Data proposal due: Monday, Feb 26th 2024 11:59pm on Canvas
Exploratory analysis due: Thursday, March 7th 2024 11:59 pm on Canvas
Project report due: Friday, March 22nd 2024 11:59 pm on Canvas

Summary:

- The minimum group size is two, and the maximum size is three. You will be forming your own groups. Only one submission of each part per group is necessary.
- Each group will submit a
  1. data proposal,
  2. exploratory analysis, and
  3. report that are outlined herein.
- The proposal, outline and report will be 10%, 10% and 40% of your final grade, respectively.
- The dataset to be analyzed should be a new dataset (i.e. must not recreate an analysis has been previously analyzed in a textbook or paper). One location you could look for data is the UCI Machine Learning Repository. Data from a previous job is a great idea, but make sure you have permission to use it. If you build your own dataset (from, say, the internet), I will be impressed and interested. You must reference where your data comes from and describe the data collection process as part of your report.
- Each document should not include in-line code. Instead, code should be included as a Python notebook in the appendix of the report. The notebook should be able to fully reproduce the analysis in the report, including creation of figures and calculation of metrics.
- Your report should explain your analysis in a way that someone with a minimal background in statistics and data science can understand your findings. At the same time, it should be detailed enough that knowledgeable readers can understand your process.

The purpose of this project is for you to:

- apply new concepts from class to conduct modelling and analysis on real data,
- practice writing a statistical report, and
- develop your scientific communication and collaboration.

More specific details on expectations:

Data proposal:

- Guidelines: The purpose of the proposal is to find a dataset to analyze and summarize the data, identifying any challenges that may exist within that dataset. The proposal will contain:
  - a statistical description of the dataset; including but not limited to data types, structures, distribution, and so on, (3.33%)

  - who, what, when, where and how the data was collected and any underlying scientific processes that may affect the data, (3.33%)
  - what scientific questions about the you will try to answer, (3.34%) and
  - if you plan to collect your own data, it must be constructed or in construction by proposal submission.

- Length: 2 pages max. Font size minimum is 10. If you have to change the font size to condense the proposal to less than two pages, the proposal is too long.

- Submission format: Proposal submissions will be in pdf file format only. In an appendix, you must include a raw extract of the dataset for review. I will be giving feedback on your proposals; the better you write the proposal, the better the feedback and higher probability of success on the future steps.

Exploratory analysis:

- Guidelines: The purpose of the exploratory analysis is investigate the data more thoroughly and then describe your planned analysis. The exploratory analysis is meant to investigate the data and decide will contain:
  - A statistically descriptive analysis of the dataset; give a detailed description about the data including the number of variables, variables types, summary statistics, graphs of data, etc., (2.5%)
  - applications of statistical analysis techniques learned in previous coursework, (2.5%)
  - what scientific questions you will try to answer, (2.5%)
  - the statistical analysis techniques you will use to answer those questions (with justification), (2.5%) and
  - (optionally) any preliminary results you would like reviewed.

- Length: 10 pages max, including figures and tables.

- Submission format: Proposal submissions will be in pdf file format only. You will also re-submit to me your dataset for quality assurance. I will be giving feedback on your projects; the better you write the analysis, the better the feedback and high probability of success on the report.

- Graphs, figures, and tables: All graphs must be labelled correctly and readable [font sizes consistent from figure to text]. Figures and tables must have numbers and descriptive captions, and must be referenced in the text with explanation.

Project report:

- Guidelines: You are expected to write a complete analysis of your dataset using techniques from class (in concert with techniques from previous courses), and write up the results in a comprehensive report. The report will contain:

  - A discussion of the scientific hypotheses you will investigate.
  - An analysis that addresses your scientific hypothesis, using model building techniques you have learned in this course, as well as other courses. Model and data diagnostics for appropriateness are expected. Plots and tables are highly encouraged, where you need to include the interpretation for each plot/table.
  - Conclusions and recommendations: give your conclusions of your outlined hypotheses based on your analysis such as important variables identified, the most appropriate model you have discovered, how statistical assumptions may be violated and how they affect your results, difficulties faced when modelling and shortcomings of the current model, interesting findings, etc..

- Graphs, figures, and tables: All graphs must be labelled correctly and readable [font sizes consistent from figure to text]. Figures and tables must have numbers and descriptive captions, and must be referenced in the text with explanation.
- Length: Maximum 16 pages including figures and tables. A concise and comprehensive analysis is ideal; a 10 page report would be a fine submission. A 16-page report that meanders through too many hypotheses would lose marks due to a lack of focus.
- Submission format: The submission will consist of the following files:

  (a) a compiled pdf,

  (b) the dataset being analyzed,

  (c) a Python notebook to re-create the analysis,

  (d) any other files I need to compile the project.

- Compiling code: I should be able to your code from my own computer using only the files provided. If I am not able to compile the project with a reasonable amount of effort (~ 5 minutes of menial debugging), you will be deducted 20% off the report's final mark. If I have to install a package to run the code or load a large dataset or calculate an analysis using your code, this is no problem. If I have to debug errors in your code, that is a big problem.
- Grading rubric: See the file `data583_projectRubric.pdf` for full details.

If you need guidance or assistance with any parts of the project, see me early and often. I am here to help you learn to write a **good** analysis of a dataset to be read by interested stakeholders. This is a skill employers (especially in data science) are looking for! **Do not wait until the last minute to ask for help!**