Youbeen Shim
Professor Jordan Rodu
STAT 4310: Data Visualization
Homework 2

Visualization on Stock.csv

Dear valued client, our firm, Visual Wizards, Co., is delighted to work with you to assist you on your concerns regarding exploring your dataset.

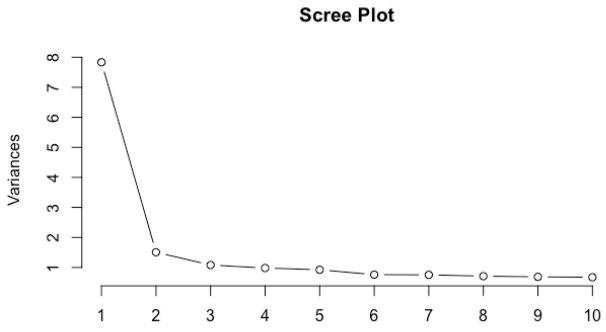
Introduction

The given data consisted of a single time variable combined with 22 stock variables that are centered, standardized, and consist of log of return value of the given stock on the given date. While it is natural to simply plot the log return values against time, the number of variables that exist makes the resulting plot messy and incomprehensible. Given the circumstances Visual Wizards, Co. has taken a dimension reduction, specifically principle component analysis (henceforth "PCA"), to distribute the variation in the given dataset across components. Through this approach, we will be able to observe patterns that may not otherwise be apparent using more common analysis and visualization techniques.

Instead of plotting the data in the traditional sense, PCA plots the data in a way that is natural for the data, and attempt to capture as much variance as possible. Each principle component, positioned orthogonally from the prior component(s) following the first exists to capture the variance not caught by the prior principle component.

