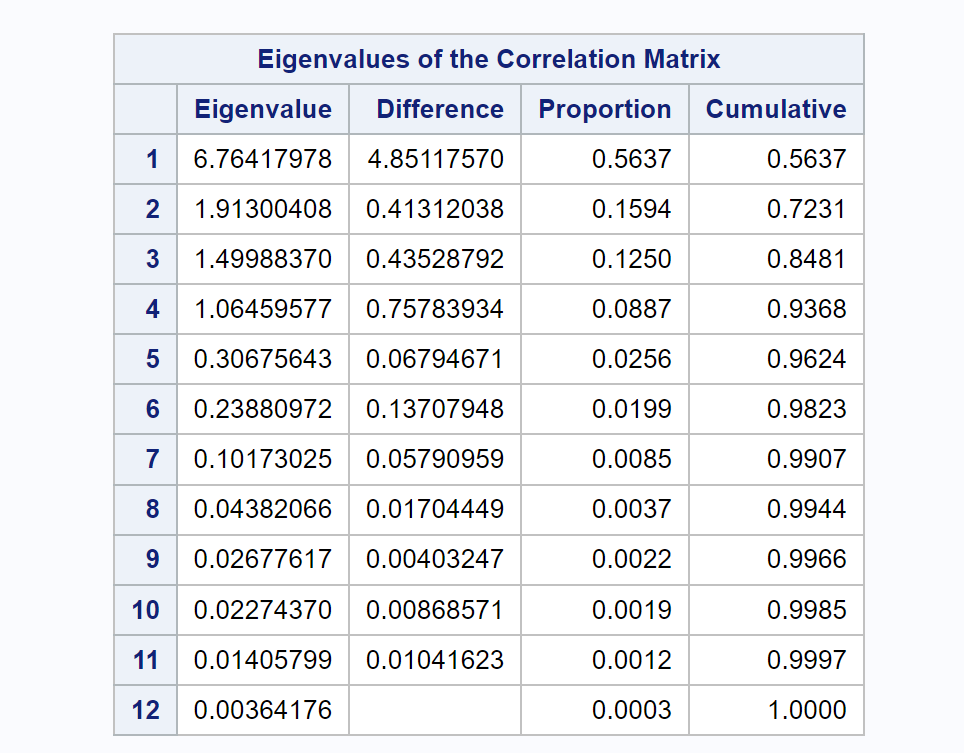
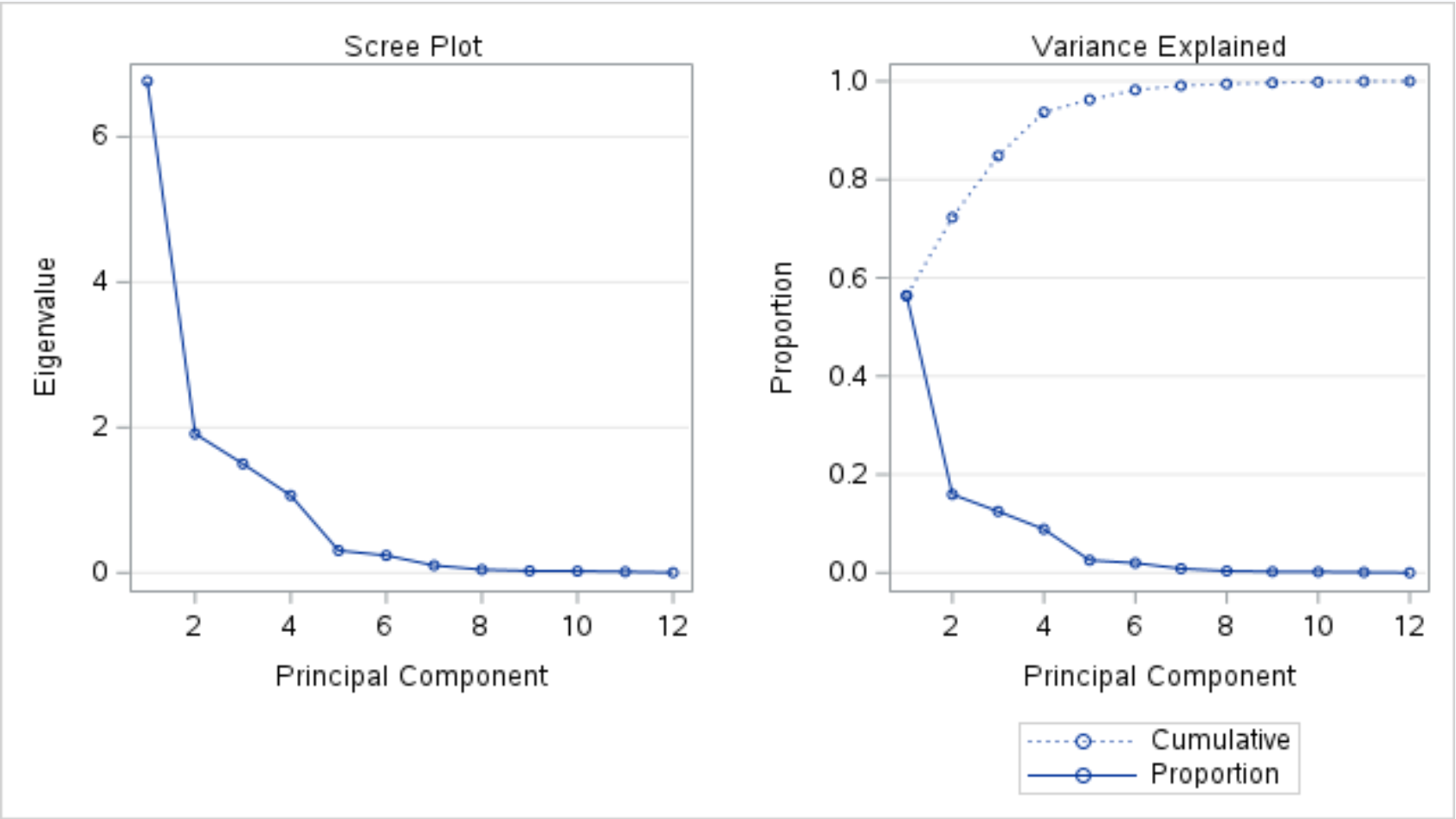
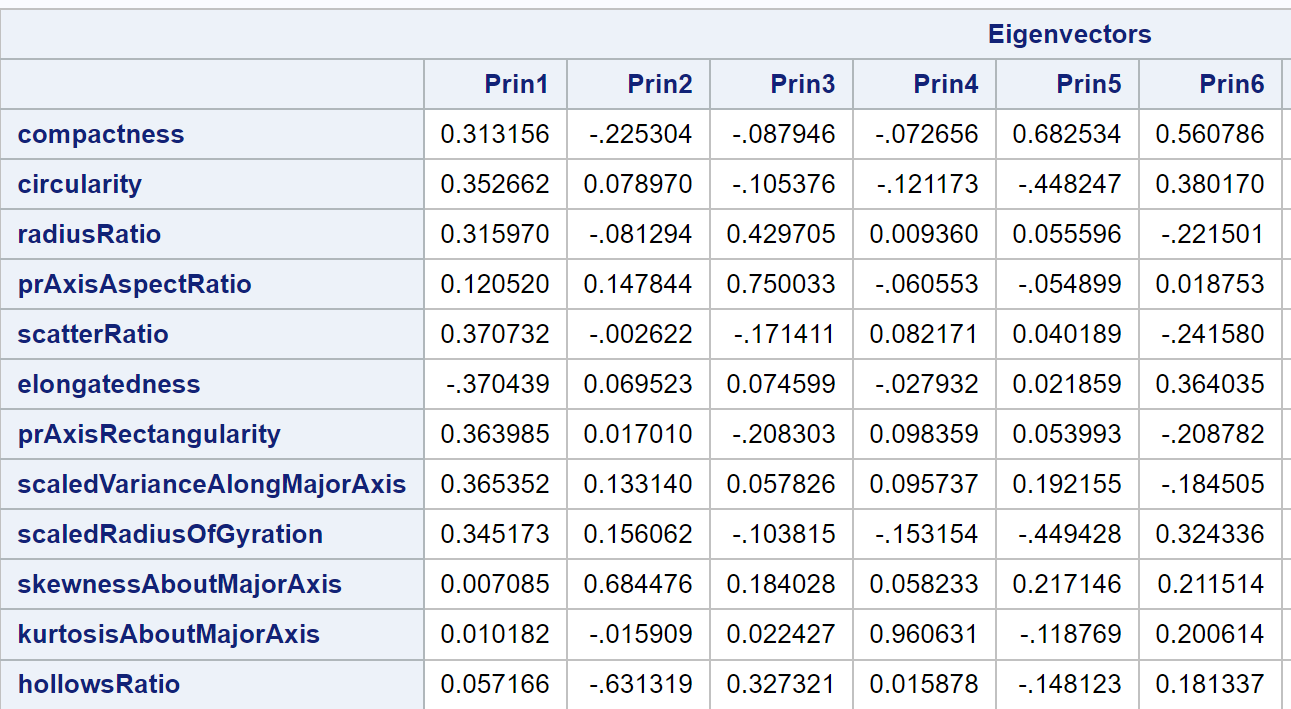
HW 6

