Capstone Project Creation - Submitted by Faseni Fakoya 18/12/2023

IBM SkillsBuild Europe Delivery - Data Analytics

Pre-requisite

- Understanding of Python, Power BI or Tableau
- Understanding of Data Cleaning
- Understanding Data Visualization

Level of Exercise: Intermediate

Duration: approximately 3 hours

Data Analytics of Airbnb Data:

Objective:

In this exericise, you will be performing Data Analytics on an Open Dataset dataset coming from Airbnb. Some of the tasks include

- Data Cleaning.
- Data Transformation
- Data Visualization.

Overview of Airbnb Data:

People's main criteria when visiting new places are reasonable accommodation and food. Airbnb (Air-Bed-Breakfast) is an online marketplace created to meet this need of people by renting out their homes for a short term. They offer this facility at a relatively lower price than hotels. Further people worldwide prefer the homely and economical service offered by them. They offer services across various geographical locations

Dataset Source

YOu can get the dataset for this assessment using the following link: https://www.kaggle.com/datasets/arianazmoudeh/airbnbopendata

This dataset contains information such as the neighborhood offering these services, room type, price,avaliabilty, reviews, service fee, cancellation policy and rules to use the house. This analysis will help airbnb in improving its services.

So all the best for your Data Analytics Journey on Airbnb data!!!

Task 1: Data Loading (Python)

1. Read the csv file and load it into a pandas dataframe.
2. Display the first five rows of your dataframe.
3. Display the data types of the columns.

In [4]:
```
## Read the csv file
import pandas as pd
import numpy as np
import warnings
warnings.filterwarnings('ignore')
pd.set_option('display.max_columns', None)
data = pd.read_csv("./Airbnb_Open_Data.csv")
```

In [5]:
```
## Display the first 5 rows
data.head()
```

Out[5]:

	id	NAME	host id	host_identity_verified	host name	neighbourhood group	neighbo
0	1001254	Clean & quiet apt home by the park	80014485718	unconfirmed	Madaline	Brooklyn	Ker
1	1002102	Skylit Midtown Castle	52335172823	verified	Jenna	Manhattan	
2	1002403	THE VILLAGE OF HARLEM....NEW YORK !	78829239556	NaN	Elise	Manhattan	
3	1002755	NaN	85098326012	unconfirmed	Garry	Brooklyn	Cli
4	1003689	Entire Apt: Spacious Studio/Loft by central park	92037596077	verified	Lyndon	Manhattan	Eas

In [6]:
```
data.shape
```

Out[6]:
```
(102599, 26)
```

In [7]:
```
## Display the data types
data.dtypes
```

Out[7]:
```
id                                 int64
NAME                              object
host id                            int64
host_identity_verified            object
host name                         object
neighbourhood group               object
neighbourhood                     object
lat                              float64
long                             float64
country                           object
country code                      object
instant_bookable                  object
cancellation_policy               object
room type                         object
Construction year                float64
price                             object
service fee                       object
minimum nights                   float64
number of reviews                float64
last review                       object
reviews per month                float64
review rate number               float64
calculated host listings count   float64
availability 365                 float64
house_rules                       object
license                           object
dtype: object
```

Task 2a: Data Cleaning (Any Tool)

1. Drop some of the unwanted columns. These include `host id`, `id`, `country` and `country code` from the dataset.
2. State the reason for not including these columns for your Data Analytics.

If using Python for this exercise, please include the code in the cells below. If using any other tool, please include screenshoots before and after the elimination of the columns.

