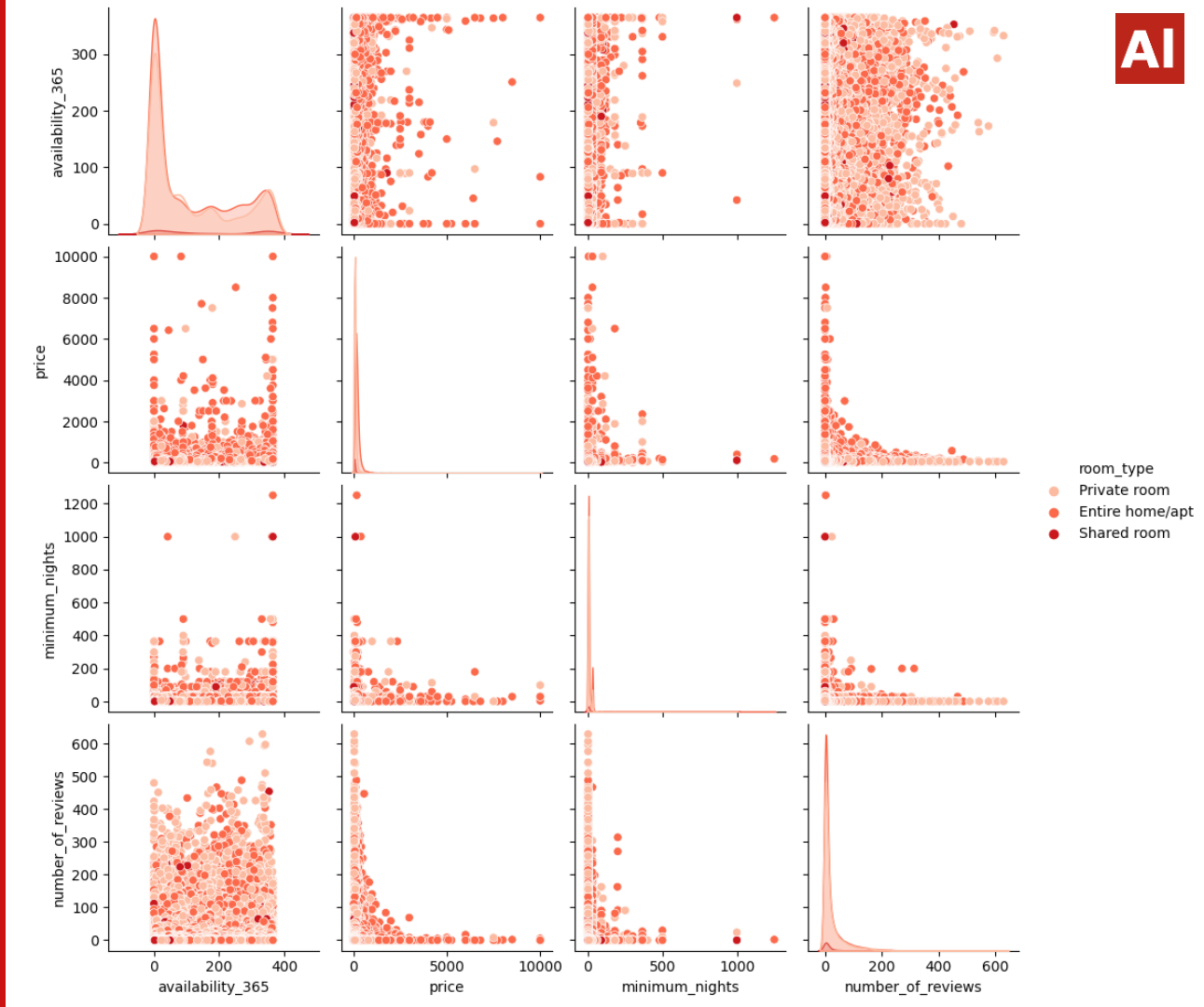
Is our data have any relations?

