Storytelling Case Study: Airbnb in NYC PPT-1

Presentation by: Swapnil Srivastava

Objective:

- To Conduct a thorough analysis of New York Airbnb Dataset
- Ask effective questions that can lead to data insights
- Process, analyse and share findings by data visualisation and statistical techniques

Background

- For the past few months, Airbnb has seen a major decline in revenue.
- Now that the restrictions have started lifting and people have started to travel more, Airbnb wants to make sure that it is fully prepared for this change.
- So, analysis has been done on a dataset consisting of various Airbnb listings in New York.

Data Preparation

- In the first phase the data captured and loaded into various environment.
- Once data is cleaned, EDA is done and new features are created.
- Then Meaningful insights are derived using various analytical methods.

Importing libraries and reading the data

In [1]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

In [65]:
```
airbnb = pd.read_csv('AB_NYC_2019.csv')
airbnb.head(10)
```

Out[65]:

	id	name	host_id	host_name	neighbourhood_group	neighbourhood	latitude	longitude	room_type	price	minimum_nights	number_of_revie
0	2539	Clean & quiet apt home by the park	2787	John	Brooklyn	Kensington	40.64749	-73.97237	Private room	149	1	
1	2595	Skylit Midtown Castle	2845	Jennifer	Manhattan	Midtown	40.75362	-73.98377	Entire home/apt	225	1	
2	3647	THE VILLAGE OF HARLEM....NEW YORK !	4632	Elisabeth	Manhattan	Harlem	40.80902	-73.94190	Private room	150	3	
3	3831	Cozy Entire Floor of Brownstone	4869	LisaRoxanne	Brooklyn	Clinton Hill	40.68514	-73.95976	Entire home/apt	89	1	
4	5022	Entire Apt: Spacious Studio/Loft by central park	7192	Laura	Manhattan	East Harlem	40.79851	-73.94399	Entire home/apt	80	10	
5	5099	Large Cozy 1 BR Apartment In Midtown East	7322	Chris	Manhattan	Murray Hill	40.74767	-73.97500	Entire home/apt	200	3	
6	5121	BlissArtsSpace!	7356	Garon	Brooklyn	Bedford-Stuyvesant	40.68688	-73.95596	Private room	60	45	
7	5178	Large Furnished Room Near B'way	8967	Shunichi	Manhattan	Hell's Kitchen	40.76489	-73.98493	Private room	79	2	
8	5203	Cozy Clean Guest Room - Family Apt	7490	MaryEllen	Manhattan	Upper West Side	40.80178	-73.96723	Private room	79	2	

Removing unimportant columns and replacing null values

Certain columns that are not efficient to the dataset can be removed

In [70]:
```
airbnb.drop(['last_review'], axis = 1, inplace = True)
```

In [71]:
```
airbnb.head()
```

Out[71]:

	id	name	host_id	host_name	neighbourhood_group	neighbourhood	latitude	longitude	room_type	price	minimum_nights	number_of_revie
0	2539	Clean & quiet apt home by the park	2787	John	Brooklyn	Kensington	40.64749	-73.97237	Private room	149	1	
1	2595	Skylit Midtown Castle	2845	Jennifer	Manhattan	Midtown	40.75362	-73.98377	Entire home/apt	225	1	
2	3647	THE VILLAGE OF HARLEM....NEW YORK !	4632	Elisabeth	Manhattan	Harlem	40.80902	-73.94190	Private room	150	3	
3	3831	Cozy Entire Floor of Brownstone	4869	LisaRoxanne	Brooklyn	Clinton Hill	40.68514	-73.95976	Entire home/apt	89	1	
4	5022	Entire Apt: Spacious Studio/Loft by central park	7192	Laura	Manhattan	East Harlem	40.79851	-73.94399	Entire home/apt	80	10	

In [72]:
```
# Now reviews per month contains lot of missing values which should be replaced with 0 respectively
airbnb.fillna({'reviews_per_month':0},inplace=True)
```

