Capstone Project
EDA on Airbnb

By - Mohd Sahil

EDA on Airbnb Dataset

1. Problem Statement
2. Data Summary
3. Data Cleaning
4. Data Wrangling
5. Visualization
6. Conclusions
7. Challenges

EDA Workflow

Business Need

Data Acquire

Data Wrangling

Analyse

Visualisation

Problem Statement

The data generated by users of Airbnb are in millions. Therefore to improve the business the data analysis should focus on both sides of the story, demand (Guests) and supply (Hosts).

- 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?

Data Summary

- The dataset contains 48895 observations and 16 columns

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 48895 entries, 0 to 48894
Data columns (total 16 columns):
 #   Column                          Non-Null Count  Dtype  
---  ------                          --------------  -----  
 0   id                              48895 non-null  int64  
 1   name                            48879 non-null  object 
 2   host_id                         48895 non-null  int64  
 3   host_name                       48874 non-null  object 
 4   neighbourhood_group             48895 non-null  object 
 5   neighbourhood                   48895 non-null  object 
 6   latitude                        48895 non-null  float64
 7   longitude                       48895 non-null  float64
 8   room_type                       48895 non-null  object 
 9   price                           48895 non-null  int64  
 10  minimum_nights                  48895 non-null  int64  
 11  number_of_reviews               48895 non-null  int64  
 12  last_review                     38843 non-null  object 
 13  reviews_per_month               38843 non-null  float64
 14  calculated_host_listings_count  48895 non-null  int64  
 15  availability_365                48895 non-null  int64  
dtypes: float64(3), int64(7), object(6)
memory usage: 6.0+ MB
```

Data Summary

```
[47] df.describe()
```

	host_id	latitude	longitude	price	minimum_nights	number_of_reviews	calculated_host_listings_count	availability_365
count	4.889500e+04	48895.000000	48895.000000	48895.000000	48895.000000	48895.000000	48895.000000	48895.000000
mean	6.762001e+07	40.728949	-73.952170	152.720687	7.029962	23.274466	7.143982	112.781327
std	7.861097e+07	0.054530	0.046157	240.154170	20.510550	44.550582	32.952519	131.622289
min	2.438000e+03	40.499790	-74.244420	0.000000	1.000000	0.000000	1.000000	0.000000
25%	7.822033e+06	40.690100	-73.983070	69.000000	1.000000	1.000000	1.000000	0.000000
50%	3.079382e+07	40.723070	-73.955680	106.000000	3.000000	5.000000	1.000000	45.000000
75%	1.074344e+08	40.763115	-73.936275	175.000000	5.000000	24.000000	2.000000	227.000000
max	2.743213e+08	40.913060	-73.712990	10000.000000	1250.000000	629.000000	327.000000	365.000000

Data Summary

Columns

- id: Unique identification code for the listing
- name: Descriptive name of the listing
- host_id: Unique identification code for the host
- host_name: First name of the host
- neighbourhood_group: Neighbourhoods are grouped into NYC boroughs
- neighbourhood: The name of neighbourhood of the listing

Data Summary

Columns

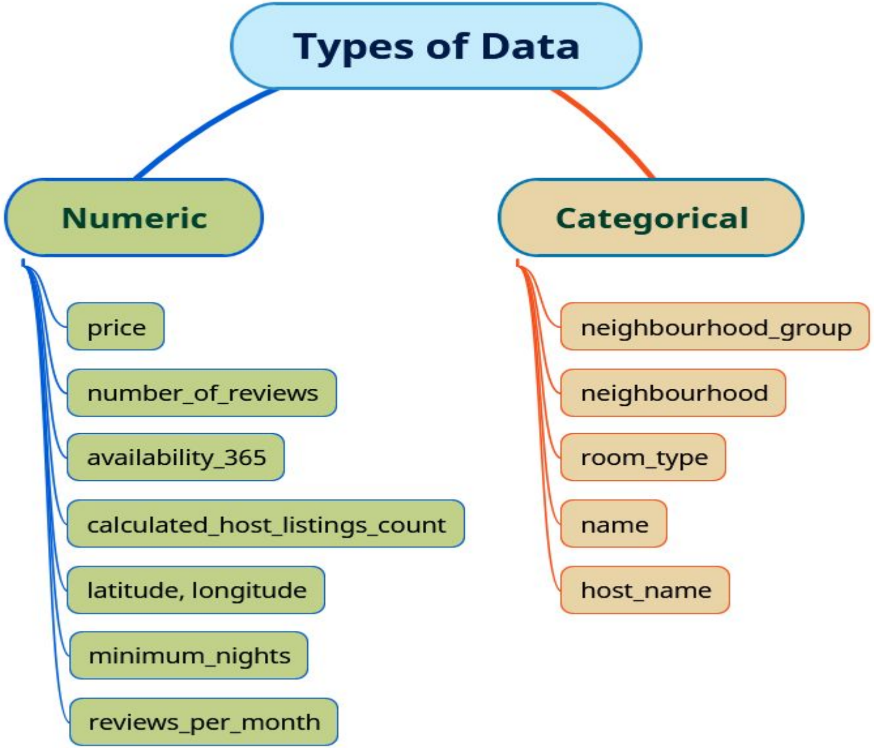
- Latitude & Longitude: Numeric variables that represents the location of the listing
- room_type: A categorical variable including Shared Room, Private Room or Entire Room/Apt
- price: The price of the listing
- minimum_nights: The minimum number of nights the host requires to book their property

Data Summary

Columns

- number_of_reviews: Number of customer reviews regarding the listing
- last_review: Date of the last review
- reviews_per_month: Number of customer reviews per month
- calculated_host_listings_count: Number of listings each host has simultaneously
- availability_365: The number of days that the listing is available in a 365 days, which is pre-defined by the host

Data Summary



Data Cleaning

Replacing Null values

