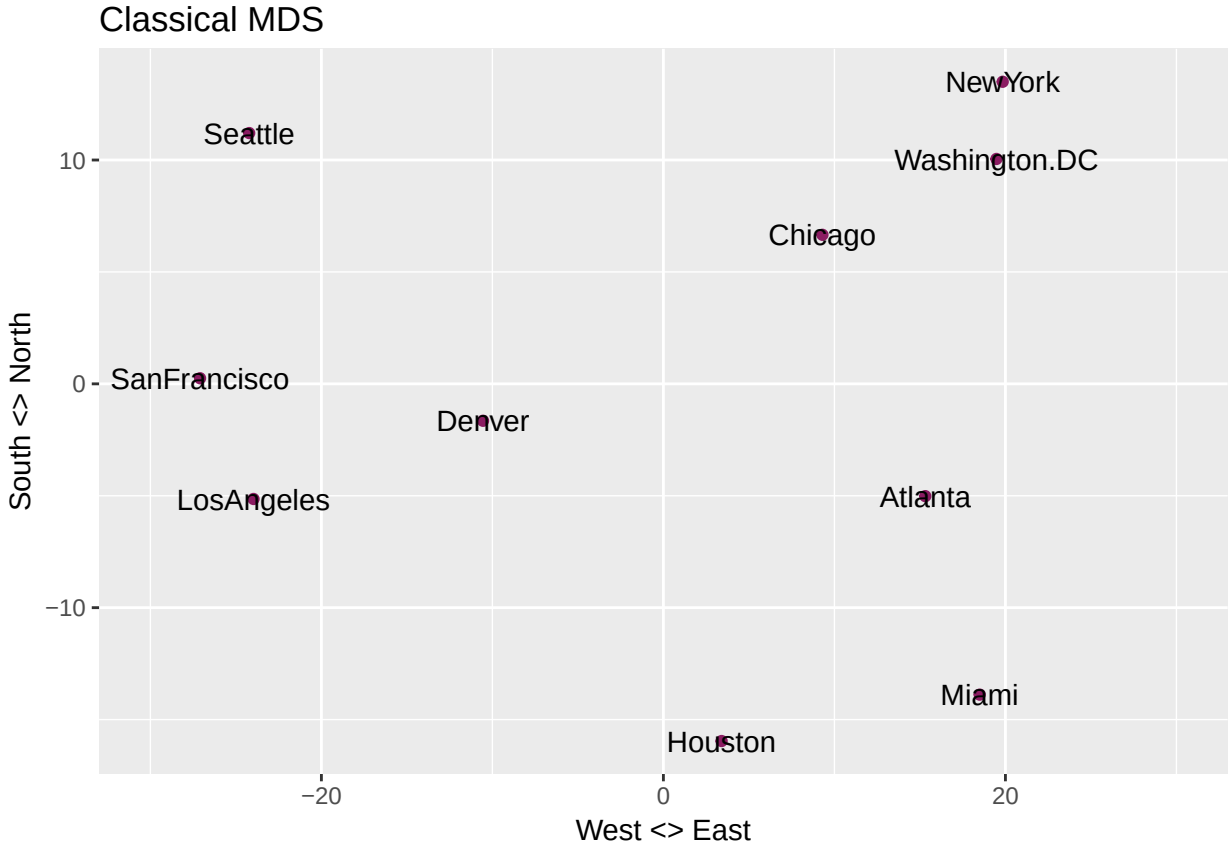
Stats 503 Homework 2

Sam Edds

2/8/2018

2. MDS-by-hand

We use multi-dimensional scaling to examine the distances between 10 cities at a reduced dimension. This technique is useful because it can handle not only quantitative, but categorical data as well. Additionally it does not make any assumptions, such as the data are from a Gaussian distribution. To analyze multi-dimensional scaling we first double-centering our data, along both the rows and the columns. With this done we can just take the corresponding eigenvectors and square-root of the eigenvalues we want. In this case we choose 2 because want to be able to graph our data in an easy to interpret and visual manner, which is easiest with two dimension. We plot these on a graph, rotating them to represent a map of the United States. We can see just as with a map of the US, New York is the Northeastern most city and Miami the southeastern most city. Chicago is more towards the center, and Seattle is the most Northwestern city. The distances are preserved and we can read our cities as a map of the US, with the relative distances clear because of the double-scaling. Because of this we also note that New York and Washington, DC are relatively close together, and closer than San Francisco and Los Angeles, just as in reality.

