**EGR 361 Final Project**

We will work on the project in class for three days (labelled below as Day 1, Day 2, and Day 3). If you need additional time to complete the project, you will need to do so outside of class.

This project is worth 10% of your grade.

All documentation for the project will be compiled into a final report that you will upload to Moodle.

Please ensure that your documentation is well organized, easy to follow, and that all questions have been answered. Screenshots should be clearly labelled, with any takeaways described in the text or a figure caption.

Please use Python for the coding sections, even if you are familiar with other software tools.

**Final Project Proposal: Dataset and Description**

Prior to the days we work on the final project in class, you will need to perform the following:

- Find an open-source dataset online that is relevant to your field and interesting to you. It needs to have at least 50 data points/examples.

    - Helpful links:
        - https://library.bu.edu/datascience_engineers/find_datasets
        - https://www.mathworks.com/help/stats/sample-data-sets.html
        - https://www.kaggle.com

    Note: if you've spent a good amount of time looking and can't find a dataset that matches your specific interest, try the opposite approach. See what datasets are available and choose one that you are interested in and would work well for the project.

- On a single page, document the following, then upload to Moodle:
    - Where you found the dataset
    - How it is relevant to your field
    - Why you find it interesting.
    - What you want to predict with your dataset
    - Which features you plan to use in the prediction

**Directions for Day 1:**

- Explore your dataset

- Document the following:
  - What format did the dataset come in? (e.g. text file, .csv file, etc.)
  - What type of data is available? Is it continuous or discrete?
  - How was the data generated? By whom? What did they control for?
  - How many samples/examples/datapoints are in the dataset?
  - What is the name and data type of the label or "true value" in the dataset?
  - What does the distribution of the label look like? What are the descriptive statistics? (Mean, median, range, standard deviation)
  - How many variables are available? What are their names? How are they relevant to the problem?
  - What do the distributions of the variables look like? (Attach screenshots)
  - What are the descriptive statistics for each variable? (Mean, median, range, standard deviation)
  - How correlated are each of the variables with the labels?
  - Are any of the variables highly correlated with each other?
  - Any code or formulas used to answer the above questions.

**Directions for Day 2:**

- Complete any outstanding tasks/questions from Day 1

- Following the pattern shown in class, format your dataset so that it can be used in a supervised linear regression problem.

  - Document the following:
    - Any standardization or normalization procedure used
    - The dimensions of your training data
    - The dimensions of your testing data
    - Any code or formulas used to answer the above questions

- Perform linear regression to create a prediction model and evaluate its performance using $R^2$ and RMSE.

  - Document the following:
    - $R^2$ and RMSE values on your training and testing sets
    - Which features were most important in creating your model? Does it make sense that those features were important?
    - Plot your model over a scatterplot of the most important feature. Use a different color to plot points used for training and testing. Attach a screenshot of this plot to your report.
    - Did any of your test datapoints fall outside of the range used in the training data? If so, how good was the prediction on those points?
    - Describe qualitatively the performance of your model.

        - What ideas do you have on how the performance of the model could be improved?
        - Any code or formulas used to answer the above questions

**Directions for Day 3:**

- Complete any outstanding tasks/questions from earlier in the week

- Discuss your setup and results with another classmate. Did they see any red flags in your setup? Make adjustments if necessary and re-run your results. What changed?

    - Document the following:
        - Who you discussed your setup with, their feedback, what adjustments you made (if any), and what impact that had on your outcomes.
        - What (albeit preliminary) conclusions can you make based off your analysis?
        - What have you learned from this whole process? Be specific and try to find something meaningful – don't just regurgitate what's already in your report.