**Assignment 1:**

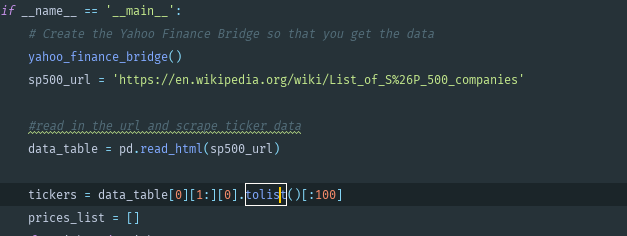
**Installation**

1. To install all the relevant python packages, use the command on the command prompt as ‘pip install –r requirements.txt’ after extracting the zip folder
2. To run the file, on a command prompt type ‘python a1.py’
3. NOTE: Sometimes yahoo finance bugs out and returns an empty dataframe. This is a known bug. Please rerun the program.

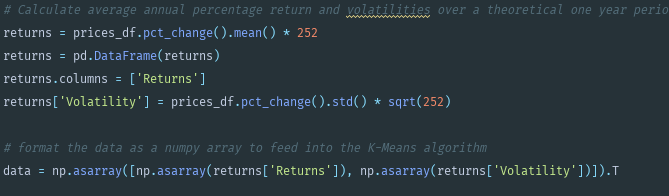
**Explanation**

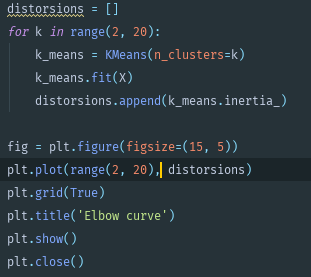
1. Scrape Data from Wikipedia to get list of stocks

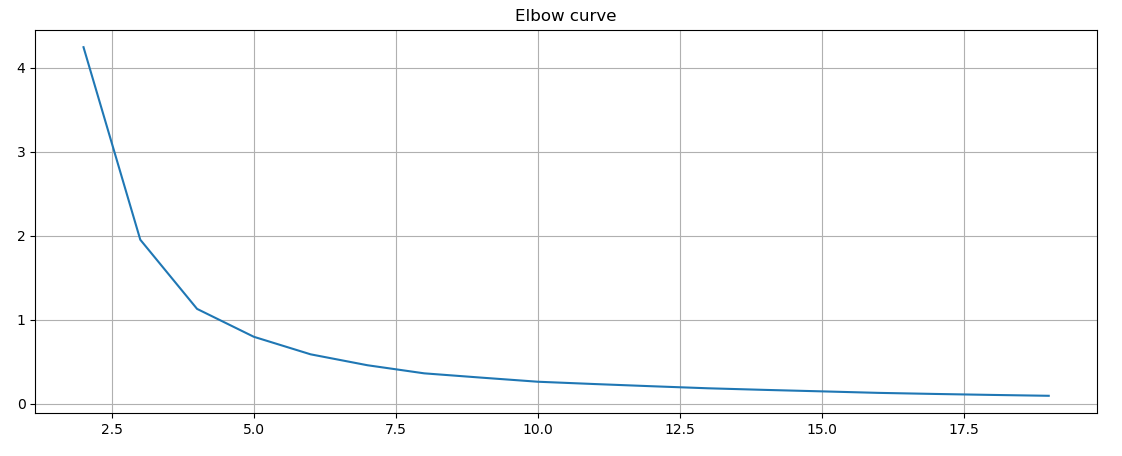
We take only 100 stocks since fetching data for 500 takes up a lot of time



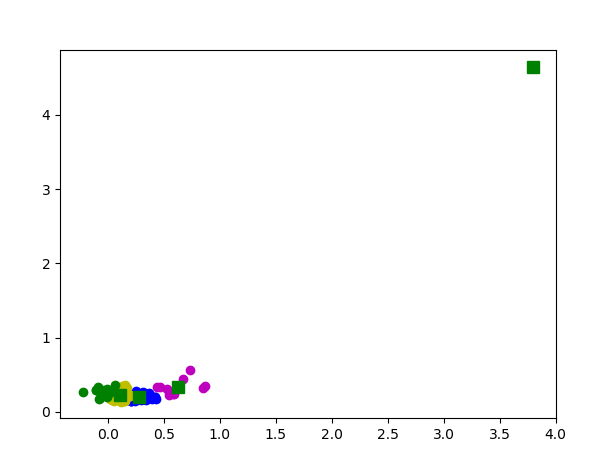
1. We choose how many clusters do we actually want to separate the data into. We can use an “Elbow Curve” to highlight the relationship between how many clusters we choose, and the Sum of Squared Errors (SSE) resulting from using that number of clusters.



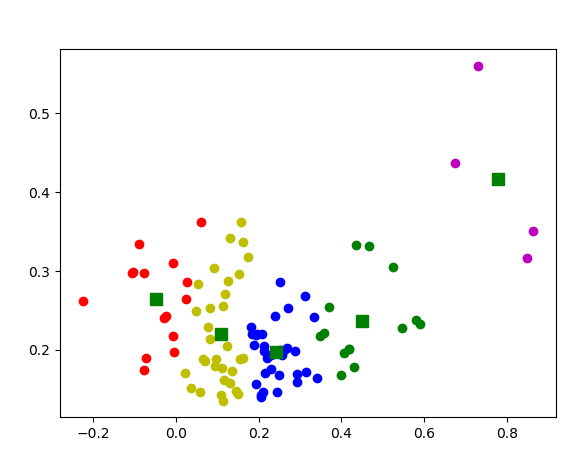




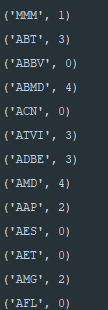
1. We can see that once the number of clusters reaches 5, the reduction in the SSE begins to slow down for each increase in cluster number. This means the optimal number of clusters for this exercise lies around the 5 mark. Hence, we use 5. This gives the output with an outlier



1. After removal of the outlier we have the cluster as follows



1. Finally, we the details of which stock is actually in which cluster



1. dividing stocks into groups with “similar characteristics” can help in portfolio construction to ensure we choose a universe of stocks with sufficient diversification between them. This would help in getting better Alpha values.