Fall 2020 Final Project

Description:
Using the skills learned in this course, you will complete the following business intelligence project and report. Note, that this project has two components: 1) Project files and 2) Report. A skeleton (outline) for how the report should be delivered has been included in this document. Additionally, these datasets can be reused in your future classes, so it is highly suggested to choose well in your data source selections.

Task:

- This task will be performed in teams (Group you signed up for on blackboard).
- You are to select **3 or more datasets** that are more valuable when joined than when separate.
  - At least one of your datasets must be related to pandemic or epidemic events in human history.
    - i.e. you could pull from the current COVID-19, SARS, Black Death, etc.
  - *While you are not limited to the following options* for your datasets, here are some possible sources:
    - Web scrapes of websites
      - **NOTE**, should you choose to do a web scrape, you need to provide a .csv of the data scraped as well as the code used to pull down the data. In your report please detail the date, location, and methodology of the scrape.
    - Kaggle datasets
    - Govermental datasets
  - Your final datasets must contain the following:
    - More than 2000 observations. If you need to go below, please see me with your reasoning. Note, more than 2000 will make your future classes easier.
    - You must have both numeric and descriptor variables
    - You must have AT LEAST 5 variables that have meaningful correlations ($x > .15$ or $x < -.15$). Binary is acceptable. Please include your correlation matrix in your report with your meaningful correlations marked for ease of reading.
  - ***NOTE***, once you have identified your datasets, it is suggested that you receive verification from me whether your datasets will be valid for the project. If you choose to do this, I would ask for a subset of your data and a brief description of "how these datasets are more valuable together than separate" and "how you plan to merge these two datasets".
    - If you choose to take advantage of this opportunity, I ask that you contact me no later than September 23rd. Also, note, I will honor these request on a first come, first served basis. If everyone ask on the 23rd, I cannot guarantee I can get to you on the 23rd. This is my way of saying, contact me earlier if you would like this opportunity.
- You are then responsible for cleaning, merging, and producing a single cohesive output of the data. (You are free to use either Pentaho, Python, or a combination of the two).
  - **Note**, due to the number of datasets, it is possible you are wanting to run a comparison between sets that cannot be directly linked. If you go this path, your output should be no more than 2 final datasets.
- You are then tasked with visualizing the data using Tableau or Python
- You are then tasked to create a report that details your work performed and gives analysis on the quality of the data.

Expectation:
It is expected that you will work on this project with your defined team, and any work submitted will have been completed solely your defined team. Any form of Academic Dishonesty on this project will be met with the most severe punishment available.

Submission:

- All code files. Note, it is the expectation that I should be able to run your code with MINIMAL edits for a different environment.
  - If you have multiple please .zip them.
- All Data Files (These should be provided in .zip files)
  - Source Files
  - Post Cleaning Files
- Report: Submitted word document using the report submission guidelines.
  - Note, I do not have a hard-set page limit for this report. It should address the questions and the project. That said, I am not looking for a dissertation. If you have questions, please ask me.

Report Specifications: Your report will be the documentation of your code and should be written in a manner that another developer could pick up your code and report and continue to develop your project. To that end, I ask that you provide the following items:

- Analysis of data – In this section you need to discuss issues related to the data you are gathering. While you are free to address any issues you see fit, I do ask that you address the following:
  - What questions can these data potentially answer?
  - What are the potential valuable data items exists within the data?
  - How might they be applied for direct business application and indirect business applications?
  - What do you suggest as potential usages for different variables within the dataset?
- Data Cleaning – In this section you need to discuss issues related to the quality of your data. While you are free to address any issues you see fit, I do ask that you address the following:
  - What is the overall quality of the data?
  - What variables contained missing data?
  - What kinds of missing value exists in the dataset and which variables are they related to?
  - What methods did you use to clean the missing data?
- Data Merging – In this section you need to discuss the methods you used to merge your datasets. While you are free to address any issues you see fit, I do ask that you address the following:
  - What were the common elements between both datasets?
  - Were there any issues with multilevel measurement in the final dataset?
  - What variables are more valuable combined than being in separate datasets?
  - In what ways has the data become more valuable since being merged? i.e. what new business insights can be generated due to the combined datasets rather than the sets being separate.
- Analysis of Visualizations – In this section you need to discuss the methods you used to visualize your datasets. While you are free to address any issues you see fit, I do ask that you address the following:
  - How well does your visualization adhere to the principles and characteristics of a good visualization?
  - How well does your visualization adhere to the concept of natural processing? Are there things in your graph that are necessary but do not have a natural processing correlate?
  - Copies of your visualizations.
  - **PLEASE NOTE**, rarely does a visualization adhere to all principles and invoke perfect natural processing. Please take the time to address how it does and how it doesn't. This isn't about having a perfect visualization, it is about conveying your knowledge about good graphs.
- Correlation matrix marked with meaningful correlations and described.
  - i.e. don't just paste a matrix, please discuss this in your paper.
- Flow diagram of your project. (draw.io would work very well for this).
  - This should illustrate all steps necessary to gather the data to visualizing the data. It should also illustrate which "files" and "language" are being used at each step.
- Instructions for code – This section should detail how to run your code.
  - Note, this should be explicit step-by-step instructions, including any variables that might need to be manipulated.
- Report Quality – It is expected that this should be written as a professional document using correct grammar and layout.

Points Breakdown: (Note, this is out of a total 50 points)

- Peer Evaluation – 10 points
- Report – 40 Points
  - Report Quality – 5 Points
  - Report content – 20 points
  - Functional Code – 15 points