PES University, Bengaluru
UE17CS324- Data Analytics
Session: Aug-Dec 2019
Assignment-2

Date of Submission: 12/09/2019 **Max. Marks: 20**

Note: *It's important to work on your approach than the final answer. Since solutions will vary from group to group, right justification for technique used will be vital. It is a group assignment, one submission per team is sufficient.*

TOPIC: DATA VISUALIZATION

About the Dataset for question 1:

fitness_data.csv: This dataset is a cleaned version of the UCI PAMAP2 dataset. It documents readings from 3 inertial measurement units (IMU), and a heart rate monitor while a subject is performing a certain physical activity (e.g. walking, cycling, playing soccer, etc.) For the purpose of this assignment, the readings from only one IMU are considered.

activities.csv: This dataset gives the mapping between the activity ID and the corresponding activity.

subject.csv: This file documents metadata of the subjects involved in the study

Question-1 **(10 marks)**

Library plotly helps make brilliant visualizations that are highly interactive and appealing.

1. Use this library to visualize all the parameters in fitness_data.csv(except activityID, subID and timestamp(s)) against timestamp(x-axis) for each subID.
2. To represent in a single graph, use dropdown menus. First dropdown menu will be used to select the subID. Second dropdown menu will be used to select the column to be analyzed.
   (For eg: User wants to visualize **heart rate activity column**(dropdown 1) for **subID 104**(dropdown 2))
3. Use a slider to control the range of timestamp.
4. What insights could you glean from this plot? (There is no single correct answer)

Question-2 **(2 marks)**

What is undersampling and oversampling? Consider the dataset subject.csv. Is there a case of undersampling or oversampling? If so, mention a technique to remedy the problem. Justify your answer.

Question-3 **(1 mark)**

There are various techniques for sampling data. Suggest a sampling technique that you think is ideal for the data in fitness_data.csv, and justify your choice.

Question-4 **(5 marks)**

In August 2018, Election Commission of India made Lok sabha 2014(**Lok Sabha-2014 data.csv**) data public so that analysts can use it for 2019 Lok Sabha election. Provide a suitable visualisation that accounts for the distribution of votes across the country.

Question-5 **(2 marks)**

Many good Bollywood movies were released in 2019, one of them being Kabir Singh. The file **tweets.txt** contains what people have tweeted about this movie. Provide suitable visualization that depicts the generals sentiment of the audience.