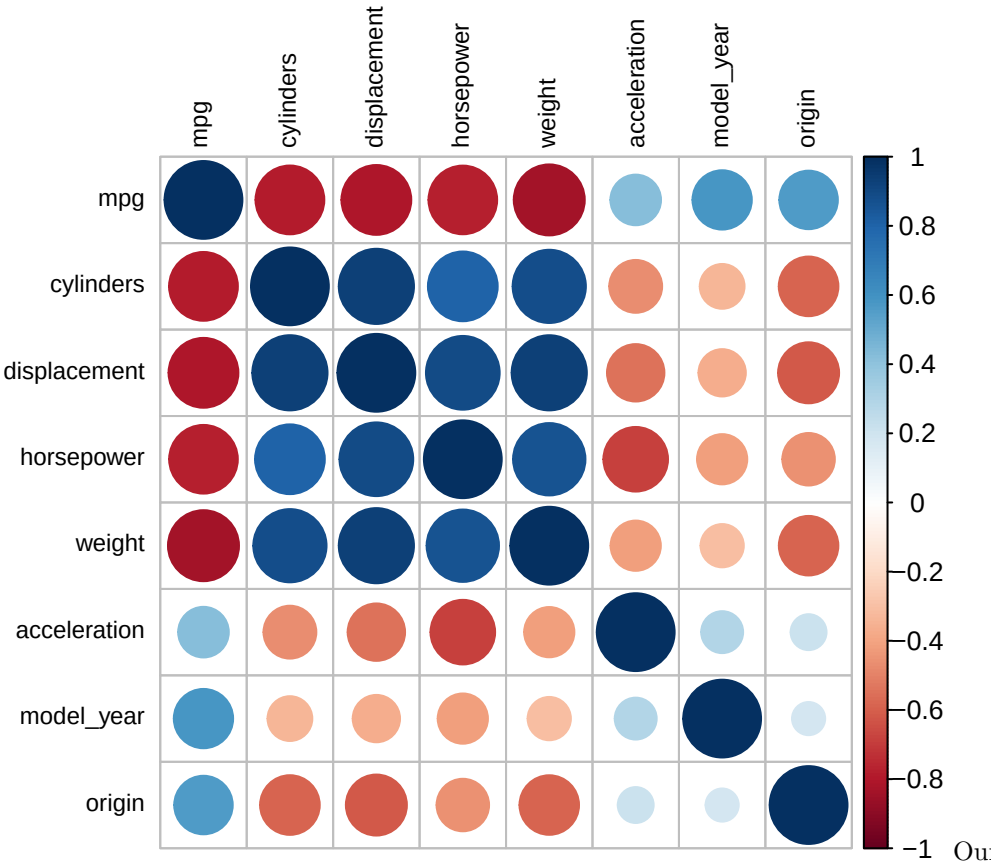
Factor Analysis and MDS on auto mpg data

