Nirav Bariya
Achievement 6
Exercise 6.1
Date: 04/28/2024

1. Summary of Airbnb listings for Toronto

Data Source: This data source is external. It was collected by insideairbnb.com and not published by Airbnb itself. The website however is not endorsing nor associated with Airbnb or its' competition. The website claims that the data compiled is for public analysis, discussion, and community website. In my opinion, I think we can consider this data set as trustworthy since it is collected from Airbnb and due transparency is shown by Inside Airbnb.

Collection Method: This dataset can be considered as usage data as this can be utilized for improving pricing of Airbnb listings. The collection method used is web scrapping as the data is collected from the Airbnb website. This is an automatic method. The data set is relatively new as it includes listings from the month of February 2024. Therefore, the time lag is minimal in this case.

Data Content: The data set has 75 features including id, listing_url, last_scrapped, bathrooms, accommodates, bedrooms, price, minimum_nights and so on. It also information about reviews including features such as review_scores_rating, review_scores_accuracy, review_scores_cleanliness, and review_scores_checkin.

Data Relevance: This data is the primary data set that I will be working with as it not only has information about pricing and features of the house, but it also has information about location and reviews.

Why this Data set?
This data set is one of the new data sets in the public domain and is relevant to the country and locality in which I am currently residing. It provides great value as decisions about how to price an Airbnb listing can be made. Moreover, the factors which impact success of the Airbnb listing can also be studied. Lastly, this data set gives ample opportunity to learn machine learning algorithms including regression and classification. It also provides time series data that may be used for time series analysis.

2. Data Profile

Cleaning the Data:

1. Missing Data:

```
toronto_listings.isnull().sum()
```

```
id                                0
scrape_id                         0
neighborhood_overview          9335
picture_url                       0
host_id                           0
host_since                        2
host_location                  5206
host_response_time             6841
host_response_rate             6841
host_acceptance_rate           5434
host_is_superhost               338
host_neighbourhood            12124
host_listings_count               2
host_total_listings_count         2
host_verifications                2
host_has_profile_pic              2
host_identity_verified            2
neighbourhood                  9334
neighbourhood_cleansed            0
neighbourhood_group_cleansed  20630
latitude                          0
longitude                         0
property_type                     0
room_type                         0
accommodates                      0
bathrooms                      5252
bathrooms_text                    7
bedrooms                       1664
beds                           5278
amenities                         0
price                          5317
minimum_nights                    0
maximum_nights                    0
minimum_minimum_nights            0
maximum_minimum_nights            0
minimum_maximum_nights            0
```

Results:

Important Notes:

1. calendar_updated column is almost missing, therefore we will drop this column.
2. Neighbourhood column has 45% missing data. This column will be dropped.
3. host_neighbourhood has 58% missing data. This column will be dropped.
4. neighbourhood_group_cleansed has most of the data missing, and therefore we will remove this column from analysis.
5. license column has 53% of the data missing. This column will also be dropped from the analysis.
6. The columns with less than 5% of the missing data will be kept as it is as the number of missing data isn't significant. For the numerical columns with more than 5% missing data, the data will be imputed if required for the analysis.

2. Handling Duplicates:

```
toronto_listings.duplicated().sum()
```

```
0
```

There are no duplicate entries in this data frame.

3. Checking for mixed type data:

Checking for Mixed type in orders

```
for col in toronto_listings.columns.tolist():
  weird = (toronto_listings[[col]].applymap(type) != toronto_listings[[col]].iloc[0].apply(type)).any(axis = 1)
  if len (toronto_listings[weird]) > 0:
    print (col)
```

```
neighborhood_overview
host_since
host_location
host_response_time
host_response_rate
host_acceptance_rate
host_is_superhost
host_verifications
host_has_profile_pic
host_identity_verified
bathrooms_text
price
has_availability
first_review
last_review
```

The columns were made consistent in the Jupyter notebook.