In [8]:
```
data.columns
```

Out[8]:
```
Index(['id', 'NAME', 'host id', 'host_identity_verified', 'host name',
       'neighbourhood group', 'neighbourhood', 'lat', 'long', 'country',
       'country code', 'instant_bookable', 'cancellation_policy', 'room type',
       'Construction year', 'price', 'service fee', 'minimum nights',
       'number of reviews', 'last review', 'reviews per month',
       'review rate number', 'calculated host listings count',
       'availability 365', 'house_rules', 'license'],
      dtype='object')
```

In [9]:
```
data.drop(['host id','id','country','country code'],axis=1, inplace=True)
```

In [10]:
```
data.head(500)
```

Out[10]:

	NAME	host_identity_verified	host name	neighbourhood group	neighbourhood	lat	
0	Clean & quiet apt home by the park	unconfirmed	Madaline	Brooklyn	Kensington	40.64749	-7
1	Skylit Midtown Castle	verified	Jenna	Manhattan	Midtown	40.75362	-7
2	THE VILLAGE OF HARLEM....NEW YORK !	NaN	Elise	Manhattan	Harlem	40.80902	-7
3	NaN	unconfirmed	Garry	Brooklyn	Clinton Hill	40.68514	-7
4	Entire Apt: Spacious Studio/Loft by central park	verified	Lyndon	Manhattan	East Harlem	40.79851	-7
...	...	...	...	...	...	...	
495	Elegant 2-BR duplex, Union Square	verified	Albert	Manhattan	Gramercy	40.73476	-7
496	Cozy private family home in Bushwick	unconfirmed	Adison	Brooklyn	Bushwick	40.69055	-7
497	Luxury 2Bed/2.5Bath Central Park View	verified	Walter	Manhattan	Upper West Side	40.77350	-7
498	Cosy Sunny 1brm in Prospect Heights	verified	Ted	Brooklyn	Crown Heights	40.67505	-7
499	East Village bedroom w rooftop	verified	Kate	Manhattan	East Village	40.72974	-7

500 rows × 22 columns

Task 2b: Data Cleaning (Python)

- Check for missing values in the dataframe and display the count in ascending order. **If the values are missing, impute the values as per the datatype of the columns.**
- Check whether there are any duplicate values in the dataframe and, if present, remove them.
- Display the total number of records in the dataframe before and after removing the duplicates.

In [11]:
```
## Check for missing values in the dataframe and display the count in ascending ord
data.isna().sum().sort_values(ascending=True)
```

Out[11]:
```
room type                            0
lat                                  8
long                                 8
neighbourhood                       16
neighbourhood group                 29
cancellation_policy                 76
instant_bookable                   105
number of reviews                  183
Construction year                  214
price                              247
NAME                               250
service fee                        273
host_identity_verified             289
calculated host listings count     319
review rate number                 326
host name                          406
minimum nights                     409
availability 365                   448
reviews per month                15879
last review                      15893
house_rules                      52131
license                         102597
dtype: int64
```

In [12]:
```
#Missing values imputation using the datatype of the columns with missing values
for col in data.columns:
    if data[col].dtype == 'O':
        data[col].fillna(value=data[col].mode()[0], inplace=True)
    else:
        data[col].fillna(value=data[col].median(), inplace=True)
```

In [13]:
```
#Confirmation of the filled values
data.isna().sum()
```

Out[13]:
```
NAME                              0
host_identity_verified            0
host name                         0
neighbourhood group               0
neighbourhood                     0
lat                               0
long                              0
instant_bookable                  0
cancellation_policy               0
room type                         0
Construction year                 0
price                             0
service fee                       0
minimum nights                    0
number of reviews                 0
last review                       0
reviews per month                 0
review rate number                0
calculated host listings count    0
availability 365                  0
house_rules                       0
license                           0
dtype: int64
```

In [14]:
```
#Total records in the original dataframe
data.shape
```

Out[14]:
```
(102599, 22)
```

In [15]:
```
## Check whether there are any duplicate values in the dataframe and if present rem
data.duplicated().sum()
```

Out[15]:
```
3461
```

In [16]:
```
#Remove dupplicate
data.drop_duplicates(inplace=True)
```

In [17]:
```
## Display the total number of records in the dataframe after removing the duplicat
data.shape
```

Out[17]:
```
(99138, 22)
```

In [18]:
```
data.to_csv('eda.sql')
```

Task 3: Data Transformation (Any Tool)

- Rename the column `availability 365` to `days_booked`
- Convert all column names to lowercase and replace the spaces in the column names with an underscore "_".
- Remove the dollar sign and comma from the columns `price` and `service_fee`. If necessary, convert these two columns to the appropriate data type.

