Assignment 1

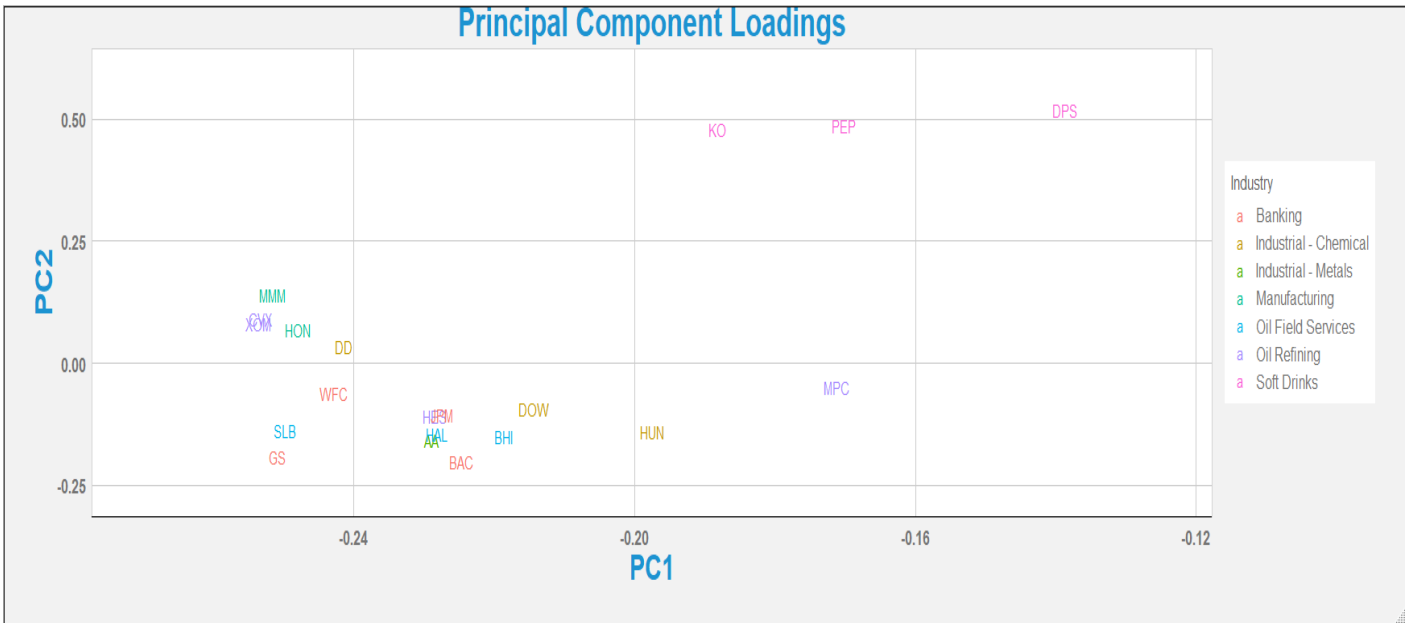
April 2021

2) We are most interested in the row of the correlation matrix that has the highest correlation to the corresponding row or value that we are looking for. We are less interested in the columns and rwos that don't have a high correlation.