Heat map showing Relations between the columns



Pairplot of –

- Availability_365
- Price
- Minimum nights
- Number of reviews

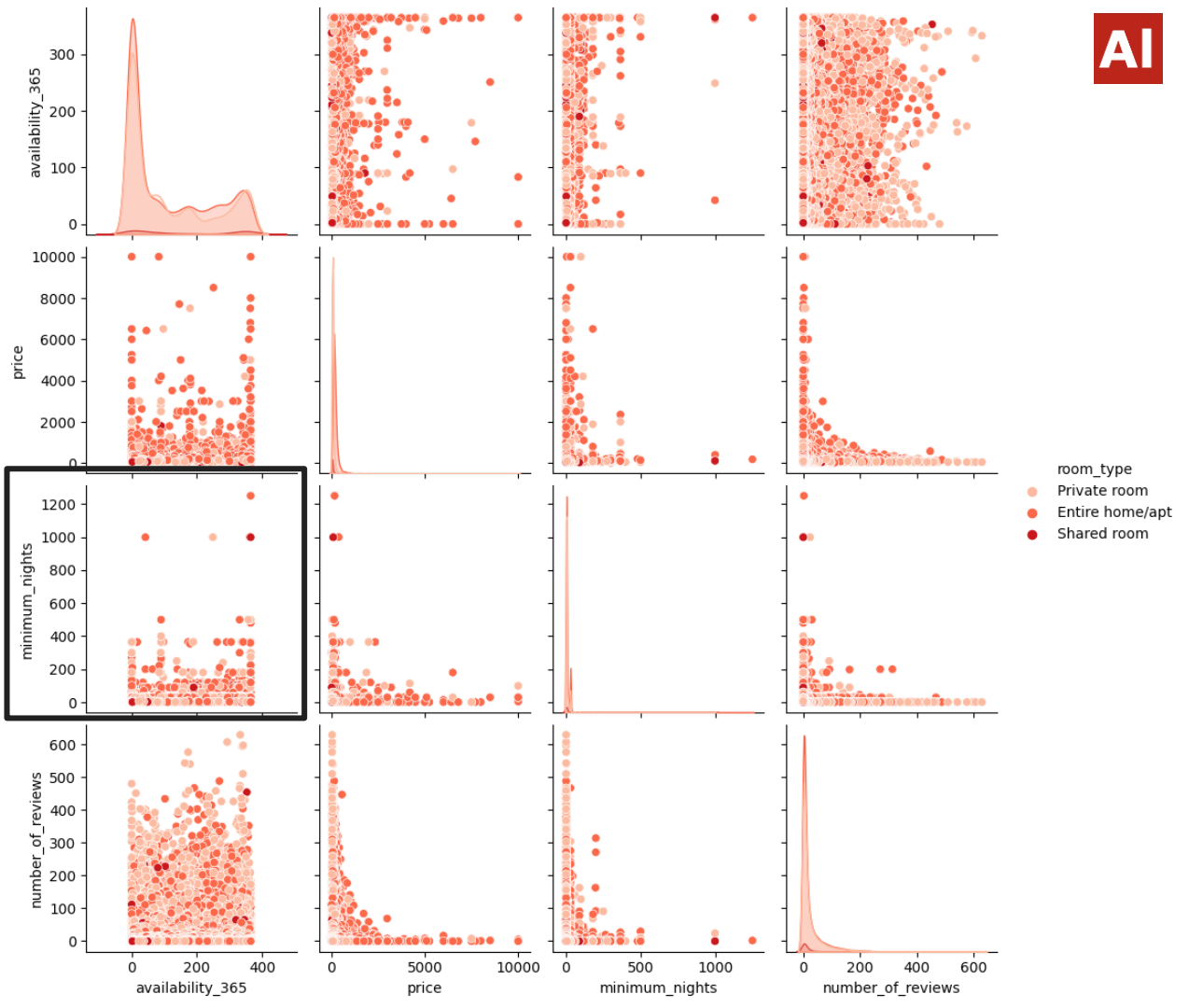


Pairplot of –

- Availability_365
- Price
- Minimum nights
- Number of reviews

Results --

- Unexpected minimum nights at 0 availability

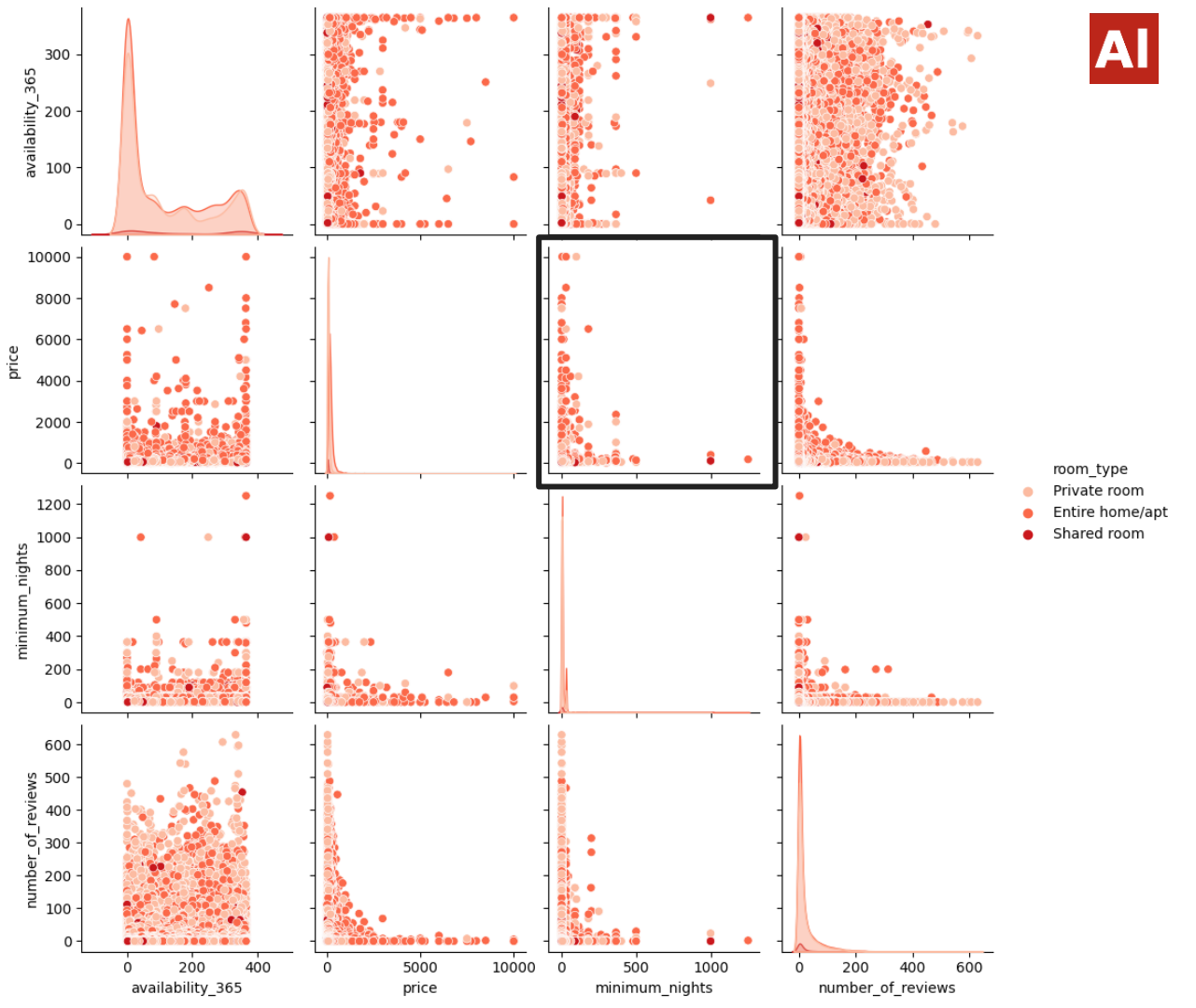


Pairplot of –

- Availability_365
- Price
- Minimum nights
- Number of reviews

Results --

- Unexpected minimum nights at 0 availability
- As minimum nights of booking increase price decreases significantly.

