Project Proposal Assignment

# Assignment Goals

A one to two page preliminary summary of the project in a Microsoft Word document format. Note that by the time you write your proposal, you should have worked with the data set you plan to use for your project. If you make major changes to your project like changing your data set, your proposal grade will be changed to a 0. If you have to make a major change to your project, it shows that you didn’t really work with the data and think about your project before you wrote your proposal. For example, you originally decided to do image recognition but after you submit your proposal, you find that the images in the data set you chose are too large for your compute environment. In this case, you can change your project but you will get a 0 grade on the proposal. Minor changes, additions, deletions, changes in direction are allowed. The goal here is to make sure that students work with the data before making a final selection.

# Overall Requirements

* Shall be a maximum of 3 pages
* The project team shall not be greater than 4 members
* Shall be a Microsoft Word Document
* High quality writing is expected with good grammar and punctuation. The document should be thoroughly proof-read and reviewed by all team members prior to submission.
* All machine learning, inference, analysis, data wrangling, and data exploration shall use spark. Visualizations can be performed using any python visualization package. In short, the data science work flow shall be performed using spark. For example, don’t pick a project that requires convolutional neural networks because convolutional neural networks are not included in spark. Don’t pick a project where the main focus of your project depends on packages outside of spark. Essentially, your projects should be focused on topics learned in class using tools used in class. An example of an acceptable use of a package outside of spark is using an NLTK package to label a data set with positive or negative sentiment for the purposes of training a natural language processing model in spark. In this example, the NLTK package is being used to support your work in spark and is not the main focus of your project. In summary, **the main focus of your project shall be to apply spark to solve a business / scientific problem**. If you have a doubt, consult with the professor. Projects that do not use spark as the primary tool will not be accepted and receive a zero grade.
* All projects shall run on databricks. If your project requires computing resources outside of databricks then it is not an appropriate project, will not be accepted, and receive a zero grade.
* The assignment shall be submitted by one (and only one) member of the team.
* The data set shall be readily available for public access and shall not require the graders to sign a non-disclosure agreement. Basically, we need to be able to download the data and the project code onto our own computers with absolutely no restrictions whatsoever.
* Do not change the file name.
* Do not add files to the repo.
* Do not remove files from the repo.
* Your project proposal shall be in the project\_proposal.docx document included in this repo.

# Project Proposal Section Requirements

The project proposal is a road map on how you will execute your project and **shall include the following sections** (and **only** the following sections) **in the order given** in the bulleted list below. Each bolded item in the list below is considered a required element of your project proposal.

* **Title page** including project title, group number, and all group member names.
* **Objective**: A concise, well written high-level objective of the project. Briefly describe what you are trying to accomplish.
* **Data Set Description**:
  + Overview / Description
  + Number of rows and cols
  + Sample predictors (does not need to be an exhaustive list)
  + A link to the dataset
  + Anything interesting or surprising about the data
* **Preliminary Data Exploration:** Explore the data, provide brief summary statistics and visualizations. Tell me something interesting and or surprising about the data. If you need to reduce the size of your data to fit on your PC or databricks, now is the time to do that. The work that you do in this section of the proposal can be directly used in your project report.
* **Predictions**

All projects shall include a prediction component. A bulleted or numbered list of very specific predictions you are trying to make. The list of items should be large enough to keep everyone on your team busy. Your business / scientific problems need to be as specific as possible. Don’t make general statements like we plan to make predictions which will aid business leaders; but rather, be more specific about the predictions you will make.

* **Inference**

All projects shall include an inference component. Provide a bulleted or numbered list of specific inference insights you would like to achieve in your project. Don’t say you plan to explore the data to gain insight; but rather, be more specific about the insights you plan to explore and how those insights add value to the project.

* **Non Spark Packages:** Include an exhaustive list of packages outside of spark, numpy, pandas, matplotlib, and seaborn that you plan to use in your project. For each package, state how the package relates to your project. Make sure that you understand that packages outside of spark, numpy, and pandas are only allowed to support work done in spark. Your project must be a spark focused project.

# Grading Rubric

* 20% Writing quality / following instructions:

High writing quality is expected. If you do not follow all instructions, you will not get an A grade. **For example, if multiple team members make submissions, don’t include the required sections, add new sections that are not in the list, change the submission file name, submit in a format other than Microsoft Word, you will not get an A grade on your proposal.**

* 20% Project Objective:

The project objective should be a concise problem statement that accurately describes the high-level goals of the project

* 20% Data set description

The data set description should provide enough information for the instructor to quickly understand key elements of the data and where it came from.

* 20% Prediction, and Inference

The prediction and inference goals should be detailed and specific.

* 20% Preliminary Data Exploration

Provide brief summary statistics and / or plots