DATA SIENCE Final Project:

For this project, you will conduct predictive analytics using at least 3 datasets of your choice and share your code and inferences through a well-documented Jupyter Notebook.

Define the Problem:

- What is the underlying problem you are trying to solve?
- What is the underlying opportunity that you are trying to develop?
    - This is usually different from the presenting problem.
    - List out the high-level questions that you will investigate during the project.

Set the Goals

- How will you measure your success? Unit of measurement
- Are there specific goals?
- Outline your plan

The Project

1. Find datasets of your choice online - Kaggle, Data.gov, and Dataverse are some resources but feel free to explore. It is important that your datasets of choice have enough samples (rows) and useful features (columns). Usually, the more, the better.
2. After choosing a datasets, come up with a question that you want to answer. For example, a question relevant to the Titanic Dataset could be 'Will a person survive or not?'.
3. Clean and manipulate the data, state your hypothesis to your question, and do the following:
    a. Create (at-least) 2 meaningful visualizations that add information or context to your project. These should be different types of visualizations from one another (i.e. don't do two scatter plots).

b. Show Statistic Features (Distribution, mean, median, etc.)

c. Build (at-least) study: If your problem is numerical use a regression study (Ex: house pricing can be predicted by using a regression study Linear regression example or another one). If your project is with discrete data then make a Comparative model or Sentiment Analysis(like we did in class).

Rubric and submission

● Mid-project check-in (5): Check-in to make sure the project is on track and check to see if any help or questions can be provided.

● Preprocessing and Manipulation (15): Any necessary cleaning and manipulation of the dataset

● Visualization (30): At least two visualizations. Visualizations are clearly visible, clean, well-labeled, and serve a clear purpose for your question(s). (No two scatter plots are allowed)

● Studies (35): At least 2 studies applied to the manipulate data. For example, run a linear regression for a classification problem. Or other study that will allow you to make a prediction or estimate future behaviors

● Write-Up (10): The methodologies and inferences are properly explained. Walk the reader through the steps and thought process you took for each step of the project. Additionally, please pre-run all the cells so I can see the output. It makes it much easier to grade.

● Creativity (10): Did you go above and beyond just satisfying the requirements? Please submit your Jupyter Notebook (that includes the link to the dataset source) and dataset in a zip file through GitHub link.