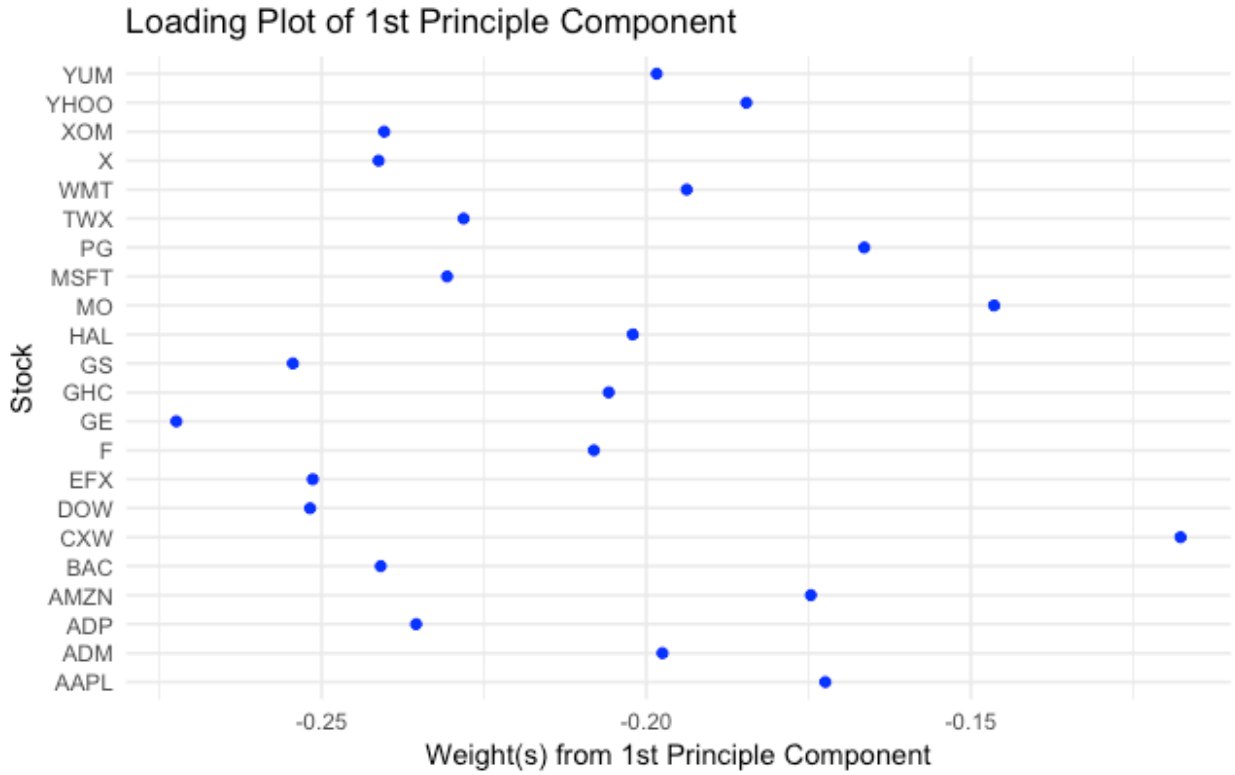
Issue 1

As explained during the introduction, each principle component is designed to capture as much variance as possible. This implies that successive principle component generally holds less and less data and only serves to complicate the model. We first plot a scree plot, Reference A, designed to display the relative variance each principle component is capturing. Through Reference A, we observed that any principle component beyond the second (perhaps beyond the third) will not provide information that overweighs the cost of overcomplicating.



Reference A: Scree Plot

Thus, given Reference A, we plot Reference B, a loading plot for the first principle component. Reference B displays the "weights" of each stock variable towards the first principle component. In essence, it can be observed that some stock variables, such as GE or GS, has high correlation with the direction of maximal variance (first principle component). This is noted by the rightward position of the variables on the loading plot. In contrast, companies such as CXW have a low correlation with the first principle component. This, similarly, is denoted my the leftward position of CXW in the loading plot. However, it is important to recognize that while the the differences in correlation exist, value of the variable loadings range from -.25 to slightly above -.15. None of the correlations are significantly strong.