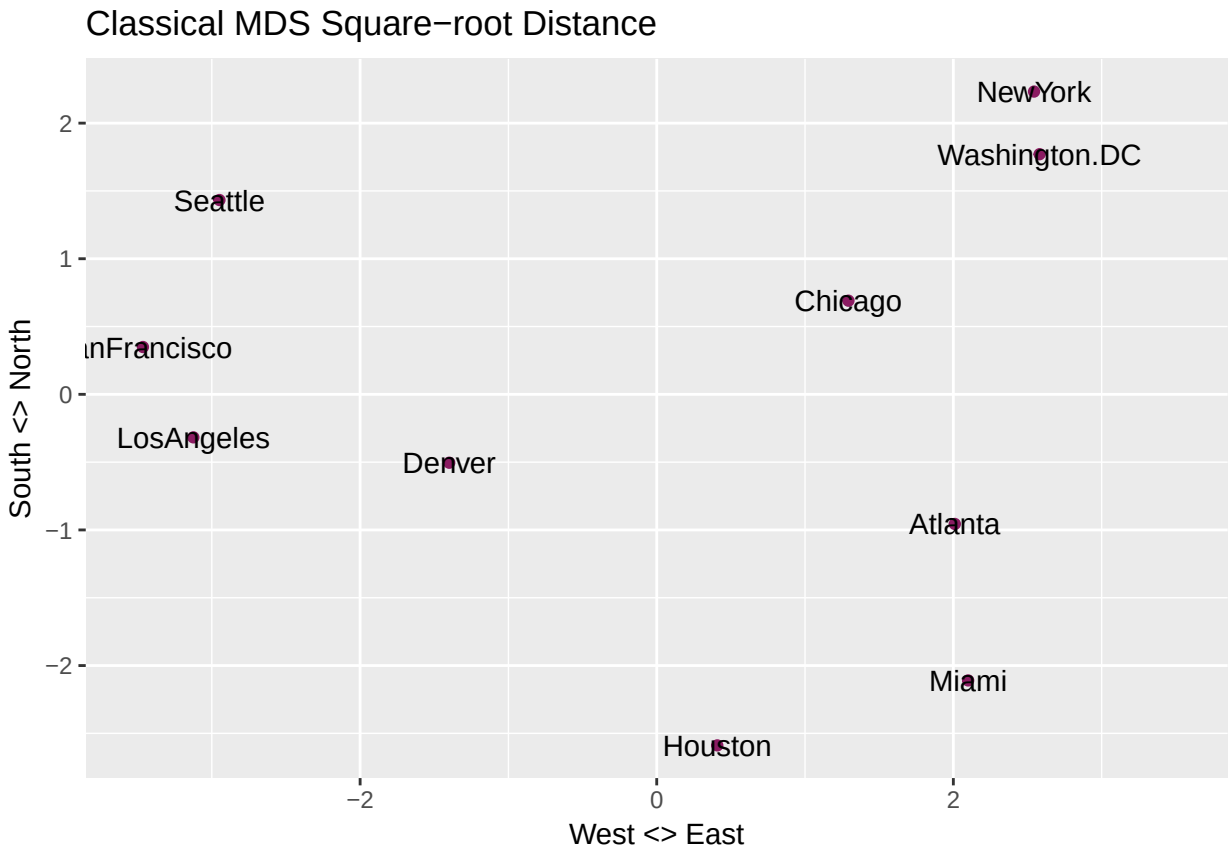


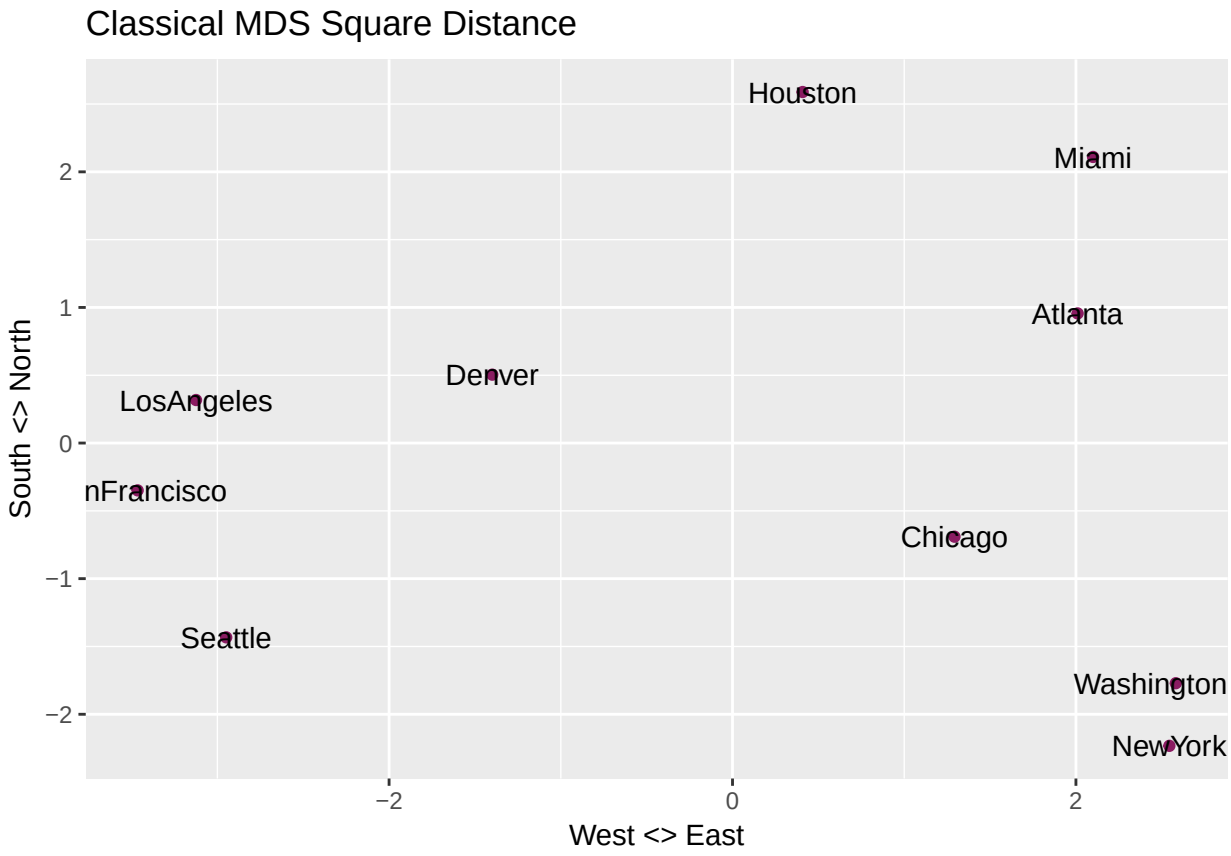
We next examine what happens to our map when we raise our distance matrix to the square root, and when we square it. Something we do in all of our graphs is to have them mimic a graph of the United States; our distances are preserved and we are just rotating our data to be more interpretable.

We notice when we square the distances our cities become much further apart. Chicago moves further North

and closer to Washington, DC, while Miami moves further South and more even with Houston. Our other cities maintain their relative positioning. This makes sense because we are squaring each component of our distance, and then rotating it back to the original map "position".

When we take the square-root of the distances our cities become much closer together, and rotate about the x-axis (again we rotate back to the original positioning to make our graph more like a map of the United States and more comparable to our results.) We notice compared to an actual map of the United States Seattle is too far East and Chicago is now too far South.




Classical MDS Square Distance
Houston
Miami
Atlanta
Denver
LosAngeles
nFrancisco
Chicago
Seattle
Washington
NewYork
2
1
0
−1
−2
South <> North
−2
0
2
West <> East

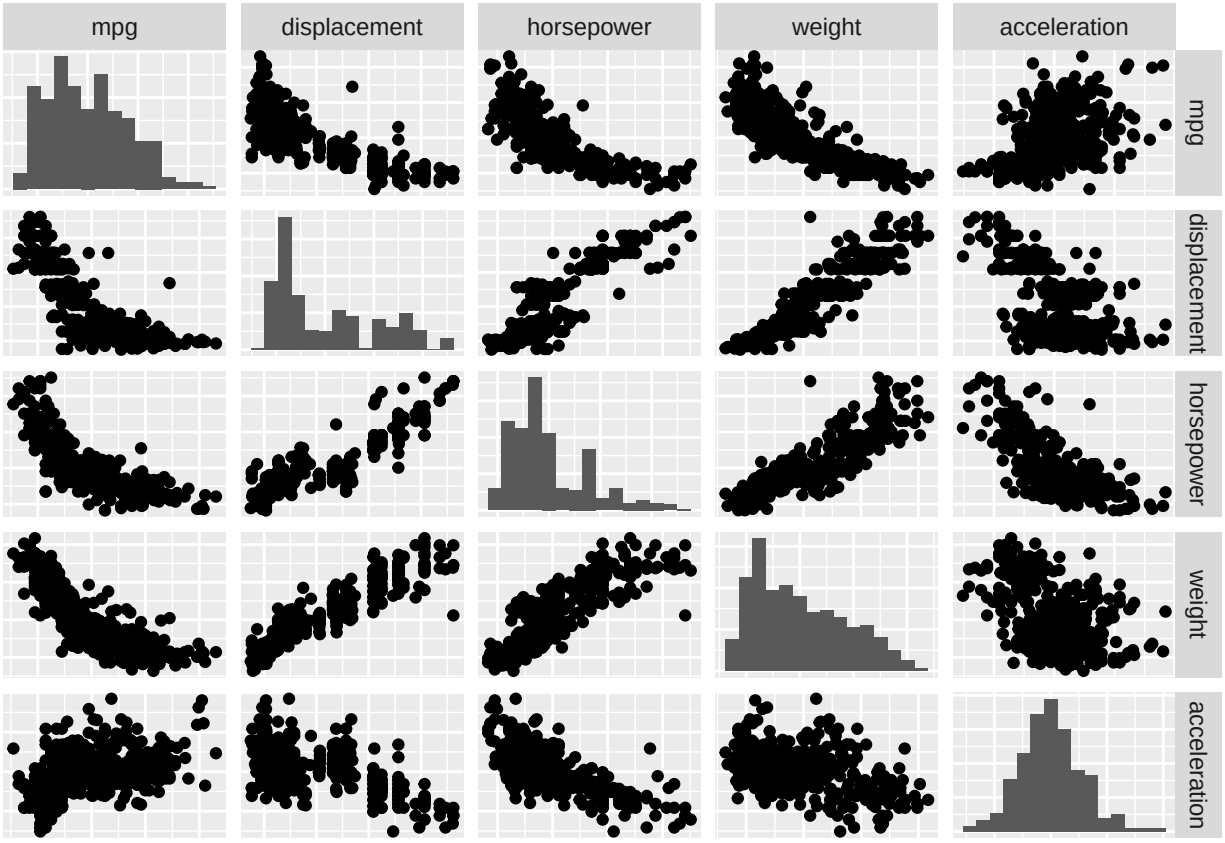
3. Report comparing Principal Component Analysis, Factor Analysis, and Multi-dimensional Scaling