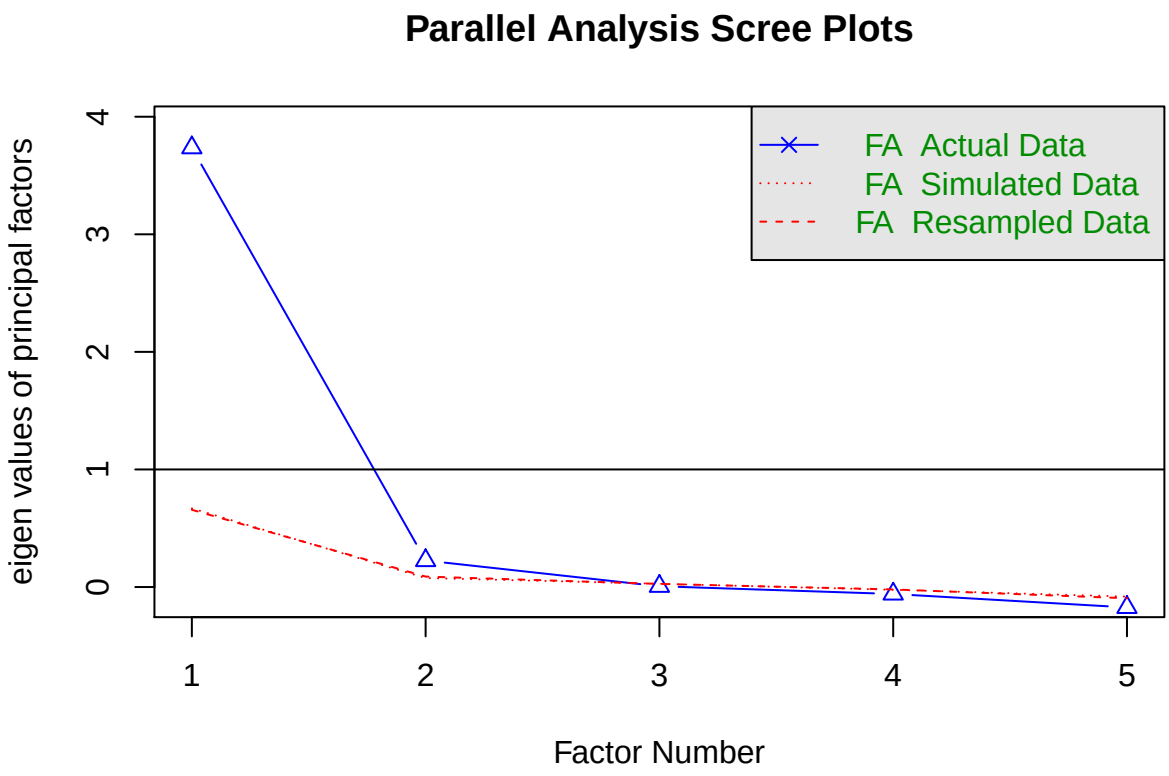
Katherine Wilkinson

2/7/2018

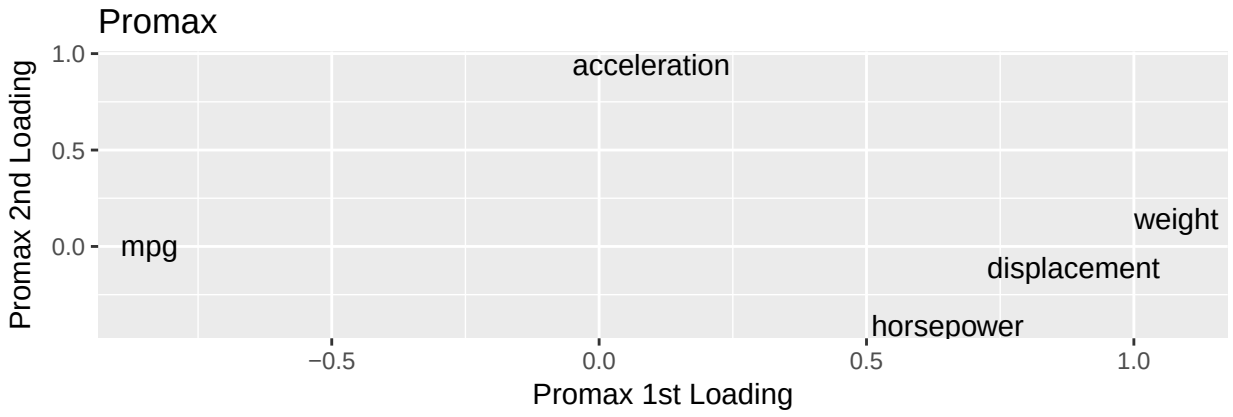
In our previous report we did Principal Componenets Analysis on our numerical automobile data to reduce dimensionality. Here, we will try a couple of other approaches, Factor Analysis and Multidimensional scaling and compare to our PCA results.

