Homework 1 Python Data Analysis and Machine Learning

楊英豪 B10803207

Why do I choose this dataset?

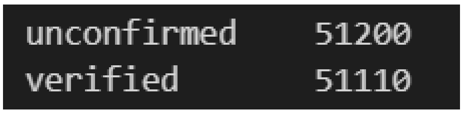
I choose the second dataset which is the Airbnb dataset, I choose this dataset over the global temperature dataset because I find that business dataset is more appealing to me. It is more relatable and closer to everyday life. Although the dataset is for the New York location, I find that the data can provide valuable insights

Questions (30% each):

1. **In the data, there are two values of host_identity_verified. Which value is larger?**

   There are two values, which is the unconfirmed and also the verified, The value of the unconfirmed one is larger, I show this values by using this command below

```
df['host_identity_verified'].value_counts(dropna= True)
```



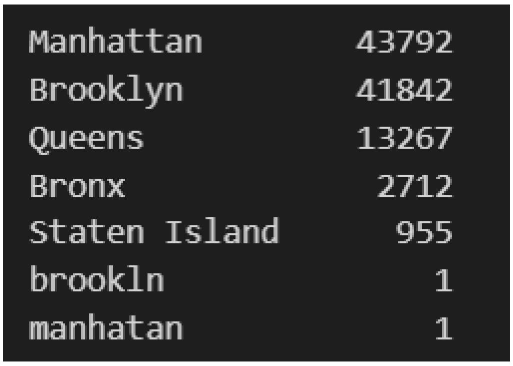
```
unconfirmed    51200
verified       51110
```

2. **What are the top 2 neighbourhood_group?**

   The top two neighbourhood group in this dataset is the Manhattan and Brooklyn with values of 43792 and 41842 respectively.

   I found this using the same function with question 1 but different parameters

```
df['neighbourhood group'].value_counts(dropna = True)
```

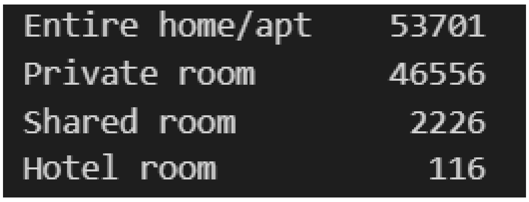


```
Manhattan        43792
Brooklyn         41842
Queens           13267
Bronx             2712
Staten Island      955
brookln              1
manhatan             1
```

3. **How many room types are in the data, and what are their proportions?**

   There are 4 type of rooms, I also get this using the same function but different parameters

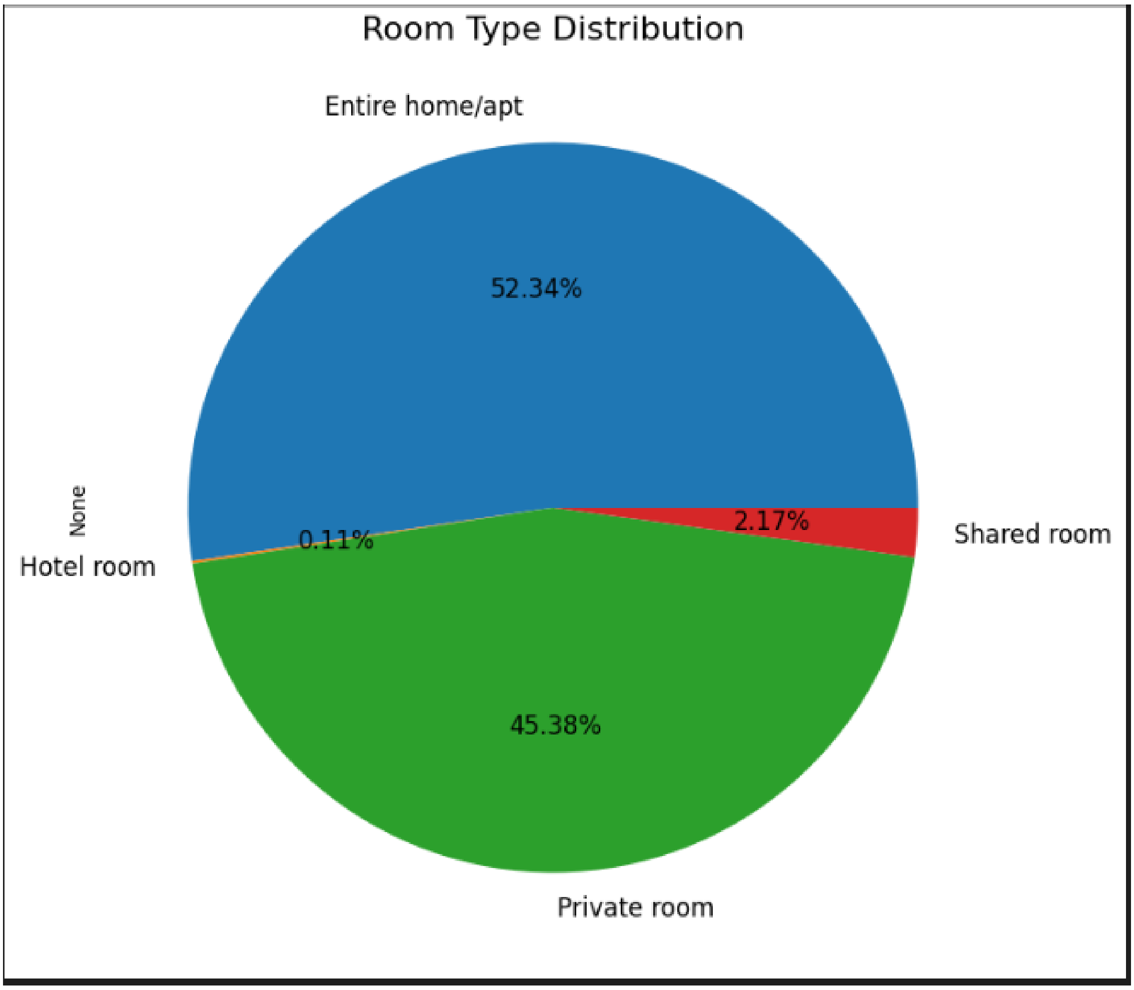
```
df['room type'].value_counts(dropna = True)
```



```
Entire home/apt    53701
Private room       46556
Shared room         2226
Hotel room           116
```