Data Summary:

Descriptive statistics

```
toronto_listings.describe()
```

	id	scrape_id	host_id	host_since	host_listings_count	host_total_listings_count	latitude	longitude	accommodate
count	2.063000e+04	2.063000e+04	2.063000e+04	20628	20628.000000	20628.000000	20630.000000	20630.000000	20630.00000
mean	4.541553e+17	2.024021e+13	2.027128e+08	2017-12-19 18:51:39.476439808	8.191584	14.012119	43.684885	-79.397742	3.12985
min	1.419000e+03	2.024021e+13	1.565000e+03	2008-08-08 00:00:00	1.000000	1.000000	43.585750	-79.623950	1.00000
25%	2.863717e+07	2.024021e+13	3.466884e+07	2015-06-06 00:00:00	1.000000	1.000000	43.646480	-79.430460	2.00000
50%	5.948689e+17	2.024021e+13	1.363575e+08	2017-06-23 00:00:00	2.000000	3.000000	43.665790	-79.397829	2.00000
75%	9.115529e+17	2.024021e+13	3.777238e+08	2020-12-01 06:00:00	4.000000	7.000000	43.713078	-79.372120	4.00000
max	1.090775e+18	2.024021e+13	5.619015e+08	2024-02-13 00:00:00	686.000000	1763.000000	43.838414	-79.128010	16.00000
std	4.527170e+17	3.941502e+00	1.883283e+08	NaN	38.129842	80.717411	0.050197	0.071932	2.00362

Observations:

1. Maximum host_listing_counts is very high! Having 686 listings on Airbnb is very large.
2. Maximum 50 bedrooms is also quite high.
3. Maximum price of 12400 is also very high. It might be a big property.
4. Standard deviation is 439.61 nights for maximum_nights!

	Data Types			
Variables	Time-variant/-invariant	Structured/Unstructured	Qualitative/Quantitative	Qualitative: Nominal/Ordinal Quantitative: Discrete/Continuous
id	Time-invariant	Structured	Qualitative	Nominal
scrape_id	Time-invariant	Structured	Qualitative	Nominal
host_id	Time-invariant	Structured	Qualitative	Nominal
host_since	Time-invariant	Structured	Qualitative	Ordinal
host_location	Time-invariant	Structured	Qualitative	Nominal
host_acceptance_rate	Time-variant	Structured	Quantitative	Continuous
host_total_listings_count	Time-variant	Structured	Quantitative	Continuous
host_has_profile_pic	Time-Invariant	Structured	Qualitative	Nominal
neighbourhood_cleansed	Time-Invariant	Structured	Qualitative	Nominal
room_type	Time-variant	Structured	Qualitative	Nominal
beds	Time-variant	Structured	Quantitative	Discrete
minimum_minimum_nights	Time-variant	Structured	Quantitative	Discrete
minimum_maximum_nights	Time-variant	Structured	Quantitative	Discrete
minimum_nights_avg_ntm	Time-variant	Structured	Quantitative	Discrete
availability_30	Time-variant	Structured	Quantitative	Discrete

Limitations and ethics:

The data is collected from the Airbnb website itself using. Web scrapping may be considered unethical and falls under grey area. The Inside Airbnb website states that no private information is used as the Name, photographs, and listings are available publicly. There is a high degree of transparency about the way the data is collected. Some of the variables are calculated by Inside Airbnb. These may not be used in the analysis and hence is not of concern to us.

Steps have been taken to anonymize location information by Airbnb itself. Considering all these, I think this data set offers great value in terms of learning and can be considered unbiased.

Questions to explore:

In a third section of your project document, define a list of questions to explore with your analysis.

1. What is the distribution of super host?
2. How are the listings distributed in Toronto? Are certain neighbourhoods have more listings than others?
3. What is the distribution of property type and room type in Toronto?
4. What are the common amenities offered in the listings?
5. Is there any relationship between bedrooms and price?
6. Are certain neighbourhoods priced more than others?
7. Is there any relationship between property type and price?
8. Is there a relationship between accommodations and price?
9. Which property types have higher price?
10. Is there a certain relationship between number of reviews and price?

11. Do customers prefer lower minimum nights or higher minimum nights?
12. Is there a relationship between price and average number of reviews per month?