## Problem 1

### 1a.)

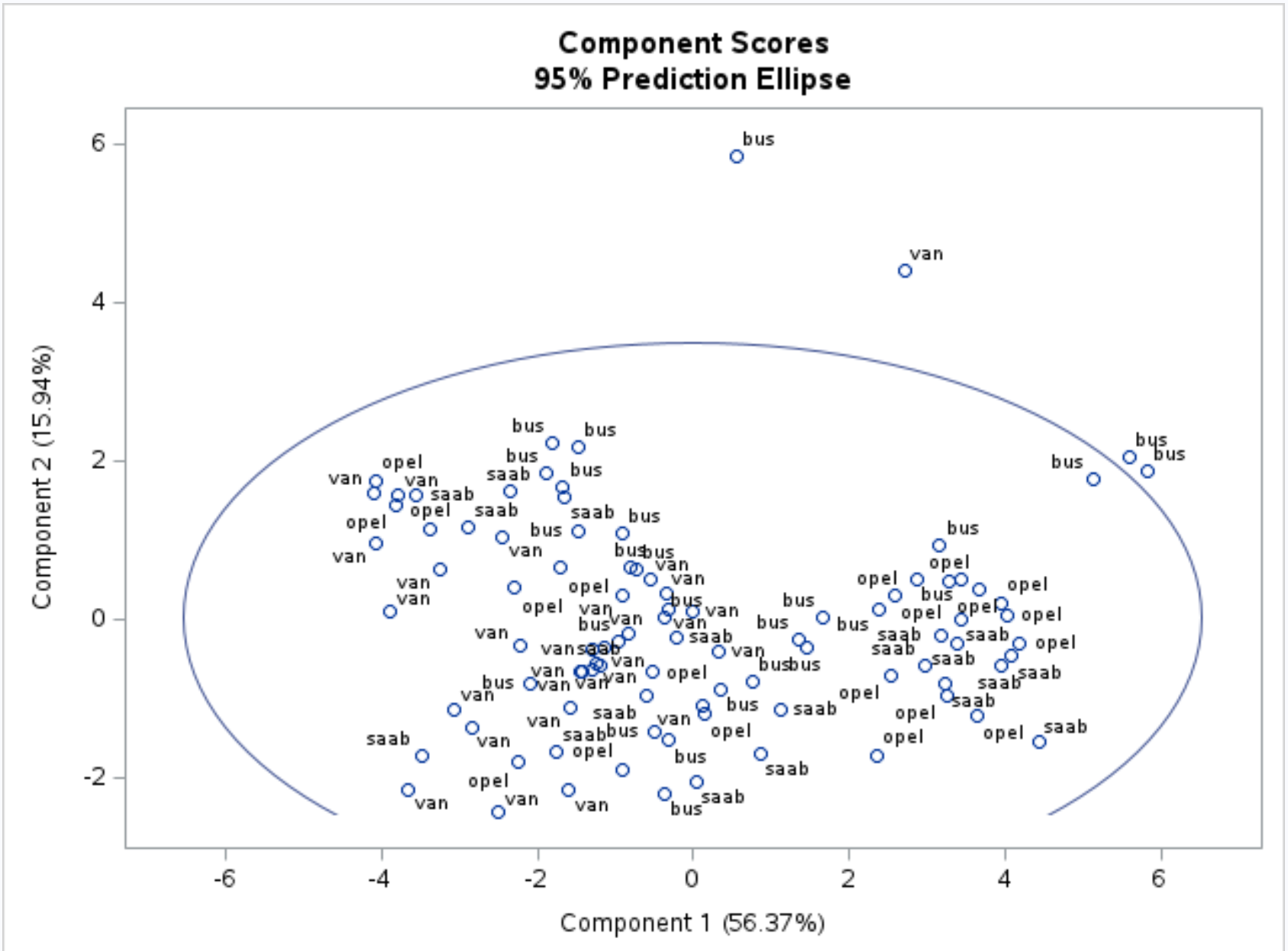
  
Too retain at least 85% of the variation in the original data, we would need 4 Principal Components which gives us about 93% of the variation in original data. However, with 3 Principal Components we get close to, but just under 85% of the variation with 84.81% of the data. So we would choose 4 Principal Components.  
Using the average eigenvalue method, we keep PCs with eigenvalue above the average which in this case is 1. So here we would also select the first 4 PCs.  
  
The Scree Plot suggests we use 5 PCs since that is where the "elbow" of the scree plot is, or where the values on the plot start to go very close to horizontal.