```
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
```

```
#replacing all NaN values
df.fillna({'reviews_per_month':0}, inplace=True)
df.fillna({'last_review':0}, inplace=True)
df.fillna({'host_name':'unknown_host_name'}, inplace=True)
df.fillna({'name':'unknown_name'}, inplace=True)

#examing changes
df.isnull().sum()
```

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

Data Cleaning

Removing unnecessary fields

```
# Removing unnecessary data

df.drop(['id','reviews_per_month','last_review'], axis=1, inplace=True)
```

- Id : Provides unique identification for a listing, similar task can be performed using host_id
- reviews_per_month & last_review : this column has more than 1000 Null values

Data Wrangling

Data Wrangling

What can we learn about different hosts and areas?

```
df_nghGrp_host = df.groupby(['neighbourhood_group'])['host_id'].count()
df_nghGrp_host
```

	neighbourhood_group	Total Airbnb Hosts
0	Bronx	1091
1	Brooklyn	20104
2	Manhattan	21661
3	Queens	5666
4	Staten Island	373

Data Wrangling

What can we learn from predictions? (ex: locations, prices, reviews, etc)

```
df_nghGrp_num_rvws = df.groupby(['neighbourhood_group'])['number_of_reviews'].sum()

df_nghGrp_num_rvws
```

	neighbourhood_group	Total Reviews
0	Brooklyn	486574
1	Manhattan	454569
2	Queens	156950
3	Bronx	28371
4	Staten Island	11541

Data Wrangling

What can we learn from predictions? (ex: locations, prices, reviews, etc)

```
df_nghgrp_price = df.groupby('neighbourhood_group')['price'].mean().

df_nghgrp_price
```

	neighbourhood_group	Mean Price
0	Manhattan	196.875814
1	Brooklyn	124.383207
2	Staten Island	114.812332
3	Queens	99.517649
4	Bronx	87.496792

Data Wrangling

What can we learn from predictions? (ex: locations, prices, reviews, etc)

```
df_price_no_of_rvws = df.groupby(['price'])['number_of_reviews'].mean().
df_price_no_of_rvws
```

	price	Avg Reviews
0	0	34.272727
1	10	14.176471
2	11	37.666667
3	12	2.000000
4	13	9.000000
...	...	...
669	7703	0.000000
670	8000	1.000000
671	8500	2.000000
672	9999	2.333333
673	10000	2.333333

674 rows × 2 columns

Data Wrangling

Which Hosts are the busiest ?

```
df_hosts_rvws = df.groupby(['host_id','host_name'])['number_of_reviews'].mean().
df_hosts_rvws = df_hosts_rvws.to_frame(name ='mean reviews').reset_index()
df_hosts_rvws
```

	host_id	host_name	mean reviews
0	47621202	Dona	602.500000
1	4734398	Jj	599.333333
2	2369681	Carol	540.000000
3	12949460	Asa	488.000000
4	792159	Wanda	480.000000
5	37312959	Maya	454.600000
6	2321321	Lloyd	454.000000
7	277379	Agnes	448.500000
8	307962	Dennis & Naoko	441.000000
9	97086824	Miss Dy	434.000000

Data Wrangling

Why the top hosts are the busiest ?

