ASSIGMENT #1

BRANDON MORETZ

TASKS

1.) *(No actionable item).*

2.) *Now that we have our data defined, let's begin our exploratory data analysis by examining the correlations. We will want to compute the complete correlation matrix, and then consider visualizing a subset of those correlations for quick and easy comparison.*

   a. *Which row or column of the correlation matrix are we most interested in?*

      i. We are interested in the row of the correlation matrix that is the corresponding index of the symbol we are looking at correlations for. We would also likely want to exclude the column that corresponds to this same index, as it will always have the value 1 (the row I, column I would be the matrix diagonal) which is the correlation of that symbol to itself.

3.) *How about we make an even fancier data visualization of the correlations? The corrplot will allow us to visualize all pairwise correlations in the data.*

   a. *Is the corrplot more useful or insightful than our simple barplot?*

      i. I would not classify one as 'better' than the other, they each offer a slightly different perspective of the data. The corrplot offers utility in the form of being able to see the entire data set in a glance, for example the last row or last column of the corrplot has the same information as the barplot, however it is in my opinion less readable. If I were doing EDA using these two graphics, I would start with the corrplot to get a general summary, and then 'dig in' to specific stocks using the barplot.

   b. *What is the difference between a 'statistical graphic' and a 'data visualization'?*

      i. A 'data visualization' is a bespoke graphical representation of a subset of a data set that emphasizes the narrative one is trying to convey. Data visualizations wouldn't generally export easily to new datasets given their highly customized nature. Statistical graphics I would classify as more standardized (less customized) graphical representations of specific characteristics of a dataset (i.e., boxplots, bar plots, histograms, scatterplots, density plots, etc.)

   c. With respect to the concept of multicollinearity, look at the corrplot and identify three stocks that should have low VIF values? Similarly, pick three stocks that should have high VIF values?

      i. Low VIF:

         1. DPS:DOW
         2. DPS:BAC
         3. DPS:HUN

ii. High VIF:

1. VV:WFC
2. HON:VV
3. CVX:XOM

4.) *In addition to statistical graphics and data visualizations we can use models as tools for exploratory data analysis. Modeling is an inherently iterative process, and we typically begin the modeling process by fitting some 'naïve models' that are nothing more than models that we believe are good starting points for the modeling process. Typically, we might start with a small model and the full model as our two initial naïve models. The full model also allows us to compute the VIF value for every predictor variable.*

a. Is multicollinearity a concern for either of these models?

i. There is a degree of concern for multicollinearity in both models, as there are several variables that have VIF scores over 2.5. Overall, I would say the concern for multicollinearity is greater in model 2 than in model 1. For model 1, this includes GS and MMM, and model 2 the variables BAC, GS, JPM, WFC, BHI, CVX, HAL, MMM, SLB and XOM.

b. What value of VIF should make you concerned about multicollinearity?

i. The numeric value of a VIF score that should warrant concern is a somewhat subject value depending on the model and the data at hand. However, in general, values over 10 are considered "very high", values over 5 are considered "high" and over 2.5 "concerning".

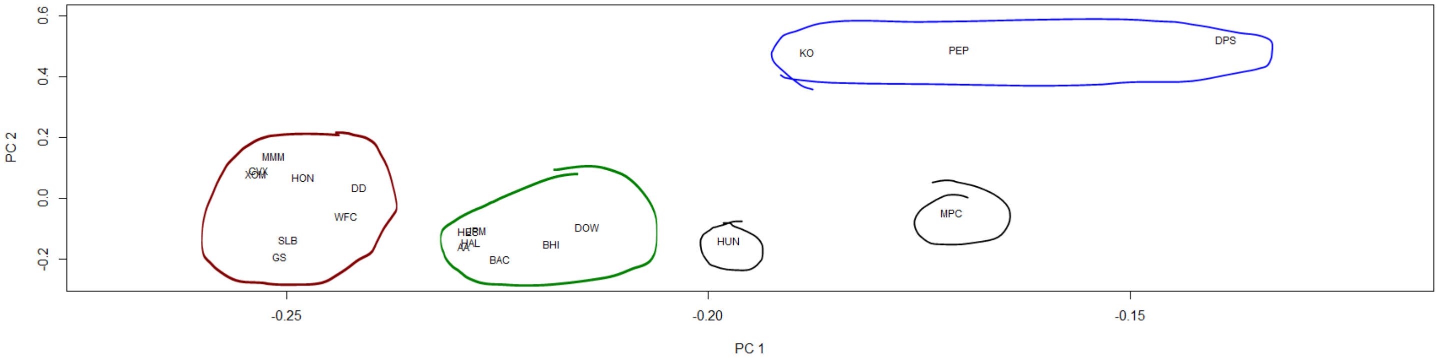
5.) *First, plot the loadings for first two principal components from the principal components analysis.*

a. What are the loadings?