We conduct a study examining the difference between principal components analysis, factor analysis, and multi-dimensional scaling using automotive data from 398 vehicles (392 with complete information). We measure their model year, origin, model, miles per gallon (mpg), number of cylinders in the engine, displacement, horsepower, weight, and acceleration. Our exploratory data analysis focuses on providing a broad summary to understand these data holistically and specific relationships between variables.

Our initial exploration shows we have vehicle data from 1970-1982, for a wide variety of vehicles coming from the United States, Western Europe, and Japan. We can see there are 3 to 8 cylinder engines (mostly 4, 6, or 8 cylinder), and a wide range in horsepower (46 to 230 hp).

We examine the histograms and scatterplots for our quantitative variables to look for potential signs of dimension reduction. We notice positive correlation between displacement, horsepower and weight and very little correlation between acceleration and mpg. Interestingly, our first three variables individually are all negatively correlated with acceleration and mpg. As a result we expect PCA, factor analysis, and multi-dimensional scaling will allow us to help reduce some dimensionality in our data.

```
##       mpg          cylinders      displacement     horsepower
##  Min.   : 9.00   Min.   :3.000   Min.   : 68.0   Min.   : 46.0
##  1st Qu.:17.50   1st Qu.:4.000   1st Qu.:104.2   1st Qu.: 75.0
##  Median :23.00   Median :4.000   Median :148.5   Median : 93.5
##  Mean   :23.51   Mean   :5.455   Mean   :193.4   Mean   :104.5
##  3rd Qu.:29.00   3rd Qu.:8.000   3rd Qu.:262.0   3rd Qu.:126.0
##  Max.   :46.60   Max.   :8.000   Max.   :455.0   Max.   :230.0
##                                                  NA's   :6
##      weight      acceleration     model_year        origin
##  Min.   :1613   Min.   : 8.00   Min.   :70.00   Min.   :1.000
##  1st Qu.:2224   1st Qu.:13.82   1st Qu.:73.00   1st Qu.:1.000
##  Median :2804   Median :15.50   Median :76.00   Median :1.000
##  Mean   :2970   Mean   :15.57   Mean   :76.01   Mean   :1.573
##  3rd Qu.:3608   3rd Qu.:17.18   3rd Qu.:79.00   3rd Qu.:2.000
##  Max.   :5140   Max.   :24.80   Max.   :82.00   Max.   :3.000
##
##           car_name
##  ford pinto    :  6
##  amc matador   :  5
##  ford maverick :  5
##  toyota corolla:  5
##  amc gremlin   :  4
##  amc hornet    :  4
##  (Other)       :369
```



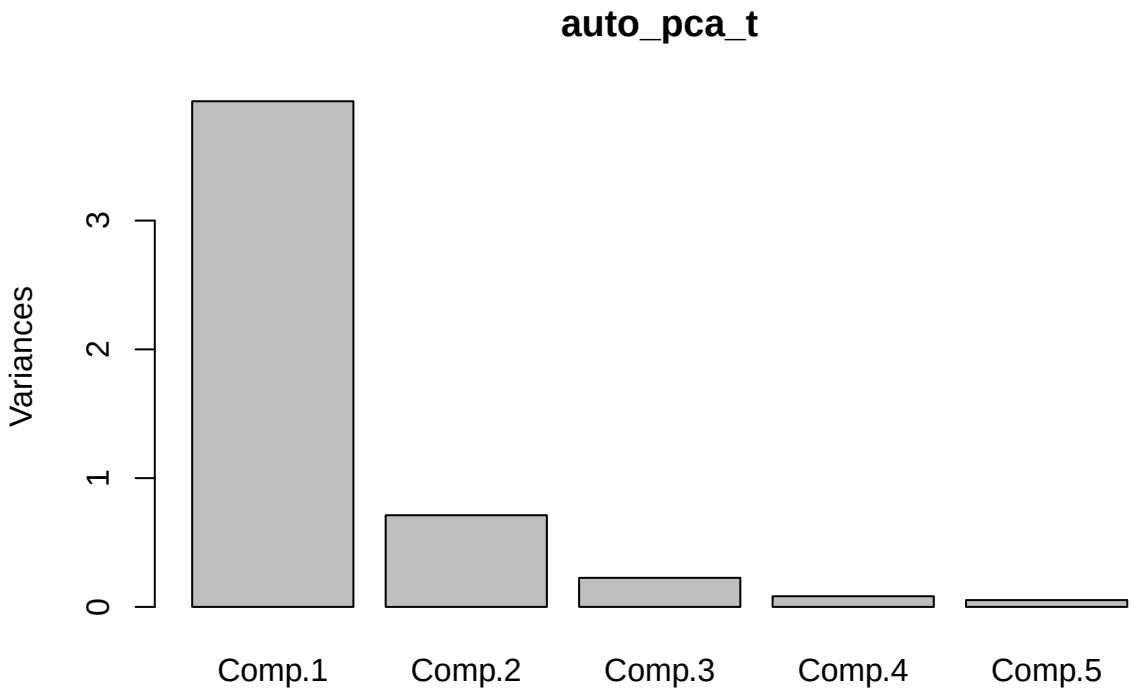

```
##   origin freq
## 1      1  249
## 2      2   70
## 3      3   79
```

```
##   cylinders freq
## 1         3    4
## 2         4  204
## 3         5    3
## 4         6   84
## 5         8  103
```