```
top_hosts_jnd = pd.merge(df_hosts_rvws, df[['host_id','price','neighbourhood_group','room_type']], how="inner")

top_hosts_jnd
```

	host_id	host_name	mean reviews	price	neighbourhood_group	room_type
0	47621202	Dona	602.500000	47	Queens	Private room
1	47621202	Dona	602.500000	47	Queens	Private room
2	4734398	Jj	599.333333	49	Manhattan	Private room
3	4734398	Jj	599.333333	49	Manhattan	Private room
4	4734398	Jj	599.333333	49	Manhattan	Private room
5	2369681	Carol	540.000000	99	Manhattan	Private room
6	12949460	Asa	488.000000	160	Brooklyn	Entire home/apt
7	792159	Wanda	480.000000	60	Brooklyn	Private room
8	37312959	Maya	454.600000	45	Queens	Private room
9	37312959	Maya	454.600000	46	Queens	Private room
10	37312959	Maya	454.600000	45	Queens	Private room
11	37312959	Maya	454.600000	45	Queens	Private room
12	37312959	Maya	454.600000	32	Queens	Private room
13	2321321	Lloyd	454.000000	39	Queens	Shared room
14	277379	Agnes	448.500000	60	Manhattan	Private room
15	277379	Agnes	448.500000	85	Manhattan	Private room
16	307962	Dennis & Naoko	441.000000	99	Queens	Entire home/apt
17	97086824	Miss Dy	434.000000	49	Queens	Entire home/apt

Data Visualization

Is there noticeable difference between traffic among different Areas ?

```
[33] df_nghGrp_num_rvws
```

	neighbourhood_group	Total Reviews
0	Brooklyn	486574
1	Manhattan	454569
2	Queens	156950
3	Bronx	28371
4	Staten Island	11541

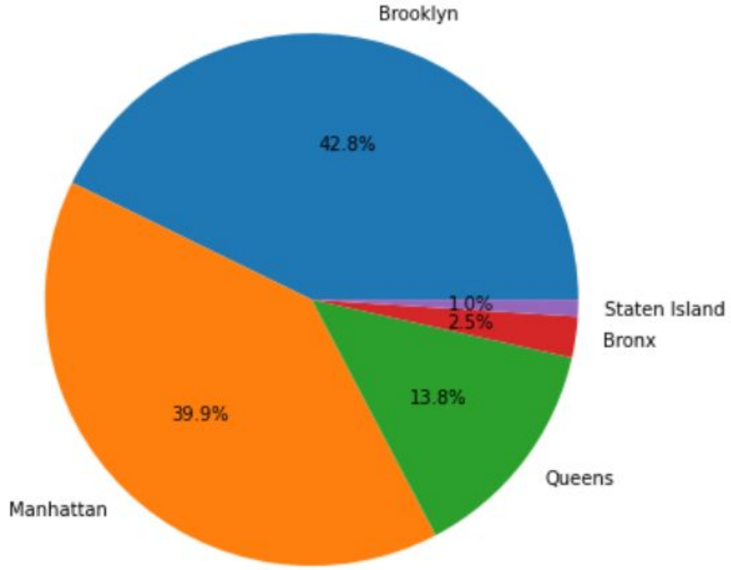
Data Visualization

Brooklyn & Manhattan share 82%(approx) of the traffic

Staten Is & Bronx are the least popular among the Guests

```
rvws = df_nghGrp_num_rvws['Total Reviews']
ngh_grp = df_nghGrp_num_rvws['neighbourhood_group']