Creating features

categorizing the "availability_365" column into 5 categories

```
def availability_365_categories_function(row):
    """
    Categorizes the "availability_365" column into 5 categories
    """
    if row <= 1:
        return 'very Low'
    elif row <= 100:
        return 'Low'
    elif row <= 200 :
        return 'Medium'
    elif (row <= 300):
        return 'High'
    else:
        return 'very High'
```

categorizing the "minimum_nights" column into 5 categories

```
def minimum_nights_function(row):
    """
    Categorizes the "minimum_nights" column into 5 categories
    """
    if row <= 1:
        return 'very Low'
    elif row <= 3:
        return 'Low'
    elif row <= 5 :
        return 'Medium'
    elif (row <= 7):
        return 'High'
    else:
        return 'very High'
```

categorizing the "number_of_reviews" column into 5 categories

```
def number_of_reviews_categories_function(row):
    """
    Categorizes the "number_of_reviews" column into 5 categories
    """
    if row <= 1:
        return 'very Low'
    elif row <= 5:
        return 'Low'
    elif row <= 10 :
        return 'Medium'
    elif (row <= 30):
        return 'High'
    else:
        return 'very High'
```

categorizing the "price" column into 5 categories

```
def price_categories_function(row):
    """
    Categorizes the "price" column into 5 categories
    """
    if row <= 50:
        return 'very Low'
    elif row <= 125:
        return 'Low'
    elif row <= 250 :
        return 'Medium'
    elif (row <= 500):
        return 'High'
    else:
        return 'very High'
```

Data Types

4.1 Categorical

```
inp0.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', 'availability_365_categories',
       'minimum_night_categories', 'number_of_reviews_categories',
       'price_categories'],
      dtype='object')
```

```
# Categorical nominal
categorical_columns = inp0.columns[[0,1,3,4,5,8,16,17,18,19]]
categorical_columns
```

```
Index(['id', 'name', 'host_name', 'neighbourhood_group', 'neighbourhood',
       'room_type', 'availability_365_categories', 'minimum_night_categories',
       'number_of_reviews_categories', 'price_categories'],
      dtype='object')
```

4.2 Numerical

```
numerical_columns = inp0.columns[[9,10,11,13,14,15]]
numerical_columns
```

```
Index(['price', 'minimum_nights', 'number_of_reviews', 'reviews_per_month',
       'calculated_host_listings_count', 'availability_365'],
      dtype='object')
```

```
inp0[numerical_columns].describe()
```

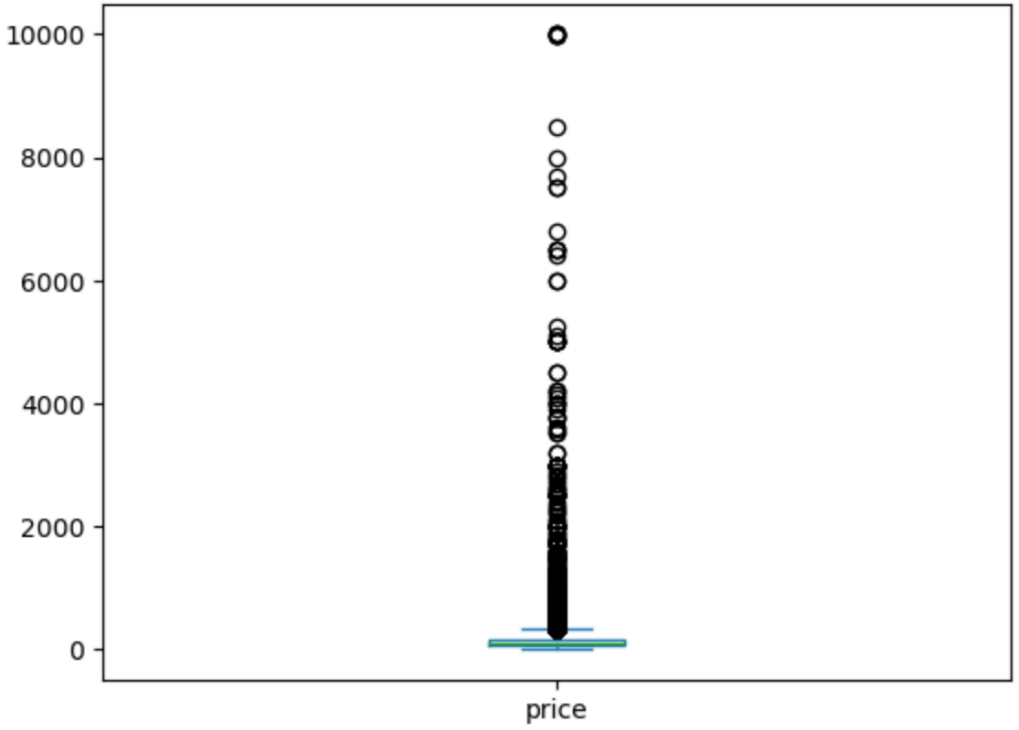
	price	minimum_nights	number_of_reviews	reviews_per_month	calculated_host_listings_count	availability_365
count	48895.000000	48895.000000	48895.000000	38843.000000	48895.000000	48895.000000
mean	152.720687	7.029962	23.274466	1.373221	7.143982	112.781327
std	240.154170	20.510550	44.550582	1.680442	32.952519	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