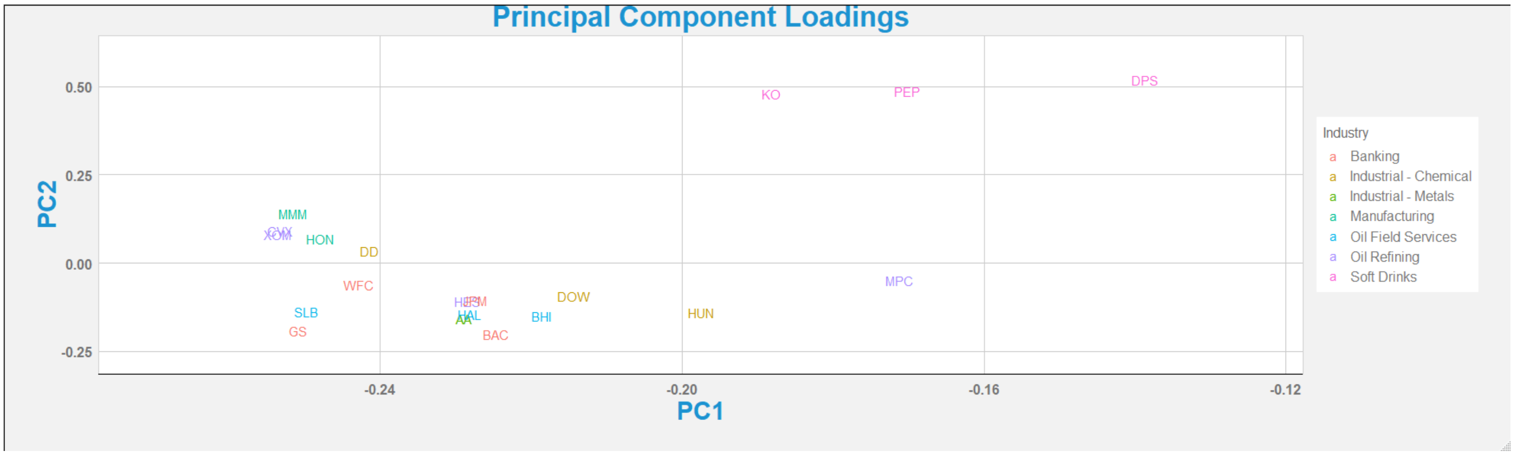
i. The loadings (or component loadings) are the correlation coefficients between the variables (rows) and factors (columns).

b. When we plot the loadings, we can see relationships in the data. What groupings (or clusters) do you see in the plot of the first two principal components? Any surprises?

i. There are definitely clusters of symbols for the first two loadings. The three major groups are circled below in red, green and blue. The black symbols are kind of in their own islands.