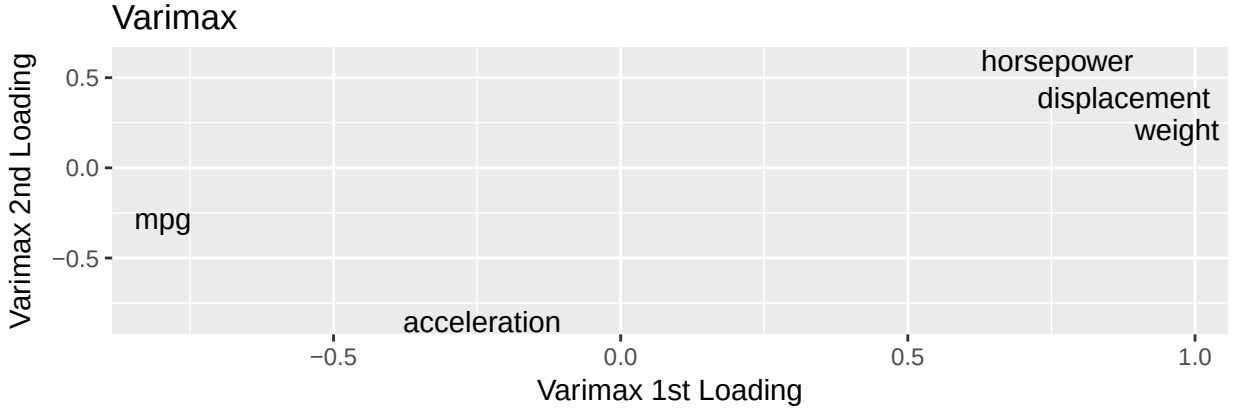


Our original dataset looking at different automobiles has 8 variables with 398 observations. There are a number of categorical variables, including the car name, origin, model_year, and cylinders. We can see from the correlation plot with all variables except car name (of which there are 305 unique names) we can see there is some significant correlation between variables. Most interesting, there is some correlation with our categorical variable cylinders and our all of our numeric varaibles except for acceleration. The number of cylinders is positively correlated with displacement, horsepower, and weight while it is negatively correlated with mpg. With the exception of acceleration, all of our numeric variables are fairly highly correlated with one another. MPG for instance is negatively correlated with displacement, horsepower and weight suggesting that the bigger the car, the less miles per gallon it will get. It may also be important to note that there are a number of NA values in the data set that we will remove to do our Factor Analysis, multidimensional scaling, and PCA.




```
## Parallel analysis suggests that the number of factors =  2  and the number of components =  NA
```





To start our Factor Analysis, we select just our 5 non-categorical variables as we did with PCA (MPG,

displacement, horsepower, weight, and acceleration). There are a number of different approaches to Factor Analysis. Initially we look at a scree plot of our scores and see that 2 should be the optimal number to use. We can view these first two loadings with two different rotation methods, Promax and Varimax and see how the loadings are rotated quite differently.

```
##                Factor1-P    Factor2-P  Factor1-V  Factor2-V
## mpg          -0.84079059  0.006663625 -0.7970997 -0.2806577
## displacement  0.88726198 -0.106632154  0.8763440  0.3893473
## horsepower    0.65164485 -0.408570412  0.7603018  0.5949123
## weight        1.07951575  0.146348624  0.9686937  0.2154295
## acceleration  0.09757409  0.943297348 -0.2410265 -0.8506209
```