plt.figure(figsize=(10,6))
plt.pie(rvws, labels = ngh_grp, autopct='%1.1f%%')
plt.axis('equal')
plt.show()
```



Data Visualization

Reasons for the variation in traffic among difference areas

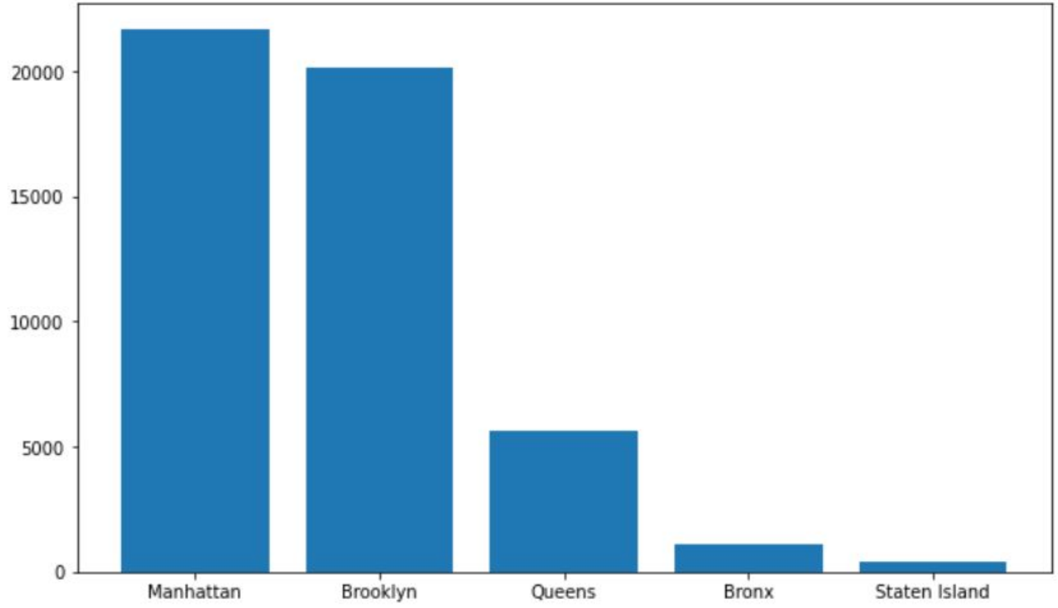
- Number of Hosts

```
df_nghGrp_host = df.groupby(['neighbourhood_group'])['host_id'].count()
df_nghGrp_host
```

	neighbourhood_group	Total Airbnb Hosts
0	Bronx	1091
1	Brooklyn	20104
2	Manhattan	21661
3	Queens	5666
4	Staten Island	373

Data Visualization