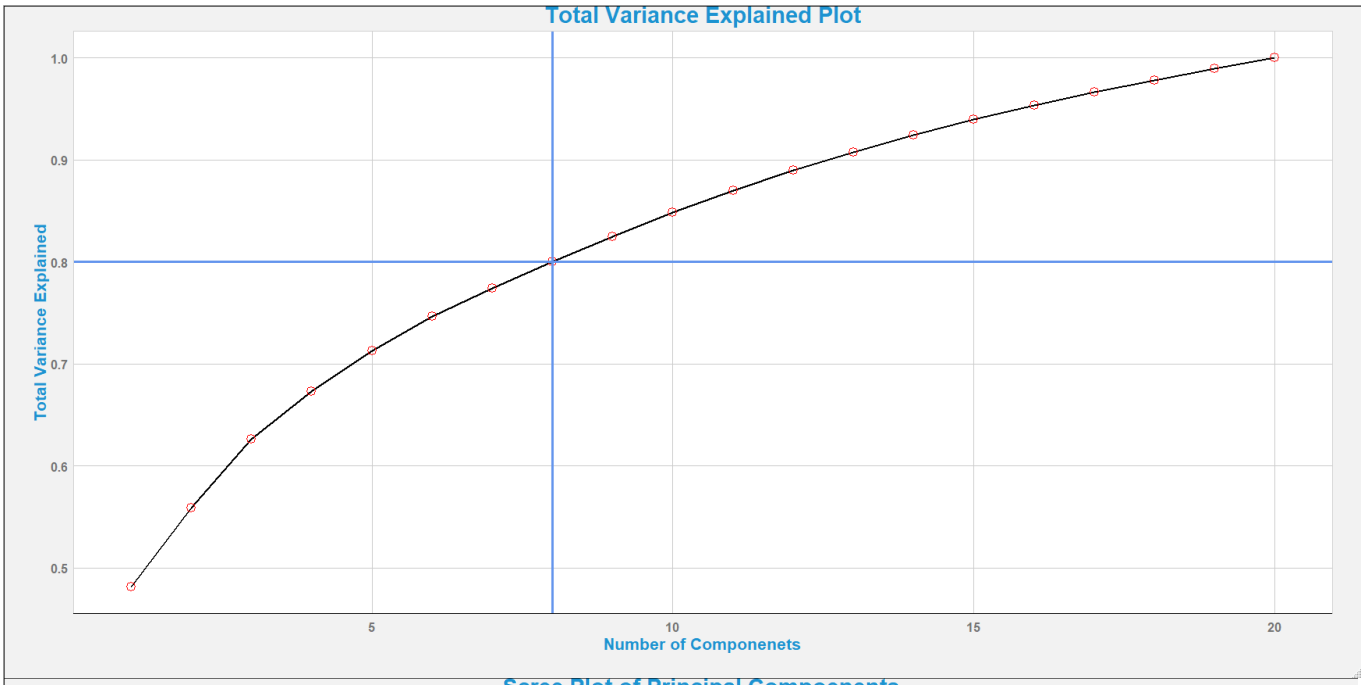
3) The corrplot is different than the bar plot but different does not necessarily mean better. They each have their strengths and weaknesses and can display different data in different ways. The difference between a statistical graphic and data visualization is that data visualizations are generally molded into the story that the user is trying to get the data to convey. It has been shaped and formatted towards the story. Statistical graphics are more straightforward visualizations based upon standard statistical calculations with more traditional visualizations like bar plots and histograms. The three I picked with low VIf are: DPS:Dow, DPS:BAC, DPS:Hun; the three with high vif are: VW.WFC, HON.VV, CVX.XOM