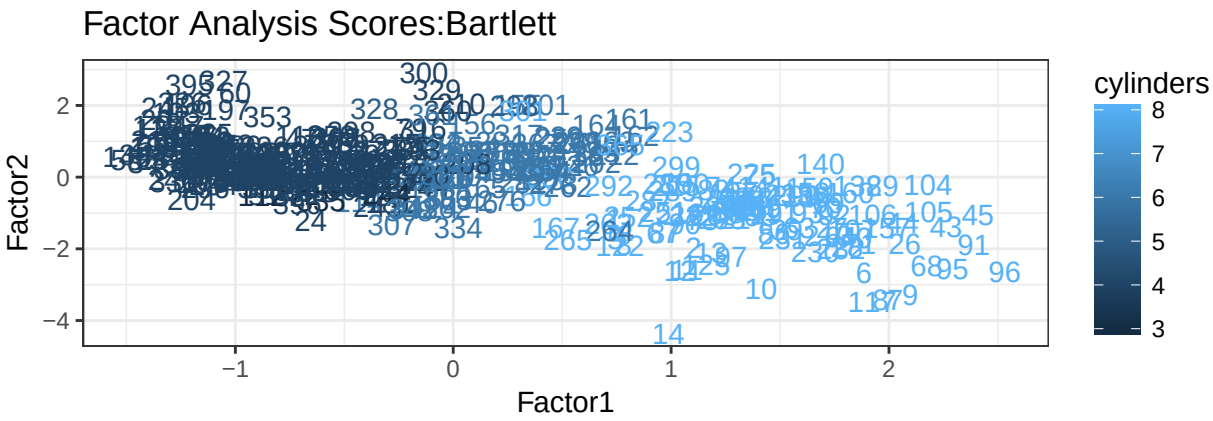
```
##
## Loadings:
##              Factor1 Factor2
## mpg          -0.841
## displacement  0.887  -0.107
## horsepower    0.652  -0.409
## weight        1.080   0.146
## acceleration          0.943
##
##                Factor1 Factor2
## SS loadings      3.094   1.090
## Proportion Var   0.619   0.218
## Cumulative Var   0.619   0.837
```

```
##
## Loadings:
##              Factor1 Factor2
## mpg          -0.797  -0.281
## displacement  0.876   0.389
## horsepower    0.760   0.595
## weight        0.969   0.215
## acceleration -0.241  -0.851
##
##                Factor1 Factor2
## SS loadings      2.978   1.354
## Proportion Var   0.596   0.271
## Cumulative Var   0.596   0.866
```

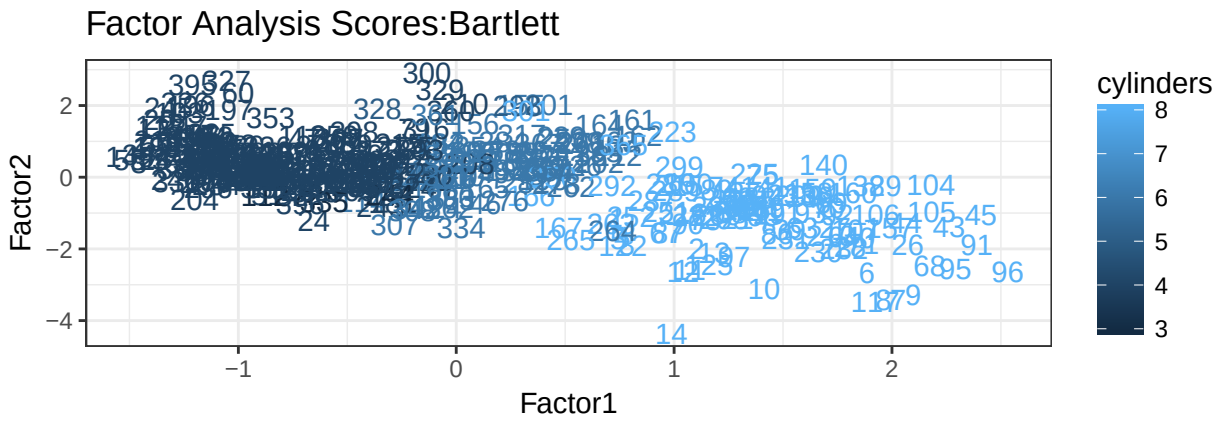
When looking at the different rotations, we can see that the Promax rotation method is slightly easier to interpret. Also, since the cumulative amount of variance explained by the 2 FActors using promax rotation is 83.7% we will use the promax results for our analysis. That is, a couple of the variables have factor loadings that are small enough to consider the effect of the factor on that variable to be negligible. More specifically, Factor 1 for acceleration is small at 0.098 and Factor 2 for MPG is small at 0.007. Displacement (0.887), horsepower (0.652), and weight (1.079) have large positive loadings on Factor 1 while mpg (-0.841) has large negative loading on Factor 1. This also follows what we can see in the first correlation plot, as mpg is negatively correlated with the other variables. For Factor 2, now acceleration (0.943) is has a very large positive loading while hosepower (-0.408) has a large negative loading. Once again, this reflects our correlation plot where we can see that acceleration and horsepower are slightly negatively correlated. It should also be noted that displacement (-0.107) and weight (0.146) have a slightly large negative and positve loading respectively on Factor 2.