### 1b.)

  
For the first Principal component, the variables that appear to have the most influence appear to be scatterRatio, elongatedness since they have the largest absolute value. While close behind are the variables prAxisRectangularity, scaledVarianceAlongMajorAxis, circularity and scaledRadiusOfGyration.

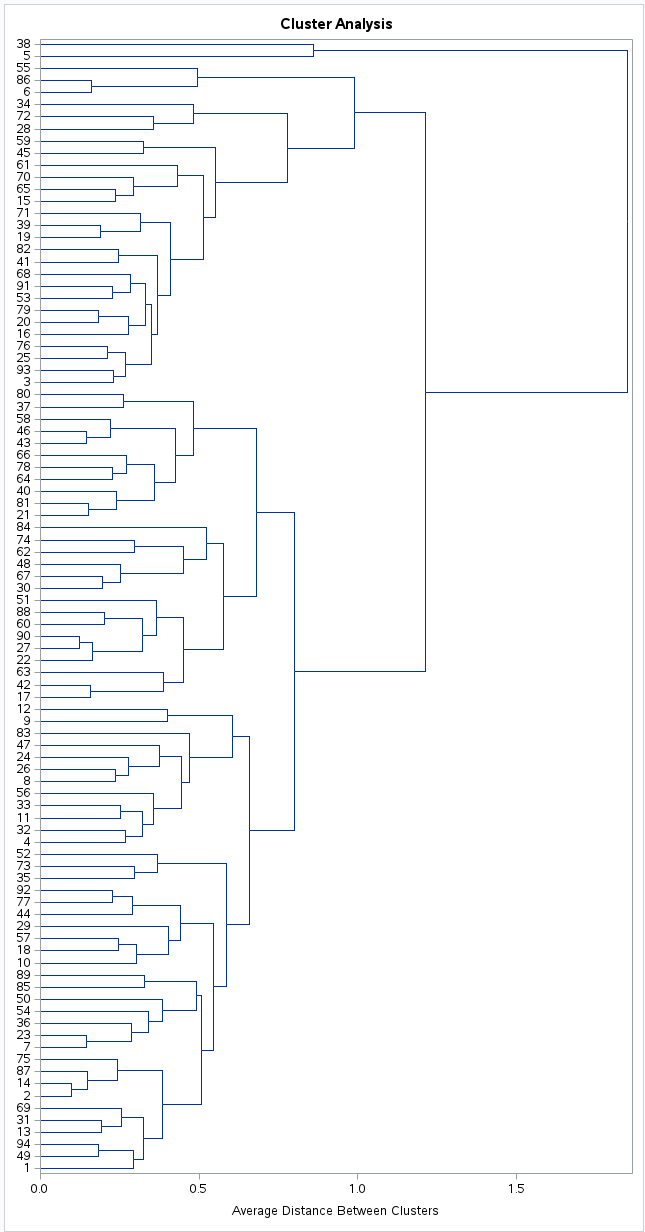
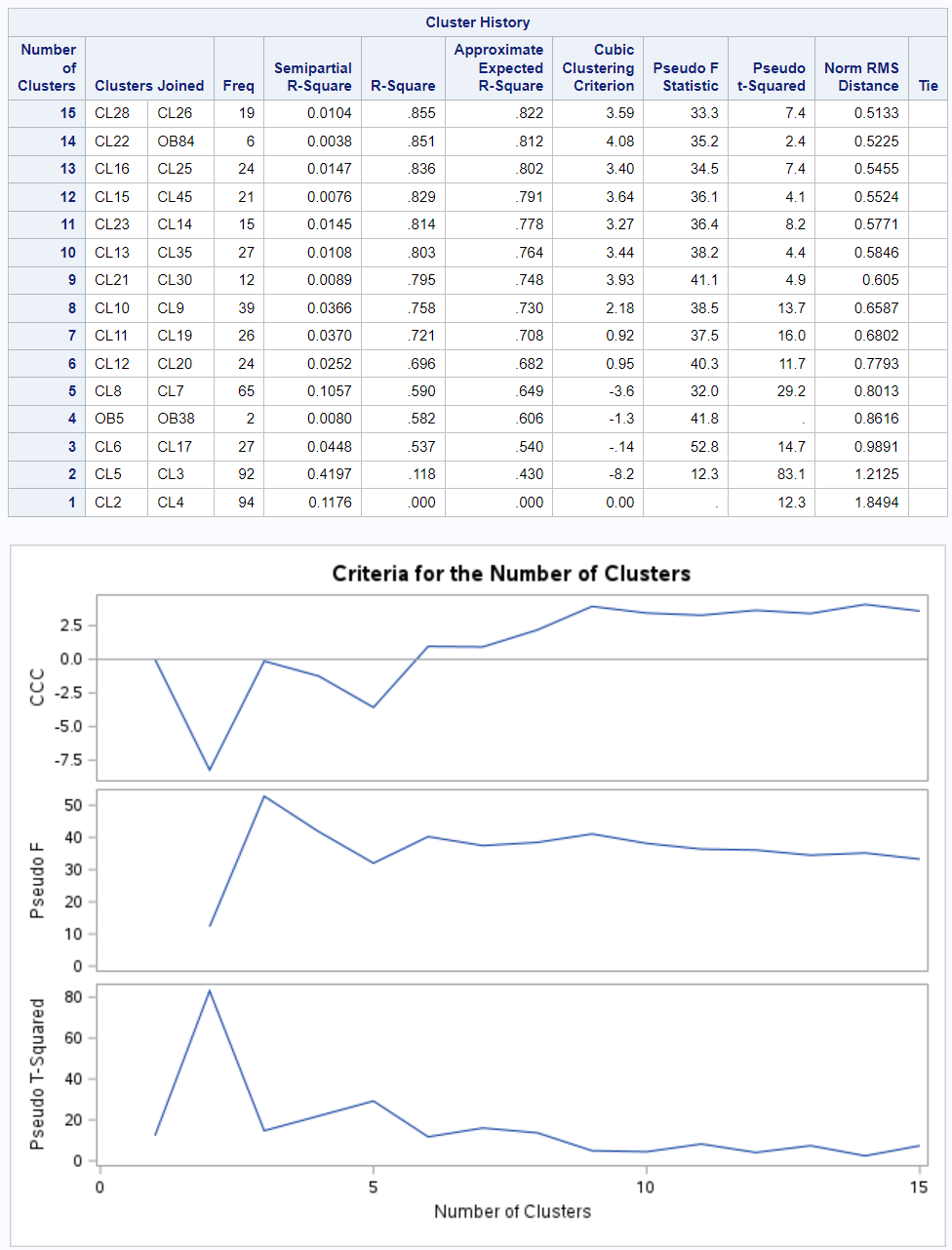
For the Second Principal component, the variable skewnessAboutMajorAxis and hollowsRatio had by far the largest influence on the component. The next closest variable is compactness which is pretty far behind in terms of absolute value.