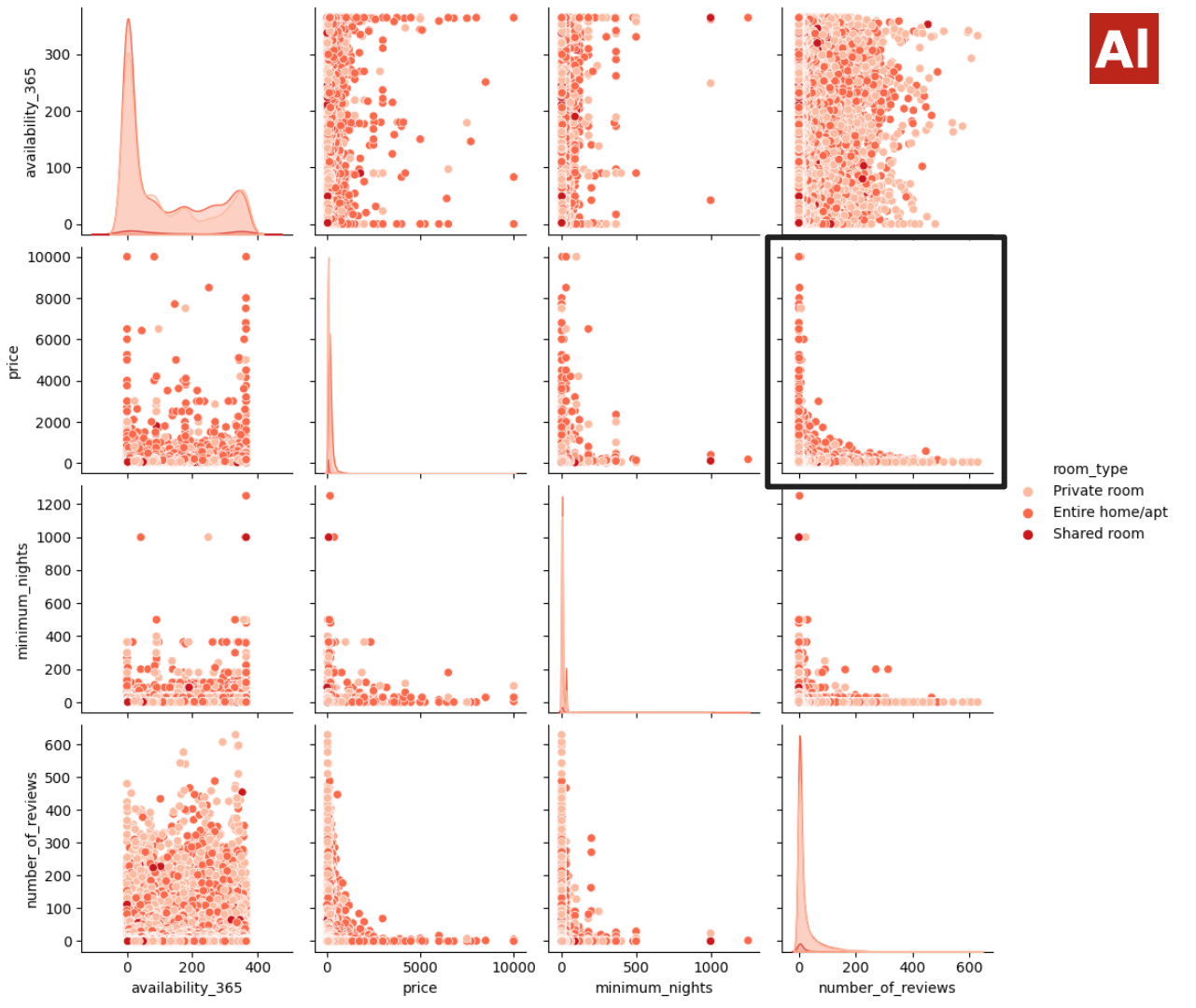


Pairplot of –

- Availability_365
- Price
- Minimum nights
- Number of reviews

Results --

- Unexpected minimum nights at 0 availability
- As minimum nights of booking increase price decreases significantly.
- There are more number of reviews where price is low



It’s time for
DATA CLEANING

We had.....

1. 10K Null values in **reviews_per_month**. → **Replaced by 0**
2. Some Nulls in name and **host_name**. → **Replaced by NA**
3. 833 listings have **greater min-nights than availability_365** , where availability_365 is not equal to 0. → **Dropped**
4. Some unwanted columns like **last reviews** → **Dropped**

And some …….

```
# Checking if there is any property having price = 0
df[df['price'] == 0].shape
```

```
(11, 15)
```

Which has been replaced by the **mean price of different room types with respect to different neighbourhood_groups**

EDA
(Exploratory Data Analysis)

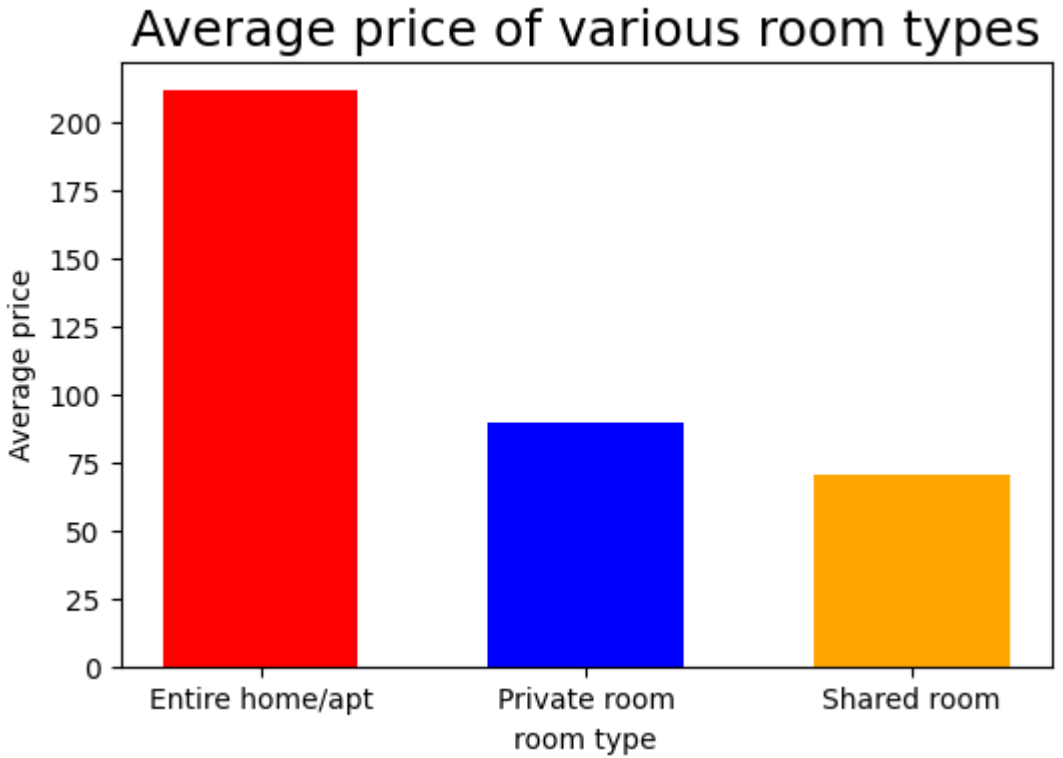
EDA
(Exploratory Data Analysis)