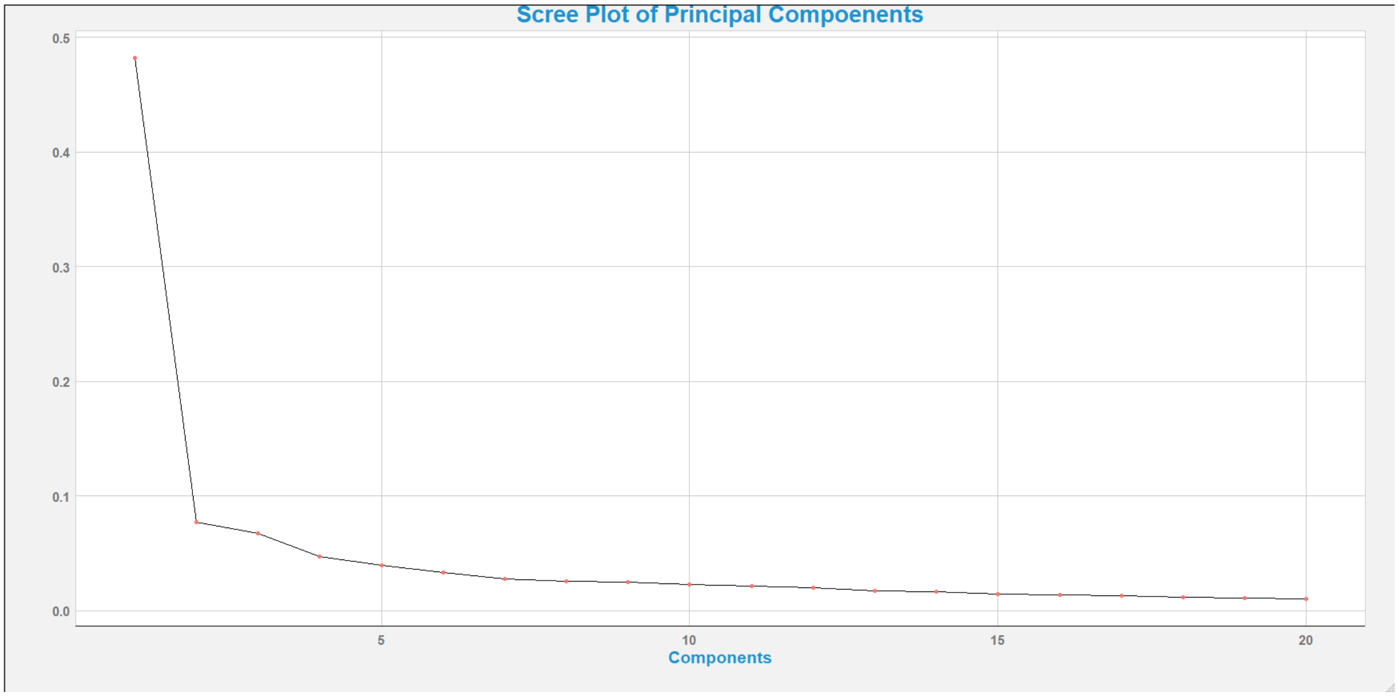
ii. Color code this plot by industry.

