CSCI2000U - Scientific Data Analysis

Final Project Instructions

This is a **group assignment**. You can find your group on Canvas/People/Project Group.

This project is student-driven, based on your chosen dataset and your project proposal. Those group members who are proven to not contribute are subject to be graded differently.

You MUST give appropriate credit to any external resources you might use in your project. Use APA for your citations.

Goals

1. Demonstrate analytical and programming skills by performing and reporting data analysis on a student-chosen dataset.
    a. Programming skills are demonstrated by a Python jupyter notebook containing the data processing.
    b. Analytical skills are demonstrated by a written report.

Introduction

At this point you have an idea of your dataset format, how you can process it, what are some research questions you are interested in. As a group you should collaborate to process the data in order to answer your proposed research questions.

For this assignment you will not be limited to specific techniques or methods, but you will still be asked to deliver specific items.

We hope you have chosen a dataset that inspires you, because you are responsible to make us perceive this dataset as interesting too.

One of the deliverables of this project is to write a report in a blog post style. This can be used in your professional portfolio with consent of your group members.

CSCI2000U - Scientific Data Analysis

Technical Report and Data Analysis Requirements

In this report (notebook) you will expand on your proposal data analysis and *it should reflect any feedback or suggestions made on your proposal*. This type of data analysis is called Exploratory Data Analysis (EDA), at this point we have attained an understanding of the main computational tools to perform EDA.

You may choose the most appropriate methods or libraries to perform your data analysis (NumPy, Pandas, Functional Programming), you might also find it convenient to mix and combined methods from each of those libraries at a time. This document should:

- Be a Python Jupyter Notebook.
- be organized with your code and documenting it using Markdown cells or comments where is pertinent to do so. One should be able to readthrough it and understand what was done.
- Outline any external resources like files, or datasets used.

The data analysis will include at least the research questions you have proposed. You are expected to find more follow up questions or more questions as you answer your initial ones. The goal is for you and your group to become experts in this data.

Document outline

1. **Introduction:** Explain why you chose the topic, the questions you are interested in studying. List team members and a description of how each contributed to the project.
2. **Description of data:** Describe the dataset, how was it collect, how you accessed it, references/credit to source.
3. **Analysis of the data:** Provide a detailed, well-organized description of data quality, including the features, any data that should be cleaned or pre-processed before you EDA.
4. **Exploratory Data Analysis:** Provide a detailed, well-organized description of your findings, including textual description, graphs, and code. Your focus should be on both the results and the process. Include, as reasonable and relevant, approaches that didn't work, challenges, the data cleaning process, etc.
5. **Potential Data Science:** Based on your data analysis and findings. Describe any potential ideas if you were to pursue a data science or machine learning project using this dataset. If you don't find any potential, explain your rationale.
6. **Conclusion:** Discuss limitations and future directions, lessons learned, maybe things you did not predict to find out or things you learned as you performed the analysis.

CSCI2000U - Scientific Data Analysis

Notes

- You are encouraged to be as intellectually honest as possible. That means pointing out flaws in your work, detailing obstacles, disagreements, decision points, etc.
- You may use the first person ("I"/"We") or specific team members' names, as relevant.
- You don't need to include all graphs, or print all data rows. You should choose what needs to be printed to provide context/information while making the technical report readable.
- This report should be clear and concise, so that anyone who took this course (your audience) can read and understand it.
- How long should it be? It should take the reader 10-15 minutes to read it.

Other resources that might be helpful

Tips for performing exploratory data analysis and reporting for technical audiences:

- https://medium.com/mlearning-ai/basic-exploratory-data-analysis-template-for-regression-problems-20ca00c58f7d
- https://towardsdatascience.com/exploratory-data-analysis-eda-a-practical-guide-and-template-for-structured-data-abfbf3ee3bd9

CSCI2000U - Scientific Data Analysis

Blog Post Report (readme.md)

This is a written report in a scientific blog style using Markdown, and it should be a short non-technical summary of the most revealing findings of your technical report analysis, but written for a non-technical audience. This will tell the story of your dataset and your findings, it will probably contain some specific tables or graphs created by you, however, the specific code used to generate those tables/graphs are in your technical Python Notebook.

The goal of this report is to communicate your research to a curious reader, that might not have taken this course, or has the technical background to be interested in the code. The reading time should be around 5-10mins. See the expected sections:

Title

- You should title your report to make it enticing for readers

About us

- You should list all authors names and respective Github names.
  - Do not list student IDs as this will be public.

Introduction

- You should introduce the topic you are reporting one.
- You should describe the dataset(s) including giving credit to their sources.
- Depending on your writing style, you may or may not add personal anecdotes to make the reading more relatable, or motivate readers to continue reading.

Discussion

- You should pick and choose the most interesting findings of you research based on your data analysis. You will offer evidence of those findings by adding selected information. This is your discussion, this should be at least 50% of your report.
- You will only present the evidence to add to the narrative. The computation and creation of graphs will reserved to your technical report.

Conclusion

- You should conclude your report.
  - In the discussion, you are presenting the findings. In the conclusion you will summarize what was discussed. To help you structure your conclusion:

        - You should have a “Reflection” (list things you learned as you worked on the project) and “Refinement” (if you were to improve the project, what would be the next steps) subsection in the conclusion.

Acknowledgements

- You should add an acknowledgements section
    - *“This project was submitted as the final course project for CSCI 2000U “Scientific Data Analysis” during Fall 2021. The authors certify that the work in this repository is original and that all appropriate resources are rightfully cited.”*

README

At the end, you should have a “traditional” readme section

- This will outline any instructions needed to run and reproduce your results.

Others resources

Some examples or reports that are more approachable and less technical:

    - https://medium.com/analytics-vidhya/how-starbucks-users-behave-on-offer-ffbc17367094
    - https://medium.com/analytics-vidhya/what-drives-the-rental-price-of-homes-and-rooms-for-guest-accommodation-496d7726d20

Your GitHub repository

You should create a GitHub repository where all team members are collaborators to.

- Your repository is public for submission.
- You can name you repository to reflect your project topic
- Your repository contains all files your project needs including the Technical report and readme outlined above.
- Your Blog Post Report is the readme.md of your repository.

CSCI2000U - Scientific Data Analysis

Submission

You will be submitting your public **Github project URL** via Canvas. More details:

- 1 submission per group
- If you have **approved uneven grading** (see section below) between the group members, you should add a comment in the submission with all the group members' names linking to their percentage of contribution to the project.

Group Grading and Collaboration Policy

All groups members are expected to fairly collaborate and work on the project.

If you are having issues contacting or getting help from any of the members, please follow these steps:

- Create a group chat on MS Teams including the instructor and all group members.
  - This will be a fair warning and an opportunity for all group members to commit with the assignment.
  - If a member still fails to fairly collaborate, this member may be graded separately.

If you don't follow the steps outlined, I will not be able to submit uneven grades.

More helpful resources

- General tips for writing and analyzing your findings into a report
  - https://towardsdatascience.com/the-ultimate-guide-to-writing-a-data-based-report-6e9703dcc095
  - https://www.kaggle.com/jpmiller/creating-a-good-analytics-report
  - https://jgscott.github.io/teaching/writeups/write_ups/
  - https://blog.datumize.com/how-to-write-a-well-structured-informative-data-analysis-report

Examples of data science project reports written in an academic paper style. This is not the style we are going for, these are given just an inspiration and examples of data science projects.

- https://cs229.stanford.edu/projects2014.html