EMIS 3309: Information Engineering

Spring 2021

Group Project

The aim of the group project is provide hands-on experience on learning information from past data. For this project, you will pick a dataset for which you believe there are interesting questions to answer. You will then apply methods we learned in this course to understand the data, and to find the best way to answer theses questions.

- Group size: ±4 students

The project consists of **three** deliverables.

1. Proposal for the project (1-2 pages, 12 pt font, single-spaced, pdf)

   Due Thursday April 1, 2021 (5 % of total points)

   (a) Members' names and student ID numbers.

   (b) Source of the dataset.

   (c) Description of the dataset including dimension, names of the features (and response), and their descriptions.

   (d) Description of the problem. Do some research to understand the background of the dataset and describe what you aim to learn from the data. Define your goals clearly.

   (e) Which types of statistical learning methods are you planning to apply? What would be the response feature?

   (f) What are some of the potential difficulties?

2. PowerPoint presentation (10 minutes)

   April 27 and April 28, 2021 (on Zoom) (5 % of total points)

   (a) Description of data and the questions that you are interested in answering.

   (b) Summary of dataset. Describe issues/findings you had for the dataset and explain how you handled them.

   (c) Summary of statistical learning approaches you tried and why you chose them.

   (d) Summary of results.

   (e) Conclusions / recommendations.

3. Project report (8 pages limit, 12 pt font, single-spaced, pdf)

   Due Monday May 10, 2021 (15 % of total points)

   (a) Title of the project
   (b) Executive summary highlighting your key findings
   (c) Problem description
   (d) Data cleaning
   (e) Data exploration
   (f) Statistical learning methods
   (g) Conclusion
   (h) References

Below are some suggestions:

- Problem description

  - Define your goals clearly and focus on the problem.

  - Setting target audience / client may help to clarify the problem.

- Data cleaning

  - Identify outliers, missing values, measurement error and abnormality.

- Data exploration

  - Give simple and appropriate statistics for features.

  - Visualize relationships between features.

  - Can you identify significant or insignificant features?

- Statistical learning methods

  - Select appropriate methods for your problem.

  - Try multiple methods to achieve your goal and compare the results.

  - Graphs help readers to understand the results.

Submission guidelines:

- Submit the proposal on Canvas as a group

- Submit the report on Canvas as a group

  – Additionally, one member of each group should email a zip file (including cleaned dataset, R code and PowerPoint slide) to `mijua@smu.edu`. Make sure to CC all your group members.

Some public data repositories:

- Government data: `https://www.data.gov/`
- Kaggle: `https://www.kaggle.com/datasets`
- The World Bank Open Data Repository: `https://data.worldbank.org/`
- University of California, Irvine Machine Learning Repository: `https://archive.ics.uci.edu/ml/datasets.html`
- More resources can be found at: `https://r-dir.com/reference/datasets.html`

Note: You may find data from other sources. Make sure to check data usage right and cite your sources!