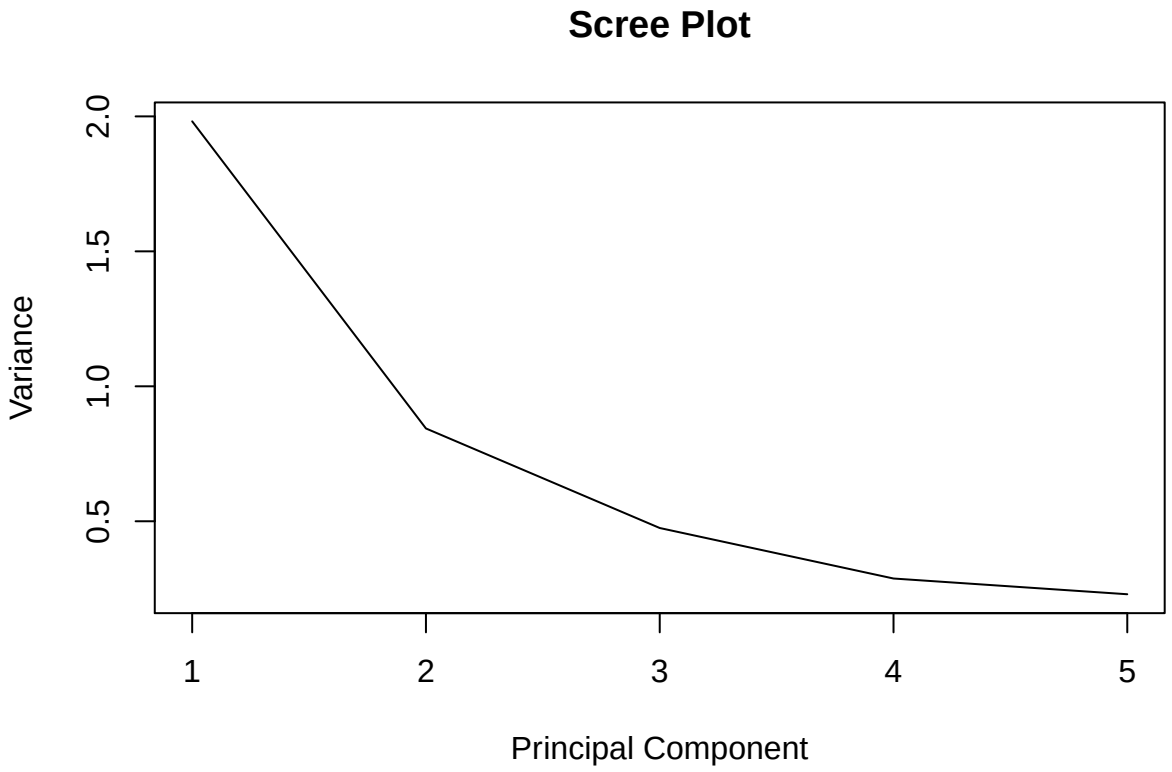
```
##   model_year freq
## 1         70   29
## 2         71   28
## 3         72   28
## 4         73   40
## 5         74   27
## 6         75   30
```

We conduct our principal component analysis on mpg, displacement, horsepower, weight, and acceleration. We center our data then compute the correlation matrix. We chose our correlation matrix, and after examining our scree plot decided to choose two principal components (of the 5 potential) which account for almost 93% of the variation in our data. I chose two over more for ease of interpretation while still accounting for most of the variance.

```
##         Comp.1     Comp.2       Comp.3      Comp.4    Comp.5
## [1,] -1.875970 -0.6485578 -0.052615303 -0.40705615 0.1353658
## [2,] -2.852657 -0.6755745  0.005838212  0.05907359 0.3668170
## [3,] -2.262764 -1.0431475  0.015093669 -0.10797706 0.2787314
## [4,] -2.188684 -0.6664257 -0.196684445  0.04445604 0.2823518
## [5,] -2.187813 -1.1464578 -0.225586703 -0.30135241 0.1080946
## [6,] -4.176246 -0.9080471  0.614989805  0.14639250 0.3073480
```




```
##    Comp.1    Comp.2    Comp.3    Comp.4    Comp.5
## 0.7853509 0.9277520 0.9728764 0.9894514 1.0000000
```



We examine the factor loadings (correlation coefficients between our variables and factors) to get a sense of what our principal components mean based on the relationship of our different variables. The first factor incorporates all of our variables close to evenly, but acceleration and miles per gallow have a positive relationship, while displacement, horsepower, and weight have a negative and move in the same direction. For our second component, acceleration matters the most in explaining variation in our data while the other variables matter much less.

```
##
## Loadings:
##              Comp.1 Comp.2 Comp.3 Comp.4 Comp.5
## mpg           0.444 -0.304  0.839
## displacement -0.483  0.135  0.371 -0.476  0.620
## horsepower   -0.484 -0.124  0.206  0.826  0.160
## weight       -0.471  0.326  0.305 -0.159 -0.744
## acceleration  0.335  0.876  0.150  0.257  0.178
##
##                Comp.1 Comp.2 Comp.3 Comp.4 Comp.5
## SS loadings       1.0    1.0    1.0    1.0    1.0
## Proportion Var    0.2    0.2    0.2    0.2    0.2
## Cumulative Var    0.2    0.4    0.6    0.8    1.0
```