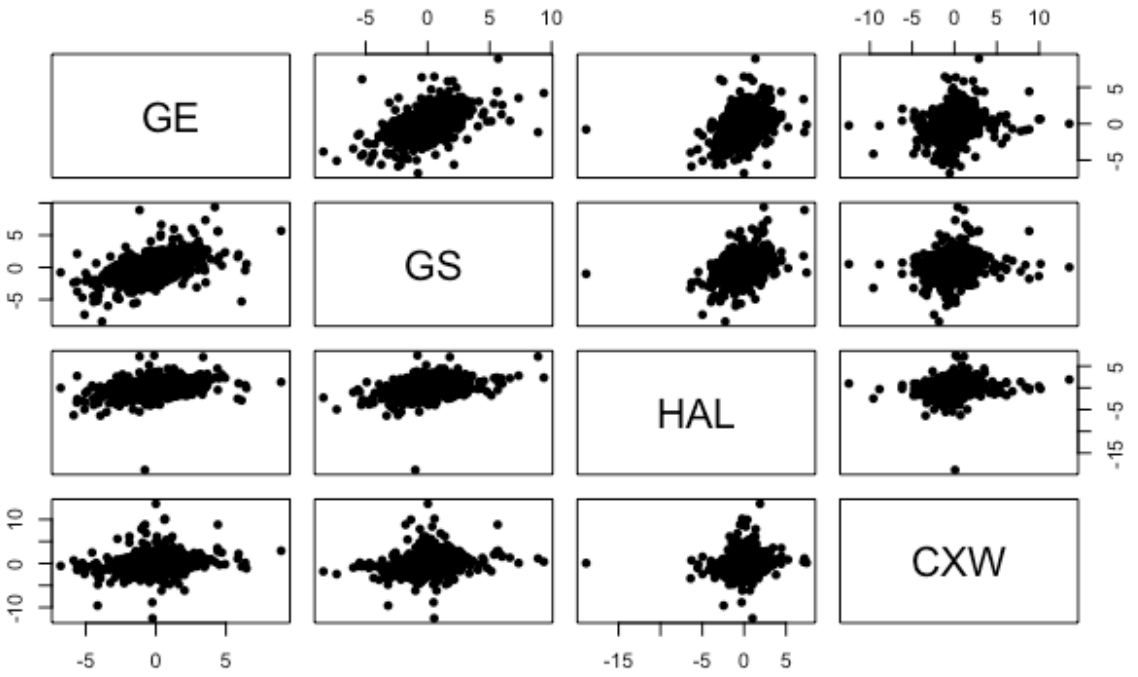


Reference B: Loading Plot (First Principle Component)

Issue 2

To address the second issue, we further explore the interactions between the specific stock variables. In order to select stock variables that will provide most information without complicating the visualization, we refer back to Reference B: Loading Plot of 1st Principle Component. As Reference B displays the stocks' correlation to the first principle component, it is expected that when stocks are located around similar weight value, that they will have higher correlation with each other.

Thus, four stock variables, EFX, DOW, ADM, and CXW are chosen in Reference C to explore the bivariate relationship between stocks in the dataset. As expected, there exists a high correlation with EFX and DOW, after which the correlations diminish until the scatterplot between EFX and CXW. It is difficult to state that stocks EFX and CXW carry any correlation, even as we intuitively note that all stock data will naturally follow the trend set by the economy.