Analysis

Airbnb Price Range:

Most listings are under price of 5000

```
In [90]: airbnb.price.plot.box()
Out[90]: <AxesSubplot:>
```

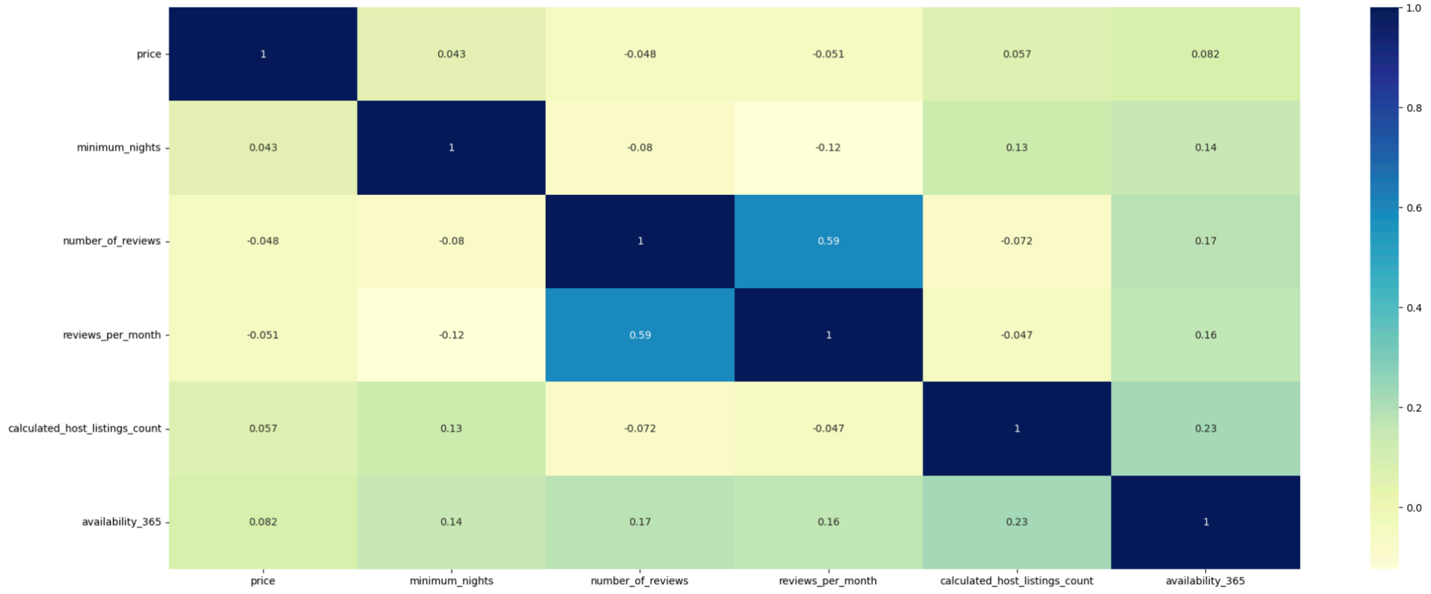


Understanding correlation between numeric columns

- High correlation between number_of_reviews and reviews_per_month.
- Significant Correlation observed between other numerical variables.
- Negative correlation observed between minimum_nights and number_of_reviews.

