Import Libaries

In [1]:
```
import numpy as np
import pandas as pd
```

Import Dataset

In [2]:
```
airbnb = pd.read_csv('Dataset/Airbnb_Open_Data.csv', low_memory=False)
```

Basic Information

In [3]:
```
pd.set_option('display.max_columns', None)
airbnb.head(5)
```

Out[3]:

	id	NAME	host id	host_identity_verified	host name	neighbourhood group	neighbourhood	lat	long	country	country cod
0	1001254	Clean & quiet apt home by the park	80014485718	unconfirmed	Madaline	Brooklyn	Kensington	40.64749	-73.97237	United States	US
1	1002102	Skylit Midtown Castle	52335172823	verified	Jenna	Manhattan	Midtown	40.75362	-73.98377	United States	US
2	1002403	THE VILLAGE OF HARLEM....NEW YORK !	78829239556	NaN	Elise	Manhattan	Harlem	40.80902	-73.94190	United States	US
3	1002755	NaN	85098326012	unconfirmed	Garry	Brooklyn	Clinton Hill	40.68514	-73.95976	United States	US
4	1003689	Entire Apt: Spacious Studio/Loft by central park	92037596077	verified	Lyndon	Manhattan	East Harlem	40.79851	-73.94399	United States	US

In [4]:
```
airbnb.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 102599 entries, 0 to 102598
Data columns (total 26 columns):
 #   Column                          Non-Null Count   Dtype  
---  ------                          --------------   -----  
 0   id                              102599 non-null  int64  
 1   NAME                            102349 non-null  object 
 2   host id                         102599 non-null  int64  
 3   host_identity_verified          102310 non-null  object 
 4   host name                       102193 non-null  object 
 5   neighbourhood group             102570 non-null  object 
 6   neighbourhood                   102583 non-null  object 
 7   lat                             102591 non-null  float64
 8   long                            102591 non-null  float64
 9   country                         102067 non-null  object 
 10  country code                    102468 non-null  object 
 11  instant_bookable                102494 non-null  object 
 12  cancellation_policy             102523 non-null  object 
 13  room type                       102599 non-null  object 
 14  Construction year               102385 non-null  float64
 15  price                           102352 non-null  object 
 16  service fee                     102326 non-null  object 
 17  minimum nights                  102190 non-null  float64
 18  number of reviews               102416 non-null  float64
 19  last review                     86706 non-null   object 
 20  reviews per month               86720 non-null   float64
 21  review rate number              102273 non-null  float64
 22  calculated host listings count  102280 non-null  float64
 23  availability 365                102151 non-null  float64
 24  house_rules                     50468 non-null   object 
```

```
 25  license                         2 non-null       object
dtypes: float64(9), int64(2), object(15)
memory usage: 20.4+ MB
```

Null Value

In [5]:
```
airbnb.isnull().sum()
```

Out[5]:
```
id                                0
NAME                            250
host id                           0
host_identity_verified          289
host name                       406
neighbourhood group              29
neighbourhood                    16
lat                               8
long                              8
country                         532
country code                    131
instant_bookable                105
cancellation_policy              76
room type                         0
Construction year               214
price                           247
service fee                     273
minimum nights                  409
number of reviews               183
last review                   15893
reviews per month             15879
review rate number              326
calculated host listings count  319
availability 365                448
house_rules                   52131
license                      102597
dtype: int64
```

Drop Columns

In [6]:
```
airbnb = airbnb.drop(['id','NAME','host id','host name','lat','long','country','Construction year','reviews per m
                      'calculated host listings count', 'country code', 'house_rules', 'license', 'last review'],
airbnb.isnull().sum()
```

Out[6]:
```
host_identity_verified    289
neighbourhood group        29
neighbourhood              16
instant_bookable          105
cancellation_policy        76
room type                   0
price                     247
service fee               273
minimum nights            409
number of reviews         183
review rate number        326
availability 365          448
dtype: int64
```

Data Preparation

Host Indentity Verified

In [7]:
```
from collections import Counter
```

In [8]:
```
print(*Counter(airbnb.host_identity_verified))
```

```
unconfirmed verified nan
```

In [9]:
```
airbnb.host_identity_verified = airbnb.host_identity_verified.fillna('unconfirmed')
airbnb.head(5)
```