1. Price Analysis
2. Listings Analysis
3. Availability Analysis
4. Profitability Analysis

1 Price Analysis

1.1 Mean price of Various Room types in NYC

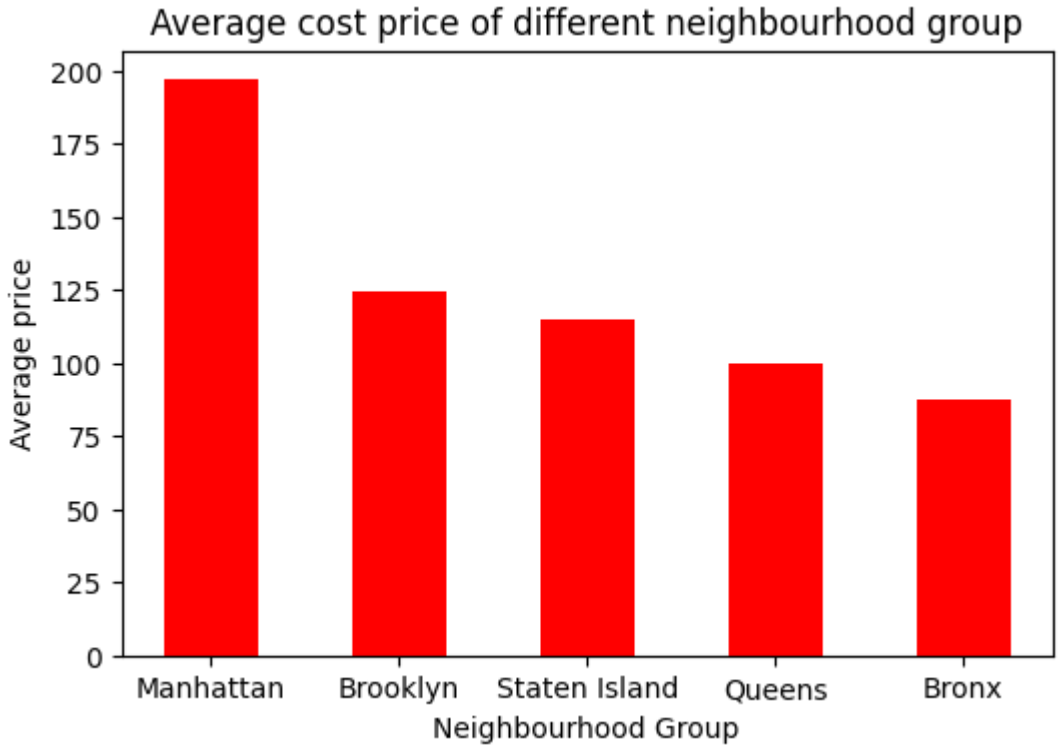
Average price of **Entire home/apt** is highest



1 Price Analysis

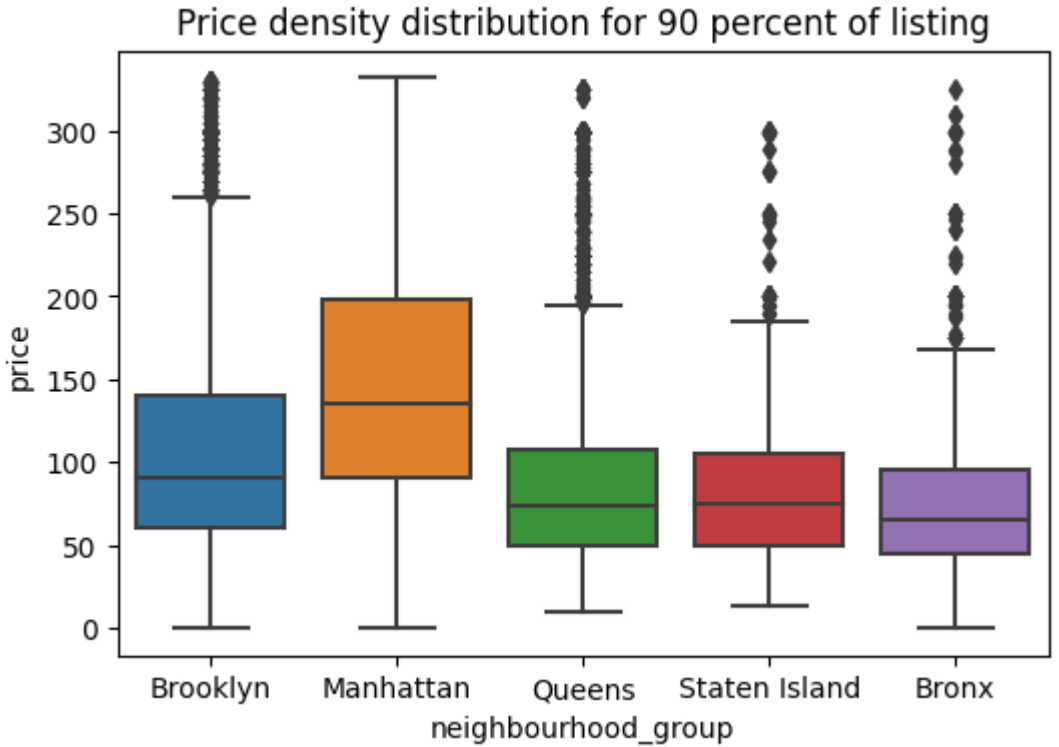
1.2 Average cost price of the different neighbourhood group