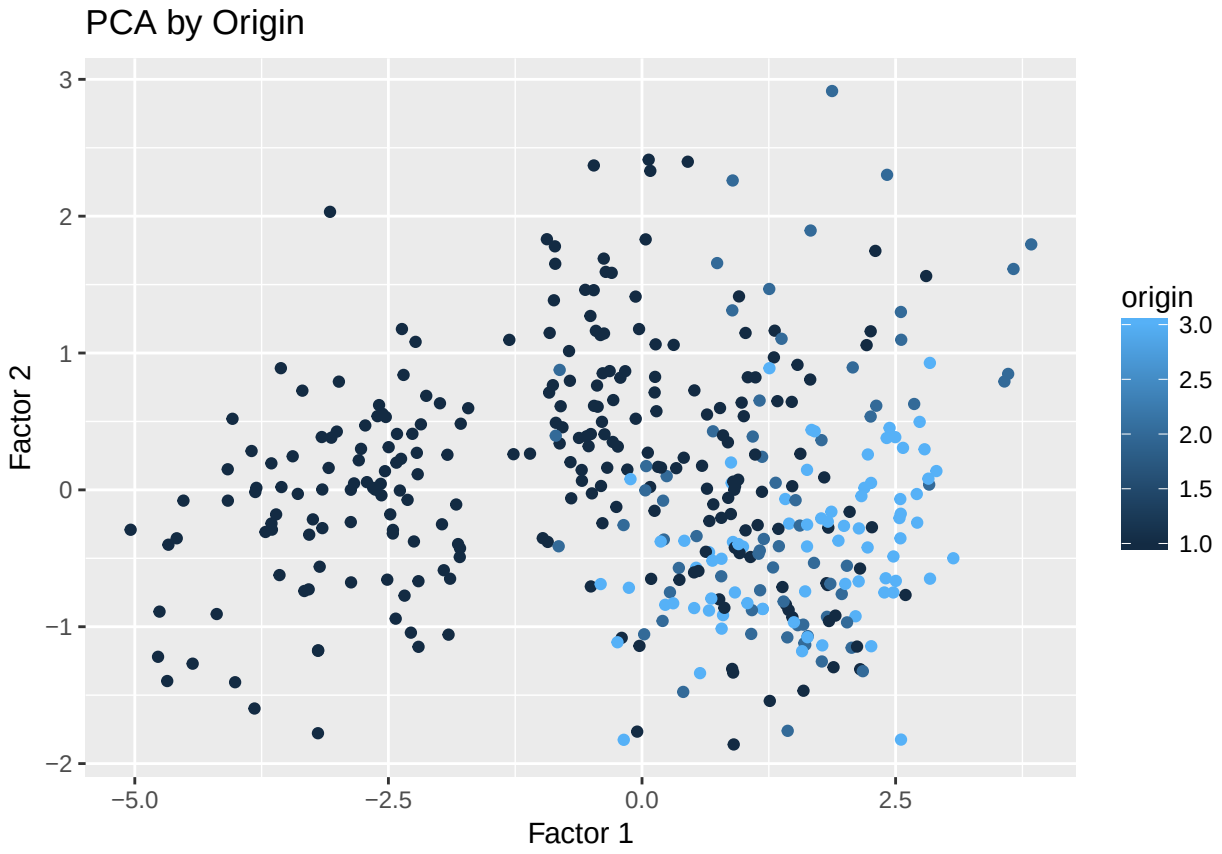
We next project our data onto our first two principal components. We notice our data are tight bound, without much variation, and a strong negative relationship between component 2 and 1, so as component 1 increases, component 2 decreases. Based on what we noticed about factor loadings we could conjecture about what this might mean. We also at this point take into account our categorical data and if there are any clear differences.

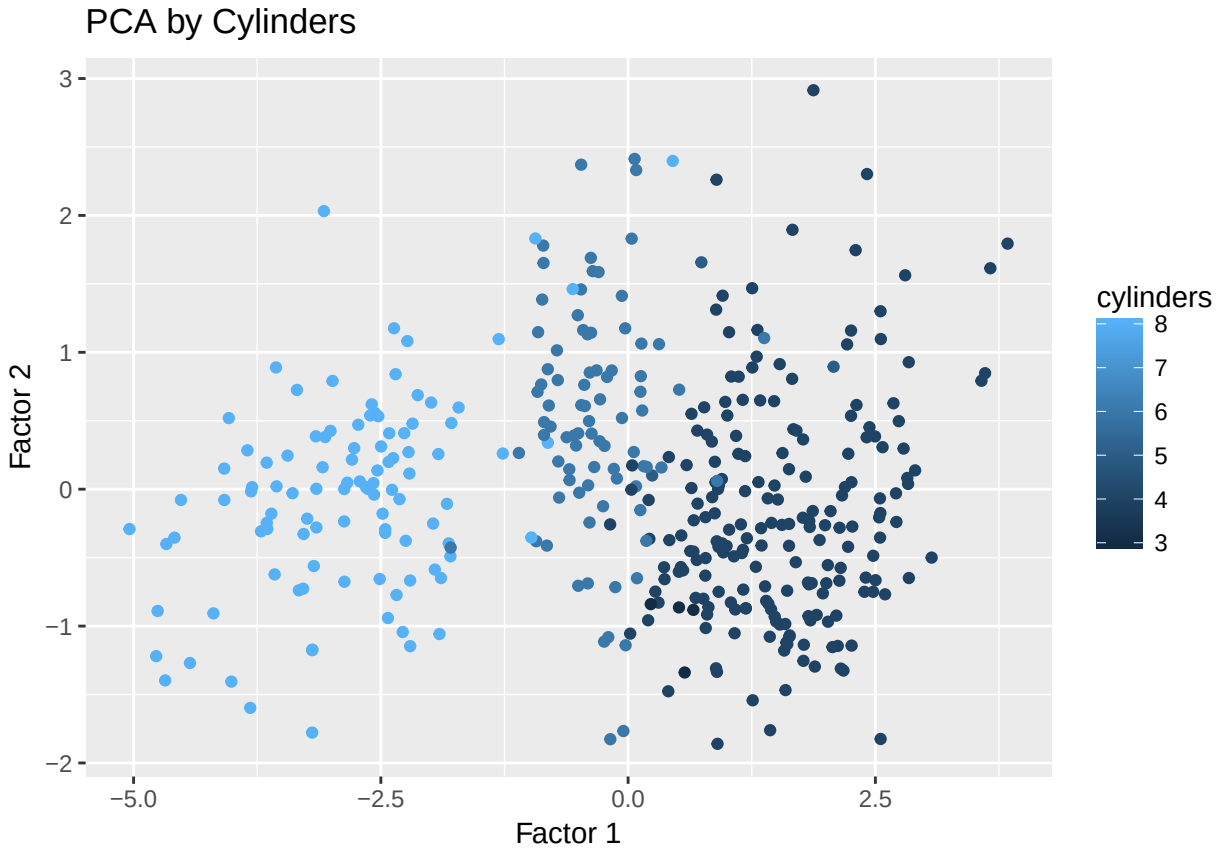
In particular we examine cylinder, model year, and origin. Overall it seems data are more distinguishable according to cylinder than the other categorical variables. For engine cylinders we see that cylinders map almost perfectly onto our principal components, with the 8 cylinder engines mapping to a high component 2, very low component 1, down to 4 cylinder engines mapping to a low component 2, very high component 1.

This could also have something to do with component 2 heavily weighting acceleration.

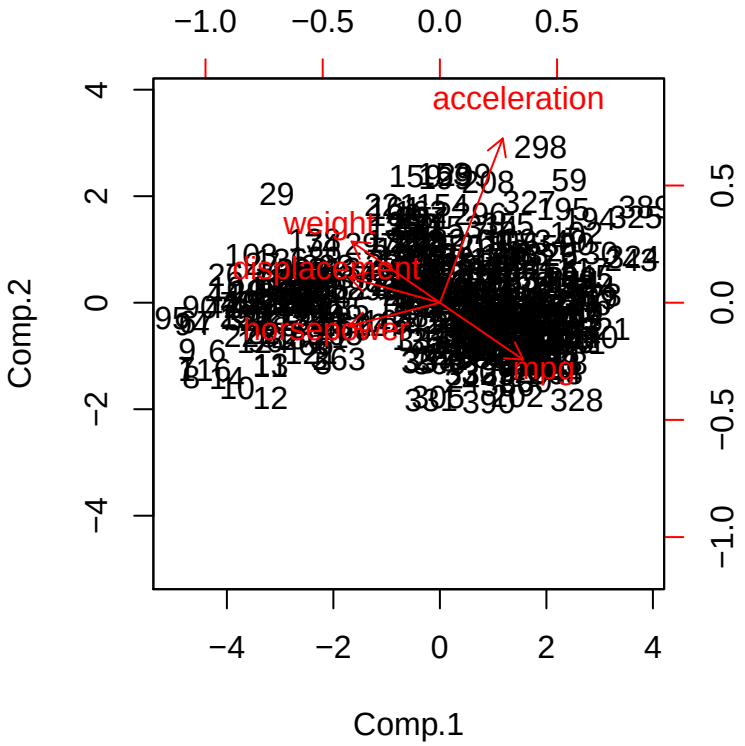
The origin data maps most vehicles from Western Europe and Japan higher in the first component, clustered low in the second component, while vehicle data from the United States spans from high in the second component, to somewhat low.

