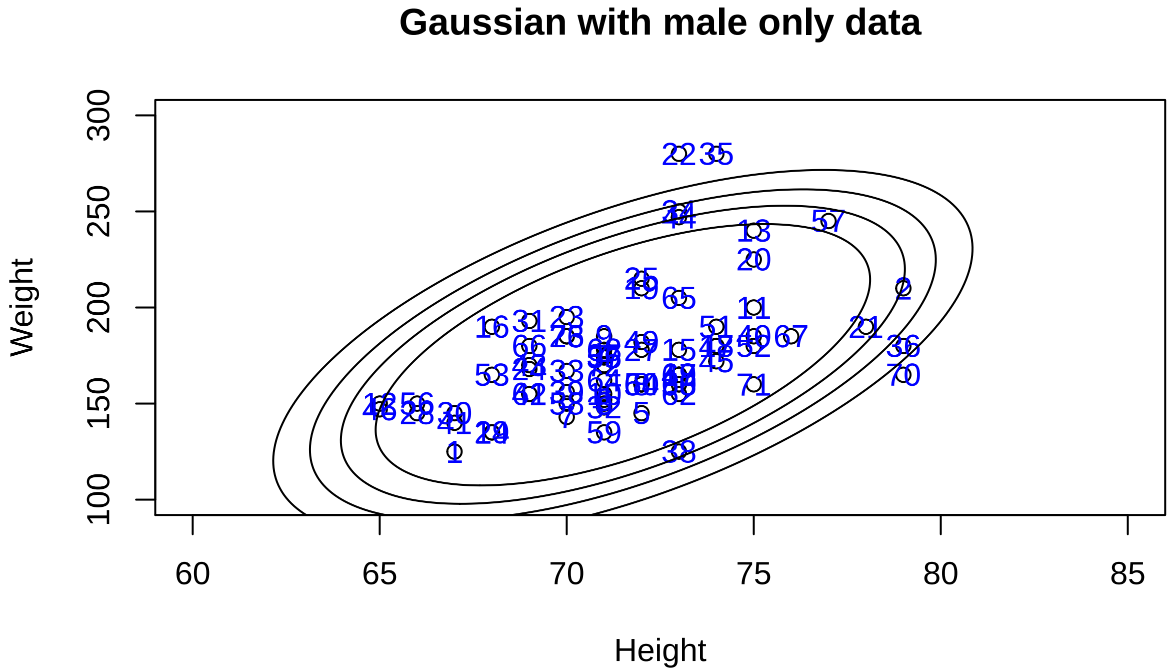
Stats 503 Homework 1

Sam Edds

1/17/2018

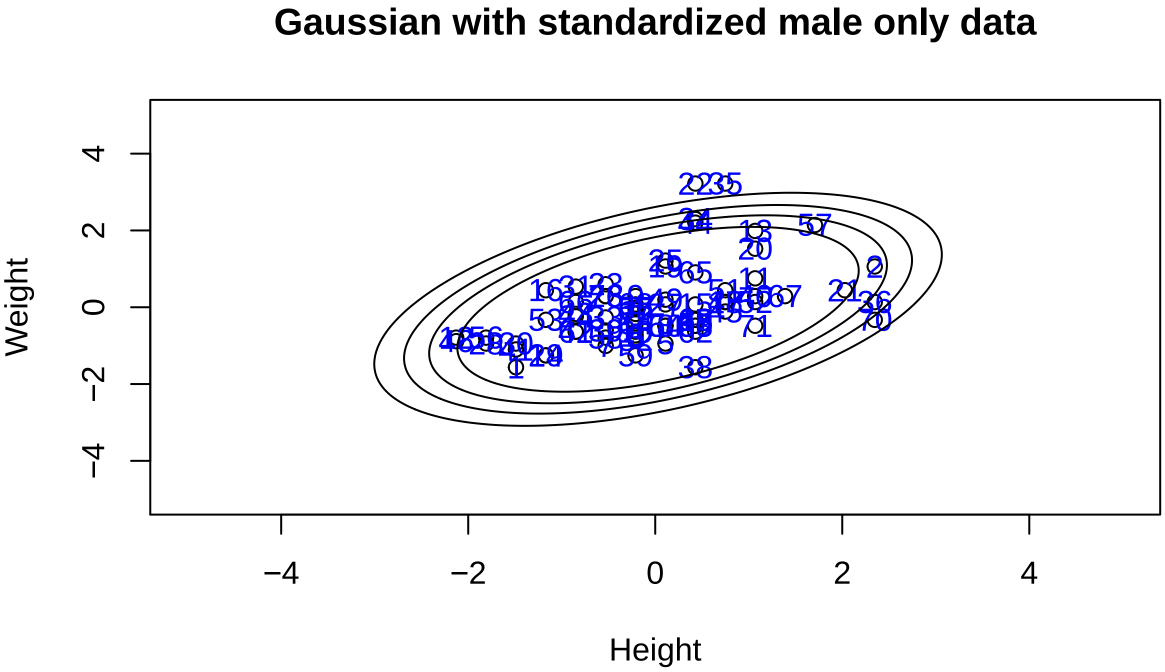
3A. Fitting Data

We first fit a 2 dimensional Gaussian to male height and weight data. We superimpose the scatter plot on these data to see which data points fall within certain confidence levels, from 90% (outermost ellipse) to 97.5% (innermost ellipse), marked by the different ellipses. We can see the majority of these data fall within the 95% confidence interval, with only two individuals fully outside of our ellipses. Most individuals are between 65 and 78 inches, weighing between 120 and 240 pounds. There are a small number of individuals that either weigh less than is normal