Final Data Project

# Instructions

This is a Group Project. You can self-assign yourself to teams of 1-3 people. Instructions will be provided in your Canvas course page. This project has 3 parts to it. Please see below for more details.

# How to select your Dataset?

1. Chose dataset that your team interested in.
2. **Classification** or **Regression** problem are preferred. However, your team can choose to do other kinds of problems too (unsupervised, Image/Video, Deep learning, etc)
3. Suggestion: you can pick from:
   * the Kaggle dataset [Kaggle](https://www.kaggle.com/datasets)sorted by “Most Votes” and filter “csv” file type. <https://www.kaggle.com/datasets>
   * [UCI Machine Learning Repository](https://archive.ics.uci.edu/ml/index.php) dataset: classification or Regression
   * <https://datasetsearch.research.google.com/>
   * other dataset repositories are welcome
4. If you **scrape and clean your own dataset**, there will be **bonus points!**

#### NOTE

* Please refer to your Canvas Course page for additional Resources/Sample projects related to the final Project.
* Please refer to your course schedule for detailed due dates for all deliverables.

# Deliverables:

## **Part 1- Project Details**

In the Part 1, you will be asked to submit your team name, team members name and email ids, your dataset and project name. You will be provided feedback by your Instructor. Thereafter, if you need to make any changes, please do so in the links provided below. Please refer to your course schedule document for detailed due dates for all deliverables.

## Submit

* Team members’ names and email IDs, Project Title, and Dataset URL in the following links. Fill out both links. Only one team member can fill out all the information. You will need a Gmail account to access these forms.
  + Use the **Check the announcement on canvas. Submit via Google Form**
  + **Due date to submit Part 1 is the end of week 9 (check canvas for date).**

## NOTE

* + Your Dataset need to be unique, you can't use the Dataset that another group already submitted
  + If you need to modify your submissions, please log in and modify your submission. Or describe it in the last question box of the form

## **Part 2- Creating Stage 1-3 of the end-to-end data process**

There are five stages of creating data processes. For Part 2, you will be working on the stages 1-3.

**Stage 1: Ask A Question: (10% points)**

* 1. Describe your project:
     + What is the dataset? Detail description.
     + What is the questions/answer you expect to find. What pattern that you expect to get from working with this data?
     + What are the benefits of this project to organization?

### **Stage 2: Get the Data (10% points)**

* 1. Preparing your dataset:
     + How you get your data? Pre-made or scrap from the web, or DB?
     + You will receive bonus points if you scrap your own data or acquire additional dataset.
     + How you clean your data?
     + How did you prepare it?
  2. Need to acquire more related dataset?

### **Stage 3: Explore the Data: (20% points)**

* 1. Descriptive analysis
  2. Make lots of charts
  3. Get to know data, develop hypotheses, patterns? anomalies?
  4. Problem with data? How to improve this dataset in the future?

## Submission Format:

* Upload your **slides** and **.ipynb** file to Canvas TBD

## **Part 3- Creating Stages 4-5 of the end-to-end data process**. (at the end of semester)

### **Stage 4: Model the Data (30% points)**

* 1. What do you want to predict from the data? machine learning
     + Explain how you select this ML model
     + Accuracy measurement, how you do train/test split, cross-validation
     + Improve your features? Features engineering, data augmentation?
     + How you do train/test split, cross-validation?
     + Model evaluations, comparing with other ML/DL algorithms.
     + Improve your model: Grid search, ensemble, hyperparameter tuning.
     + Alternative approaches?

### **Stage 5: Communicate the Data (30% points)**

* 1. Presentation of your project
     1. What is already there, what did you add? It is very important to reference all existing work, and clearly stated your contribution.
     2. Conclude, summarize, discussions
  2. Prescriptive analysis.
     1. What action can you make after seeing the result of this project?
     2. What are the potential applications for this project?
     3. What could benefit the direct benefit to an organization?

## **How to present, submit the work? (updated)**

* 1. Make a 10-30 minutes video presentation.
     1. presentation topics including:
        1. The whole project: updated of state 1-5
        2. Walkthrough your code. Local Jupiter notebook is colab are preferred.
     2. Videos format:
        1. It should be a voice-over screen recording video.
        2. You can show your face in a portion of your video too, it will make your presentation more engaging (optional).
        3. up to 30 minutes.
        4. submit as .mp4 file on Canvas (or upload on youtube public/unlisted and submit the url)
  2. **What to Submit:** required all as following
     1. your .ipynb with link to the dataset
     2. Your presentation slides as .pptx or .pdf or Google Sheet.
     3. Your .mp4 video file (or youtube url)
  3. **Due date is firm on Monday, May 3rd, 6:00 pm.**

## **How is this Final data project graded?**

## Graded according to the % of each stage

* + 3-person group will have a higher expectation of the quality of the work than the 2 person group
  + 1-person team – expected size of the project: comparable to **two** lab assignments
  + 2-person team – expected size of the project: comparable to **four** lab assignments
  + 3-person team – expected size of the project: comparable to **six** lab assignments
  + Total 100% for this Final Project. (instead of having a final exam)

