Project Requirements

Applied Statistics and Data Analytics

April 8, 2022

Objectives

This project is intended to equip you with the following skills:

1. Learn to apply proper data analysis skills to practice;
2. Be able to use R to analyze spreadsheet data;
3. Be able to present the results of data analysis in a concise and well-structured report.

Grouping

You will be grouped by yourselves and each group consists of **three students at most**. Each group picks one student as the team leader. The team leader is responsible for handing in **the team member list** and **the final report and poster**. You can divide up your work by group discussion but please notice that **each team member should responsible for some work of data analysis instead of only devoting to the report and poster**.

Requirements

1. Dataset: Each group chooses one dataset for further analysis and you can choose from the datasets we provide. You are also welcome to use your own dataset too, but you are supposed to discuss with TA about the feasibility.
2. Methodology: The data analysis task involves finding relationships between multiple factors. The provided datasets can be used for a prediction/classification/cluster purpose. Your analysis may include (but is not limited to) the following steps.
    - Description of data, including statistical description of each factor (better supplied with graphical representation);
    - Preprocessing of data (e.g., changing format of dataset to make it readable by R, dealing with missing values, transforming data if necessary, etc.);
    - Finding pair-wise correlation of all factors, and identifying potential colinearity problems;
    - Applying data analysis methods (e.g., linear regression, LASSO, ridge regression, logistic regression, decision tree, random forest, support vector machine, k-means, Gaussian Mixture Model, etc.) for prediction/classification/cluster;
    - Performing model diagnostics;
    - Visualising data analysis results.

3. Deliverables: (1) a clear and complete report; and (2) a concise and well-structured poster. Both of the report and poster need to present your work briefly, including a narrative, data analysis methods, and your key findings. **Each team member needs to clearly identify the data analysis work they are responsible for in the report.**

    - Other format requirements of the report includes:
        - The report should be written in English;
        - Pages: (a) the main content $\leq$ 8 pages; (b) the appendix except for references and division of labor $\leq$ 3 pages (the appendix is only used as supplementary materials to understand the main context when grading, not as the basis for grading);
        - Font: Times New Roman with size 12;
        - Line Spacing: single;
        - Page Margins: Top 2.54cm, Bottom 2.54cm, Left 2.54cm and right 2.54cm;
        - Document Format: .docx or .pdf.
    - Other format requirements of the poster includes:
        - The poster should be written in English;
        - Tool: you can use PS/PPT or other poster makers to make your posters;
        - Size: 594mm (width) $\times$ 840mm (height);
        - Resolution: 300dpi;
        - Color Channel (if PS): CMYK;
        - Document Format: .jpg or .pdf;
        - Content & Layout: no specific requirement, but please refer to good academic posters.

Evaluation

1. Review process: (1) the evaluation by the professor and two teaching assistants; and (2) peer review by students. The peer review process includes two parts:

    - **Inter-group review.** We will hold an exhibition to include our project posters from the class at the end of the term. Each student is responsible for reviewing a poster of the other team. The TA will randomly generate assignment of teams. **Average quality** of the review reports by team members accounts for 5% of your project grade.
    - **Intra-group review.** The team leader needs to handle in the division of labour together with the final report. Each student needs to evaluate the performance of both yourself and other team members.

2. Scores calculation:

    - Basic team score = 60%*Report + 35%*Poster + 5%*Inter-group review + ($\leq$ 5%) Bonus

    - Final individual score = 80%*Basic team score + 20%* Intra-group review

Timeline

The key milestone dates are listed as the following table. The TA will also remind you and publish notice in the WeChat group and Web Learning when appropriate.

Table 1: Tasks and due dates

Key Tasks	Due Date
Release of project requirements	April 8, 2022 (Week 7)
Grouping and choice of dataset	April 15, 2022 (Week 8)
In-process inspection	May 14-15, 2022 (Week 12)
Submission of poster	June 17, 2022 (Week 16)
Submission of report	June 25, 2022 (Week 17)

List of optional projects

- 01-Song Popularity
- 02-Superstore
- 03-FIFA2018
- 04-Personal Key Indicators of Heart Disease
- 05-Human Stress Detection in and through Sleep
- 06-Risk Score for Unique Postcode
- 07-TikTok Trending Tracks
- 08-World Energy Consumption

Document List

- Project Requirements;
- Eight Candidate Datasets: each dataset includes .csv data file and the data description;
- Report Guideline.