Average price in **Manhattan** is highest

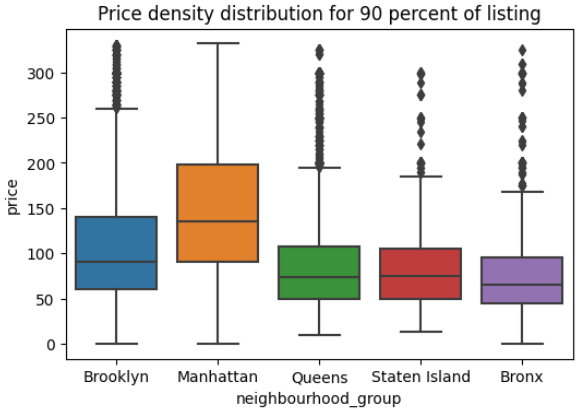


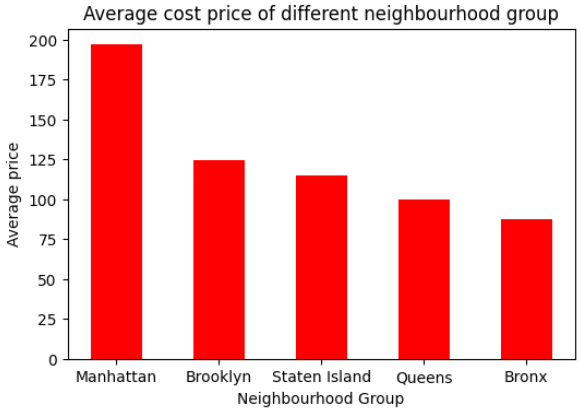
1.3 Price density distribution for 90 percent of listing

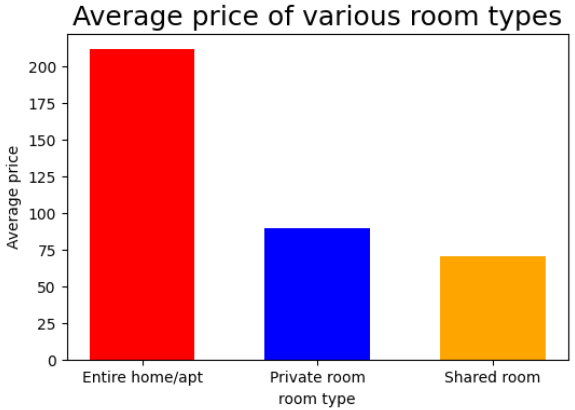
Average price in **Manhattan** is highest and lowest outliers



1 Price Analysis





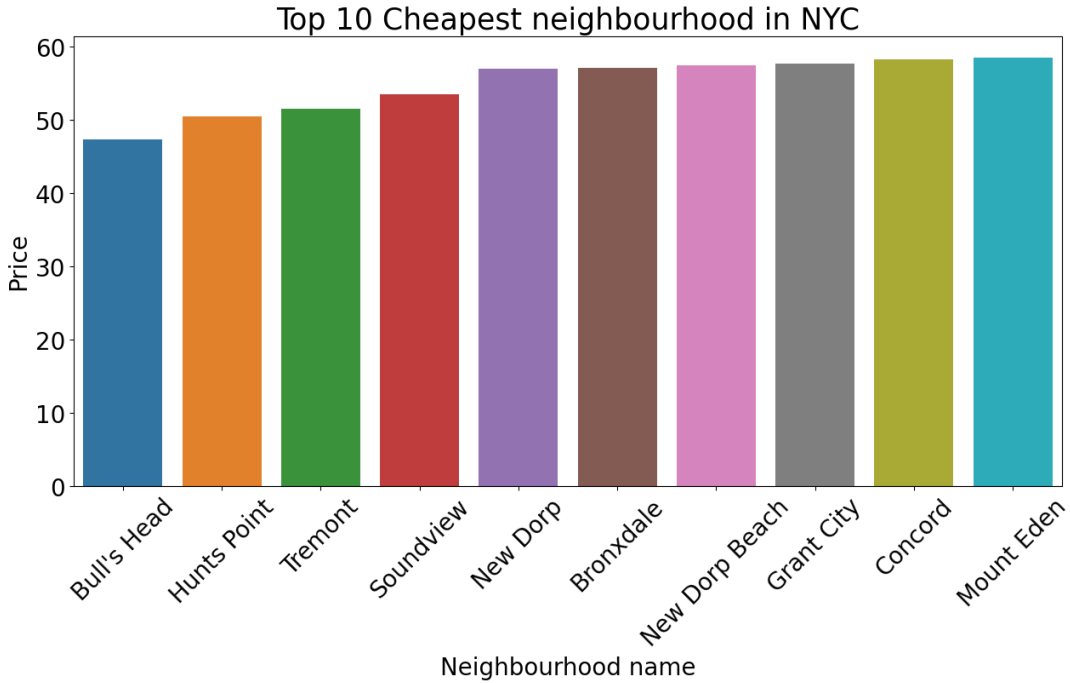


- Order of costliest neighbourhood group on basis of average price and price distribution => Manhattan > Brooklyn > Staten Island > Queens > Bronx.

- Staten Island was having least no. of listing but it is not the cheapest.