If using Python for this exercise, please include the code in the cells below. If using any other tool, please include screenshoots of your work.

In [19]:
```
data.columns
```

Out[19]:
```
Index(['NAME', 'host_identity_verified', 'host name', 'neighbourhood group',
       'neighbourhood', 'lat', 'long', 'instant_bookable',
       'cancellation_policy', 'room type', 'Construction year', 'price',
       'service fee', 'minimum nights', 'number of reviews', 'last review',
       'reviews per month', 'review rate number',
       'calculated host listings count', 'availability 365', 'house_rules',
       'license'],
      dtype='object')
```

In [20]:
```
## Rename the column availability 365 to days_booked
data.rename(columns={'availability 365': 'days_booked'},inplace=True)
```

In [21]:
```
data.columns
```

Out[21]:
```
Index(['NAME', 'host_identity_verified', 'host name', 'neighbourhood group',
       'neighbourhood', 'lat', 'long', 'instant_bookable',
       'cancellation_policy', 'room type', 'Construction year', 'price',
       'service fee', 'minimum nights', 'number of reviews', 'last review',
       'reviews per month', 'review rate number',
       'calculated host listings count', 'days_booked', 'house_rules',
       'license'],
      dtype='object')
```

In [22]:
```
#Convert all column names to lowercase and replace the spaces in the column names w
data.columns = [col.lower().replace(" ", "_") for col in data.columns]
```

In [23]:
```
data.head()
```

Out[23]:

	name	host_identity_verified	host_name	neighbourhood_group	neighbourhood	
0	Clean & quiet apt home by the park	unconfirmed	Madaline	Brooklyn	Kensington	40.647
1	Skylit Midtown Castle	verified	Jenna	Manhattan	Midtown	40.753
2	THE VILLAGE OF HARLEM....NEW YORK !	unconfirmed	Elise	Manhattan	Harlem	40.809
3	Home away from home	unconfirmed	Garry	Brooklyn	Clinton Hill	40.685
4	Entire Apt: Spacious Studio/Loft by central park	verified	Lyndon	Manhattan	East Harlem	40.798

In [24]:
```
## Remove the dollar sign and comma from the columns. If necessary, convert these t
def remove_signs(value):
    return float(value.replace("$", "").replace(",", "").replace(" ", ""))

data["price"] = data["price"].apply(lambda x: remove_signs(x) if pd.notnull(x) else
data["service_fee"] = data["service_fee"].apply(lambda x: remove_signs(x) if pd.not
```

In [25]:
```
data.head()
```

Out[25]:

	name	host_identity_verified	host_name	neighbourhood_group	neighbourhood	
0	Clean & quiet apt home by the park	unconfirmed	Madaline	Brooklyn	Kensington	40.647
1	Skylit Midtown Castle	verified	Jenna	Manhattan	Midtown	40.753
2	THE VILLAGE OF HARLEM....NEW YORK !	unconfirmed	Elise	Manhattan	Harlem	40.809
3	Home away from home	unconfirmed	Garry	Brooklyn	Clinton Hill	40.685
4	Entire Apt: Spacious Studio/Loft by central park	verified	Lyndon	Manhattan	East Harlem	40.798

In [26]:
```
data.isnull().sum()
```

Out[26]:
```
name                              0
host_identity_verified            0
host_name                         0
neighbourhood_group               0
neighbourhood                     0
lat                               0
long                              0
instant_bookable                  0
cancellation_policy               0
room_type                         0
construction_year                 0
price                             0
service_fee                       0
minimum_nights                    0
number_of_reviews                 0
last_review                       0
reviews_per_month                 0
review_rate_number                0
calculated_host_listings_count    0
days_booked                       0
house_rules                       0
license                           0
dtype: int64
```

Task 4: Exploratory Data Analysis (Any Tool)

- List the count of various room types avaliable in the dataset.
- Which room type has the most strict cancellation policy?
- List the average price per neighborhood group, and highlight the most expensive neighborhood to rent from.

If using Python for this exercise, please include the code in the cells below. If using any other tool, please include screenshoots of your work.