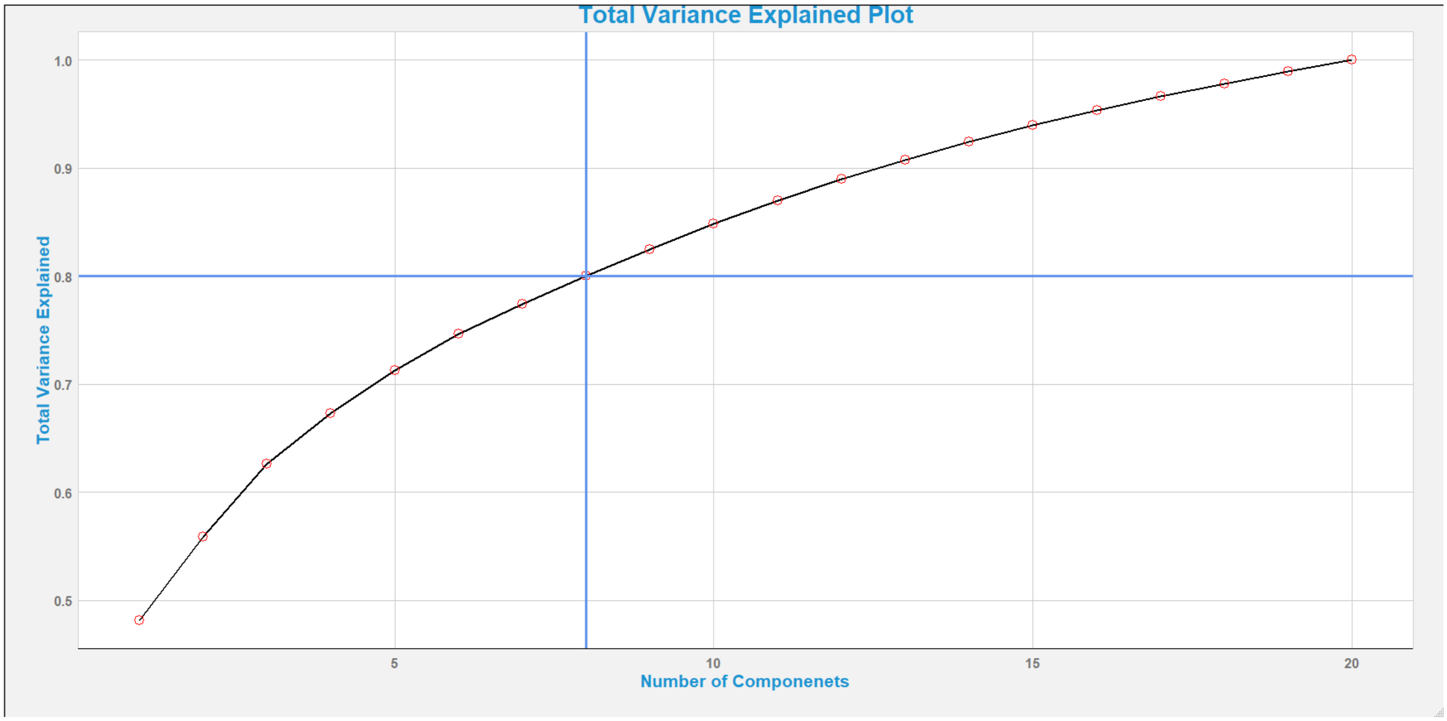


6.) In using PCA for dimension reduction, after we compute the principal components we need to decide how many principal components to keep or retain. How many principal components do you think that we should keep? Why? (Hint: What decision rules should we use to determine the number of principal components to keep?)

a. There are no hard and fast objective rules to follow when selecting the number of principal components to keep, this greatly depends on our objective. Here our objective is explain the most variance with the fewest possible components. In general, we should keep the minimal number of components that meet the cutoff value for our percentage of explained variance (typically between 80-90%, depending on situation).

The typical scree plot here is not particularly useful, so we'll look to the cumulative variance explained plot.

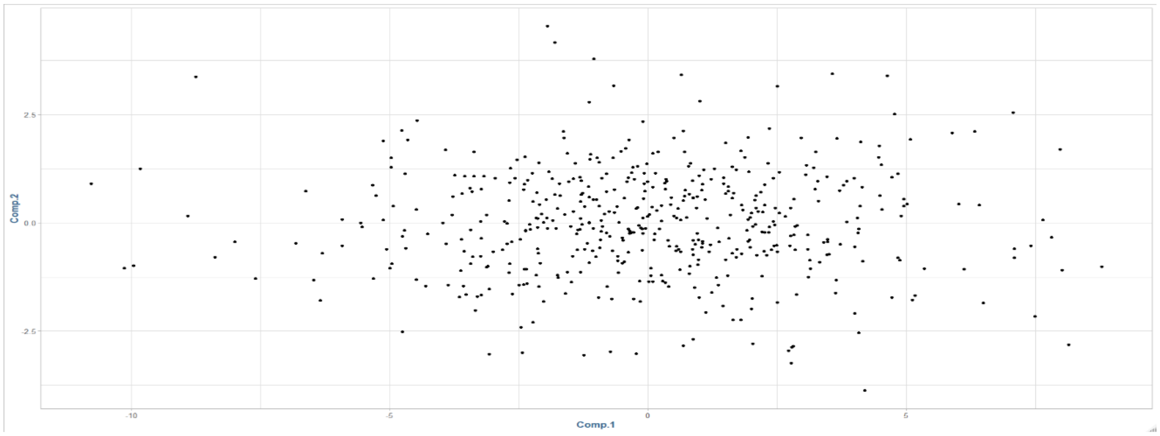




Following the rough guidelines for explained variance, we should keep the first eight components if we want to account for approximately 80% of the total variance explained.

7.) Now let's use principal components in predictive modeling.

a. The predictor variables for our predictive model will be the PCA scores. What are the scores?