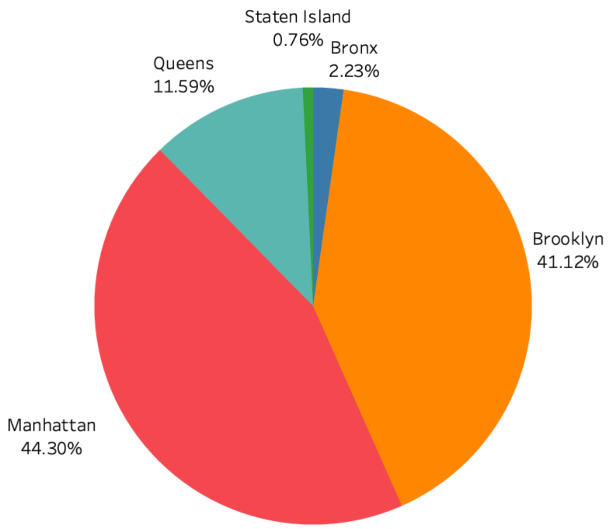
```
In [93]: cor=airbnb[['price','minimum_nights','number_of_reviews','reviews_per_month','calculated_host_listings_count','avail

         plt.figure(figsize=(25,10))
         sns.heatmap(cor, cmap="YlGnBu", annot = True)
         plt.show()
```



NEIGHBOURHOODS WITH MOST AIRBNB LISTINGS

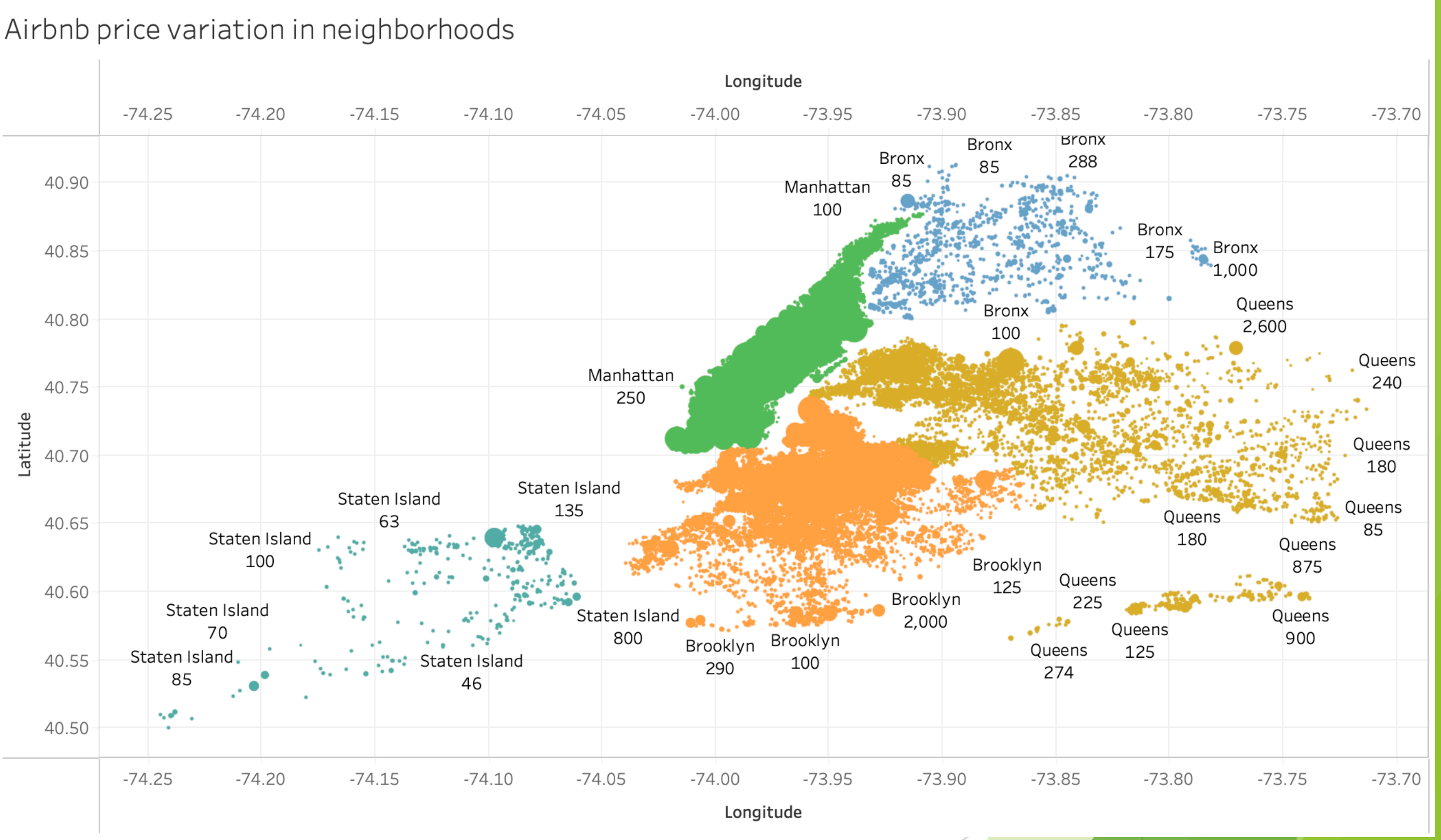
Neighborhoods Groups participation chart