### 1c.)

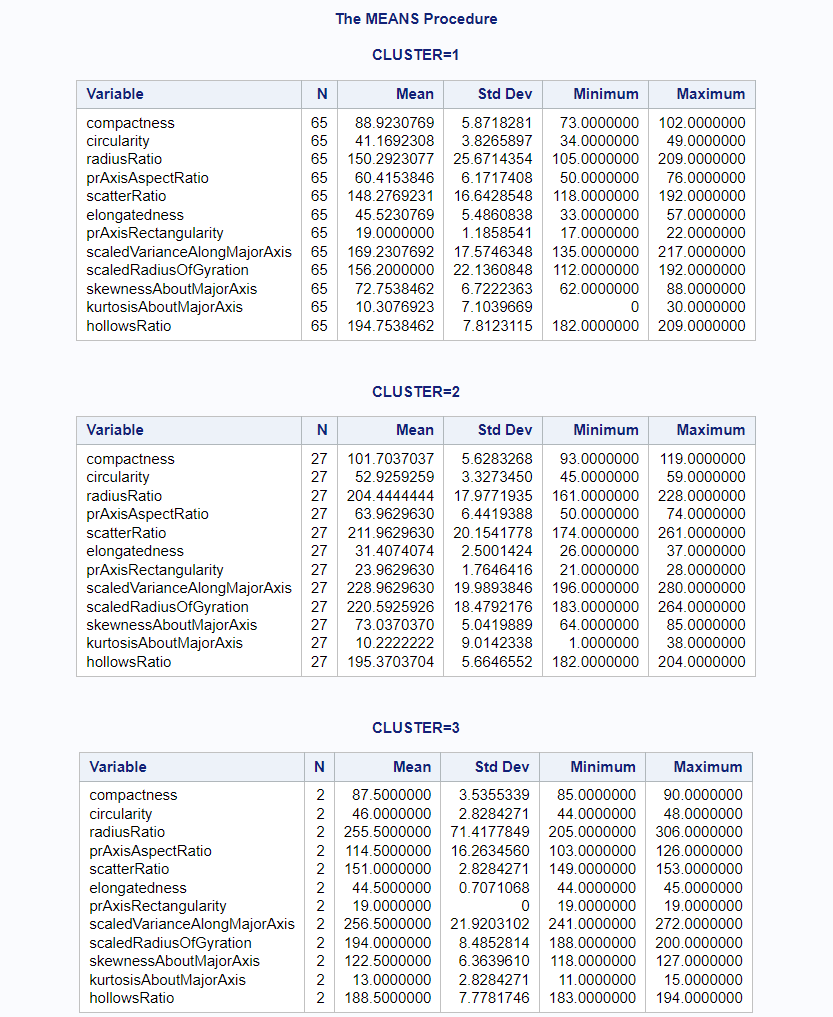
  
We see 3 bus and 1 van extreme observation in this plot. We would have a rather poor separation of vehicle types based on the grouping from this plot. We might do a good job of separating Opel & Saab observations from Van & Bus observations but we would struggle to distinguish the observation types within these two subgroups since on the right there is a cluster containing the majority of saab & opel observations and to the left of that cluster is most of the vans & bus with a few opel & saab observations.