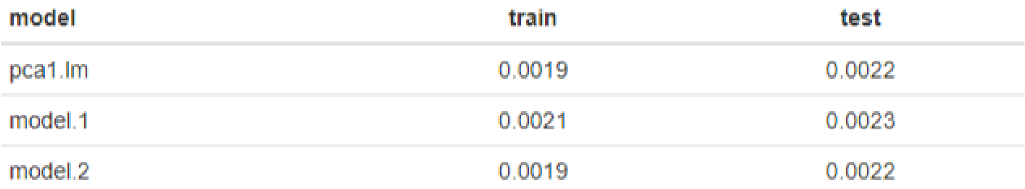


i. The score for an observation is the distance from the origin (as opposed to the direction, which is defined by the loading vector) that is defined for a given component. We can look at the first two component scores for the stock data visually in the graph below.

b. The VIF values associated with every predictor variable in any principal components regression model should all be one. Why?

i. The VIF value is an indicator of multicollinearity in a linear model; the components derived by principal component analysis are derived orthogonally in linear space (no two components go in the same linear direction). This makes multicollinear relationships with the components impossible.

8.) Present the MAE values from models pca1.lm, model.1, and model.2 in a table so that they are easily compared and discussed in your paper. Discuss these results. Which model is the best model? Why?



model	train	test
pca1.lm	0.0019	0.0022
model.1	0.0021	0.0023
model.2	0.0019	0.0022

These three models are relatively close in terms of MAE scores. The fact that pca1 and model 2 have identical performance is the biggest surprise here due to model 2 using all 20 individual stocks, and the pca model is using the first 8 principal components, which explain around 80% of the overall variance of the 20 individual stocks. I would consider the pca.1 the "best" model in this scenario in terms of being "production" quality, given that it uses the fewest variables to make predictions (simpler is always better, all other parameters held constant).

However, the only drawback could be the interpretability of the coefficients given they are derived from principal components. A relatively small concern, however it should be noted.

The performance of model 1, which is derived directly from the returns of 10 of the large cap stocks to predict the performance of the larger tracking ETF, is a bit surprising as it is using far less information than the other two models and performance only degrades slightly.

9.) In this assignment we have looked at PCA from the traditional unsupervised learning perspective. Now we want to look at PCA in a supervised learning perspective, which will be more useful when using principal components in predictive modeling.

a. How many principal components does the variable selection approach suggest that we keep? Which ones?

b. How does this differ from our previous discussions about the number of principal components to keep?

c. Create a new table where you add these new MAE results to the previous MAE results.

d. How does our backward.lm model compare to all the other models that we have fit in this assignment?

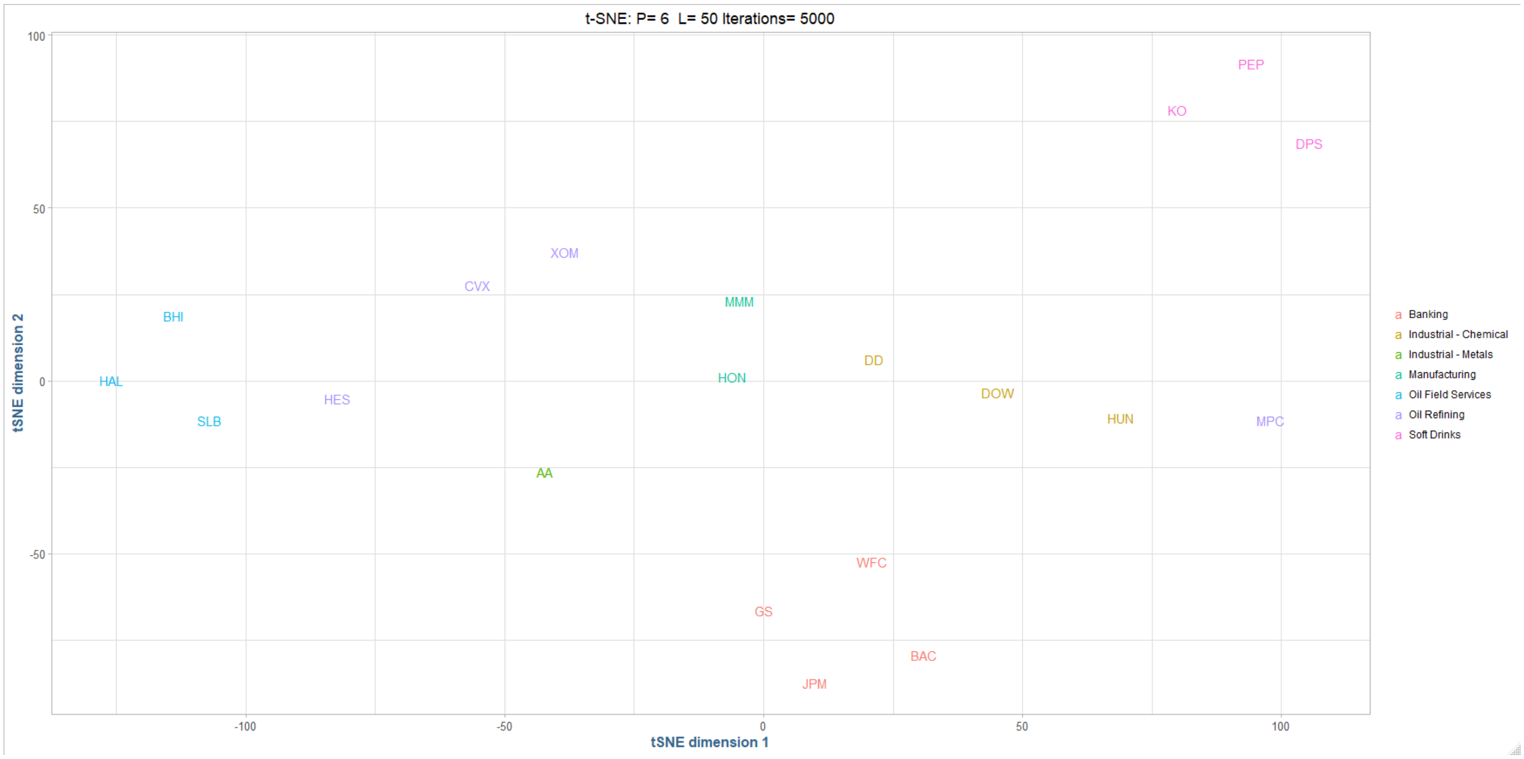
e.