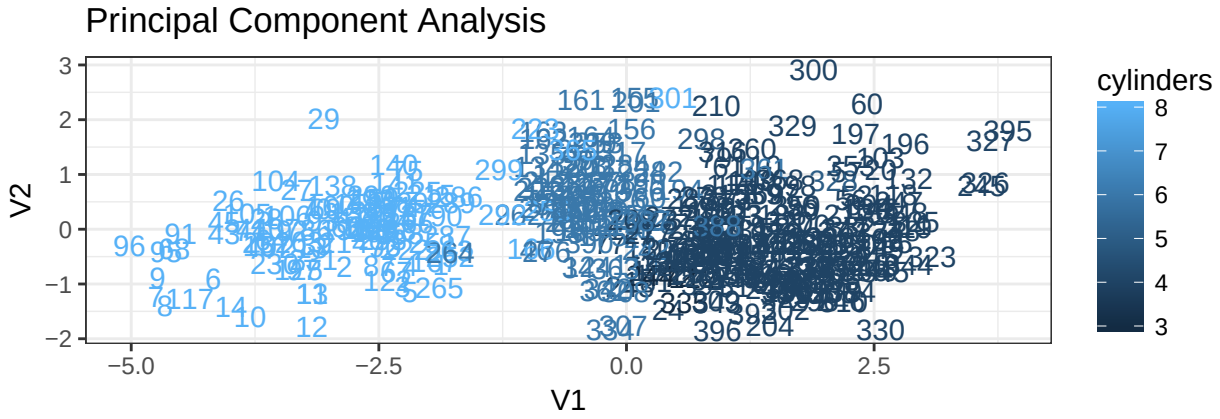
```
##                    PC-1        FA-1       PC-2         FA-2
## mpg           0.4442640 -0.84079059 -0.3038692  0.006663625
## displacement -0.4832332  0.88726198  0.1347900 -0.106632154
```

```
## horsepower   -0.4844417  0.65164485 -0.1242676 -0.408570412
## weight       -0.4712207  1.07951575  0.3263218  0.146348624
## acceleration  0.3352350  0.09757409  0.8761089  0.943297348
```



Next, we can look at a couple of different methods to calculate our scores for our Factor Analysis; Regression and Bartlett. Again, we can see in our plots that there are some definite differences between the two methods. From our previous analysis using PCA, we saw that there are some distince attributes from the cylinder variable. With this knowledge and from our plots, we will use the Bartlett method as it gives us a more distinctive and interpretable separation by cylinder.