## Problem 2

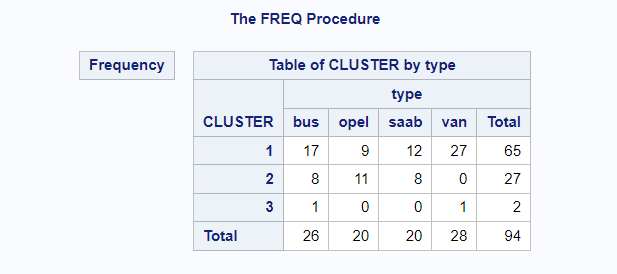
### 2a.)

  
Based on the dendrogram, a good value for number of clusters would be either 4 or 3 clusters as that keeps the clusters pretty well separated.  
  
Based on the ccc, we would choose either 1 or 3 clusters since that is where we have a significant drop in ccc in the G + 1 cluster, with 1 having the largest such drop after it.  
Based on the pseudo F we would choose 3 clusters since that is where the peak occurs in the in the graph and after 3 clusters we would have the most significant drop in pseudo F values.  
Based on the Pseudo t squared, we would pick 2 clusters since that is where the most prominent peak then dip are.

### 2b.)

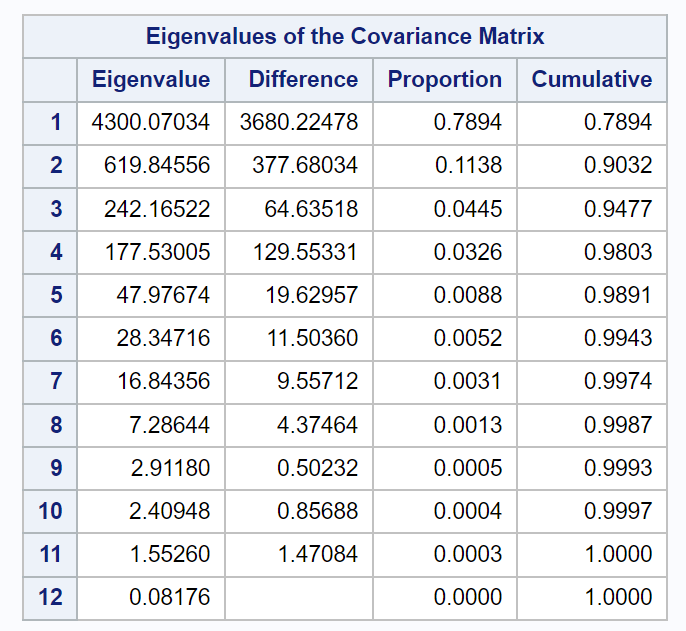
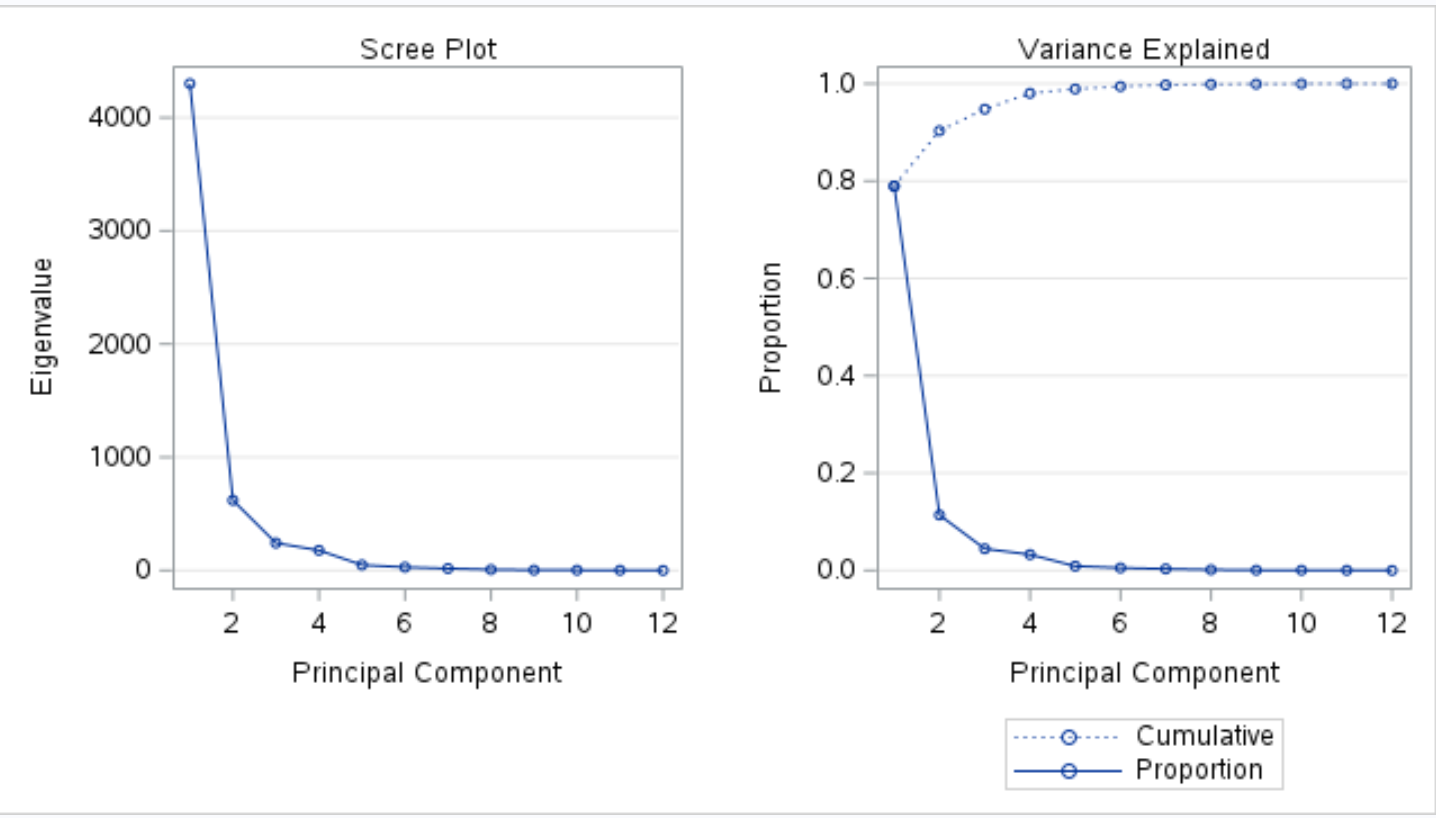
I chose 3 clusters as that appears to the most common number of clusters among all the methods of deciding potential clusters, here are the descriptive statistics:  
  