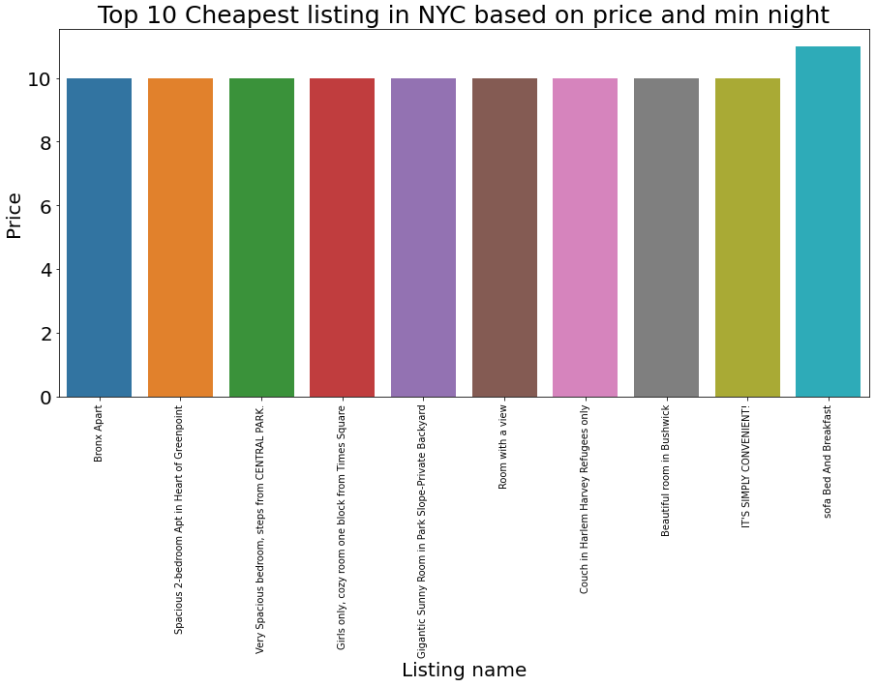
1 Price Analysis

1.4 Top 10 cheapest neighbourhood



1.5 Cheapest listing on the basis of min nights and price

	name	min_price_to_stay	min_nights
34446	Bronx Apart	10	1
27972	Spacious 2-bedroom Apt in Heart of Greenpoint	10	1
31066	Very Spacious bedroom, steps from CENTRAL PARK.	10	1
24100	Girls only, cozy room one block from Times Square	10	1
32810	Gigantic Sunny Room in Park Slope-Private Back...	10	1
33505	Room with a view	10	1
21700	Couch in Harlem Harvey Refugees only	10	1
47218	Beautiful room in Bushwick	10	1
22835	IT'S SIMPLY CONVENIENT!	10	1
35005	sofa Bed And Breakfast	11	1



EDA
(Exploratory Data Analysis)

1. Price Analysis
2. Listings Analysis
3. Availability Analysis
4. Profitability Analysis

2 Listings Analysis

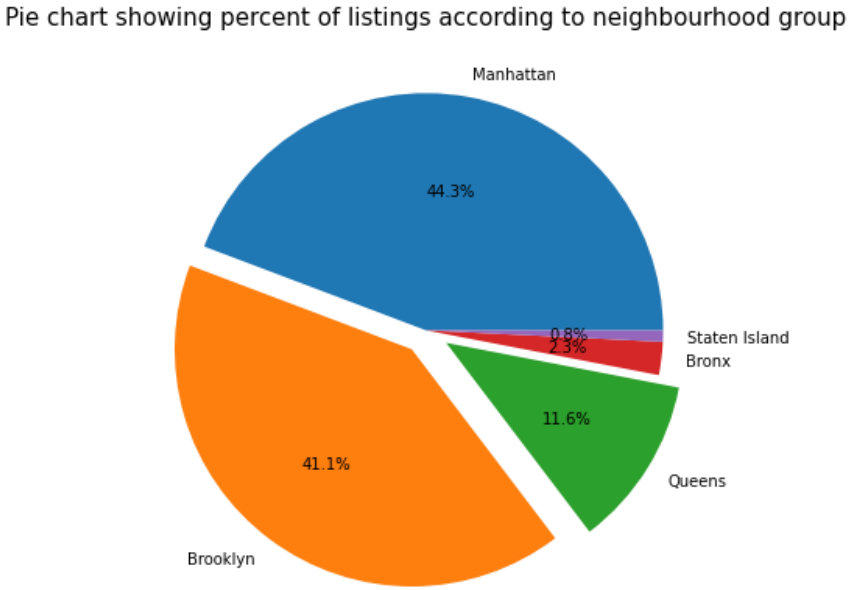
2.1 Airbnb listings neighbourhood_group wise

Result:

1. Manhatten & Brooklyn are having high no. of listing.
2. Staten island and Bronx have low no. of listing.

Inference:

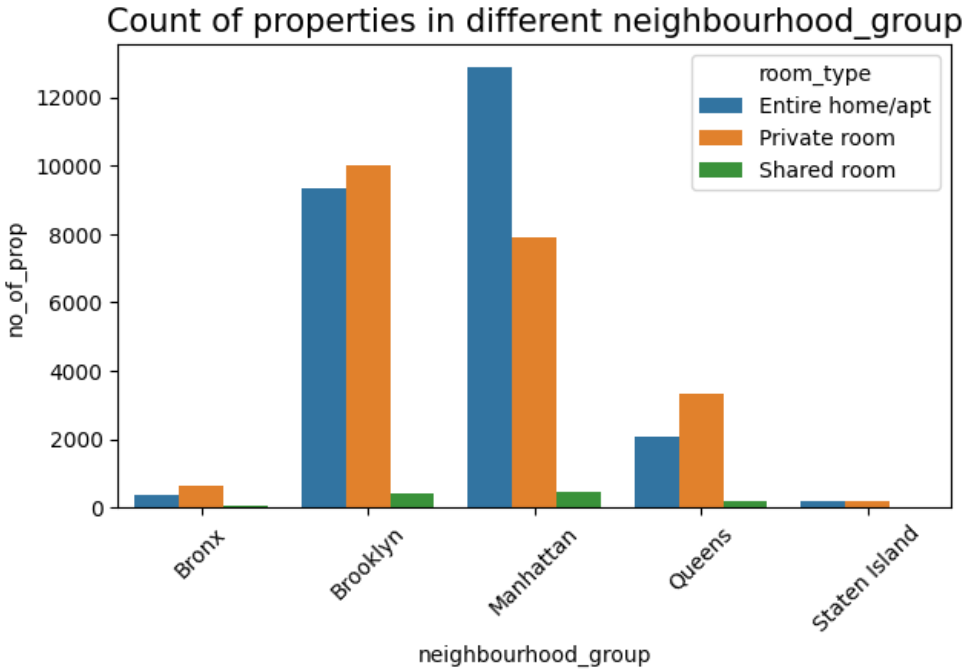
If we want to do advertisement or marketing for selling flats we should focus on Manhattan and Brooklyn.