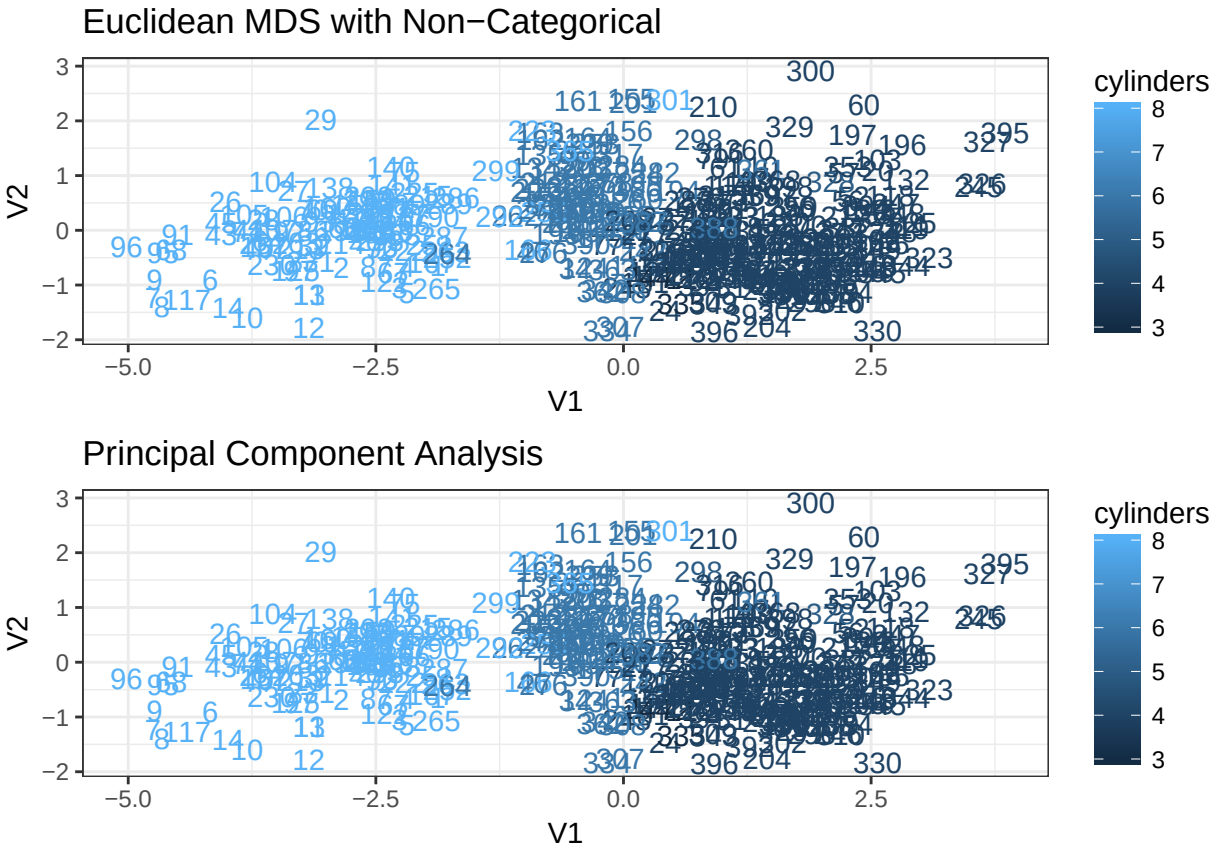




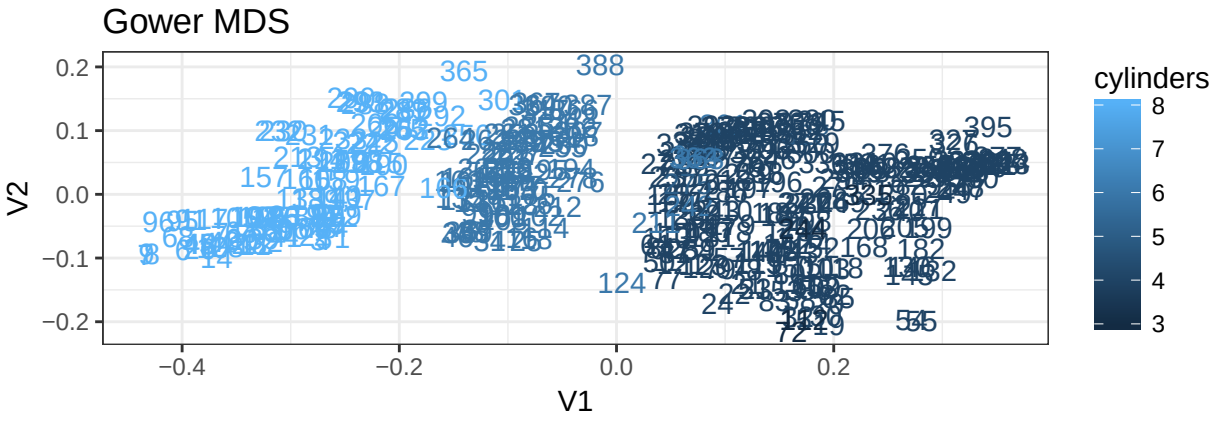

```
##                 Comp.1     Comp.2     Factor1      Factor2
## mpg           0.4442640 -0.3038692 -0.84079059  0.006663625
## displacement -0.4832332  0.1347900  0.88726198 -0.106632154
## horsepower   -0.4844417 -0.1242676  0.65164485 -0.408570412
## weight       -0.4712207  0.3263218  1.07951575  0.146348624
## acceleration  0.3352350  0.8761089  0.09757409  0.943297348
```