What's interesting in that cluster one has significantly more data points compared to the last two clusters. Additionally, most of the means appear to be pretty close to each other, however for radius ratio there is a large difference between the means among the 3 clusters. There is a large difference in scatter ratio between cluster 2 and clusters 1 & 3.

### 2c.)

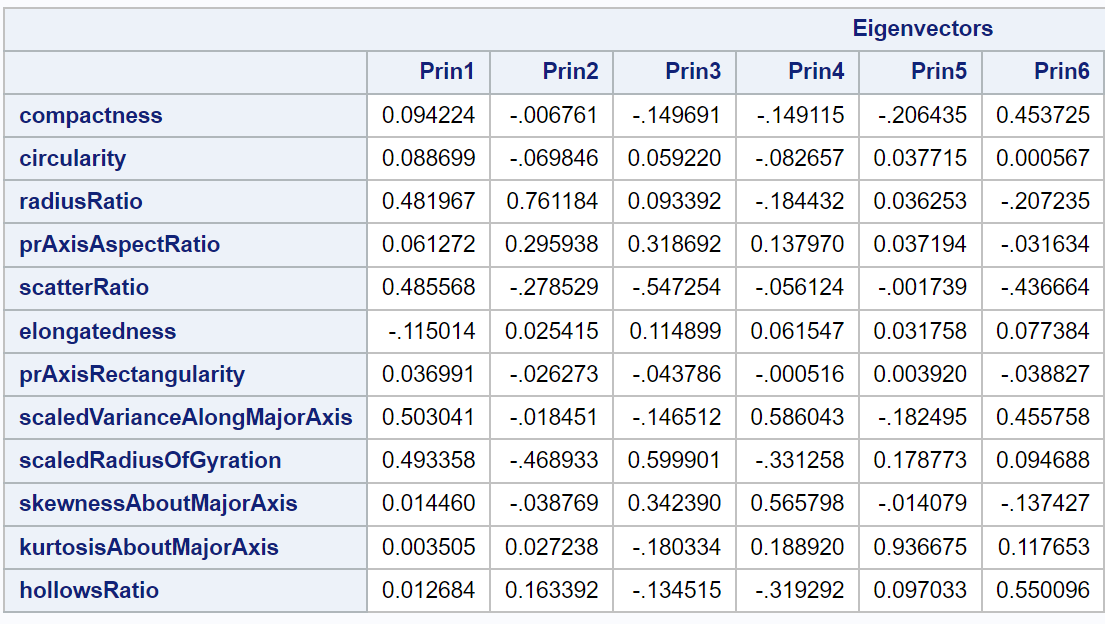
  
Cluster 1 does best when predicting Bus or Van while cluster 2 is better for distinguishing Opel & Saab. However, overall these cluster do pretty poorly in classifying the type of the observation as we see a lot of overlap.