In [27]:
```
## List the count of various room types avaliable with Airbnb
data["room_type"].unique()
```

Out[27]:
```
array(['Private room', 'Entire home/apt', 'Shared room', 'Hotel room'],
      dtype=object)
```

In [28]:
```
## Which room type adheres to more strict cancellation policy
data_strict_cancellation_policy = data[data["cancellation_policy"] == "strict"]
room_type_with_strict_policy = data_strict_cancellation_policy["room_type"]
```

In [29]:
```
room_type_with_strict_policy.value_counts()
```

Out[29]:
```
Entire home/apt    17238
Private room       14936
Shared room          718
Hotel room            34
Name: room_type, dtype: int64
```

In [30]:
```
#data.to_csv('Semi_cleaned_data_for_EDA.csv')
```

In [31]:
```
data.head(50)
```

Out[31]:

	name	host_identity_verified	host_name	neighbourhood_group	neighbourhood	
0	Clean & quiet apt home by the park	unconfirmed	Madaline	Brooklyn	Kensington	4
1	Skylit Midtown Castle	verified	Jenna	Manhattan	Midtown	4
2	THE VILLAGE OF HARLEM....NEW YORK !	unconfirmed	Elise	Manhattan	Harlem	4
3	Home away from home	unconfirmed	Garry	Brooklyn	Clinton Hill	4
4	Entire Apt: Spacious Studio/Loft by central park	verified	Lyndon	Manhattan	East Harlem	4
5	Large Cozy 1 BR Apartment In Midtown East	verified	Michelle	Manhattan	Murray Hill	4
6	BlissArtsSpace!	unconfirmed	Alberta	Brooklyn	Bedford-Stuyvesant	4
7	BlissArtsSpace!	unconfirmed	Emma	Brooklyn	Bedford-Stuyvesant	4
8	Large Furnished Room Near B'way	verified	Evelyn	Manhattan	Hell's Kitchen	4
9	Cozy Clean Guest Room - Family Apt	unconfirmed	Carl	Manhattan	Upper West Side	4
10	Cute & Cozy Lower East Side 1 bdrm	verified	Miranda	Manhattan	Chinatown	4
11	Beautiful 1br on Upper West Side	verified	Alan	Manhattan	Upper West Side	4
12	Central Manhattan/near Broadway	verified	Michael	Manhattan	Hell's Kitchen	4
13	Lovely Room 1, Garden, Best Area,	verified	Darcy	brookln	South Slope	4

	name	host_identity_verified	host_name	neighbourhood_group	neighbourhood	
	Legal rental					
14	Wonderful Guest Bedroom in Manhattan for SINGLES	verified	Leonardo	Manhattan	Upper West Side	4
15	West Village Nest - Superhost	verified	Daniel	Manhattan	West Village	4
16	Only 2 stops to Manhattan studio	unconfirmed	Heather	Brooklyn	Williamsburg	4
17	Perfect for Your Parents + Garden	verified	Ryan	Brooklyn	Fort Greene	4
18	Chelsea Perfect	verified	Alberta	manhatan	Chelsea	4
19	Hip Historic Brownstone Apartment with Backyard	unconfirmed	Martin	Brooklyn	Crown Heights	4
20	Huge 2 BR Upper East Cental Park	verified	Audrey	Manhattan	East Harlem	4
21	Sweet and Spacious Brooklyn Loft	verified	Alissa	Brooklyn	Williamsburg	4
22	CBG CtyBGd HelpsHaiti rm#1:1-4	verified	Mary	Brooklyn	Park Slope	4
23	CBG Helps Haiti Room#2.5	unconfirmed	William	Brooklyn	Park Slope	4
24	CBG Helps Haiti Rm #2	unconfirmed	Charlotte	Brooklyn	Park Slope	4
25	MAISON DES SIRENES1,bohemian apartment	unconfirmed	Miranda	Brooklyn	Bedford-Stuyvesant	4
26	Sunny Bedroom Across Prospect Park	unconfirmed	Carlos	Brooklyn	Windsor Terrace	4