2.2 Number of properties in different neighbourhood_group

Inference:

1. **Manhatten** have highest no. of listings.
2. In all the three types of room, the **Shared room is least in every neighbourhood**. This shows that in New York people dosent like to share room.
3. **InManhattenn mostly Entire homes and apartment are listed** this means people in this area people are leaving with their families.



EDA
(Exploratory Data Analysis)

1. Price Analysis
2. Listings Analysis
3. Availability Analysis
4. Profitability Analysis

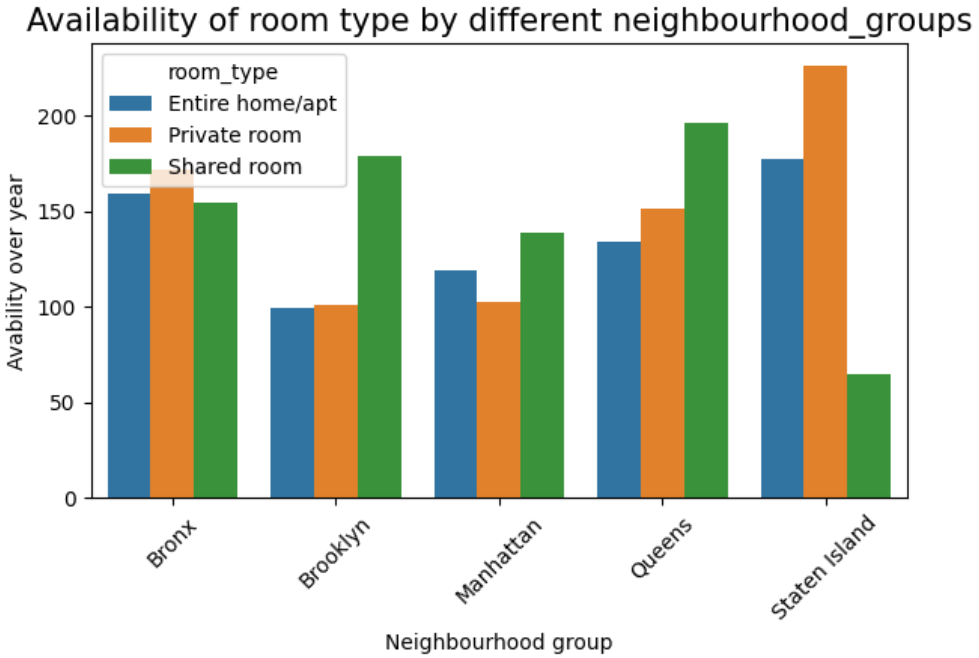
3 Availability Analysis

3.1 Checking availability of different room and different neighbourhood group.

Inference:

1. From previous analysis we got to know that Manhatten and Brooklyn have highest count of property, this graph shows Manhatten and Brooklyn have less availibility then compared to other neighbourhood groups which is good for the host having these propeties.

2. Having property in staten island and bronx is a loss making business for the host as they are empty half of year.

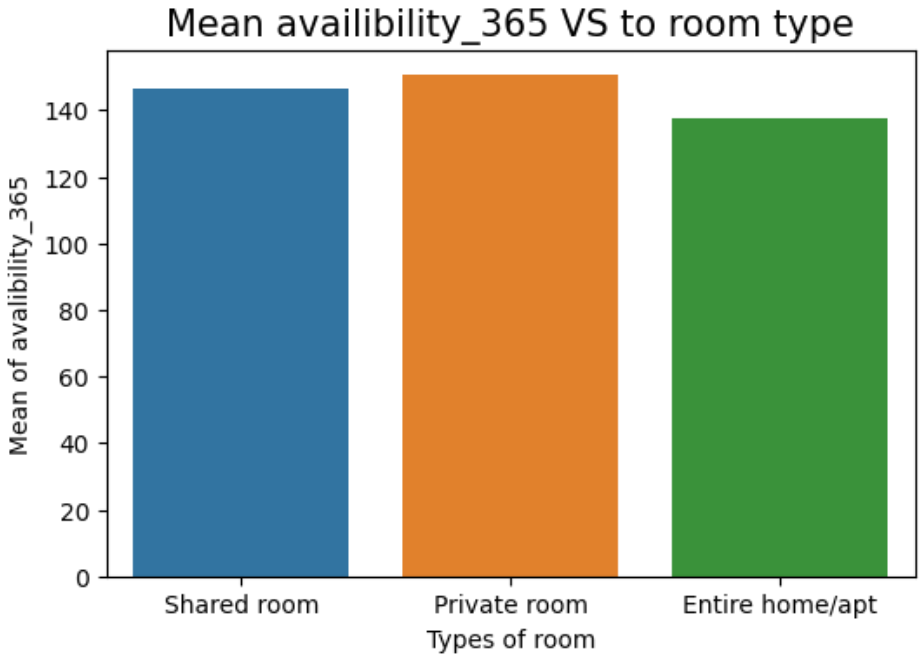


3.2 Average availability on the basis of room type

Inference:

1. Private room has highest mean availibility.

2. Entire home has least mean availibility.



EDA
(Exploratory Data Analysis)