Out[9]:

	host_identity_verified	neighbourhood group	neighbourhood	instant_bookable	cancellation_policy	room type	price	service fee	minimum nights	number of reviews	rev
											num
0	unconfirmed	Brooklyn	Kensington	False	strict	Private room	$966	$193	10.0	9.0	
1	verified	Manhattan	Midtown	False	moderate	Entire home/apt	$142	$28	30.0	45.0	
2	unconfirmed	Manhattan	Harlem	True	flexible	Private room	$620	$124	3.0	0.0	
3	unconfirmed	Brooklyn	Clinton Hill	True	moderate	Entire home/apt	$368	$74	30.0	270.0	
4	verified	Manhattan	East Harlem	False	moderate	Entire home/apt	$204	$41	10.0	9.0	

Neighbourhood and Neighbourhood Group

In [10]:
```
print(*Counter(airbnb['neighbourhood group']))
```

```
Brooklyn Manhattan brookln manhatan Queens nan Staten Island Bronx
```

In [11]:
```
airbnb['neighbourhood group'] = airbnb['neighbourhood group'].replace('manhatan', 'Manhattan')
airbnb['neighbourhood group'] = airbnb['neighbourhood group'].replace('brookln', 'Brooklyn')
```

In [12]:
```
airbnb = airbnb.dropna(subset=['neighbourhood group', 'neighbourhood'])
```

In [13]:
```
airbnb.isnull().sum()
```

Out[13]:
```
host_identity_verified      0
neighbourhood group         0
neighbourhood               0
instant_bookable          102
cancellation_policy        73
room type                   0
price                     245
service fee               273
minimum nights            407
number of reviews         183
review rate number        324
availability 365          436
dtype: int64
```

Instant Bookable

In [14]:
```
print(*Counter(airbnb['instant_bookable']))
```

```
False True nan
```

In [15]:
```
airbnb['instant_bookable'] = airbnb['instant_bookable'].fillna('False')
```

In [16]:
```
airbnb.isnull().sum()
```

Out[16]:
```
host_identity_verified      0
neighbourhood group         0
neighbourhood               0
instant_bookable            0
cancellation_policy        73
room type                   0
price                     245
service fee               273
minimum nights            407
```

```
number of reviews         183
review rate number        324
availability 365          436
dtype: int64
```

Cancellation Policy

In [17]:
```
airbnb['cancellation_policy'].unique()
```

Out[17]:
```
array(['strict', 'moderate', 'flexible', nan], dtype=object)
```

In [18]:
```
airbnb['cancellation_policy'].describe()
```

Out[18]:
```
count       102481
unique           3
top       moderate
freq         34330
Name: cancellation_policy, dtype: object
```

Since moderate is the top value, the NaN value will be replaced with moderate

In [19]:
```
airbnb['cancellation_policy'] = airbnb['cancellation_policy'].fillna('moderate')
airbnb.isnull().sum()
```

Out[19]:
```
host_identity_verified      0
neighbourhood group         0
neighbourhood               0
instant_bookable            0
cancellation_policy         0
room type                   0
price                     245
service fee               273
minimum nights            407
number of reviews         183
review rate number        324
availability 365          436
dtype: int64
```

Price

In [20]:
```
airbnb['price']
```

Out[20]:
```
0           $966
1           $142
2           $620
3           $368
4           $204
           ...
102594      $844
102595      $837
102596      $988
102597      $546
102598    $1,032
Name: price, Length: 102554, dtype: object
```

In [21]:
```
airbnb['price'].describe()
```

Out[21]:
```
count     102309
unique      1151
top         $206
freq         137
Name: price, dtype: object
```

In [22]:
```
airbnb['price'] = airbnb['price'].replace('[\$,]', '', regex=True).astype(float)
```

In [23]:
```
airbnb['price']
```

Out[23]:
```
0          966.0
1          142.0
2          620.0
3          368.0
4          204.0
           ...
102594     844.0
102595     837.0
102596     988.0
102597     546.0
102598    1032.0
Name: price, Length: 102554, dtype: float64
```

In [24]:
```
airbnb['price'].describe()
```

Out[24]:
```
count    102309.000000
mean        625.269693
std         331.678071
min          50.000000
25%         340.000000
50%         624.000000
75%         913.000000
max        1200.000000
Name: price, dtype: float64
```

In [25]:
```
mean = airbnb['price'].mean()
airbnb['price'] = airbnb['price'].fillna(mean)
airbnb.isnull().sum()
```

Out[25]:
```
host_identity_verified      0
neighbourhood group         0
neighbourhood               0
instant_bookable            0
cancellation_policy         0
room type                   0
price                       0
service fee               273
minimum nights            407
number of reviews         183
review rate number        324
availability 365          436
dtype: int64
```