```
# Visualization
plt.figure(figsize=(10,6))
plt.bar(df_nghGrp_host['neighbourhood_group'], height = df_nghGrp_host['Total Airbnb Hosts'],width = 0.8)
plt.show()
```



Data Visualization

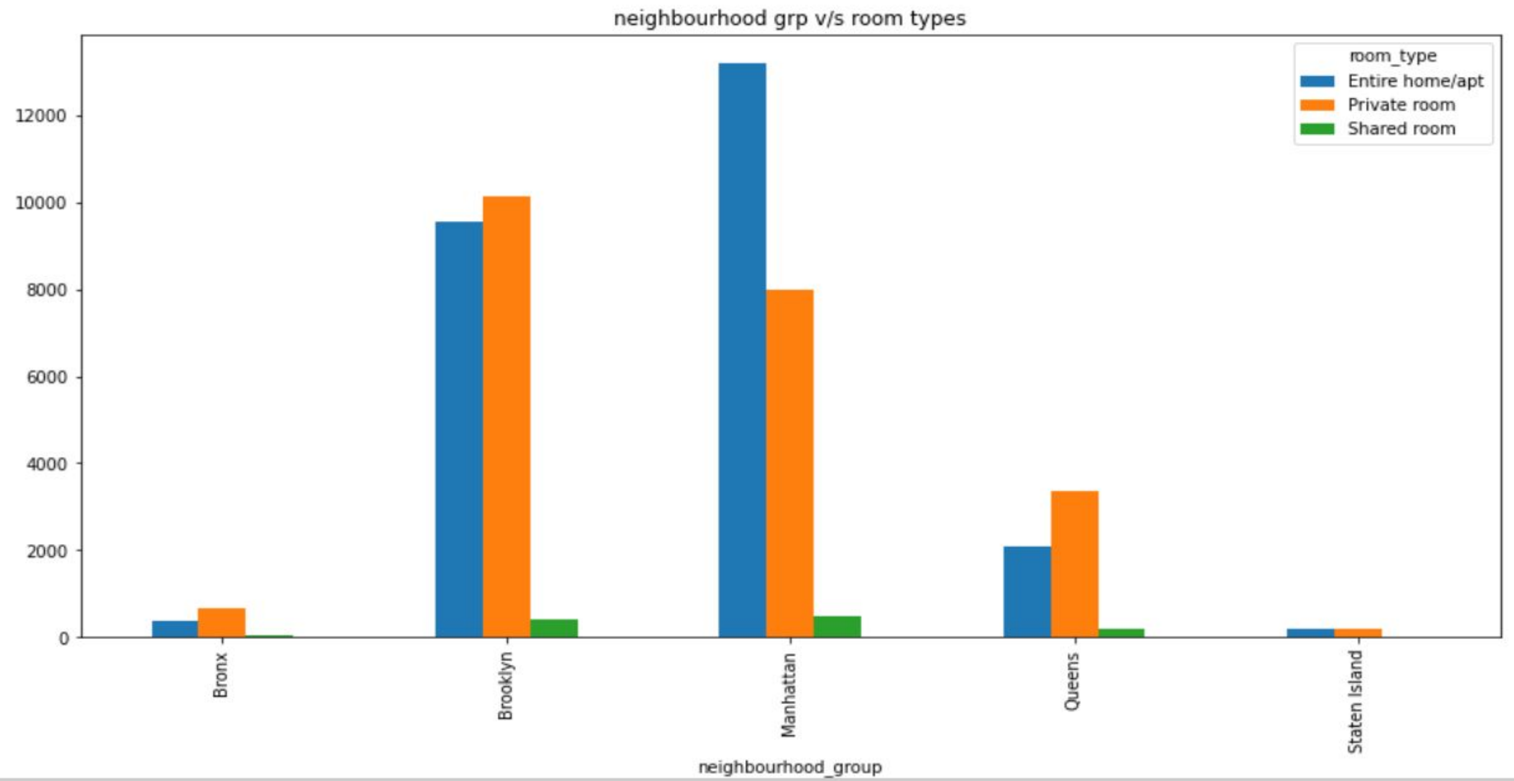
Reasons for the variation in traffic among difference areas

- Types of Room

[38]

	neighbourhood_group	room_type	count
0	Manhattan	Entire home/apt	13199
1	Brooklyn	Private room	10132
2	Brooklyn	Entire home/apt	9559
3	Manhattan	Private room	7982
4	Queens	Private room	3372
5	Queens	Entire home/apt	2096
6	Bronx	Private room	652
7	Manhattan	Shared room	480
8	Brooklyn	Shared room	413
9	Bronx	Entire home/apt	379
10	Queens	Shared room	198
11	Staten Island	Private room	188
12	Staten Island	Entire home/apt	176
13	Bronx	Shared room	60
14	Staten Island	Shared room	9

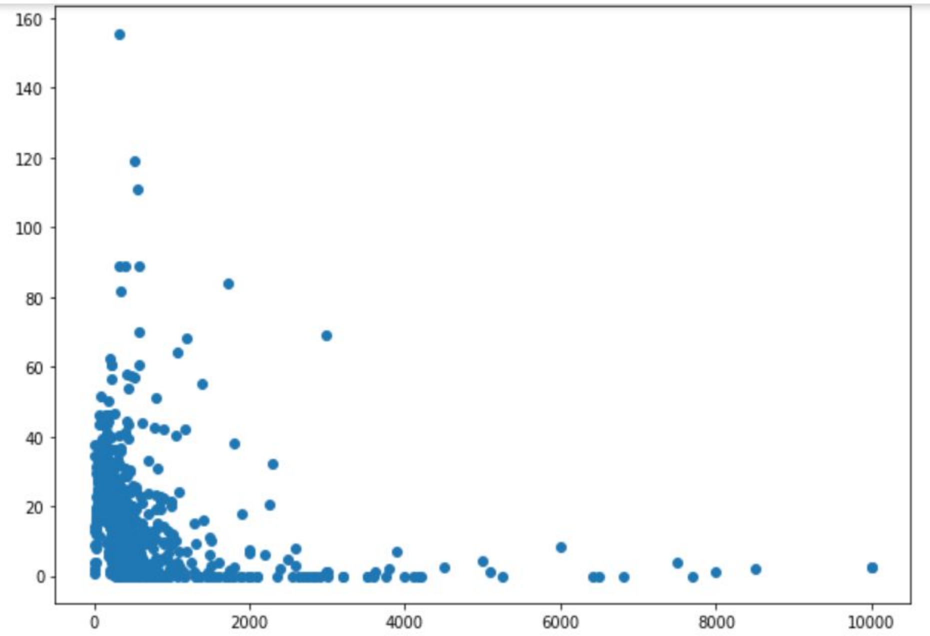
Data Visualization



Data Visualization

Relationship between price and number of reviews

	price	Avg Reviews
0	0	34.272727
1	10	14.176471
2	11	37.666667
3	12	2.000000
4	13	9.000000
...	...	...
669	7703	0.000000
670	8000	1.000000
671	8500	2.000000
672	9999	2.333333
673	10000	2.333333



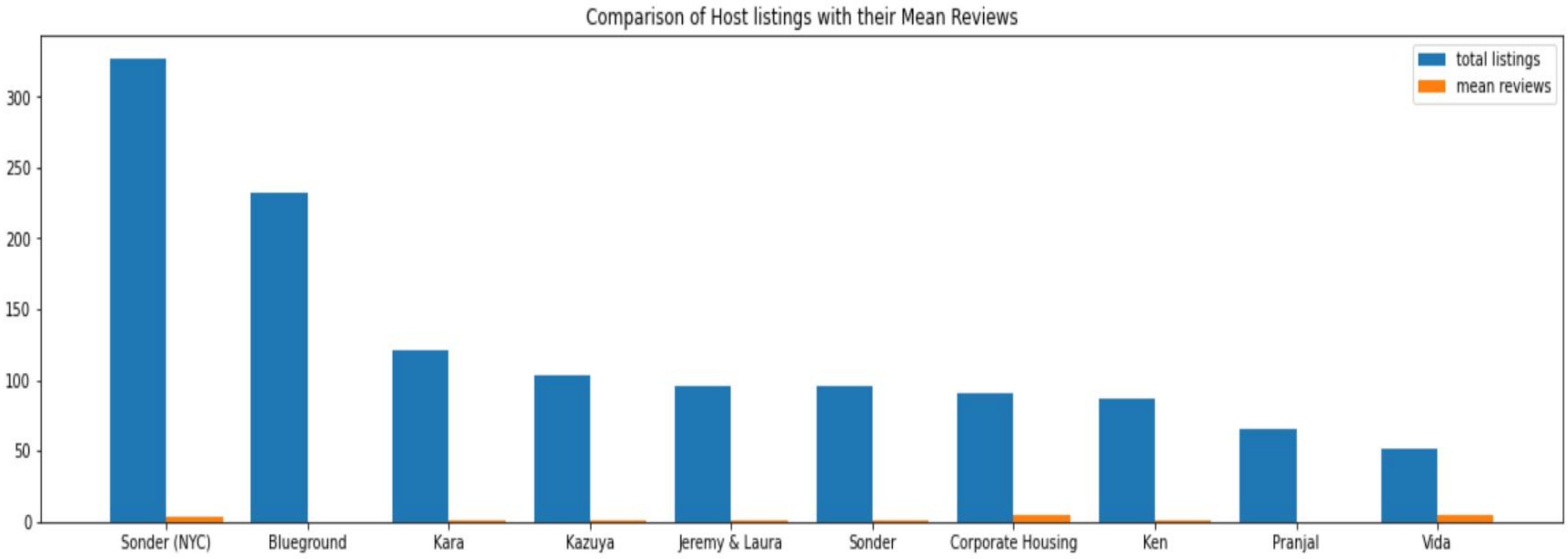
Data Visualization

For the top host category, more Airbnb listings are not resulting in more reviews

	host_id	host_name	total listings	mean reviews
0	219517861	Sonder (NYC)	327	3.917431
1	107434423	Blueground	232	0.125000
2	30283594	Kara	121	0.537190
3	137358866	Kazuya	103	0.844660
4	16098958	Jeremy & Laura	96	1.437500
5	12243051	Sonder	96	0.447917
6	61391963	Corporate Housing	91	4.582418
7	22541573	Ken	87	0.632184
8	200380610	Pranjal	65	0.015385
9	7503643	Vida	52	4.653846

Data Visualization

More number of Airbnb listings not resulting into more popularity among guests



Conclusion

1. Number of Hosts available in a location affects the traffic - Areas where number of hosts are more have higher reviews
2. Majority of Guests prefer to pay a lesser price.
3. Types of Room offered affects the traffic (shared room type is the least popular among guests, whereas Private Room is preferred by more than half of the total Guests).
4. Areas where the availability of private rooms and entire home/ apartment are maximum, the traffic is more.
5. For the top host category, more Airbnb listings are not resulting in more reviews.

Challenges

- For better data exploration, additional features would be quite helpful, such as positive and negative numeric (0-5 stars) reviews for each listing. This can help us to determine the best-reviewed hosts for NYC along with 'number_of_review' column that is provided.
- Missing values in certain columns like reviews per month and last review, hinders in data analysis process.
- Dropping unnecessary fields without compromising on data insights.

Thank You