DATA ASSESSMENT - GROUP 6

AIRBNB PRICE PRIDICTION IN CANADA

We have chosen 6 different Airbnb datasets from different cities in Canada. We have also grouped the data based on the population of the cities into 3 segments namely: **Mega** (Toronto and Montreal), **Mid** (Vancouver and Winnipeg), and **Small** (Quebec city and Victoria)

```
# Importing libraries:

import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import plotly.express as px
from statsmodels.graphics.gofplots import qqplot
from mpl_toolkits.mplot3d import Axes3D
from sklearn.preprocessing import StandardScaler
import os
```

Loading Datasets

```
Toronto = pd.read_csv("Toronto_Ontario_mega.csv")

Montreal = pd.read_csv("Montreal_Quebec_mega.csv")

Vancouver = pd.read_csv("Vancouver_BritishColumbia_mid.csv")

Winnipeg = pd.read_csv("Winnipeg_Manitoba_mid.csv")

Quebec = pd.read_csv("Quebec_Quebec_small.csv")

Victoria = pd.read_csv("Victoria_BritishCoumbia_small.csv")
```

Adding the column in each datasets

```
Toronto['City'] = 'Toronto'

Montreal['City'] = 'Montreal'

Vancouver['City'] = 'Vancouver'

Winnipeg['City'] = 'Winnipeg'

Quebec['City'] = 'Quebec'

Victoria['City'] = 'Victoria'
```

Segmenting cities into Mega, Mid, and Small with respect to population

```
Toronto['City_type'] = 'Mega'

Montreal['City_type'] = 'Mega'

Vancouver['City_type'] = 'Mid'

Winnipeg['City_type'] = 'Mid'

Quebec['City_type'] = 'Small'

Victoria['City_type'] = 'Small'
```

We observed that all the dataset have the same variables. Hence we have merged the datasets into one and rename it as: Airbnb

```
Airbnb =  pd.concat([Toronto, Montreal, Vancouver, Winnipeg, Quebec, Victoria])
```

```
#Exporting Airbnb to file

Airbnb.to_csv('Airbnb.csv', index=False)
```

Assessing the dataset

```
Airbnb.head()
```

	id	name	host_id	host_name	neighbourhood_group	neighbourhood	latitude	longitude	roc
0	2818.0	Quiet Garden View Room & Super Fast WiFi	3159	Daniel	NaN	Oostelijk Havengebied - Indische Buurt	52.36435	4.94358	
1	20168.0	Studio with private bathroom in the centre 1	59484	Alexander	NaN	Centrum-Oost	52.36407	4.89393	
2	27886.0	Romantic, stylish B&B houseboat in canal district	97647	Flip	NaN	Centrum-West	52.38761	4.89188	
3	28871.0	Comfortable double room	124245	Edwin	NaN	Centrum-West	52.36775	4.89092	
4	29051.0	Comfortable single room	124245	Edwin	NaN	Centrum-Oost	52.36584	4.89111	

```
Airbnb.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 34694 entries, 0 to 4161
Data columns (total 20 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   id                              34694 non-null  float64
 1   name                            34687 non-null  object
 2   host_id                         34694 non-null  int64
 3   host_name                       34694 non-null  object
 4   neighbourhood_group             4162 non-null   object
 5   neighbourhood                   34694 non-null  object
 6   latitude                        34694 non-null  float64
 7   longitude                       34694 non-null  float64
 8   room_type                       34694 non-null  object
 9   price                           34694 non-null  int64
 10  minimum_nights                  34694 non-null  int64
 11  number_of_reviews               34694 non-null  int64
 12  last_review                     28928 non-null  object
 13  reviews_per_month               28928 non-null  float64
 14  calculated_host_listings_count  34694 non-null  int64
 15  availability_365                34694 non-null  int64
 16  number_of_reviews_ltm           34694 non-null  int64
 17  license                         12143 non-null  object
 18  City                            34694 non-null  object
 19  City_type                       34694 non-null  object
dtypes: float64(4), int64(7), object(9)
memory usage: 5.6+ MB
```

```
#Checking for null values in the dataset

Airbnb.isnull().sum()
```

```
id                                  0
name                                7
host_id                             0
host_name                           0
neighbourhood_group             30532
neighbourhood                       0
latitude                            0
longitude                           0
room_type                           0
price                               0
minimum_nights                      0
number_of_reviews                   0
last_review                      5766
reviews_per_month                5766
calculated_host_listings_count      0
availability_365                    0
number_of_reviews_ltm               0
license                         22551
City                                0
City_type                           0
dtype: int64
```