Finally, our model year data does not appear to follow as clear a pattern as origin and cylinder because data from old and new model years index high for each component.





We make a biplot to check our intuition from the factor loadings and notice that weight and mpg move in the opposite directions, while weight, horsepower, and displacement all move in similar directions. Acceleration moves in yet another direction, orthogonal to weight and mpg. This is the most interesting plot because we can visual the different directions in which our variables load for the two principal components. Weight, horsepower, and displacement seem related in that horsepower and displacement are typically higher when vehicles weigh more. Gas mileage on the other hand is inversely related to these metrics, while acceleration is typically less dependent on these metrics. Finally, acceleration and horsepower are inversely related, which seems surprising.



While PCA optimizes for the greatest variance, Factor Analysis ensures that residuals are uncorrelated so the error and loadings are indepedent. With factor analysis we are examining the latent factors that impact our automotive data. For this analysis we test different rotations and examine our factor loadings in order to decide which rotation to use in order to help us determine the number of factors. From PCA we lean towards two, and then examine our factor loadings to create more intuition. We test the promax and varimax rotations, as well as no rotation. Since promax is a more extreme version of varimax (exacerbating the differences to make smaller values smaller and larger ones larger), it makes for the easiest interpretation of the factor loadings because our weights are most clear. With this in mind we choose promax, and two factors, again for ease of intepretability.

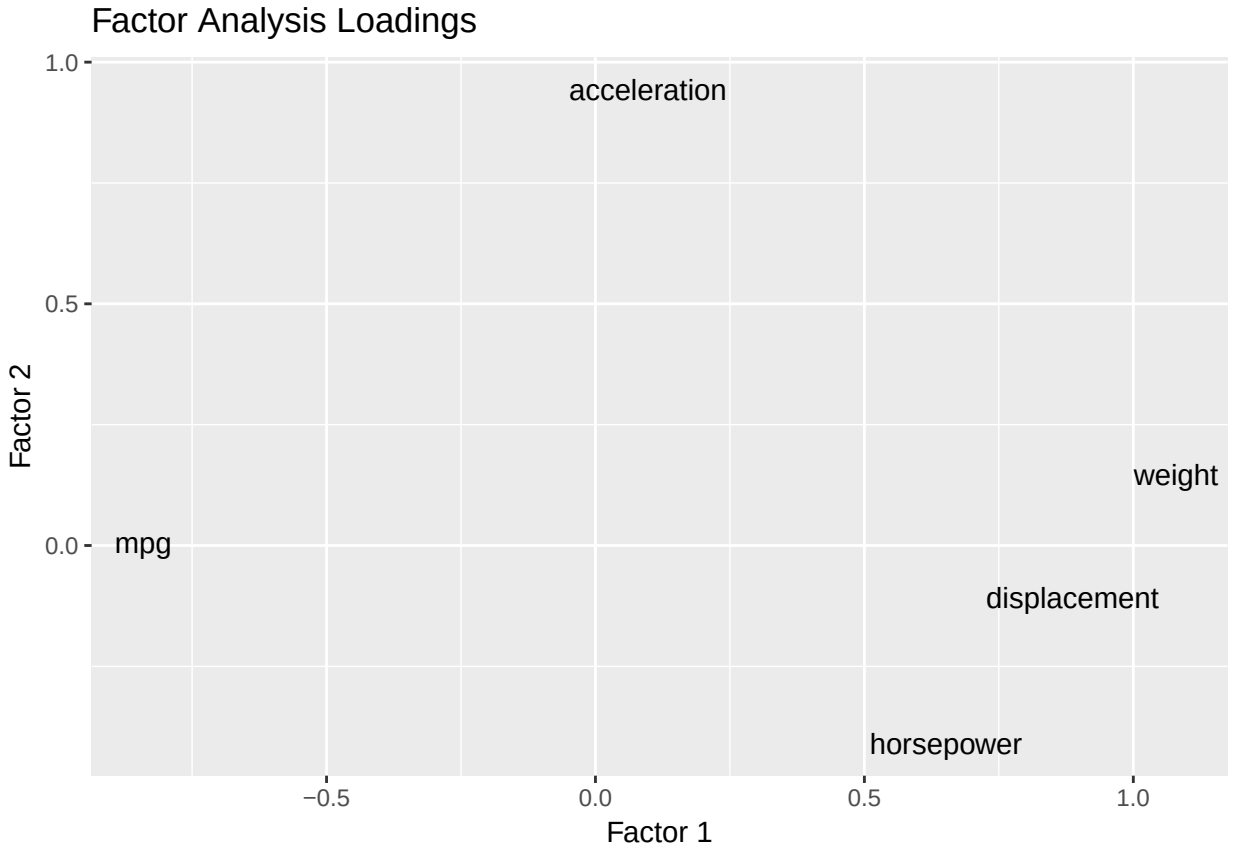
We can see that our first factor weights most heavily on weight, and heavily on displacement, as well as mpg (which is directionally different from displacement and weight). Interestingly, acceleration is essentially a non-factor, and horsepower somewhat less important that the three big variables. Our second factor places almost all of the weight on acceleration and the other variables are essentially ignored. The second factor is similar in essence to the PCA loadings which had most of the weight placed on acceleration, while the first factors are a bit different. For PCA these were loaded almost evenly, and with displacement, horsepower, and weight directionally the same and opposite weight and acceleration. For FA however, we mpg moves directionally different from the other variables and weight is more heavily weighted compared to the other variables. This may be due to the different optimization methods mentioned above. This is most easily seen in our graph, showing the different loadings plotted, with horsepower displacement and weight moving in the same direction, which is very different from mpg and acceleration.

```
##                     mpg displacement horsepower     weight acceleration
## mpg           1.0000000   -0.8051269 -0.7784268 -0.8322442    0.4233285
## displacement -0.8051269    1.0000000  0.8972570  0.9329944   -0.5438005
## horsepower   -0.7784268    0.8972570  1.0000000  0.8645377   -0.6891955
## weight       -0.8322442    0.9329944  0.8645377  1.0000000   -0.4168392
## acceleration  0.4233285   -0.5438005 -0.6891955 -0.4168392    1.0000000
```

