Proposed Project Name: Stock Prices Predictor
Due Date: October 24, 2018

Problem Statement:
The field of machine learning has become a pedestal of innovative technology for millions of data scientists across the globe and thus, an increasing number of business organizations, especially the financial conglomerates, are being greatly motivated to spend their resources on developing software that can possess the ability to monitor and analyze key company performances, as well as to anticipate the prices of various stocks. One such software is the stock prices predictor that can be defined as a system with the facility to learn about the performance of a company and implement the learned concept to predict future stock prices. With the availability of a large variety of information on stock markets, this project has created a hot-bed of opportunities for data scientists with a predisposition for finance. In this project, a combination of supervised learning techniques will be employed to build the backbone of the stock price predictor, while simultaneously comparing its performance with respect to the implemented methods.

Source(s) of Datasets:
Stock prices data is very granular and furthermore, different varieties of datasets are involved, such as market volatility indices, global macroeconomic indicators, fundamental indicators, etc. However, financial markets usually have shorter feedback cycles, which make it more convenient for data scientists to legitimize their work on new datasets. Hence, for the purpose of this project, variations in stock prices over a 6-month period can be predicted by using one or more organizations' quarterly financial reports in the form of datasets which will be collected from the following repositories:

- https://www.quandl.com/
- https://www.quantopian.com/
- http://www.nasdaq.com/

Implementation Plan:
The Stock Price Predictor project will consist of two important phases: **1) Data Collection** and 2) **Feature Extraction**. Below is a brief description of the aforementioned phases that will form the foundation of the stock price predictor:

1. Data Collection:
   The first part of the data collection process will be to transform the datasets found in the repositories into clean datasets. This technique is widely known as data-preprocessing. The datasets will be pre-processed in the following ways:
   a. Data reduction, while maintaining their numerical equivalents
   b. Data normalization followed by the recovery of missing values
   c. Integration of transformed data into respective data files

   After the completion of data pre-processing, the *cleaned* dataset(s) will be divided into training and test datasets for evaluation. The more recent values of data will be taken as training values and about 15-20 percent of the total dataset(s) will be set aside as test values.
2. Feature Extraction:
   New features will be generated from base features that will provide a better idea of the data, such as 30-day moving averages, previous day differences, etc. During the Feature Selection process, the unnecessary features will be reduced as per the k highest scores, which will be achieved by applying a linear model to analyze the effects of a linear regressor. If time allows, an attempt will be made to add the Twitters Daily Sentiment Score as an attribute for individual companies based on the users' tweets regarding that specific company and also the tweets on the corresponding company's website.

Team Members and Task Allocation:
Ayushi Chaturvedi
Abrar Alam
Shreyansh Sharma
The following tasks will be equally split between the team members:

- Literature review
- Supplemental research
- Coding algorithm
- Project report write-up