- 85% of the listings are Manhattan and Brooklyn neighbourhood groups
- Staten Island has the lowest contribution of less than 1%.

Price variation in different Neighbourhoods

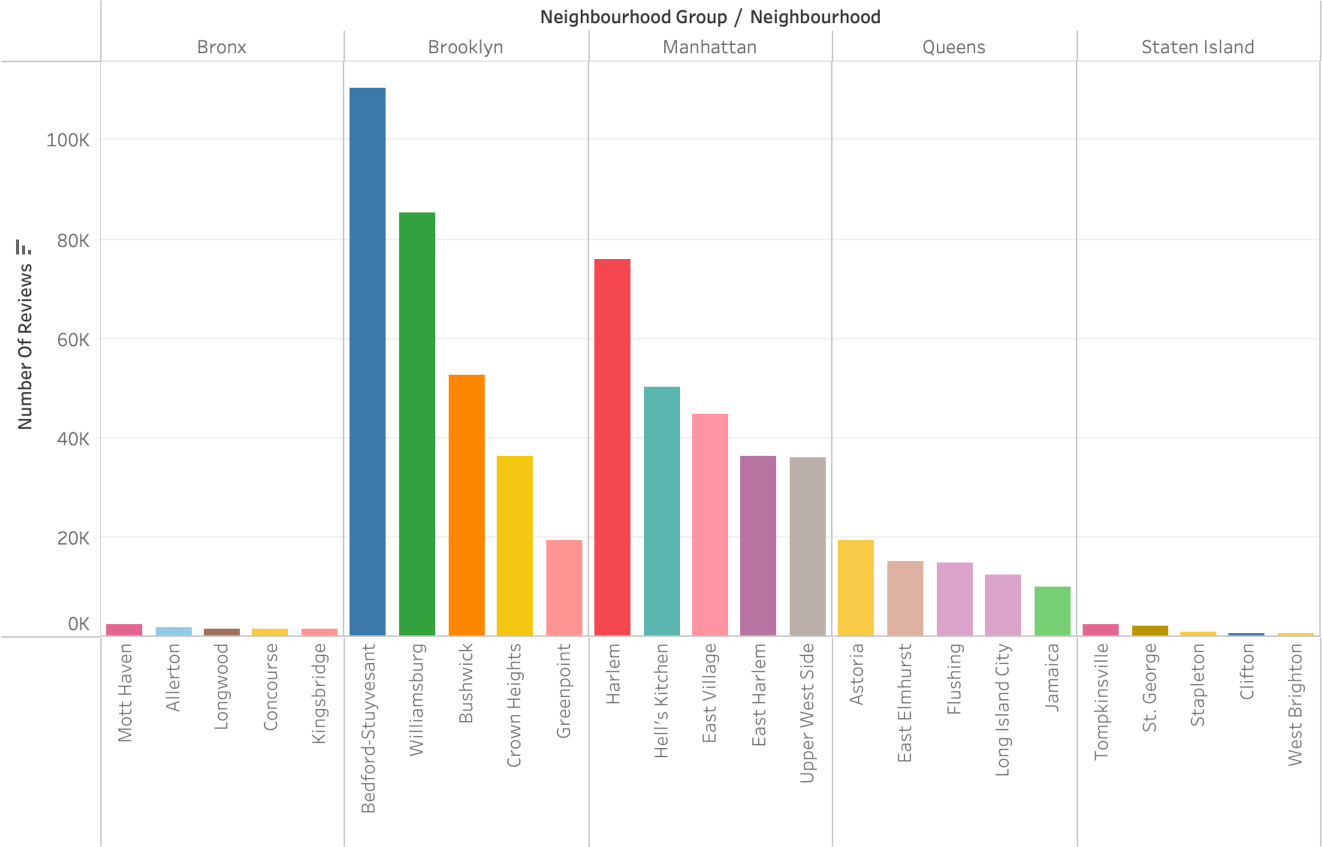
- We see that, Airbnb has high prices in Manhattan, Brooklyn & Queens.
- Prices are highest in Manhattan & Brooklyn owing to the high population density and it being the financial and tourism hubs of NYC. Staten Island has the least prices, due to its low population density and very few tourism destinations.