f. Which model is best and why?

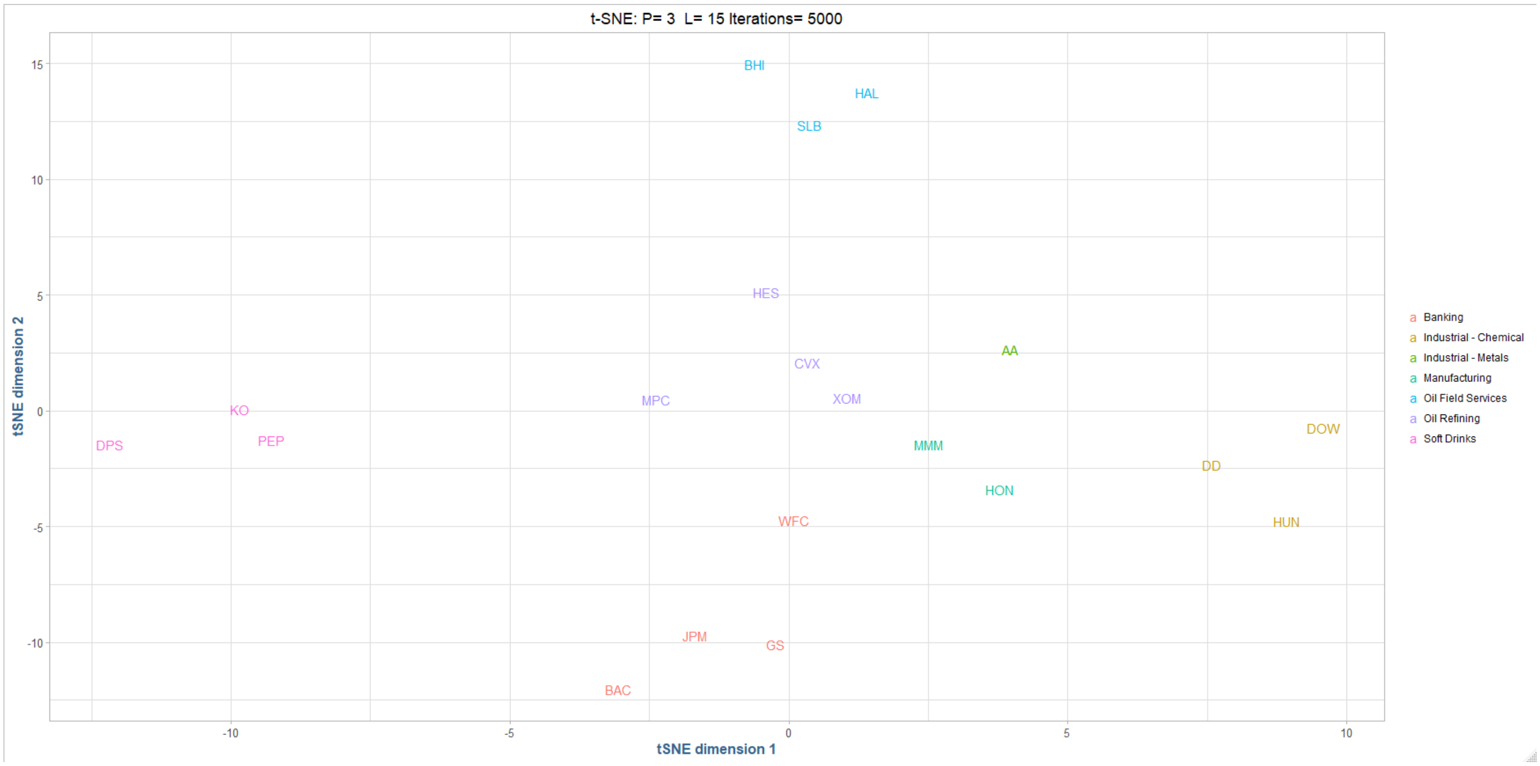
10.) Let's go back to the beginning with this data set, and use the t-SNE approach to reduce and visualize this data.