	name	host_identity_verified	host_name	neighbourhood_group	neighbourhood	
27	Magnifique Suite au N de Manhattan - vue Cloitres	verified	Adrianna	Manhattan	Inwood	4
28	Midtown Pied-a-terre	unconfirmed	Andrew	Manhattan	Hell's Kitchen	4
29	SPACIOUS, LOVELY FURNISHED MANHATTAN BEDROOM	verified	Daryl	Manhattan	Inwood	4
30	Modern 1 BR / NYC / EAST VILLAGE	unconfirmed	Tyler	Manhattan	East Village	4
31	front room/double bed	unconfirmed	Byron	Manhattan	Harlem	4
32	Spacious 1 bedroom in luxe building	verified	Mary	Manhattan	Harlem	4
33	Loft in Williamsburg Area w/ Roof	unconfirmed	John	Brooklyn	Greenpoint	4
34	back room/bunk beds	verified	Alfred	Manhattan	Harlem	4
35	Large B&B Style rooms	unconfirmed	Jared	Brooklyn	Bedford-Stuyvesant	4
36	Lovely room 2 & garden; Best area, Legal rental	verified	Brad	Brooklyn	South Slope	4
37	Clean and Quiet in Brooklyn	verified	Arthur	Brooklyn	Bedford-Stuyvesant	4
38	Cute apt in artist's home	verified	Joyce	Brooklyn	Bushwick	4
39	Country space in the city	verified	Deanna	Brooklyn	Flatbush	4

	name	host_identity_verified	host_name	neighbourhood_group	neighbourhood	
40	LowerEastSide apt share shortterm 1	unconfirmed	Clark	Manhattan	Lower East Side	4
41	ENJOY Downtown NYC!	unconfirmed	Byron	Manhattan	East Village	4
42	Beautiful Sunny Park Slope Brooklyn	verified	Alina	Brooklyn	South Slope	4
43	1bdr w private bath. in lofty apt	unconfirmed	Charlie	Brooklyn	Fort Greene	4
44	West Side Retreat	unconfirmed	Alford	Manhattan	Upper West Side	4
45	BEST BET IN HARLEM	unconfirmed	Chester	Manhattan	Harlem	4
46	Entire apartment in central Brooklyn neighborh...	unconfirmed	David	Brooklyn	Prospect-Lefferts Gardens	4
47	1 Stop fr. Manhattan! Private Suite,Landmark B...	unconfirmed	Victoria	Queens	Long Island City	4
48	Charming Brownstone 3 - Near PRATT	verified	Jared	Brooklyn	Bedford-Stuyvesant	4
49	bright and stylish duplex	verified	Chloe	Brooklyn	Bedford-Stuyvesant	4

In [32]:
```
## List the prices by neighborhood group and also mention which is the most expensi
```

Task 5a: Data Visualization (Any Tool)

- List the count of various room types avaliable with Airnb
- Which room type adheres to more strict cancellation policy
- List the prices by neighborhood group and also mention which is the most expensive neighborhood group for rentals
- List the top 10 neighborhoods in the increasing order of their price with the help of a horizontal bar graph. Which is the cheapest neighborhood.
- List the neighborhoods which offer short term rentals within 10 days. Illustrate with a bar graph

- List the prices with respect to room type using a bar graph and also state your inferences.
- Create a pie chart that shows distribution of booked days for each neighborhood group .Which neighborhood has the highest booking percentage.

If using Python for this exercise, please include the code in the cells below. If using any other tool, please include screenshoots of your work.

Task 5b: Data Visualization (Any Tool)

- Does service price and room price have an impact on each other. Illustrate this relationship with a scatter plot and state your inferences
- Using a line graph show in which year the maximum construction of rooms took place.

If using Python for this exercise, please include the code in the cells below. If using any other tool, please include screenshoots of your work.

In [ ]:

In [ ]:

Task 5c: Data Visualization (Any Tool)

- With the help of box plots illustrate the following
  - Effect of Review Rate number on price
  - Effect of host identity verified on price

If using Python for this exercise, please include the code in the cells below. If using any other tool, please include screenshoots of your work.

In [ ]:

In [ ]:

In [ ]: