# Problem 2

Uses sp500\_ohlcv.csv.

The csv file sp500\_ohlcv.csv contains several years of daily data on almost all of the 500 stocks in the S&P 500 stock index. For each equity we observe daily records of the open, high, low, and closing prices, as well as trading volume. Many of the stocks in the index tend to “move together” meaning there is correlation between values over time.

1. Read in the datafile and count the number of features in the data.
2. Perform a principal components analysis on the data (exclude dates and scale the data), determine how many components are required to capture at least 85% of the market’s variation (you’ll need to calculate the cumulative PVE).
3. How many principal components should be kept according to the “eigenvalue / variance greater than 1” criterion?
4. How many principal components should be kept according to the PVE greater than uniform criterion?
5. How many principal components would you keep based on the “Elbow” method? (*you’ll have to create a scree plot and will need to use some creativity or ingenuity to find a way to visualize things because of how many components there are*)
6. Remember that each row in the data represents a day’s worth of trading in the markets. Using the 1st and 2nd principal components, we can define two new coordinate axes (i.e., x-axis (1st PC) and y-axis (2nd PC)) and then plot each day’s score for the 1st and 2nd principal components to represents its location relative to these axes and the rest of the days in the data. Create a scatter plot using PC1 scores on the x-axis and PC2 scores on the y-axis. Identify any patterns and give an interpretation of what the location of a data point means in the plot.

# Problem 3

[based on Shmueli2018 4.1]: Uses Cereals.csv

Consumer Reports ranks cereals on the “best buy” among cereals when value is determined by nutritional contents. As a result, a lot of “empty” sugary cereals tend to rank quite low. You work for a company which is considering a small set of different recipes for a new cereal they would like to market. You wonder if you can learn something about the rankings of cereals based on their rating and contents. Before building a predictive model, you want to analyze the data and understand something about the nature of data features across cereals. A lot of cereals have similar profiles. Use the Cereals.csv data to answer the following questions and perform related tasks.

1. Which variables are quantitative/numerical? Which are ordinal? Which are nominal?
2. Plot a histogram for each of the quantitative variables. Based on the histograms and summary statistics, answer the following questions:
   1. Which variables have the largest variability?
   2. Which variables seem skewed?
   3. Are there any values that seem extreme?
3. Compute the correlation table for the quantitative features. In addition, generate a matrix plot for these feature correlations showing, in a heatmap style way, where the highest correlations are.
   1. Which pair of variables is most strongly correlated?
   2. How can we reduce the number of variables based on these correlations?
   3. How would the correlations change if we normalized the data first?
4. Perform PCA on the scaled quantitative features of the Cereals dataset and create a scree plot. Choose a method for determining the number of principal components to keep and justify your choice.
5. Compute the coordinates of each of the quantitative features in the space spanned by the principal components (see the Spotify PCA Part 2 videos for details) and identify which features are most correlated with the first two principal components.
6. Compute the **cosine squared** measure of the quality of feature representation by the principal components. For each feature, identify the principal component that most represents it. (See spotify video)
7. Using the matrix of component vectors, create the table (See Spotify video) which shows the contribution of each feature to the direction of each principal component.

# Problem 4

[Shmueli2018 4.4]\*\*: Uses Wine.csv

*Chemical Features of Wine*. Create PCA output on data (non- normalized) in which the variables represent chemical characteristics of wine, and each case is a different wine.

1. The data are in the file Wine.csv. Perform PCA on the quantitative variables **without scaling**. Explain why the value for PC1 is so much greater than that of any other column.
2. Comment on the use of normalization (standardization) in part (1).
3. After normalizing the quantitative variables, perform PCA again and use your preferred method to identify how many components might be kept.
4. With the principal components you’ve decided to keep identify which wine chemical features contribute most to those principal directions and comment on which features in your data are represented well by those principal components and which are not.