a. Conduct a t-SNE analysis on the Stock Portfolio data.

i. I conducted a t-SNE analysis on the correlations of the stock returns, and then labeled the stock tickers by industry as we did in a previous task. The results look like the following:



Additionally, if we transpose the raw return data, we can get similar clustering results:



b. Interpret the results, then compare and contrast the t-SNE results with that of the Principal Components analysis you've conducted previously.

    i. There are quite a few differences between PCA and t-SNE. The greatest difference being how the analysis technique works, where PCA is a linear method and t-SNE is a probabilistic (hence, stochastic) method. The linear framework of PCA makes it substantially more interpretable than t-SNE, even though PCA itself it is difficult to interpret.

       Another big difference is the results of the two techniques. PCA derives N number of components based upon the data passed in, and the summation of all the components is equal to 100% of the variance in the data, whereas t-SNE strictly projects a N dimensional space into strictly two or three dimensions for the purposes of visualization.

       Additionally, PCA derives components that can be used for further analytics with regression models for example, whereas I don't see how this would be possible in t-SNE given we will only ever have 2 or 3 coefficients and an unknown amount of explained variance in the data.

       The t-SNE analysis found the correct clusters of companies by industry using both the raw returns and the correlations, which I will admit is a bit surprising. I expected it more out of the preprocessed correlation data, however, the fact that the same results can be produced with the raw data (same clusters, just different coordinates), with arguably a bit clusters that are a bit more accurate is impressive.

CONCLUSION

This assignment was an interesting one in that we looked at stock returns, which is one of my favorite types of data to work with (financial / economic time-series). The log returns were used as the normalization procedure, which is industry standard due to both the lognormal distribution of the resulting transformation, but the additional mathematical properties of being additive over time (as opposed to multiplicative with simple returns).

The topic of 'Statistical Graphic' vs 'Data Visualization' was approached, and I thought it was an appropriate topic for this type of cluster analysis, given that most of our work is to find statistical relationships in high-dimensional data and the majority of our workflow ends with a custom data visualization instead of a standard statistical graphic.

The PCA analysis itself was an interesting task, and I was admittivity a bit surprised on how accurately the PCA analysis picked up on the correlations of the returns between companies in the same industry and the resulting regression model derived from the first 8 components (8 = 85% explained variance), resulted in the same MAE as the "full" regression model built with all the available parameters in the data set, while also guaranteeing that there is no multicollinearity in the underlying data (orthogonality in the components).

The t-SNE was interesting. This is my first introduction to this technique, and I can see why it is a popular method. The raw stock data did not inherently lend itself to this dataset as far as I could tell, as the wide data-format needed to be either 1.) transposed to column major (columns 2:N -> returns, rows = companies), or 2.) pre-processed into a correlation matrix (which, again gives us the companies as rows) in order to correctly label the data so that we can interpret the results. Running the t-SNE on the wide data format produced just a scatter plot, not even split out into clusters no matter what parameters I used.