THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

Visualization of Complex Data

DATS 6401

Final Term Project (FTP)

[display all numbers with 2-digit decimal precision]

Title = [font: 'serif', color:'blue, size: large enough]

X, Y label = [font: 'serif', 'color', darkred] size large enough

The objective of FTP is to apply the python visualization techniques covered in the course to a real-world dataset. Develop a python web-based dashboard (app) for interactive information visualization. Deploy the developed App through Google Cloud Platform (GCP) for production. The final term project has three phases:

Phase I: Static graphs & Tables[35pts]. You need to explore the dataset and reveal underlying information by plotting static plots for the following feature type. You may need to clean the dataset, if necessary, in this phase.

- Categorical features [must be ordered and numbers displayed on the graph if applicable]
- Numerical features

You are free to use any package that you feel comfortable with for this phase. The plots in this phase are static which means a plot cannot be updated without re-execution of the code. The following list of plots needs to be included in this phase for exploring the selected dataset.

  - Line-plot
  - Bar plot: stack, group
  - Count plot
  - Pie chart
  - Dist plot
  - Pair plot
  - Heatmap with cbar
  - Histogram plot with KDE
  - QQ-plot
  - KDE plot will fill, alpha = 0.6, chose a palette, chose a linewidth
  - Im or reg plot with scatter representation and regression line
  - Multivariate Box or Boxen plot
  - Area plot
  - Violin plot
  - Joint plot with KDE and scatter representation
  - Rug plot
  - 3D plot and contour plot
  - Cluster map
  - Hexbin
  - Strip plot

- Swarm plot

Plotting a static graph & tables for this phase is necessary but not sufficient. The **sufficient condition is the explanation of each plot and table** (pretty table or tabulate). A plot or table without explanation or inferences will receive zero credit. In this phase you also need to include subplots (at least 4) for storytelling i.e., multiple pie/bar plots in one subplot that tells a story about features inside the dataset then write your observations. You need to include legend (hue) in your static plot for comparison. All figures must include title, legend, x-label, y-label, and grid and they must be customized with assorted color, font size, line width, ... (see the top). Make sure the information on the graph is not blocked.

Phase II: Interactive web-based dashboard[35pts]. In this phase you will need to use Dash framework for interactive web-based dashboard. All the plots in the previous phase need to become interactive through a dashboard in this phase. In an interactive dashboard a user can update a plot without re-executing the code. A user must be able to switch the input, and the graph should be updated accordingly. There is no need to re-execute the python code to update a plot in this phase. You can use any items in the core component & HTML component (see https://dash.plotly.com/dash-core-components). The bare minimum items are.

a. Check list.
b. Dropdown
c. Graph
d. Loading
e. Download
f. Radioitems
g. RangeSlider
h. Slider
i. Tab
j. Textarea
k. Tooltips
l. Br
m. Div
n. Figure
o. Select H1-H6 as needed.
p. Header
q. Img
r. Label
s. Title

You need to design an interactive python-based front end [dashboard] that interactively plots the data. You need to start with the draft of the dashboard that is submitted in the proposal. You need to design the dashboard with multiple tabs that display the followings on the front end and be interactive: data cleaning [different methods that a user can select] , outlier detection and removal [different methods that a user can select], dimensionality reduction[PCA], normality tests [different methods that a user can select], data transformation[different methods that a user can select], loading data, plots of numerical features [user must be able to select the feature(s) and type of the plot], plots for categorical features[user must be able to select the feature(s) and type of the plot] , statistics .

Phase III: Deployment [10pts]: In this phase you need to dockerize the develop app and deploy it through goggle cloud platform for production. The GCP needs to be used for deployment. [if you want to use different cloud platform, you need to get a pre-approval]. You need to include the working world web address (www) of your term project at the top of the report.

If you need more clarification or need pre-approval, please come to talk to me. The required software for FTP is python only and the required IDE is PyCharm [only]. You can use Tableau in parallel for the verification of the results, but all graphs must be generated in python.

A formal report [latex or word in the .pdf form], presentation and demo of the created app using Google Cloud Platform (GCP) is required by the deadline. API format for citations.

SPECIFIES

The final formal report must be typed and should contain the following sections. You are free to place phase I, II and III requirements under an appropriate section in the report as shown below.

1- **Cover page.** Name of the school, author, instructor, date, course name and number.
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 [5pts]**: You need to provide a description of 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. (phase II)
7- **Pre-processing dataset [5pts]**: Data cleaning for missing samples, NAN's. Explain which method was used for data cleaning. You will need to display the first few observations of the cleaned dataset and the corresponding statistics. (phase II)
8- **Outlier detection & removal [5pts]:** Use one of the methods explained in class for the outlier detection and removal from the raw dataset. Write down your observations. (phase II)
9- **Principal Component Analysis (PCA) [5pts]:** Perform a complete PCA analysis of the cleaned dataset for a feature dimension reduction. Include the complete explanation in your report. Check the condition number and the singular values of the reduced dimension features. You can replace PCA with a random forest method. Write down your observations. (phase II)
10- **Normality test [5pts]:** Use one of the tests explained in class to see if the dataset comes from the Gaussian distribution or not. Write down your observations. (phase II)
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. (phase II)
12- **Heatmap & Pearson correlation coefficient matrix [5pts]:** Display the Pearson correlation coefficient between variables using heatmap and scatter plot matrix. Explain your observations through a table. (phase II)
13- **Statistics [5pts]:** 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. (see phase II)

14- **Data visualization:** Visualize the dataset using the static plots in phase II and discuss what can be observed from each plot. You need to write down your observations for each plot. (see phase II)
15- **Subplots:** see phase II.
16- **Tables:** see phase II.
17- **Dashboard [35pts]:** Take a screen shot of the designed dashboard [each tab separate screen shot] and place it into your report. The functionality of the design dashboard needs to be demonstrated during the presentation. see phase III.
18- **Observations:** For every graph in your report, you need to provide your observations. Plots without observations will receive zero points. The main objective is to use python to understand datasets and you must reveal underlying information using visualization tools. (phase II)
19- **The conclusion** is an important section of your final report which could include the following:

   a. What did you learn from various graphs created in this project?
   b. How does the created python dashboard help users to gain information from the selected dataset?
   c. Is the created dashboard user friendly? You can put the created app through professional social media [LinkedIn,] and ask people for comments on the created app and then add people comments inside your report. Make sure to get permission from the owner of the commenters.
   d. Functionality: How functional is the created dashboard?

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

Submission Notes

- 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. 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.
- Include a **readme.txt** file that explains how to run your python code. All the results in your report must be regenerated to grant the grade.
- The FTP is defined to be an individual or group [max two persons]. All the coding must be done individually, and it must be genuine. Copying code from the internet without proper citation will be considered **plagiarism** and FTP grade will be disregarded. Make sure to write your own code to avoid future complications.
- 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. For a proper title to the table developed inside your report.
- **Final presentation & Demo:** The presentation weighs 20% of the term project grade. You have 20 minutes to present your term project and demonstrate the functionality of the developed dashboard to the class. Your name will be called in a random fashion during the presentation day. If your name is not called (due to time constraint) then you need to record your presentation [need a PowerPoint] at home and submit the recorded presentation and the PowerPoint through canvas by the deadline. The due date by **Wednesday, April 23rd** .
- **Final report** [the pdf and the python file] is due by **Wednesday, April 30th.**
- Upload the **final report (as a single pdf)** plus **the .py file(s)** through Canvas by the due date.