## Problem 3

### 3a.)

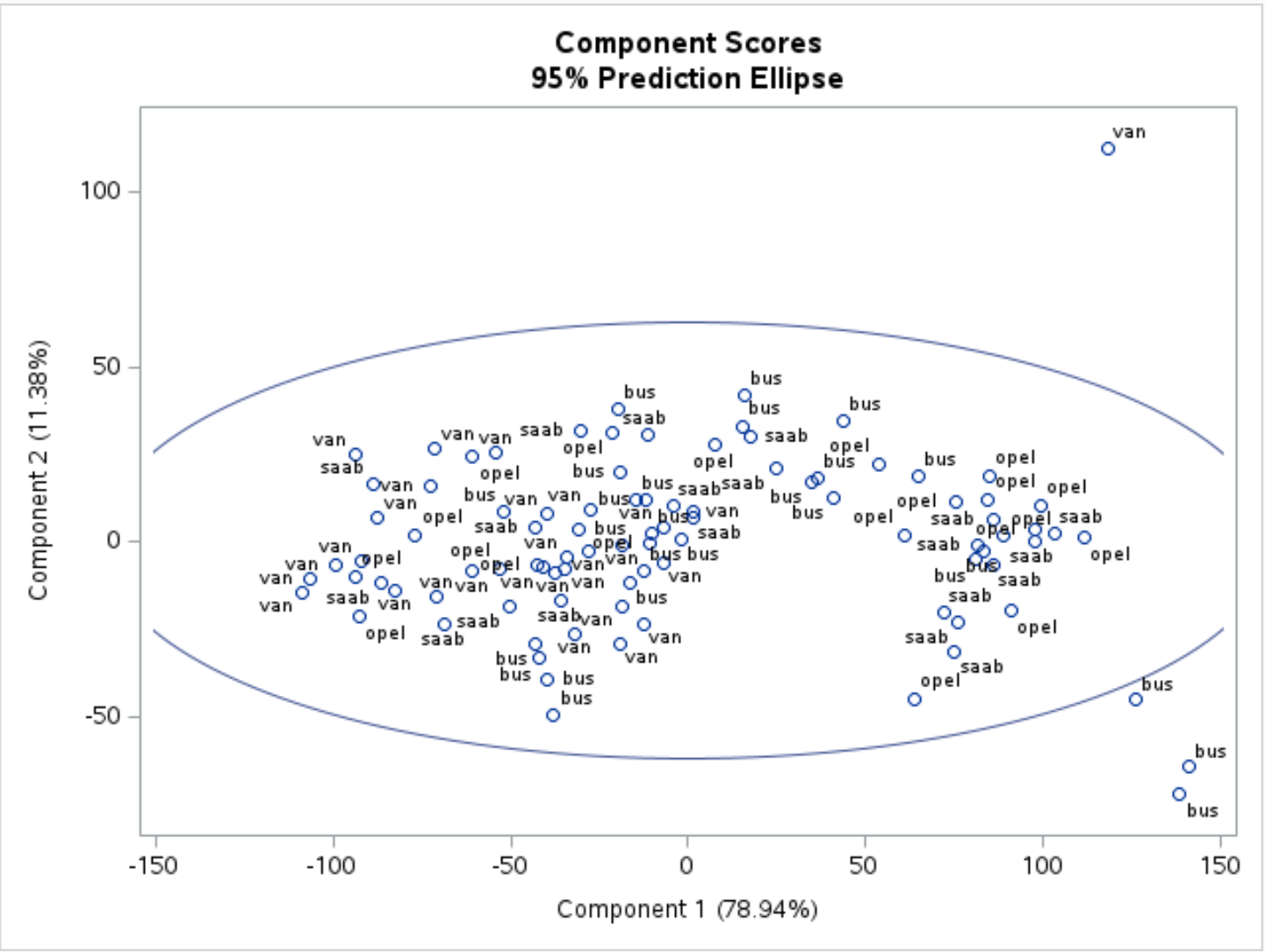
  
To get at least 85% of the variation we would pick 2 PCs based on our analysis.  
For average eigenvalues, the average is 453.91so we would pick the first 2 PCs since they are the only ones with an average higher than  
  
From the scree plot, the elbow is between 2 and 5 PCs. Picking 3 PCs would be a good estimate from the plot since after 3 the plot values are generally pretty horizontal.

### 3b.)

  
For the first PC, the most influential predictors appear to be scaledVarianceAlongMajorAxis, scaledRadiusOfGyration, radiusRatio and scatterRatio.

For the Second PC, the most influential predictors appear to be radiusRatio followed by scaledRadiusOfGyration. While all the other predictors are much less influential.

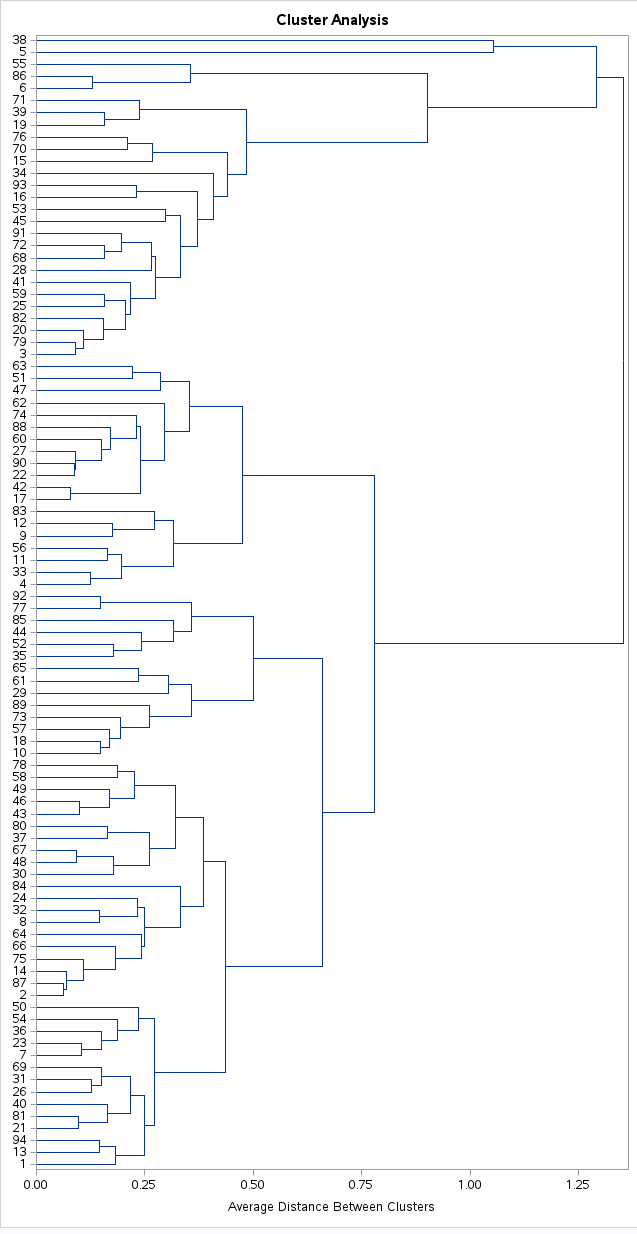
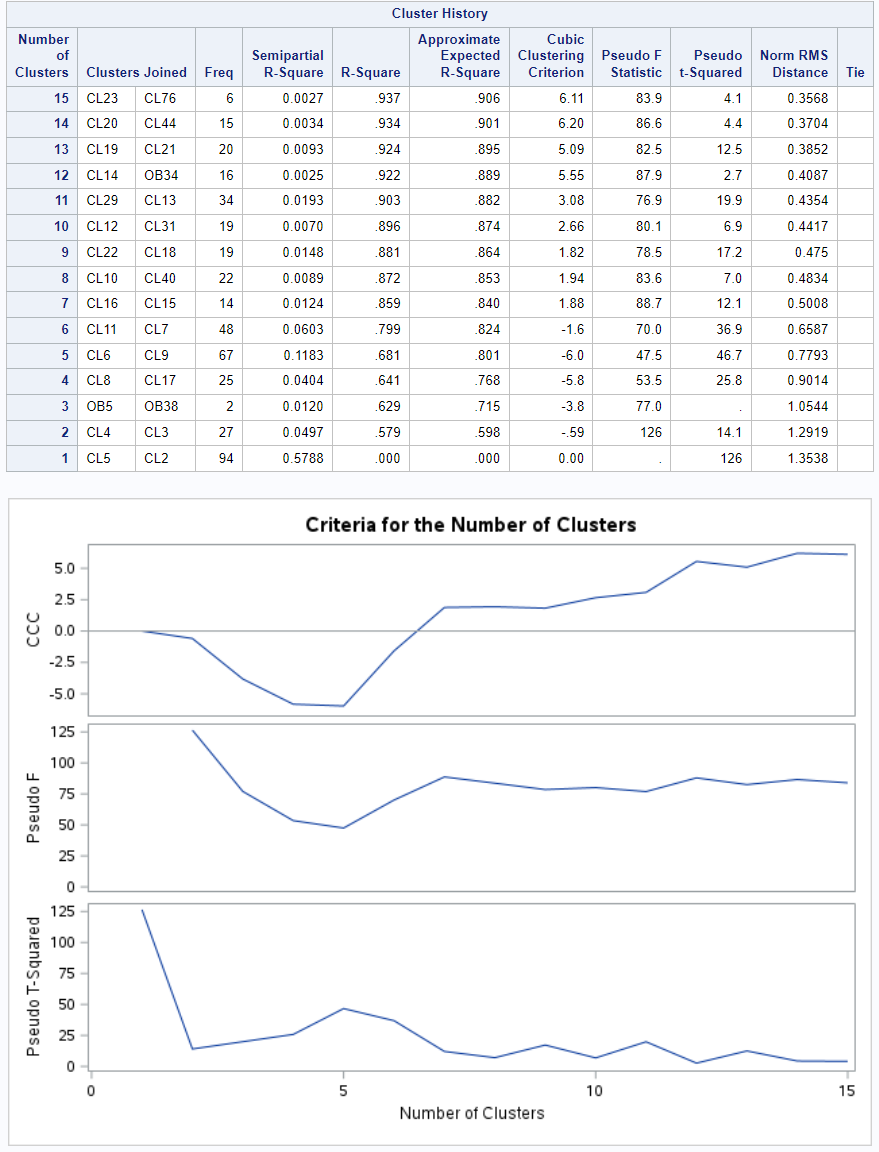
### 3c.)

