Programming for Data Analytic
SOFT8032

Lecturers: Dr Farshad Ghassemi Toosi

Nov 2020

1 Second Assessment. First Project

This project contributes 30% in your final mark. This is an individual project and has to be all done by yourself. You may be called for a zoom meeting to explain different parts of your submission, if needed.

Any question regarding the project should be communicated with farshad.toosi@cit.ie or Canvas message.

1.1 Dataset Overview

For this project we are going to perform a number analytic tasks on the US airBnB dataset which is shared on Canvas with you. This dataset contains the details of a large number accommodations (houses or rooms). Each accommodation has the following details described in 16 columns:

1. id: Accommodation ID
2. name: Accommodation name
3. host_id: Accommodation's host ID
4. host_name: The name of the host
5. neighbourhood
6. latitude
7. longitude
8. room_type
9. price: Price per night
10. minimum_nights: Minimum number of nights required to book.
11. number_of_reviews: Total number of reviews
12. last_review

13. review
14. availability_365: Number of days in year the BnB is available for rent
15. city: The city of the accommodation.

1.2 Project Specification

The objective of this project is to provide an insight into some of the relationships and trends that exist within this dataset. Please note you can use Pandas, Numpy and Matplotlib as means of analysing data within this dataset. You may as well use math, random modules and etc if needed.

1.2.1 Details of the project specified as a number of tasks

1. Use an appropriate visualization technique to visually analyse the size of 4 different groups of accommodations (group A, group B, group C and group D). The groups are created as follows:

   (a) The average of all prices in the dataset is found and all accommodation equal or less than average are categorized as group 1 and all accommodation with price higher than average are categorized as group 2.

   (b) Find the average price in group 1 and create group A and group B: Group A contains all the accommodation that their price is equal or less than the average price in group 1 and Group B contains all the accommodation that their price is higher than the average price in group 1.

   (c) Find the average price in group 2 and create group C and group D: Group C contains all the accommodation that their price is equal or less than the average price in group 2 and Group D contains all the accommodation that their price is higher than the average price in group 2.

   Highlight/emphasise the group with the smallest size visually. Note that appropriate visualization features are needed to be embedded in the plot (e.g., percentages, label and etc).

2. In the dataset, there are some hosts that own more than one accommodation. Analyse the data set and find out the name and id of the first 20 hosts that own the highest number of accommodation. Use an appropriate visualization technique (See Fig 1) to visualize the data. Use appropriate visualization features. Use comment and specify the host_id and host_name of the host who owns the highest number of accommodation. Note: There might be different hosts with the same host_name but there are no two different hosts with the same host_id.

3. Use an appropriate visualization approach and analyse accommodation prices cheaper than 500. You are required to investigate and find out what prices are more common and what prices are less common. Use appropriate visualization feature, e.g., label etc. Use comment and analyse your finding, (e.g., there are more accommodations with price around *X1* and there are few accommodations with price around *X2*).

4. Use an appropriate visualization approach and depict whether or not there are outliers among prices in the range [150, 500] or not. Use comment and discuss where the outliers reside.

5. There are four different accommodation types (in the column with header *room_type*) in the dataset. Use an appropriate visualization technique and compare the average price of these four different accommodation types. Use appropriate visualization features and make a comment and specify what accommodation type (*room_type*) has the lowest average price.

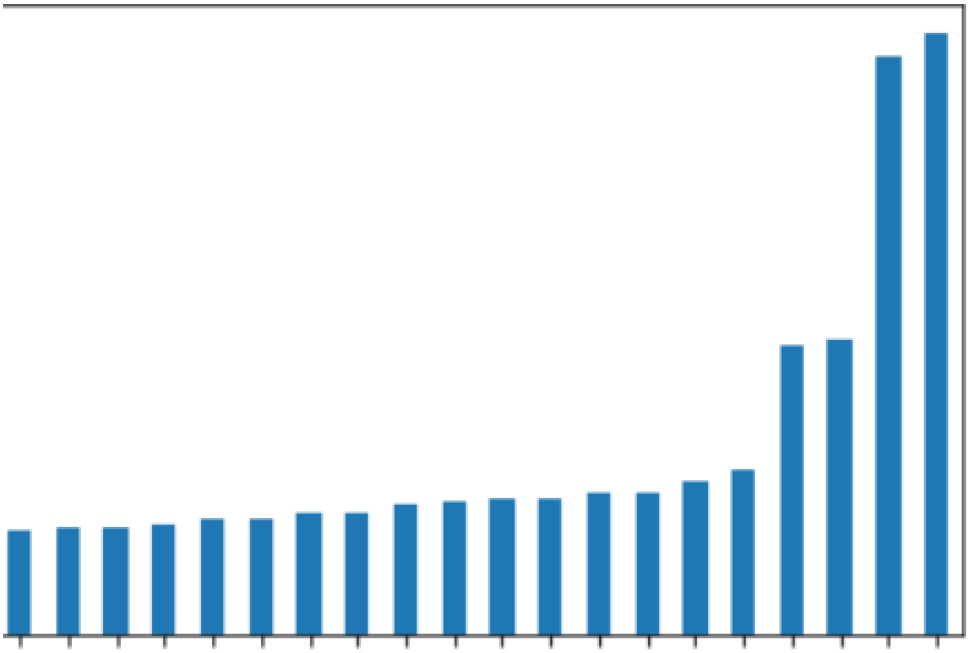

Figure 1: Example

2 Rubric

This rubric is subject to change.

1. Correct task implementation (visualization) with meaningful labels, annotation (if needed) legend, comment and etc. (100%)
2. Correct task implementation (visualization) with less/minimum meaningful labels, annotation, comment and etc. (80%)
3. Partly correct task implementation (visualization) with partly meaningful labels, annotation (if needed) legend, comment and etc. (50%)
4. Wrong task implementation (visualization). (0%)

3 Submission

There is template file (Python file) provided for you on Canvas. You are required to complete your project in that file. Each task needs to be implemented as a separated function with one line interpretation as a comment below the function.

Please write your name, student ID and your course name as a comment in the designated area in the provided python file.

The template file should be re-named at the end using your student ID followed by letter s, for example if your student ID is: 1234567 the the python file should be named: s1234567.py

The deadline for this project is 4th of December 2020 at 23:59. Late submission is accepted with 10 marks penalty and the deadline for the late submission is 10th of December at 23:59.

Please submit your project via Canvas.