I also plot a pie graph using the function from matplotlib

```
room_counts = df.groupby('room type', dropna = True).size()

room_counts.plot.pie(figsize = (8,8), autopct= '%.2f%%', fontsize = 12)

plt.title('Room Type Distribution', fontsize= 16)

plt.show()
```



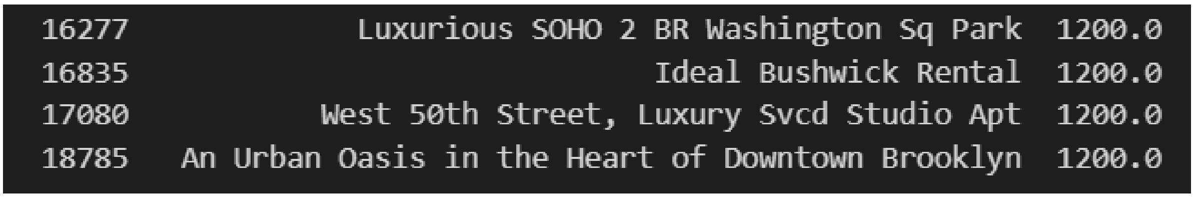
Please describe your insight and finding (if have) from the cleaned data and visualization figures

1. I found several insights and finding in the data

The most expensive ones are $1200, I use the code below to find this

```
#convert to string
df['price'] = df['price'].astype(str)
```

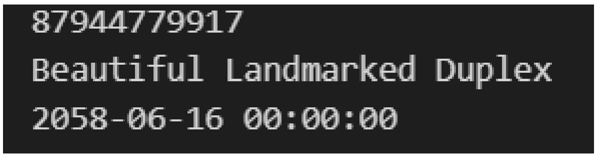
```
#clean data, remove dollar sign and comma
df['price'] = df['price'].str.replace('$', '').str.replace(',', '')

#convert to float
df['price'] = df['price'].astype(float)

#take the 10 largest within the column price
most_expensive = df.nlargest(10, 'price')

#show it by name
print(most_expensive[['NAME', 'price']])
```



```
16277            Luxurious SOHO 2 BR Washington Sq Park  1200.0
16835                             Ideal Bushwick Rental  1200.0
17080          West 50th Street, Luxury Svcd Studio Apt  1200.0
18785  An Urban Oasis in the Heart of Downtown Brooklyn  1200.0
```

2. I also find that the latest review date may have some error it is shown that the latest review was in 2058-06-16, which is wrong data

```
df['last review'] = pd.to_datetime(df['last review'], format='%m/%d/%Y')
df.dropna(subset=['last review'], inplace=True)
df = df.sort_values(by='last review', ascending=False)
latest_review = df.iloc[0]
airbnb_name = latest_review['NAME']
host_id = latest_review['host id']
last_review_date = latest_review['last review']
print(host_id)
print(airbnb_name)
print(last_review_date)
```



```
87944779917
Beautiful Landmarked Duplex
2058-06-16 00:00:00
```