Methodology Document – Airbnb Case study for 2019 data NYC

Business Understanding:

Airbnb is a marketplace where people share private spaces all over the world to be rented for the short-term. The spaces are mostly apartments, villas, houses or castles for a unique travel experience available at various prices. The present case study aims to help Airbnb prepare in the best way for the post-pandemic changes and help increase their revenue. To analyze key findings based on customer preference for business growth, expansion and strengthen the business foundation for upcoming endeavors.

Data Cleaning using Python:

Importing warnings and libraries:

```
import warnings
warnings.filterwarnings("ignore")

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

Reading the CSV data file:

```
df = pd.read_csv('AB_NYC_2019.csv')
df
```

	id	name	host_id	host_name	neighbourhood_group	neighbourhood	latitude	longitude	room_type	price	minimum_nights	numb
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	

Checking data types and other info:

```
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
```

```
df.describe()
```

	id	host_id	latitude	longitude	price	minimum_nights	number_of_reviews	reviews_per_month	calculated_host_listings
count	4.889500e+04	4.889500e+04	48895.000000	48895.000000	48895.000000	48895.000000	48895.000000	38843.000000	48895.
mean	1.901714e+07	6.762001e+07	40.728949	-73.952170	152.720687	7.029962	23.274466	1.373221	7.
std	1.098311e+07	7.861097e+07	0.054530	0.046157	240.154170	20.510550	44.550582	1.680442	32.
min	2.539000e+03	2.438000e+03	40.499790	-74.244420	0.000000	1.000000	0.000000	0.010000	1.
25%	9.471945e+06	7.822033e+06	40.690100	-73.983070	69.000000	1.000000	1.000000	0.190000	1.
50%	1.967728e+07	3.079382e+07	40.723070	-73.955680	106.000000	3.000000	5.000000	0.720000	1.
75%	2.915218e+07	1.074344e+08	40.763115	-73.936275	175.000000	5.000000	24.000000	2.020000	2.
max	3.648724e+07	2.743213e+08	40.913060	-73.712990	10000.000000	1250.000000	629.000000	58.500000	327.

Checking for 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
```

Checking percentage of major null values and dropping the column ‘last_review’:

```
(10052/48895)*100
```

```
20.55833929849678
```

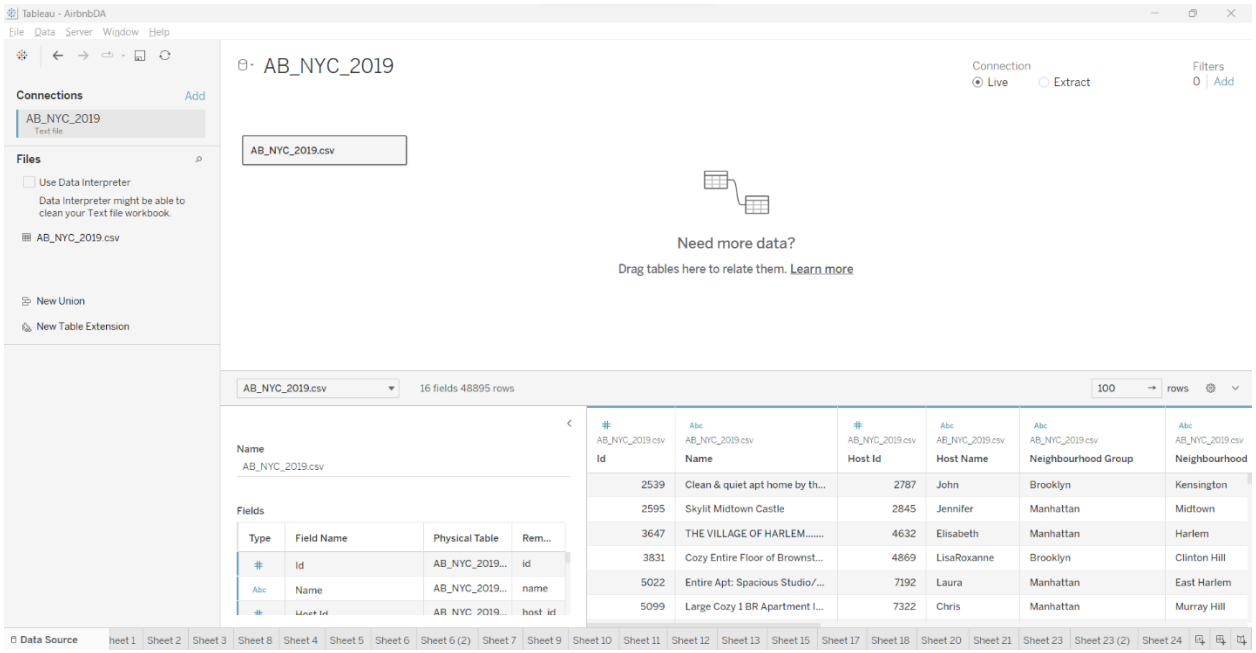
```
df1 = df.drop(columns=['last_review'])
```

```
df1
```

host_id	host_name	neighbourhood_group	neighbourhood	latitude	longitude	room_type	price	minimum_nights	number_of_reviews	reviews_per_month	cal
2787	John	Brooklyn	Kensington	40.64749	-73.97237	Private room	149	1	9	0.21	
2845	Jennifer	Manhattan	Midtown	40.75362	-73.98377	Entire home/apt	225	1	45	0.38	
4632	Elisabeth	Manhattan	Harlem	40.80902	-73.94190	Private room	150	3	0	NaN	

Data Visualisation using Tableau:

- Based on the data csv file various visualisations are created with different parameters such as neighbourhood, neighbourhood groups, reviews per month, total reviews, price, host name and number of listings and their availability of days for booking.



Inferences drawn from analysis and visualisation:

- Entire home/Apartment types of rooms in the area Manhattan and Brooklyn are the most reviewed types of listings, consistently per month as well as overall to go for by the customers, followed by Private room type in the same area.
- Similar inferences drawn for availability of booking.

- Highest average of prices has been in neighbourhoods of Staten island, which is also the lowest reviewed neighbourhood of all.
- Williamsburg and midtown have the highest prices for listings, Williamsburg being in the Brooklyn region and midtown in Manhattan. Although specifically the neighbourhoods of Bedford-Stuyvesant and Williamsburg show the highest prices overall which come under Brooklyn area.
- Highest average minimum nights spent for a booking has shown regions in Manhattan.
- The analysis of price vs number of reviews shows that higher amounts of reviews have been given to the lower price ranges, which shows customer preferences.
- The neighbourhood with the highest calculated host listings is shown as financial districts.
- The highest reviewed listing host names have less than average availability, simultaneously highest reviewed listings names are of the lowest prices although some of the highest pricey neighbourhoods still show above average reviews.

Tools used for Data cleaning: Python

Tools used for Data Visualisation: Tableau