THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

Visualization of Complex Data

DATS 6401

Final Term Project (FTP)

The objective of FTP is to apply the course learning objectives to visualize complex data. A python web-based dashboard (app) must be created for the interactive visualization of the selected dataset. The required software for the FTP is python and you are allowed to use packages of interest in python to accomplish the FTP objectives. You can use Tableau in parallel for the verification of the results, but all graphs must be generated in python.

The first step in FTP is the data selection. The dataset must satisfy the following criteria:

- Pick and an interesting, applied real world dataset set from industry.
- It must be a multivariate dataset with at least 50K observations.
- It must contain categorical data with at least 2 categories.
- It must contain numerical data with at least 2 columns of numerical data.
- It could be a time series or non-time series.
- It must come from non-classified (public) database.
- A good example of dataset that satisfies above criterial is 'diamonds' dataset. This dataset can NOT be picked for your FTP.

There are several resources available to acquire dataset i.e.

- https://www.kaggle.com/
- https://archive.ics.uci.edu/ml/index.php

A formal report, presentation and demo of the created app using Google Cloud Platform (GCP) is required by the deadline.

SPECIFIES

The final formal report must be typed and should contain the following sections:

1- **Cover page.**
2- **Table of content.**
3- **Table of figures and tables.**
4- **Abstract.**
5- **Introduction**. An overview of the procedures to accomplish the FTP objectives and an outline of the report.
6- **Description of the dataset**: You need to provide a description on the selected dataset and how the dataset satisfies the dataset criteria. You need to specify which variable in the selected dataset will serve as dependent variable and which ones serve as independent variables. You will need to explain the importance of the selected dataset in industry.

7- **Pre-processing dataset**: Data cleaning for missing samples, NAN's. Explain which method was used for the data cleaning. You will need to display the first few observations of the cleaned dataset and the corresponding statistics.

8- **Outlier detection & removal:** Use one of the methods explained in class for the outlier detection and removal from the raw dataset.

9- **Principal Component Analysis (PCA):** Perform a complete PCA analysis of the cleaned dataset for a possible feature dimension reduction. Include the complete explanation into your report. Check the condition number and the singular values of the reduced dimension features.

10- **Normality test:** Use one of the tests explained in class to see if the dataset comes from the Gaussian distribution or not.

11- **Data transformation:** If transformation of the dataset is needed, you need to explain which method was used. For example: non-gaussian to gaussian distribution transformation.

12- **Heatmap & Pearson correlation coefficient matrix:** Display the Pearson correlation coefficient between variables using heatmap and scatter plot matrix.

13- **Statistics :** You need to statistically analyze the dataset and write down your observations accordingly. Use the statistics tools discussed in class. You will need to display the estimated multivariate kernel density estimate.

14- **Data visualization:** Visualize the dataset using the following plots and discuss what can be observed from each plot. You need to write down your observations for each plot bellow. You need to plot data using seaborn package and use hue for the categorical data.

   a. Line-plot
   b. Bar-plot : stack, group
   c. Count-plot
   d. Cat-plot
   e. Pie-chart
   f. Displot
   g. Pair plot
   h. Heatmap
   i. Hist-plot
   j. QQ-plot
   k. Kernal density estimate
   l. Scatter plot and regression line using sklearn
   m. Multivariate Box plot
   n. Area plot (if applicable)
   o. Violin plot

15- **Subplots:** You need to provide subplots in your report that tell a story to a reader. Pick a method discussed in class.

16- **Dashboard:** You need to create a web-based app that represent the dataset interactively. The app must be created using **Dash package** in python and then deploy to the Google Cloud Platform (GCP). You will need to provide **a link** in your report that displays the final dashboard worldwide. You need to have the following html and core components in your app:

   a. Multiple Division
   b. Multiple Tabs
   c. Range slider

d. Drop down menu
e. Button
f. Input field
g. Output field
h. Text area
i. Check box
j. Radio Items
k. DatePickerSingle (if applicable)
l. DataPickerRange (if applicable)
m. Upload component
n. Download component
o. Graphs : refer to 14 for the list of plots

17- **Recommendations:** This section of your FTP report provide a summary and recommendations after visualizing the dataset. Recommendation is an important section of your final report which could include the followings:
   a. What did you learn from various created graph in this project?
   b. How does the created python dashboard help users to gain information from the selected dataset?
   c. Is the created app user friendly? You can put the created app through a LinkedIn and ask people for comments on the created app and then add people comments inside your report. Make sure to get permission from owner of the commenters.
   d. Functionality: How functional is the created App?

18- A **separate appendix** should contain supporting python codes that is developed for this project.

19- References

20- The **soft copy of your python programs** needs to be submitted separately as a .py to verify the results in the report. Make sure to include the dataset in your submission. <u>Make sure to run your code before submission. If the python code generates an error message, 50% of the term project points will be forfeited.</u>

21- Include a **readme.txt** file that explains how to run your python code. All the results in your report must be regenerated to grant grade.

22- The FTP is defined to be individual unless an approval is granted for collaboration. All the coding must be done individually, and it must be genuine. Copying a code from internet without proper citation will be considered as a **<u>plagiarism</u>** and FTP grade will be disregarded. <u>Make sure to write your own code to avoid future complications.</u>

23- All figures in your report must include a proper x-label, y-label, title, and a legend [if applicable]. Pick an appropriate theme or style for the plotted graphs. If you have a table inside your report, then make sure to include a proper title. Including grid is optional.

24- **<u>Final presentation & demo :</u>** You will be given 20 minutes to present your term project and show to the classroom that the created web-based app works. The presentation weighs 20% of the term project grade. You need to create a power point for your presentation. Due date by **Wednesday April 27th**

25- **<u>Final formal report submission</u>** weighs 80% of the FTP and is due by **Wednesday, May 4th**.

Upload the **final <u>report (as a single pdf</u>**) plus **<u>the .py file(s)</u>** through BB by the due date.