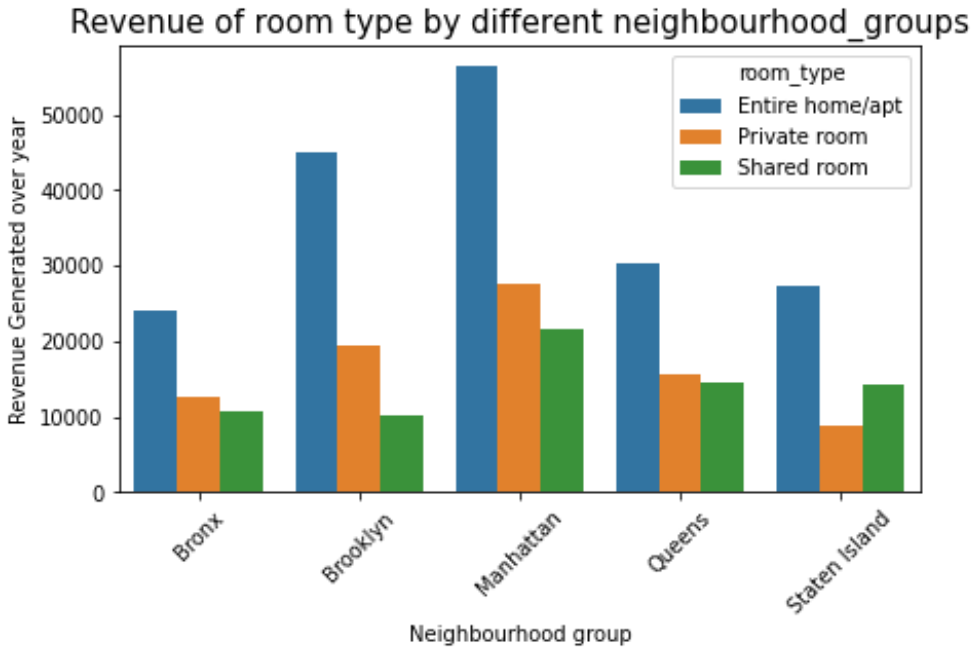
1. Price Analysis
2. Listings Analysis
3. Availability Analysis
4. Profitability Analysis

4 Profitability Analysis

4.1 Checking revenue generated by different property.

Inference:

As we can see, in every neighbourhood “Entire home” has generated the most revenue. So buying a property "Entire home" and renting it, is a profitable business irrespective of neighbourhood type.

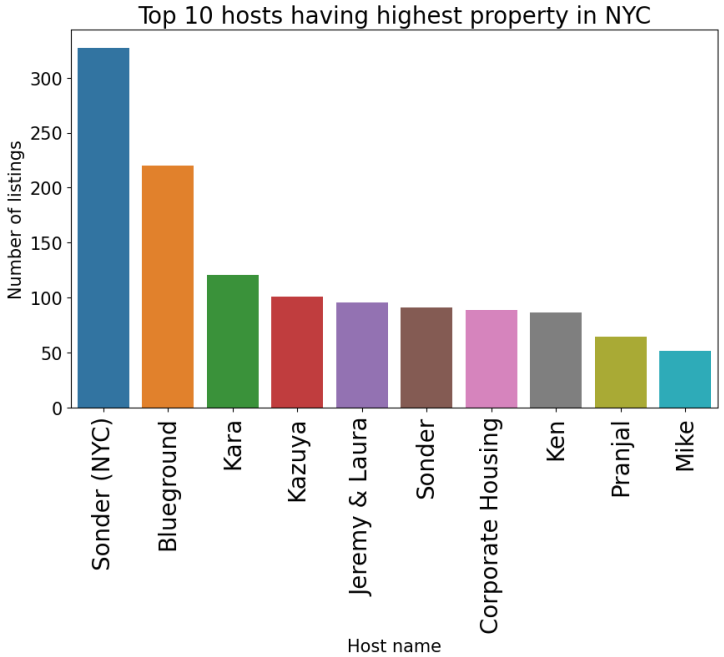


4 Profitability Analysis

4.2 Finding host having highest no. of properties

Inference:

As we can see, in every neighbourhood “Entire home” has generated the most revenue. So buying a property "Entire home" and renting it, is a profitable business irrespective of neighbourhood type.

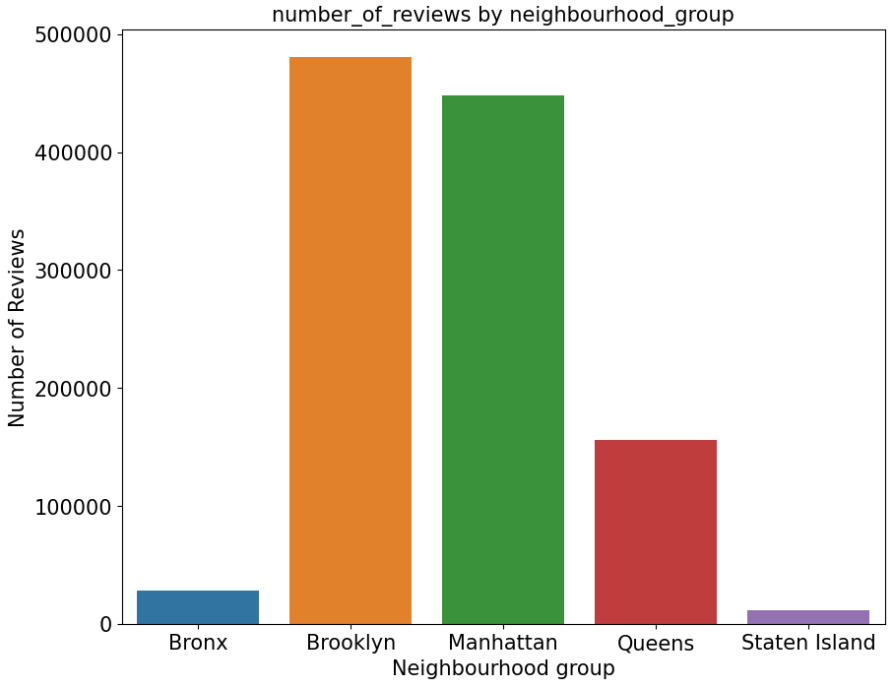


4 Profitability Analysis

4.3 Which neighbour hood group is most reviewed

Inference:

Brooklyn and Manhattan have the highest no. of reviews.



Conclusion

Conclusion

1. Entire home/apt is highly expensive.
2. Manhattan living cost is highest, Bronx living cost is lowest.
3. Cheapest neighbourhood is Bulls head.
4. The Cheapest listing is the Bronx apart.
5. Manhattan have the highest no. of listing.
6. In Manhattan entire home is mostly preferred but in Brooklyn ratio between the entire home and private room is 50:50.
7. Private room has the highest availability; the Entire home has least availability.
8. Revenue generated by Entire home is highest irrespective of neighbourhood group.
9. Sonder have a maximum property in New York.

Q & A