CPSC380 Introduction to Data Science

Term Project

1 About Term Project

The course term project will be implemented individually. You will need to find out dataset online, use your data science knowledge and skills to do data cleaning and processing, do exploratory data analysis, and present your work. The project is worth 30% of your course grade, which will be split into three parts: (1) project proposal (2) project presentation (slides and program demo) and (3) final source code and report.

Your project will be evaluated by two main criteria:

- **Significance of your project**: How significant of your project? How complex of your project? How many aspects, angles, variations did you explore? This will mainly be evaluated through source code (jupyter notebook).
- **Presentation**: How well did you explain what you did, your results, and interpret the outcomes? Did you use the good graphs and visualizations? How clear was the writing? This will be evaluated through your oral presentation and final report.

2 Important Dates

Task	**Deadline**	**Note**
Proposed project	03/31/2023 (Friday)	1 page description of your project, including brief introduction of your project, and the dataset link (i.e., URL) you will investigate.
Presentation	04/24(Monday) 04/26 (Wednesday) 04/28 (Friday)	- 10-minute presentation of your work, through slide presentation and jupyter notebook demonstration, - 2 minutes Q&A
Final report + source code	05/05/2023 (Friday)	- 5-6 page summary report of your project - Source code

3 Project Dataset

3.1 Data Sources

Below is a list of websites you may find the dataset for your course project:

1. Data.Gov: `http://www.data.gov`. This website is the home of the U.S. Government's open data.
2. UN Data: `http://data.un.org/Explorer.aspx`. This website is the home of the United Nations (UN) open data.
3. World Bank: `https://datacatalog.worldbank.org/collections`. This website is the home of the World Bank open data.
4. The Global Open Data Index : `https://index.okfn.org/place/us/`
5. Any other interesting datasets applicable for data analysis.

3.2 General suggestions on dataset selection

When you select dataset for your project, you may consider the followings things:

- Dataset size should be reasonably large, in terms of number of rows and columns.
- Multiple related CSV files may be collected (instead of one), unless the dataset is very large (contains large numbers of rows and columns).
- You may need to data cleaning and data merge if necessary. For instance, the following dataset needs data cleaning:

  `https://catalog.data.gov/dataset/covid-19-blueprint-for-a-safer-economy-data-chart-archived`

3.3 Exploratory Data Analysis

When you do data analysis, you may do the following tasks:

1. Data cleaning and data preparation
2. Data merge
3. Date grouping and aggregation
4. Data transformation
5. Statistical Analysis
6. Data visualization

Note: If you analyze the dataset that are involved with the techniques not covered in course lectures, you definitely get higher points for the project.