Reference C: Pairs Plot

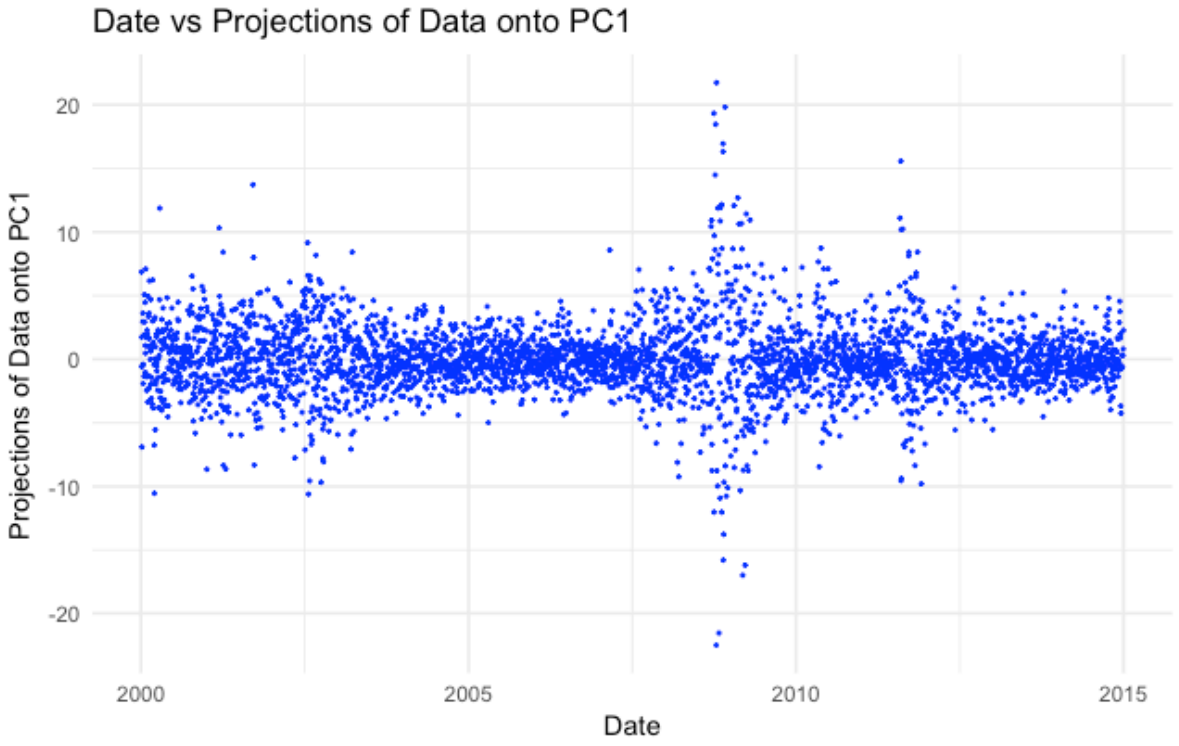
Issue 3

Issue 3 dives deeper into the subject of the relationship between the data and the principle component. In order to preserve most of the information in the data while reducing the dimensionality, we opt to keep the value of the project of the data onto the first principle component. In short, Reference D preserve the value of the perpendicular distance from the datapoint to the first principle component, and plot that projection against date. The resulting plot, Reference D, shows a thick band centered around 0 with some major vertical disruption at specific time values.

The closer the projections are to zero, more successful the first principle component was in capturing the variance. This visualization is constant with the historical expectation, as we observe that the areas of high uncaptured variance are the time periods in which there was great economic disruption -most notably the Great Recession of 2008.

Intuitive approach to divide such data simply looking at just the general shape of Reference D would be to capture the areas of high vertical disruption in some form. A short range of x value containing a high range of y value should be classified using a metric to distinguish the unique effects that are occurring.

In this manner, principle components are very instrumental in providing some type of classification that is not easily noted in more typical statistical methods. The use of principle component, however, should not be the singular method of classification and must always be used as a tool for data exploration, not characterization.