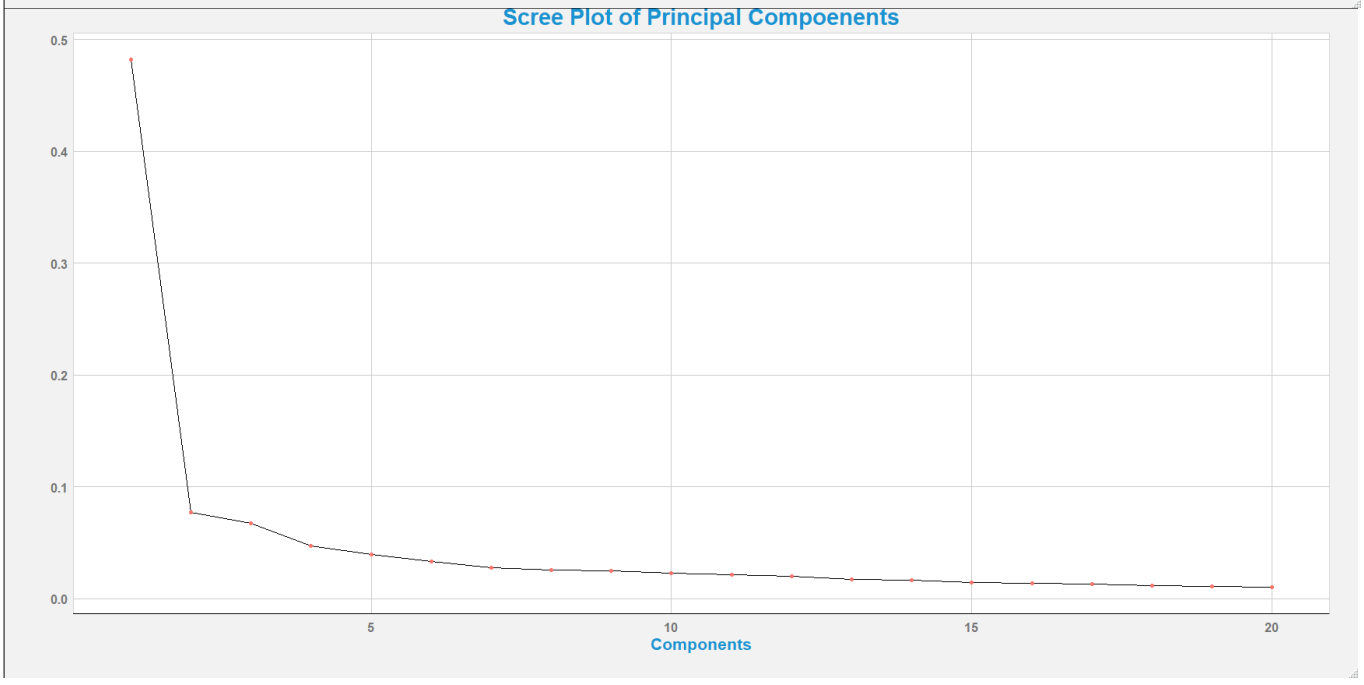
4) Multicollinearity is a concern for both models because several variables have VIF scores above 2.5 with more variables in model 2 of concern. The value of VIF that makes me concerned about multicollinearity are those ranging from 2.5 to 10 with anything about 2.5 being concerning as it grows to 10.

5) The loadings are correlation coefficients between the variables and factors. We see that the first two loadings seem to be the most grouped together. The surprises are that there are some values that cannot be grouped.

