**Project 4 Group 2 Proposal**

**Project Data Information (Data we will work with and the field we’re interested in):** We will be using a Kaggle dataset about the S&P 500 daily stocks. <https://www.kaggle.com/datasets/yash16jr/s-and-p500-daily-update-dataset>

**Questions we will look to answer:**

1. Visualizations:
   1. What are the best and worst stock(s) – in quartiles?
      1. Is this by share, by percentage, by dollar amount(s)?
   2. Which stocks are the most volatile?
2. Model
   1. Regression
      1. Random Forest
   2. K-means model
   3. Classification
      1. Model stocks by % of growth in the last 5 or 10 years with predictability factor for future growth

**Target variable in the dataset:** Evaluate stocks and their open, close, high, low prices to determine the optimal stock to purchase shares to gain the largest return on investment by % growth over the past 5 or 10 years focusing on close prices or open/close percentage differences.

**Link to GitHub repository:** [**https://github.com/Nalfrey/Project\_4.git**](https://github.com/Nalfrey/Project_4.git)

Project Timeline/Expectations:

* Exploration and cleanup of Kaggle csv
* Analysis – Creating questions, and interpreting data to present in a clear, concise manner.
* Use data model(s) that initializes, trains, and evaluates the data
* Data is retrieved from SQL or Spark
* Model demonstrates meaningful predictive power of at least 75% classification accuracy or .80 R-squared
* Present the model optimization and evaluation process showing iterative changes made to the model and resulting changes in model performance that is documented in either CSV or Excel table (or in the Python script)
* Git hub has .gitignore and README including all content of the project