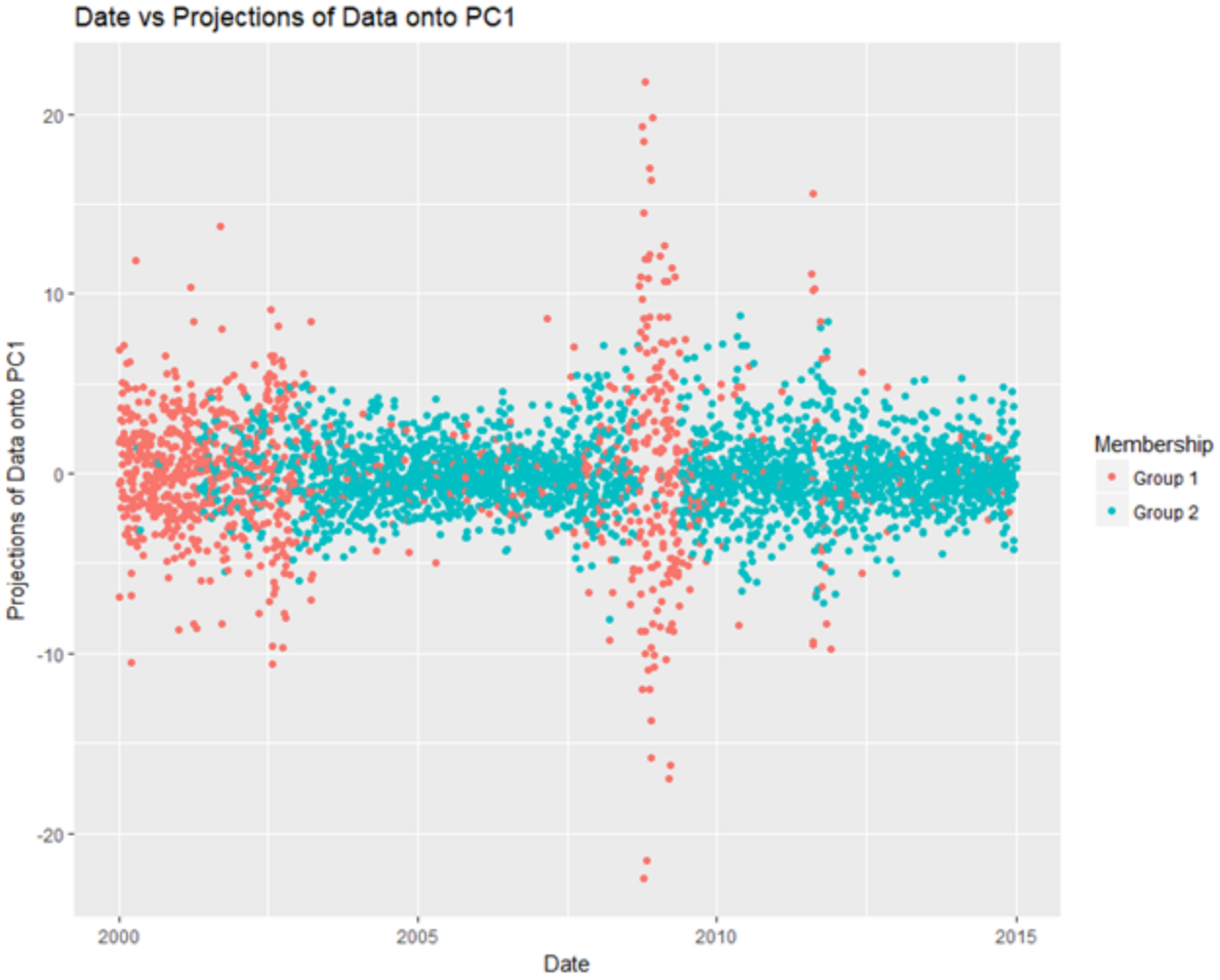


Reference D: Projections of Data to 1st PC against Date

Issue 4

To further explore the dataset, we infer that the data carries a multivariate normal mixture distribution, where each individual data point carries a probability of coming from a specific gaussian. Reference E displays the same general plot as Reference D, but categorizes the individual data points into one of two groups. Group 1 are the individual data points that had a posterior probability (of pertaining to a gaussian distribution of interest) value of greater than or equal to .5. Group 2 are the individual data points that had a posterior probability value of less than .5.

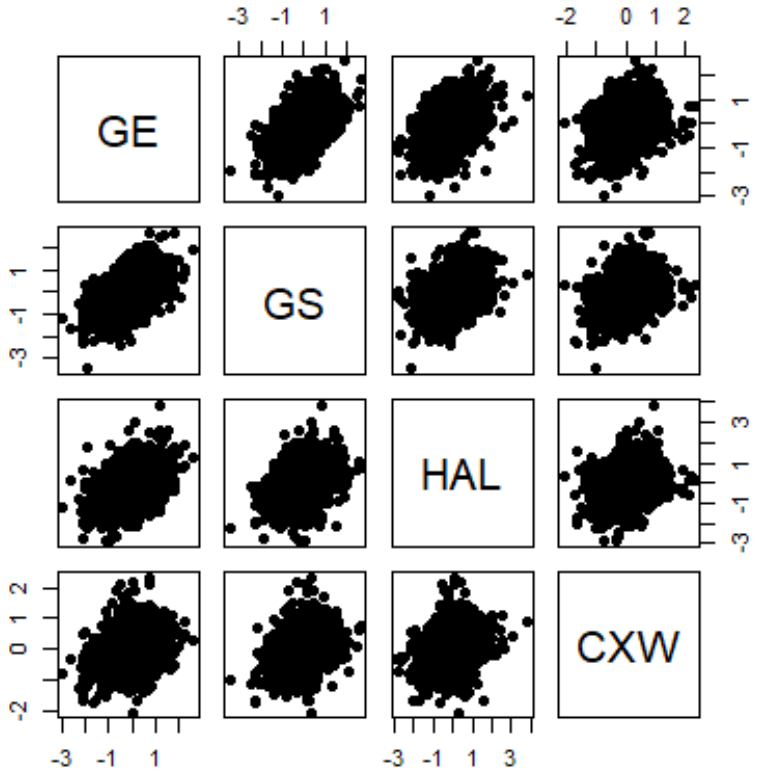
This way, Reference E: Multivariate Gaussian Mixture Model, is able to identify the times of high variance and better convey the length and inclining and declining variability over the course of time. It also more accurately displays the expected(captured) variance, making it significantly more useful than Reference D.