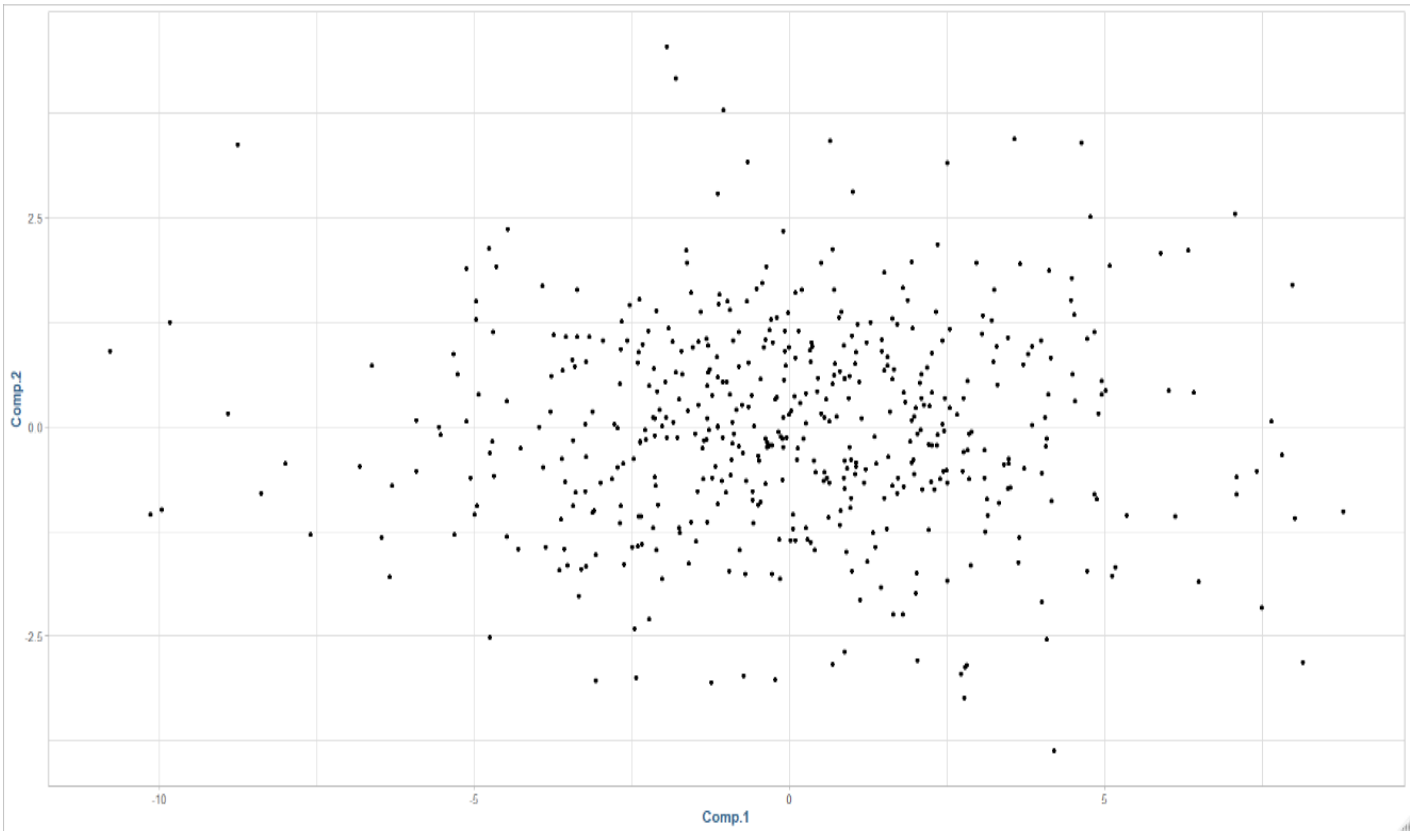


6) I think we should drop the vast majority of components because we are using the PCA to try to explain variance so we should only keep those components that help explain the variance. The decision rule of variance will help explain which components to keep and which to drop. We will keep the 8 components later, because after viewing the plot for variance we see that the top 8 components cover 80% of the variance, a very high amount, and should thus be included.


Total Variance Explained Plot
Total Variance Explained
1.0
0.9
0.8
0.7
0.6
0.5
5
10
15
20
Number of Componenets


Scree Plot of Principal Compoenents
0.5
0.4
0.3
0.2
0.1
0.0
5
10
15
20
Components

7) The PCA scores are very varied with the most being between -5 and 5 tapering off towards -10 and 10. They define the differnce from the origin for each component. The VIF values should all be one because it is an indicator of multicollinarity and thus should not stray too far.



8)