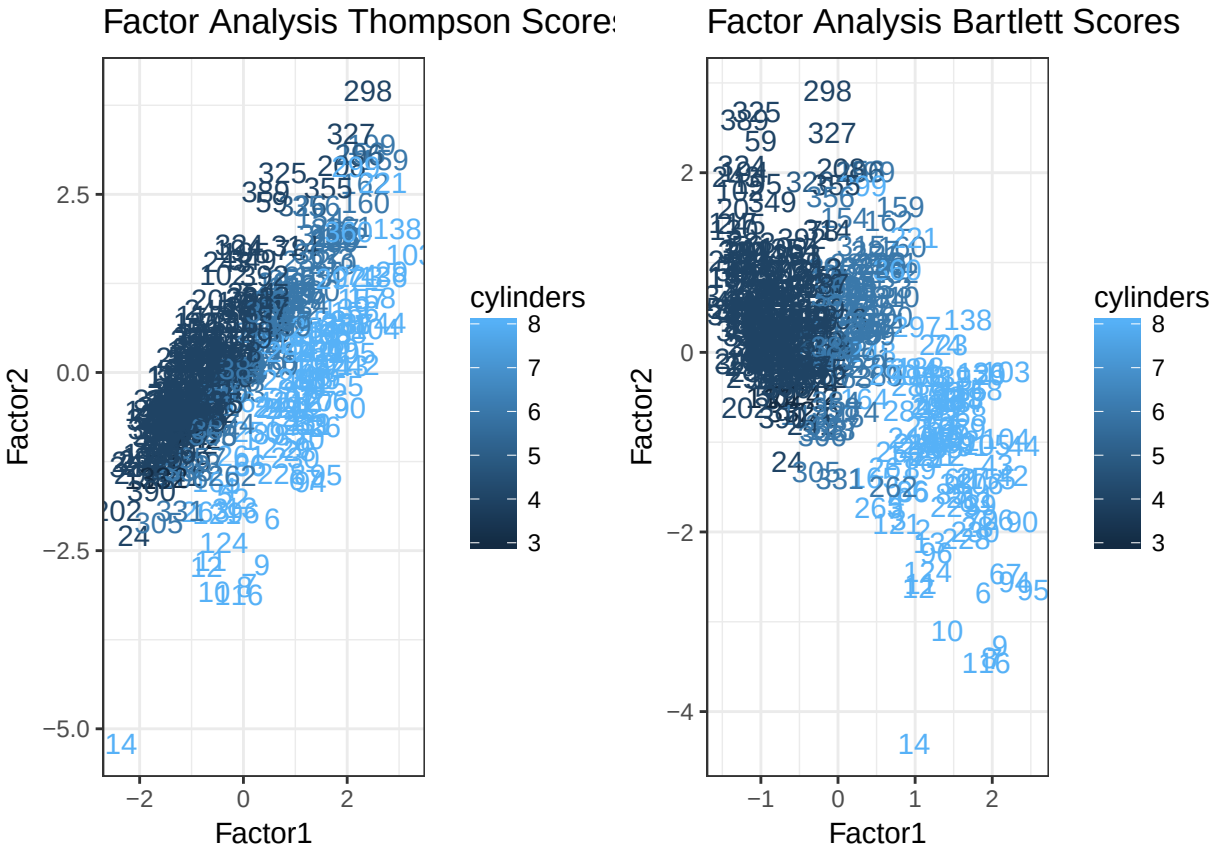
```
##                    [,1]       [,2]        [,3]
## mpg          -0.8450173 -0.7970997 -0.84079059
## displacement  0.9550356  0.8763440  0.88726198
```

```
## horsepower    0.9113555  0.7603018  0.65164485
## weight        0.9864738  0.9686937  1.07951575
## acceleration -0.5020569 -0.2410265  0.09757409
```

```
##                     [,1]       [,2]         [,3]
## mpg           0.00908116 -0.2806577  0.006663625
## displacement -0.08647020  0.3893473 -0.106632154
## horsepower   -0.31845033  0.5949123 -0.408570412
## weight        0.10791990  0.2154295  0.146348624
## acceleration  0.72772836 -0.8506209  0.943297348
```



We also examine scores, testing both Thompson and Bartlett's scoring methods with our promax rotation. Our graphs are color coded by cylinders because we want to be able to see the most separation in our variables. We chose cylinders based on our previous PCA results, which suggested that cylinders are the most clear variable upon which our factors are split. Our Thompson's results (regression based) show a positive, almost linear relationship between vehicles with the same number of cylinders. We see that a low number of cylinders are higher in factor 2, and corresponding data with more cylinders are shifted to the right to create three clear groups, those with the least to the most cylinders. Vehicles with the most cylinders end up loading most heavily in factor 1 compared to similar vehicles with a different number of cylinders. The Bartlett scores are more clustered compared to the linear Thompson results, but these too show a clear distinction by number of cylinders. Vehicles with more cylinders have higher scores for factor 2, compared to those with lower cylinders, again moving left-to-right in clusters. Both of these results show similar outcomes, so we could proceed with either set of scores.

