Semester Project

Project Specifications:

- The instructor will place you into a group of 2-3 students
- Pick a **data set** that you and your group find interesting. (Example source: UC Irvine Machine Learning Repository. Feel free to select your data from any other source as appropriate.)
- Form a **research question**
- Perform **data pre-processing**, data cleaning, outlier removal, and so on to sanitize your data as necessary.
- Save your data in a .csv file (or other format as appropriate for your data set and project scenario).
- Read in data to your program from the data file.
- (*Optional* - do as appropriate) Process the data or perform any calculations or statistics on it before storing the data into a data frame (see next step).
- Save the data into one or more data frames (or other structures as appropriate)
- Once you have stored your data, **explore your data** to reveal interesting/useful information based on your project scenario.
- Write at least two unit tests. For example, these might be short tests to show that two different functions work as intended.

Deliverables:

1. WRITTEN REPORT (no more than 10 pages) containing:

- **Abstract**: Paragraph outline describing your question, what you did, and what you learned
- **Introduction:** Describe your project scenario. Starting out, what did you hope to accomplish/learn?
- **Data description:** Describe your data set and its significance. Where did you obtain this data set from? Why did you choose the data set that you did? Indicate if you carried out any preprocessing/data cleaning/outlier removal, and so on to sanitize your data.
- **Data processing methodology**: Describe briefly your process, starting from where you obtained your data all the way to means of obtaining results/output.
- **Results:**
- - Show at least two visualizations
- - Display and discuss the results. Describe what you have learned and mention the relevance/significance of the results you have obtained.
- **Testing:** Describe what testing you did. Describe the unit tests that you wrote. Show a sample run of 1 or 2 of your tests (screen captures or copy-and-paste is fine).
- **Conclusions:** Summarize your findings, explain how these results could be used by others (if applicable), and describe ways you could improve your program. You could describe ways you might like to expand the functionality of your program if given more time.

2. PRESENTATION

- Each group will give a presentation not to exceed 10 minutes
- The presentation should briefly include:
  - research question
  - data summary
  - data processing methodology
  - visualizations
  - results
  - conclusions

- The presentation file format should be powerpoint or pdf
- The file name should begin with GroupName[n]_ where [n] is your group number

Be sure to practice beforehand, and time yourselves.

3. CODE

- Clearly document, organize, and name your code file or files
- The files can be in Jupyter Notebooks or Python scripts

Submission:

- **In one Zip file submit through Collab**: (1) written report (2) presentation (3) code files

Rubric

Description	**Possible Points**	**Earned Points**	**Comments**
Paper includes abstract	10		
Paper includes introduction	10		
Paper discusses data source and provides data summary	10		
Paper discusses data preprocessing	10		
Paper includes at least two visualizations	10		
Paper includes results, clearly shown	10		
Code presents/discussed unit tests	10		
Code is clear and well-documented	10		
Presentation skills and video • All group members	20		

spoke in the video - The live session presentation was under 10 minutes - The presentation was of good quality, clear and easy to understand			
Total Points	100		