Data Analytics Term Project

1 Description

The objective of this project assignment is to create a comprehensive data analytics project utilizing a real-world dataset. Your primary responsibilities will include selecting a suitable public dataset, cleaning and preparing the data, and applying one or more data analytic techniques from the course to address your research questions.

Your project can focus on various data analytics models such as simple data Manipulation questions or a Hypothesis Test, or any other relevant techniques that we have learned in class.

To successfully complete this assignment, it is crucial to document each step of the data analysis process and to provide clear and insightful explanations of your project results. This will not only showcase your technical skills in data analytics but also demonstrate your ability to communicate complex data-driven insights to a broader audience.

2 Public Dataset List

First select a data set from the available public datasets.

Some of the important data collection links are:

- Machine Learning Data sets https://archive.ics.uci.edu/
- https://www.kaggle.com/datasets
- https://www.data.gov/

Notes:

- The data set must not be very Large.
- You can reuse one of the assignment data sets.

3 Define a Project (2 Points)

Define a data science research question based on the selected dataset. You can select one of the listed datasets or search on the web, and pick up one of the available public datasets.

You should describe the following items

1. What is your data set about?
2. Clean up your data and reduce it to no more than 2000 observations if your data set is very large. The size of the dataset should be a dataset that you cab process on your laptop.

3. What is exactly your research question? What do you want to learn from data? What is your learning model ,e.g., a Data Wrangling/Data Manipulation question, or a Hypothesis Test for example a t-test?

4. What is your current expectations about the results? Why?

5. Describe how you evaluate your project? How to access the correctness of your model? How well would you expect that the model will work?

4 Implementation (6 Points)

- You need to implement your project in R or Python. You are allowed to use any Libraries. You are also allowed to implement your project without using any libraries.
- You code should be completed and be compilable without any errors. We should be able to read your documents, and be able to run your project.
- Run your implementation (on your Laptop) and generate the results.
- Provide the interpretations of your results.
- What can you do to improve your results? Apply your ideas to improve your results.
- Provide any references that you have used.

5 Prepare a Presentation of your Project (2 Points)

- Create a document or a presentation to describe your project and results
- Describe your code.
- Describe the results of your project in a professional way.
- You may want to visualization diagrams and describe the results based on some diagrams - but having diagrams is not a MUST have to get the full credit.
- Describe the model and results of your project in a way that every person in this field can read, enjoy and understand.

Grading will be based on quality of your presentation, and correct describing of algorithms or concepts.

5.1 Turnins

- Submit your Jupyter Notebook or Python code
- Submit your documentation (It can be part of your Jupyter Notebook or a separate file like PPT or Docs, or other formats)
- Submit your Video Presentation. Upload your video file to Canvas Assignment. We will later download it and upload to Panopto Video Gallery for the entire class.

5.2 Grading Criteria

Your term project will be graded based on the following rubrics:

1. Correct description of the selected dataset.
2. Correctness of your data preparation and cleanup approach.
3. General Correctness of data analysis approach.
4. Correctness of your implementation and results.
5. Correctness and completeness of the interpretations of your results.
6. completeness of all provided references including all code and idea sources.