Popular Neighbourhoods

- We see that Bedford-Stuyvesant from Brooklyn is the highest popular with over 100K no of reviews in total followed by Williamsburg with over 80K reviews.
- Harlem from Manhattan got the highest no of reviews followed by Hell’s kitchen.
- The higher number of customer reviews imply higher satisfaction in these localities.

Top 5 most popular neighborhoods in each Borough



CONCLUSION

- Strong significant insights are derived based on various attributes in the dataset.
- Data collection team should collect data about review scores so that it can strengthen the later analysis.
- Based on the insights, a clustering machine learning model can be made to identify groups of similar objects in datasets with two or more variable quantities.
- Brooklyn and Manhattan emerged to be the boroughs with highest number of listings and have higher prices than the others, owing to the high population density and it being the financial and tourism hubs of NYC. This makes them suitable for business in Airbnb market.

APPENDIX -DATA SOURCES

Column	Description
id	listing ID
name	name of the listing
host_id	host ID
host_name	name of the host
neighbourhood_group	location
neighbourhood	area
latitude	latitude coordinates
longitude	longitude coordinates
room_type	listing space type
price	
minimum_nights	amount of nights minimum
number_of_reviews	number of reviews
last_review	latest review
reviews_per_month	number of reviews per month
calculated_host_listings_count	amount of listing per host
availability_365	number of days when listing is available for booking

APPENDIX -DATA METHODOLOGY

- Conducted a thorough analysis of New York Airbnbs Dataset.
- Cleaned the data set using python.
- Derived the necessary features.
- Used group aggregation, pivot table and other statistical methods.
- Created charts and visualisations using Tableau for generating insights.

APPENDIX -DATA ASSUMPTIONS

```
Categorical Variables:
    - room_type
    - neighbourhood_group
    - neighbourhood

Continous Variables(Numerical):
    - Price
    - minimum_nights
    - number_of_reviews
    - reviews_per_month
    - calculated_host_listings_count
    - availability_365
- Continous Variables could be binned in to groups too

Location Varibles:
    - latitude
    - longitude

Time Varibale:
    - last_review
```