The objective of the course project is to apply all the skills and techniques learned during the course to a real-world dataset. In this project, you will apply the machine learning skills covered in the class by training a machine learning model on a real-world dataset. You will:

1. Find a real-world dataset of your choice online
2. Parse, clean and analyze data
3. Create visualization
4. Ask and answer interesting research questions about the data
5. Use any ML algorithms to tarin your model and comment on the performance of the model.
6. Suggest the method for improvement of the performance of the model.
7. Prepare a final report about the findings.

Evaluation criteria:

* Dataset must contain at least 5 columns and few thousand rows of data
* You must ask and answer at least 5 questions about the dataset
* EDA must include some statistics and visualization
* Must include explanations apart from code
* Must not be plagiarized i.e., copy-pasted for somewhere else. Can take some part of code and try to modify and create your own.

**Target Audience**

Be sure to target your report at people who may not know your area/domain well. Remember that you want to convey that you have done a lot of analytical work and that rigor should be conveyed, but not necessarily the calculations. It is important to start off so that the reader is interested. What aspect of this analysis interested you? You could have chosen anything, but you chose this particular topic. Why?

**Breadth of Analysis**

From a grade/score point of view, it is important to touch upon as many different aspects as you can. These include the premise, asking the right questions, data manipulation, analysis, visualization, and conclusion.

**Length of your report**

Typically, your report will be about **5-7** pages long. You can always include an appendix if you have some supplementary material that you wish to include.

Note that several components have come together to make a good final project. It starts with obtaining data (it always starts with this!), cleaning the data (Data Quality), manipulating the data (shaping) to suit your analysis. Then you must decide on different analysis options, apply those techniques. Make sure you are creating different graphs but present only the ones that are informative. Make sure that your conclusions can be supported by your analysis, and that the questions that you raised are answered appropriately.

**Deliverables**

You must submit the following:

1. The Final Report (text with graphs)
2. Your code (Ipython Notebook). Notebook must include all the steps listed in the project guide lines above. Notebook must be error free output for all the cells. Must be documented extensively using markdown cells
3. Notebook must include references, to relevant sites, documentation, data set or link

Follow the step-by-step guide to work on your project

Step 1: Select a real-world dataset.

Step 2: Perform data preparation and cleaning

* Load the dataset into Jupyter notebook
* Explore the basic information about the data, e.g., size of data ranges, some basic statistics etc.
* Handle missing, incorrect and invalid data
* Perform any additional steps (parsing dates, feature engineering, create additional columns etc.)

Step 3: Perform exploratory analysis and visualization

* Compute statistics for numeric columns
* Explore distribution of numeric columns using histograms etc.
* Explore relationship between various columns
* Make a note of interesting insights from the exploratory analysis

Step 4: Train and evaluate two machine learning models, tune hyperparameters and reduce the overfitting to improve the model. Evaluate the performance of the models using the metrics and diagnostics we used in class.

Step 5: Summarize your inferences and write a conclusion

* Write a summary of what you’ve learned from the analysis
* Include interesting insights and graphs from previous sections
* Share ideas for future work
* Share the links to resources used

Step 6: Make a submission on Canvas.