Dropping columns with few or negligible observations

```
#Dropping the variable: neighbourhood_group because the column has few and negligible observation

Airbnb.drop(["neighbourhood_group"], axis=1, inplace=True)
```

```
#Dropping the variable: license because the column has few and negligible observation

Airbnb.drop(["license"], axis=1, inplace=True)
```

There are three columns with null variable: name, last_review, and reviews_per_month. we will clean the name columns as it seems important for analysis.

```
#Filling missing values in column "name" with 0

Airbnb['name'].fillna(0, inplace = True)
```

```
#Dropping missing values in "name column"
Airbnb_name = Airbnb[Airbnb['name'] == 0 ].index
Airbnb.drop(Airbnb_name, inplace = True)

Airbnb
```

	id	name	host_id	host_name	neighbourhood	latitude	longitude	room_type	p
0	2.818000e+03	Quiet Garden View Room & Super Fast WiFi	3159	Daniel	Oostelijk Havengebied - Indische Buurt	52.364350	4.943580	Private room	
1	2.016800e+04	Studio with private bathroom in the centre 1	59484	Alexander	Centrum-Oost	52.364070	4.893930	Private room	
2	2.788600e+04	Romantic, stylish B&B houseboat in canal district	97647	Flip	Centrum-West	52.387610	4.891880	Private room	
3	2.887100e+04	Comfortable double room	124245	Edwin	Centrum-West	52.367750	4.890920	Private room	

Checking the shape of Airbnb. We now have 18 variables and 34652 observations as shown below:

```
#Shape of the dataset

Airbnb.shape
```

```
(34652, 18)
```

```
Airbnb.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 34652 entries, 0 to 4161
Data columns (total 18 columns):
 #   Column                          Non-Null Count  Dtype  
---  ------                          --------------  -----  
 0   id                              34652 non-null  float64
 1   name                            34652 non-null  object 
 2   host_id                         34652 non-null  int64  
 3   host_name                       34652 non-null  object 
 4   neighbourhood                   34652 non-null  object 
 5   latitude                        34652 non-null  float64
 6   longitude                       34652 non-null  float64
 7   room_type                       34652 non-null  object 
 8   price                           34652 non-null  int64  
 9   minimum_nights                  34652 non-null  int64  
 10  number_of_reviews               34652 non-null  int64  
 11  last_review                     28889 non-null  object 
 12  reviews_per_month               28889 non-null  float64
 13  calculated_host_listings_count  34652 non-null  int64  
 14  availability_365                34652 non-null  int64  
 15  number_of_reviews_ltm           34652 non-null  int64  
 16  City                            34652 non-null  object 
 17  City_type                       34652 non-null  object 
dtypes: float64(4), int64(7), object(7)
memory usage: 5.0+ MB
```

```
#Checking for duplicate values

Airbnb.duplicated().sum()
```

```
0
```

There are no duplicate values in the Airbnb dataset

Ethical Principles

Checklist

Airbnb Data Policies We have maintained the Data Policies of the Airbnb (http://insideairbnb.com/data-policies) - Inside Airbnb is a mission-driven activist project with the objective to provide data that quantifies the impact of short-term rentals on housing and residential communities, as well as create a platform to support advocacy for policies to protect our cities from the impacts of short-term rentals.

Consent: We will Only take the data that we need from the website as stipulated in the website. The data is open and available on the inside airbnb website and it can be used for analysis purpose.

Clarity: We are very clear about what we are doing with the data which includes analysis that will portray different prices across 6 cities in Canada namely: Toronto, Montreal, Quebec City, Winnipeg, Victoria, and Vancouver in addition to predicting the prices

Consistency: Our analysis will give an insight into possible cost of Airbnb by intending users within the 6 cities we have chosen to analyze, this will guide users in selecting their prefered Airbnb in those cities.

Control: The analysis is available for users to decide and make their choices withing the 6 cities. Users data will not be required before they have access to the prediction analysis.

Consequences: There is absolutely no harmful consequence for the users of our analysis as it is not designed to collect personal information from users.

References

Dataset: http://insideairbnb.com/get-the-data.html

Airbnb Website: https://www.airbnb.ca/

Airbnb Disclaimer: http://insideairbnb.com/about.html