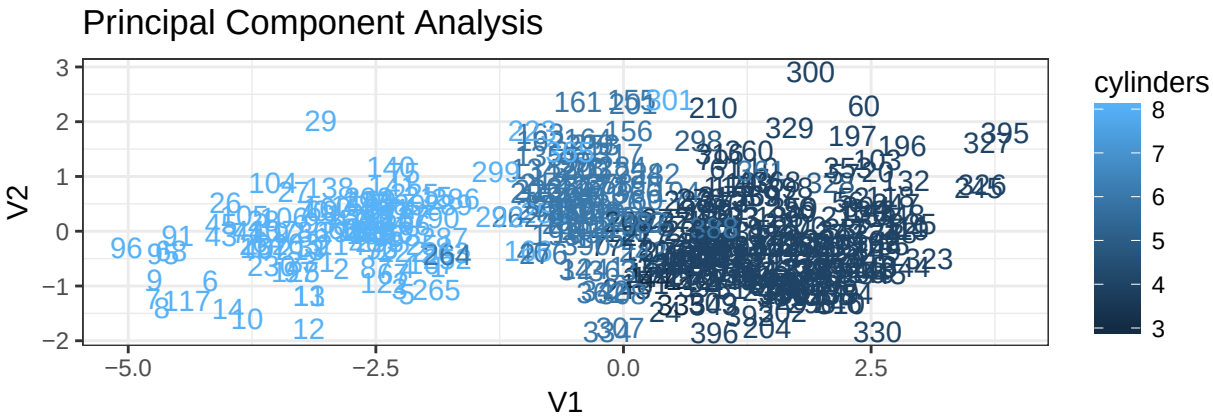
When comparing our Factor Analysis to our Principal Component Analysis, we can see some similarities, most distinctly the data separating by number of cylinders. The rotation between the two analysises is nearly opposite, which also reflects the differences between the loadings values themselves. That is, the greater the number of cylinders, the larger Factor1 is for the Factor Analysis, but the smaller the first Principal Component is. The ranges in each of the graphs are also quite different, with the principal components ranging much further than the Factor loadings.

We can now look at how multidimensional scaling effects our data. Initially, we will do MDS, with a Euclidean distance measure, only on our non-categorical variables. From our plot here we can see both our MDS results as well as the data projected onto the first two Principal Components. As we would have expected, these two plots appear to be nearly identical. Once again, we use cylinders to show the distinct groupings in the data.



Unlike with PCA and FA however, we can include some of our categorical variables and apply MDS to more of our data set. In order to do this, we choose the Gower method to help weigh these categorical variables. Our Gower MDS thus includes every varaible except car name, which is almost unique for each data point. Once again, for consistency, we can choose to view our data and see how it all groups by number of cylinders. With our Gower MDS we can see even more unique grouplings by cylinders, which once again reflects the correlation between the variables that we saw in our introduction to the data. We can see however, that it does not align as well with our principal components. This does make sense as our principal components used just the non-categorical variables.





Overall, each of our methods are slightly different but do give us similar results. Both PCA and FA can be done with just two components and factors respectively. MDS, if done with just non-categorical variables is the same as our PCA. Throughout all of our analysis, we can see quite clearly that all the non-categorical variables are correlated with cylinders and have clear groupings by the number of cylinders.

By # of Cylinders