Python End Term Assignment

[60 Marks]

Problem

Imagine how lengthy the process of data exploration may become especially when you are dealing with datasets that contains a large number of variables! I hope you have already started experiencing this tedious process. For every numerical variable you identify in your dataset, you generate histogram and boxplot. For every categorical variable you identify in your dataset, you create bar chart and maybe pie chart. This is how the first few steps of your data exploration look like.

Generating graphs one at a time may be okay if you have a dataset with less number of columns. However, this may be an extremely tedious task when you are working on a large dataset with lots of columns in it. Not only that this process can also become extremely time consuming. In this assignment, you will create functions to reduce this everyday task of yours to a considerable extent and save hours and hours of your time. So, here's what you will be doing…

1. You will create a function named '`Histogram`' which will take the entire dataset as input and return the histograms for all the numerical variables in your dataset as .png files in your working directory. Make sure that all the generated graphs have proper titles and axis labels. (10 marks)

   ```
   Histogram(data)
   -> returns histograms for all numerical variables in data.
   ```

2. Make an improvement on the function you have created in 1. Create a function names '`Graphs`' which will take a dataset as input and return histograms and boxplots for all the numerical variables and bar plots for all categorical variables. (10 marks)

   ```
   Graphs(data)

   -> returns histograms and boxplots for all numerical variables and
   bar plot for all categorical variables in data.
   ```

3. Often, we are not required to plot the graphs for all the variables in our dataset. Add an additional argument to the function in 2 named '`var`'. This will take a list containing the variable index and return the graphs for only those variables. By default, (i.e. if the list of variable index is not provided), then it must return the required graphs for all the variables in data. (10 marks)
   Example:

   ```
   Graphs(data, var=c(1,3,4))

   -> Will generate the graphics for only the variables 1,3 & 4 in the
   data.
   ```

4. Sometimes, we do not want to mess up our working directories with so many image files. Create an additional argument for the function "`dir`" (directory), such that the function exports all the files to that specified folder (which need not necessarily be your working directory). (10 marks)

   Example:

   ```
   Graphs(data, Variable=c(1,3,4), dir=".../Praxis/Graphs")
   ```

   ```
   -> will generate all the necessary graphics for the variables 1, 3
   and 4 in the specified location in your system which is ".../Praxis/
   Graphs"
   ```

Do any one from question 5 and question 6

5. **Further Improvements:** Create at least 2 more features of this function which you think can benefit you in your EDA process. (20 marks)

6. **New Improvements:** Design a new function which you feel can help you a lot in your EDA process. (20 marks)

Team

For this assignment you need to work in a team with **two** members.

Submission

Following items are needed to be submitted.

1. A short YouTube video demonstrating each function (separately for each problem).
2. A SINGLE Jupyter notebook containing all the solution.
3. A GitHub link containing these functions.

Make sure you mention the names of your team members and their roll number in every submission documents. *Your notebooks must have the nomenclature – Roll1-Roll2-PY-ENDTERM (where Roll1 and Roll2 are the roll numbers of the students who worked on the problem).*

Submission Link: https://forms.gle/8HBPpBtX3DyN9nta9

Submission Deadline: 22.11.2020 11:59:59 PM