  
There are 3 bus and 1 van extreme points in this chart. Once again based on these PCs we can separate Opel & Saab from Bus & Van observations pretty well as there is a cluster on the right of opel & saab datapoints and then bus & van datapoints are towards the left. However there is an amount of opel and saab observations on the left so this distinct won't be perfect.

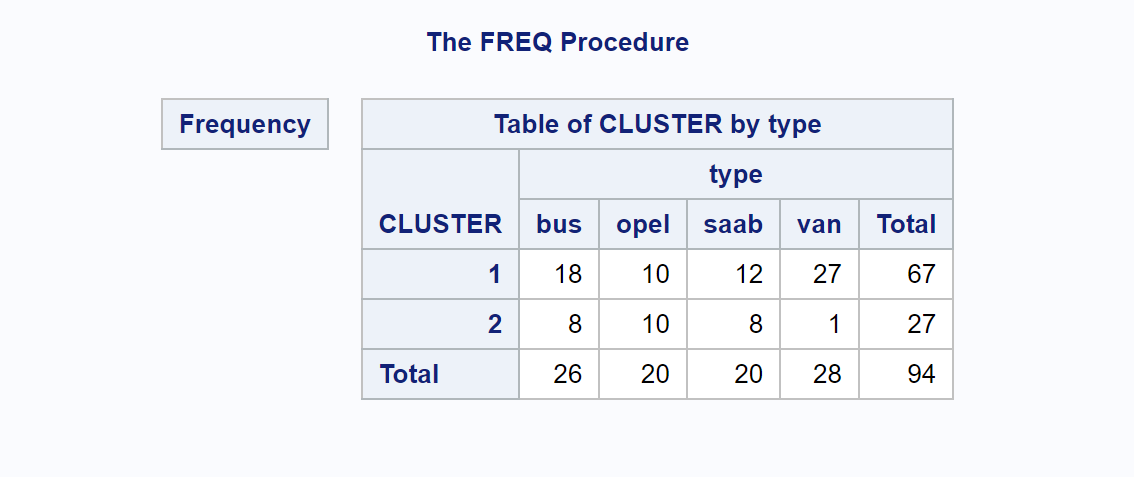
The covariance based approach needed less PCs to account for more variation in the data however looking at just the first two PCs, both appear to be about equal in separating out the different types. Both models have a grouping of opel and saab observations on the right of the score plot and then a larger more dispersed grouping of van and bus observations on the left.

## Problem 4

### 4a.)

  
Based on the dendrogram, a good amount of clusters would be 2 or 3 since that keeps a goo average distance between clusters.  
  
For the CCC, it suggests using 2 clusters.  
The pseudo F statistics suggests using 2 clusters.  
The pseudo t squared suggests using 1 cluster.

### 4c.)

Since 2 was the most common cluster value, I chose to use 2 as the number of clusters.  
  
As we can see this does a poor job overall of separating types, however we see that most of the van and bus observations end up in the first clusters, while opel and saab observations are generally evenly distributed among the 2 clusters.

Overall the analysis suggests that we aren't able to classify vehicle type extremely well using cluster analysis, neither of the two (standardize vs unstandardized) did a better job at classifying vehicle types when only look at all 4 types. If classifying buses and vans combined into one group, our models might do a better job.