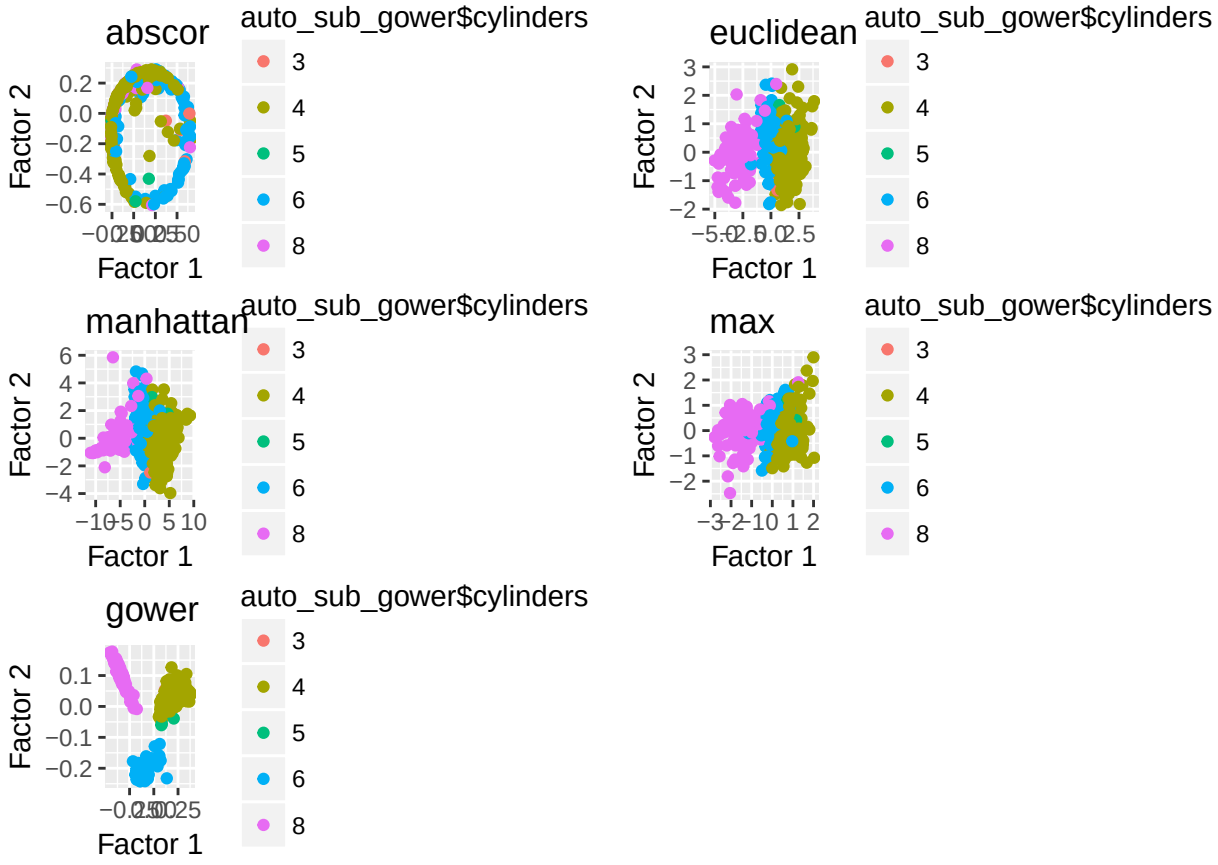


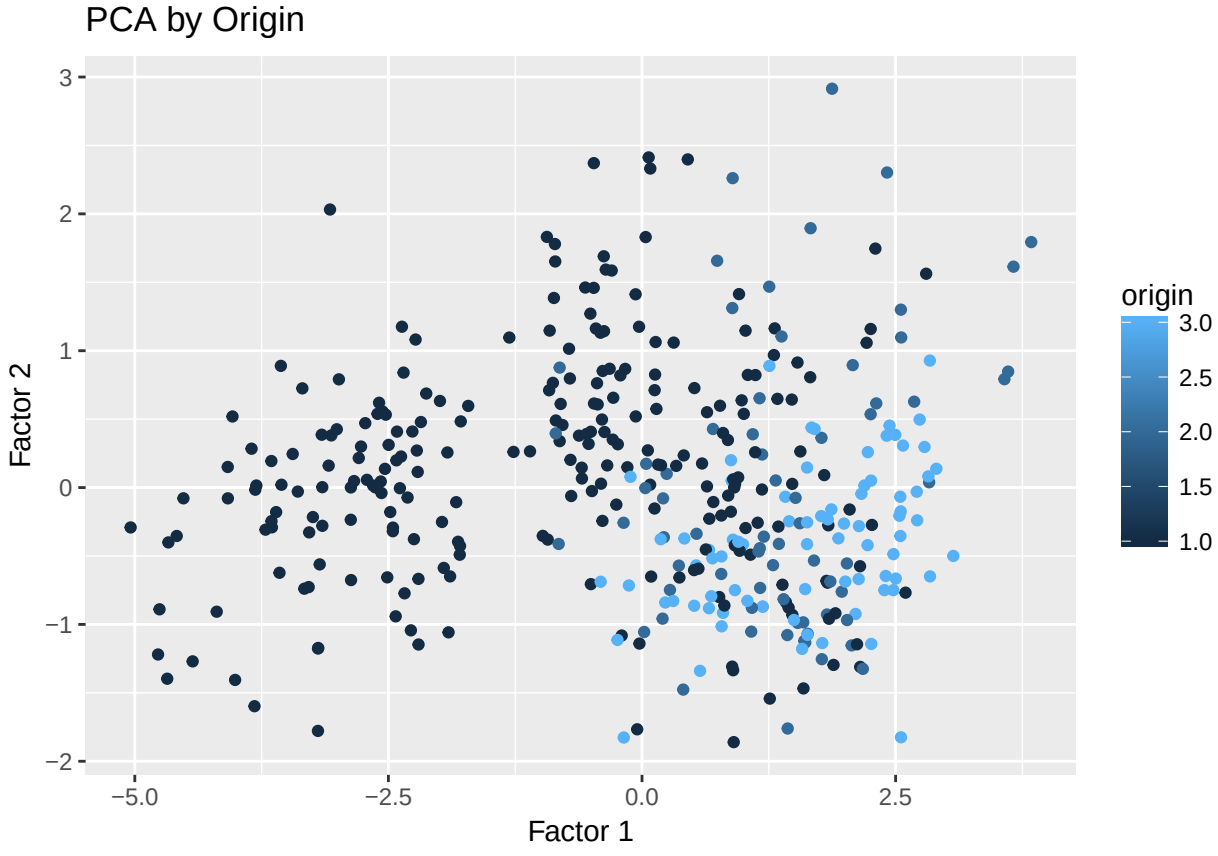
Finally we compare these results to MDS, which is different from PCA and Factor Analysis in that we are using distances and also able to sometimes include categorical variables as well. We create a centered and scaled dataset on which to compute distances other than Gower, and a full dataset that includes origin, model_year, and cylinders as factors (car_name is already considered a factor) on which to compute Gower distances. For Gower our factor variables are compared only in such that x = y in level (the same category), not that say 1 and 3 are closer together and meaningfully so than say 1 and 10.

We test a number of different distances including Abscor, Euclidean, Maximum, Manhattan, and Gower, overlaying the cylinder coloring to see if we have clear differentiation between different groups. Because we have already done this for PCA and Factor Analysis, while they are different, we still expect to see similar separation in these data based on cylinder.

From our various plots we can see different amounts of separation according to number of cylinders. Since we are looking for the most cleanly separated data. Euclidean seems quite similar to Maximum distance, and even though the dispersion is slightly different, similar to Manhattan in terms of amount of separation. Gower, which as we noted treats factor variables through in-category comparison, seems to most cleanly separate our data. This seems reasonable given that cylinder is one of the categorical variables upon which Gower computes a comparison. We notice our Euclidean distance score results match our PCA results, which we expect to be consistent.

Overall our results are fairly consistent between PCA, FA, and MDS. While PCA is maximizing variance, FA is ensuring the residuals and loadings are orthogonal, and MDS is perserving distances and has no underlying distributional assumption, we can clearly see that the number of cylinders is a way to separate our automative data (and our PCA results match our MDS results). Through PCA and FA we can also see the general relationship between the quantitative variables, namely that weight, displacement, and horsepower move in the same direction, distinct from acceleration and mpg. Depending on our differing primary goal we may choose any of these.


abscor
euclidean
manhattan
max
gower
auto_sub_gower$cylinders
3
4
5
6
8
Factor 1
Factor 2


PCA by Origin
origin
3.0
2.5
2.0
1.5
1.0
Factor 1
Factor 2