Service Fee

In [26]:
```
airbnb['service fee']
```

Out[26]:
```
0         $193
1          $28
2         $124
3          $74
4          $41
           ...
102594    $169
102595    $167
102596    $198
102597    $109
102598    $206
Name: service fee, Length: 102554, dtype: object
```

In [27]:
```
airbnb['service fee'] = airbnb['service fee'].replace('[\$,]', '', regex=True).astype(float)
```

In [28]:
```
airbnb['service fee']
```

Out[28]:
```
0         193.0
1          28.0
2         124.0
3          74.0
4          41.0
          ...
102594    169.0
102595    167.0
102596    198.0
102597    109.0
102598    206.0
Name: service fee, Length: 102554, dtype: float64
```

In [29]:
```
airbnb['service fee'].describe()
```

Out[29]:
```
count    102281.000000
mean        125.022135
std          66.327633
min          10.000000
25%          68.000000
50%         125.000000
75%         183.000000
max         240.000000
Name: service fee, dtype: float64
```

In [30]:
```
mean = round(airbnb['service fee'].mean())
airbnb['service fee'] = airbnb['service fee'].fillna(mean)
airbnb.isnull().sum()
```

Out[30]:
```
host_identity_verified      0
neighbourhood group         0
neighbourhood               0
instant_bookable            0
cancellation_policy         0
room type                   0
price                       0
service fee                 0
minimum nights            407
number of reviews         183
review rate number        324
availability 365          436
dtype: int64
```

Minimum Nights

In [31]:
```
airbnb['minimum nights'].describe()
```

Out[31]:
```
count    102147.000000
mean          8.133249
std          30.551705
min       -1223.000000
25%           2.000000
50%           3.000000
75%           5.000000
max        5645.000000
Name: minimum nights, dtype: float64
```

In [32]:
```
airbnb['minimum nights'] = airbnb['minimum nights'].fillna(3)
airbnb.isnull().sum()
```

Out[32]:
```
host_identity_verified      0
neighbourhood group         0
neighbourhood               0
instant_bookable            0
cancellation_policy         0
room type                   0
price                       0
service fee                 0
minimum nights              0
number of reviews         183
```

```
review rate number        324
availability 365          436
dtype: int64
```

Number of Reviews

In [33]:
```
airbnb['number of reviews'].describe()
```

Out[33]:
```
count    102371.000000
mean         27.460072
std          49.470095
min           0.000000
25%           1.000000
50%           7.000000
75%          30.000000
max        1024.000000
Name: number of reviews, dtype: float64
```

In [34]:
```
airbnb['number of reviews'] = airbnb['number of reviews'].fillna(0)
airbnb.isnull().sum()
```

Out[34]:
```
host_identity_verified      0
neighbourhood group         0
neighbourhood               0
instant_bookable            0
cancellation_policy         0
room type                   0
price                       0
service fee                 0
minimum nights              0
number of reviews           0
review rate number        324
availability 365          436
dtype: int64
```

Review Rate Number

In [35]:
```
airbnb['review rate number'].describe()
```

Out[35]:
```
count    102230.000000
mean          3.279096
std           1.284626
min           1.000000
25%           2.000000
50%           3.000000
75%           4.000000
max           5.000000
Name: review rate number, dtype: float64
```

In [36]:
```
mean = round(airbnb['review rate number'].mean())
airbnb['review rate number'] = airbnb['review rate number'].fillna(mean)
airbnb.isnull().sum()
```

Out[36]:
```
host_identity_verified      0
neighbourhood group         0
neighbourhood               0
instant_bookable            0
cancellation_policy         0
room type                   0
price                       0
service fee                 0
minimum nights              0
number of reviews           0
review rate number          0
availability 365          436
dtype: int64
```

Availability 365

In [37]:
```
airbnb['availability 365'].describe()
```

Out[37]:
```
count    102118.000000
mean        141.117139
std         135.434207
min         -10.000000
25%           3.000000
50%          96.000000
75%         269.000000
max        3677.000000
Name: availability 365, dtype: float64
```

In [38]:
```
mean = round(airbnb['availability 365'].mean())
airbnb['availability 365'] = airbnb['availability 365'].fillna(mean)
airbnb['availability 365'] = abs(airbnb['availability 365'])
airbnb.isnull().sum()
```

Out[38]:
```
host_identity_verified    0
neighbourhood group       0
neighbourhood             0
instant_bookable          0
cancellation_policy       0
room type                 0
price                     0
service fee               0
minimum nights            0
number of reviews         0
review rate number        0
availability 365          0
dtype: int64
```

In [39]:
```
airbnb.to_csv('Dataset/Modified Airbnb.csv')
```

End