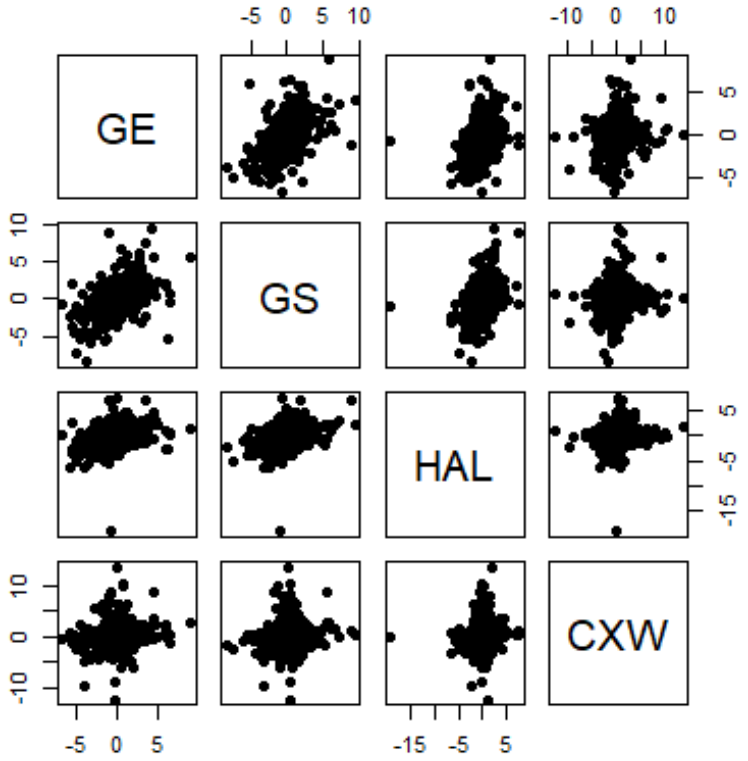
Reference E: Projections of Data to 1st PC against Date (Multivariate Gaussian)

Issue 5

If the points in Reference D can be categorized into Group 1 and 2 based on the posterior probability of coming from a specific gaussian, it makes intuitive sense that same can be applied to Reference C: Pairs Plot. In the data exploration process, it is possible that effects pertaining to one group mask the trend created by another group. Thus, we take the four stocks that we covered in Reference C -GE, GS, HAL, and CXW- and only create two pairs plot. Each pairs plot -Reference F-G1 and Reference F-G2- contains only the individual plots that pertain to Group 1 and Group 2 category established in Issue 4.



Reference F-G1: Pairs Data, Group 1 Points only



Reference F-G2: Pairs Data, Group 2 Points Only

As expected, the pairs plot displayed in Reference F-G2 lack any correlation beyond GE and GS. The plots also carry high variance, where the the value of the points range from approximately -10 to 10. In contrast, pairs plot displayed in Reference F-G1 shows a thick, yet straightforward linear pattern in a relatively small range (approx. -3 to 2). The combination of the two plots prove that the effect of points pertaining to Group 2 were masking the trend generated by the points pertaining to Group 1.

This also shows that the stock data do all follow a general trend set by the economy. However, this is frequently and easily masked by the significant and uncontrollable variables that impact the economy non-cyclically.

Conclusion

We have explored the stock data with the combination of typical and non-traditional multidimensional methods in hopes of capturing various patterns of the stock data and in a higher sense, the economy (principle influencer of stock data). In using multidimensional methods, we have successfully identified and categorized sources of high variance.

Also, we have noted possible ways in which high dimensional data may obscure the trends and how dimension reduction tools may aid in uncovering information that are of interest to the client.

We sincerely enjoyed working with the data, and would like to take a moment to thank the client for their generous cooperation. We look forward to working with the client again if any other concerns arise.