DSO-599 Final Project (due by 3/1 at 6pm)

Learning Objectives:

- Articulating questions of interest before an analysis and establishing the relevance to the associated business.
- Assembling a data set and describing it at a high level using summary statistics and plots.
- Clearly describing the steps of an analysis and explaining how they address the questions of interest.
- Writing code using Python and Pandas correctly, efficiently, and in a readable way.
- Appropriate use of data analytics techniques with Pandas, including plotting, data assembly, handling of missing data, tidying data, analyzing text data, and using apply and groupby operations.
- Communicating effectively a technical analysis to a non-technical business audience.

Description

Each team will submit on Blackboard a Jupyter notebook that analyzes your chosen dataset following your proposal. The target audience is a senior manager of a business or organization who might be interested in the result of the analysis, and you should communicate in appropriate and convincing language why your analysis is relevant to the business or organization.

The Jupyter notebook should include the following five components (note that A, B, C, and E are analogous to the components of the proposal):

A. Motivation: a list of questions that can be (at least partially) addressed by analyzing the dataset and an explanation of why these questions are relevant to a business or organization.

B. Description of Dataset: a description of what is in the dataset and what each field means, along with code that showcases the data using basic summary statistics and plots.

C. Methodology: a description of how you propose to analyze the data, outlining the overall sequence of steps from a conceptual perspective, so that a trained analyst unfamiliar with Python will be able to follow along with your analysis and understand what you are doing.

D. Analysis: Analysis using Python code following the methodology above. Your code should be efficient, in the sense that if there is something that can be done with 1 or 2 lines, you shouldn't be trying it to do it in a roundabout way with 20 lines. The code should be substantial, meaning that it should demonstrate understanding of at least 4 out of 6 topics covered in last 6 sessions of the course (not counting the review and the exam). The code should be readable, meaning that you use appropriate variable names, not have overly long lines, and use comments or markdown as appropriate to explain what you are doing as you are doing it.

E. Discussion: a critical examinination of why your methodology and analysis is appropriate for your chosen questions of interest, and which parts of the analysis you think are the weakest links. If you had more time or resources, what additional data would you collect and why would they be helpful?

Your Jupyter notebook should be formatted in a way that looks professional, with appropriate headers and bullets, and free from spelling or grammar mistakes. You should limit tables, figures, and code outputs only to those that help you make your point, and avoid long and unnecessary outputs that may be distracting to a non-technical audience.

Grading Rubric

The project report will have a total of 42 points, with 3 points coming from each of the following 14 categories. The below description outlines what it would take to obtain full grade from each category. There will be a 0.5 point deduction for every minor issue and a 1 point deduction for every major issue.

1. Questions of Interest: Clearly describing questions that may be addressed by analyzing the data and phrasing each question in as precise language as possible.

2. Establishing Relevance: Clear and convincing explanation of why the above questions of interest are relevant to the business or organization. The explanation should demonstrate an understanding of the business application and what are the most important issues for the business.

3. Description of Dataset: Clearly and systematically describing what is contained in the dataset and what is not contained. If the data has many files or column headings, you should describe what each file and column contains.

4. Summary Statistics and Plots: Code that correctly loads the data and show appropriate basic statistics about the dataset, such as how many entries is in the data, what time periods does the data cover, what are the averages of the numerical columns, and how many unique elements is in each column. You should also give at least 3 appropriate plots that give a high level view of the data.

5. Clarity of Methodology: Clear description in Markdown cells of how you are analyzing the data step by step and why you are taking these steps. The descriptions should be understandable by a data analyst who is unfamiliar to Python.

6. Correctness of Methodology: Your methodology addresses the question of interest, with no signficant errors or gaps in the logic.

7. Correctness of Code: Your code is free from syntax or logical errors, and you should be using Pandas functions and Python constructs correctly.

8. Efficiency of Code: Your code accomplishes the desired analysis in the simplest way possible: if something can be done with one command or a few lines, you should not be doing it using 10 or 20 lines.

9. Substantial Code: Your code demonstrates understanding of at least 4 out of the following 6 topics covered in the Pandas portion of the course:

- Session 9: Plotting.
- Session 10: Concatenating or merging data.
- Session 11: Handling missing data.
- Session 12: Tidying data or transforming data types.
- Session 13: Analyzing text data using vectorized operations.
- Session 14: Using apply or groupby operations.

10. Readability of Code: Your code uses appropriate variable names and avoid overly long lines. You use comments and Markdown cells to explain what you are doing so someone unfamiliar with Python and Pandas can follow along.

11. Critical Examination: Your discussion section examines your methodology and analysis, and justifies why it is appropriate, while conceding the weak points.

12. Additional Data: A clear description of what additional data would be helpful for the analysis and why.

13. Appropriate Communication: Avoiding technical jargons and explaining technical words if you choose to use them. Keep in mind the target audience is a senior manager with a non-technical background.

14. Polish: Appropriate use of headers, bolding, and bullets inside Markdown cells. Report is organized in a easily readable way and not overly lengthy. Proper spelling and grammar. Avoiding long tables or code outputs that hamper readability.