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

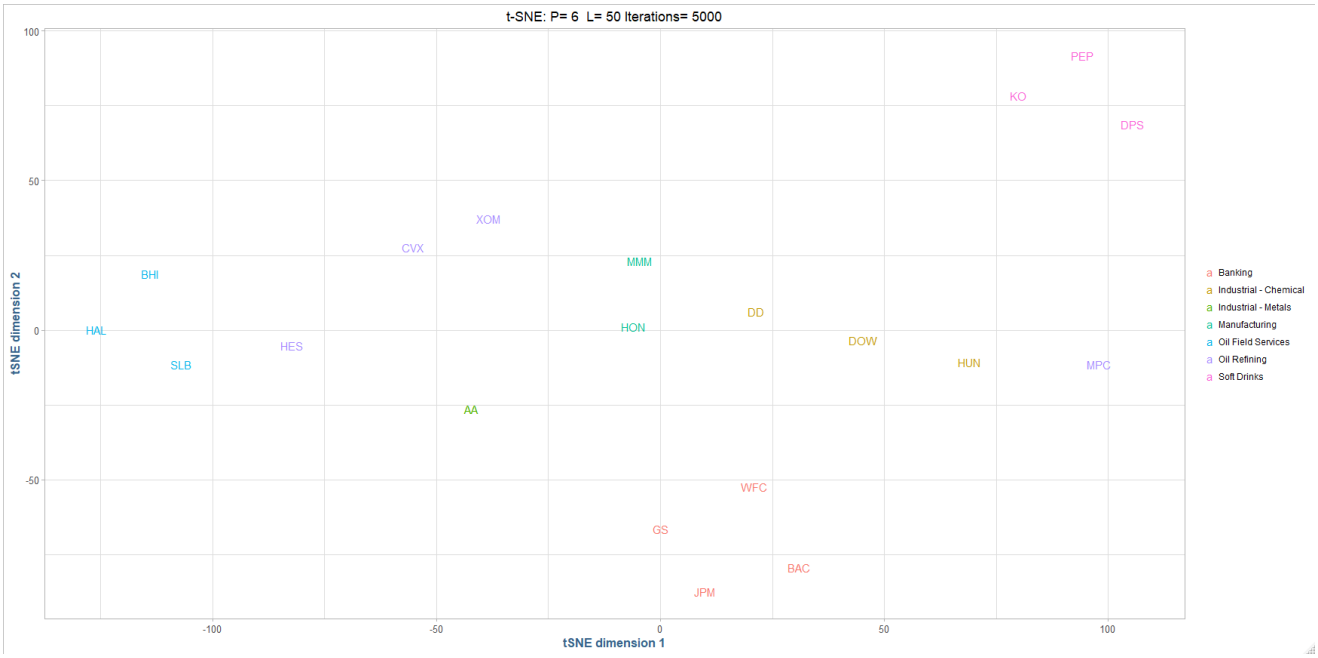
The model seems to be model 1. It has the best combined train and test scores and the highest test score. Although the pca and model 2 also tested comparably, it seems like model 1 gets the slight edge. It is interesting to note that the pca and model 2 had the same scores after components were laminated

down to eight for the pca, so the other components seem to have minimal affect on the outcome. Model1 1, on the other hand, is working on a smaller dataset so it's a possibility for overfitting.

9)

The variable selection approach suggests keeping eight components – specifically 1, 8, 10, 14, 2, 3 , and 7. This differs from our previous discussion because the components that are to be kept are different. Previously we kept numbers 1-8, but now we are keeping 10 and 14 in lieu of the others.

10)



The t-SNE model correctly grouped the data and found correct clusters based upon data provided and correlations. The model performed well despite limitations and could be favorably compared to the pca model and other models used in this assignment. The results are different than the pca model we conducted previously because the models are different in how hey are set up. The pca is a more linear while the t-SNE analysis is more stochastic. These difference show up in the results as well where the pca model took is components from the dataset while the t-SNE analysis projects dimensions into tow or three dimensional spaces. The pca model opens up potential further analysis because of the mulit-component aspect of the model, which t-SNE cannot because it si limited to two or three dimensions.

11)

This assignment was an interesting study in unsupervised learning for a numerous amount of reasons. It was good to note the main differences between statistical graphics and data visualizations and how we show the results of our data science and analysis. The pca model was an important task and is an important part of data science. Most time we will have to edit the amount of components applied to a model, so it was interesting to see how it would affect the outcome. Here, we only used the top 8, and

although it did not have much of an impact, it was interesting to see the results. Lastly, the t-SNE model also showed interesting results. It correctly found the clusters with improved accuracy from our previous models.