for their height, or because they weighed more than is normal for their height. For example the two individuals completely outside our widest confidence intervals; these individuals are much heavier than other individuals with similar height, and are therefore outliers.



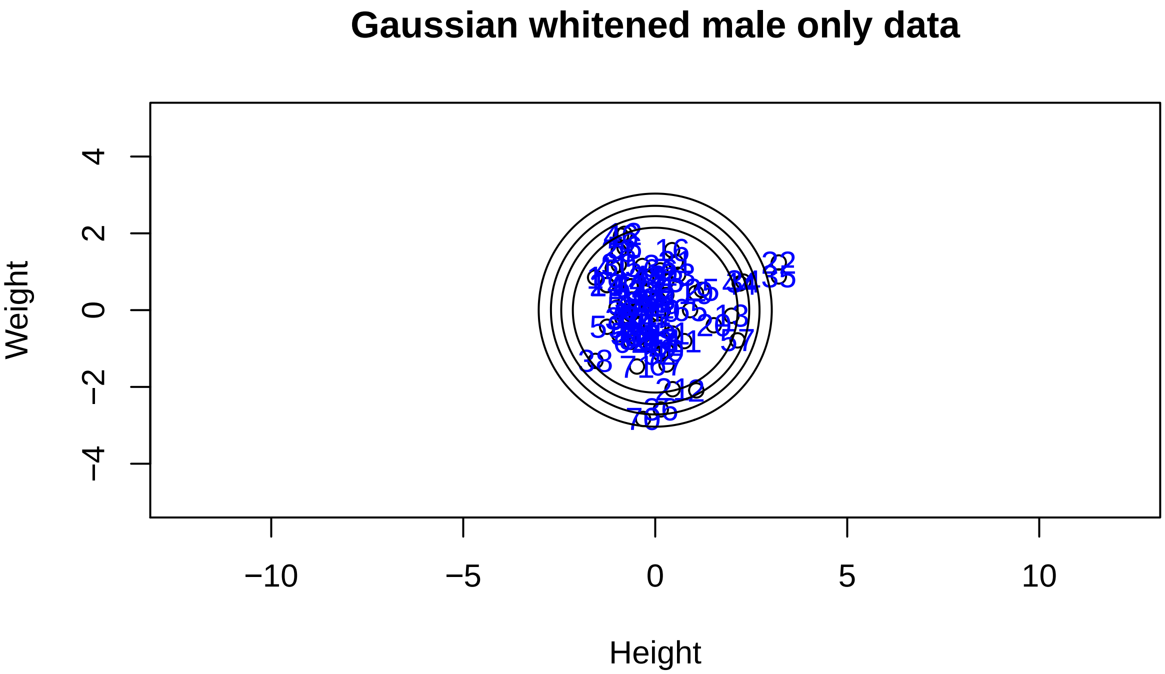
3B. Standardizing

We next standardize our results, subtracting off the mean to center and then scale these data before repeating our analysis. Now these data are centered around 0 for both height and weight. Our results remain the same, and let us see these data in terms of standard deviations so it is easier to judge relative distance even though the unit interpretation is now unclear.



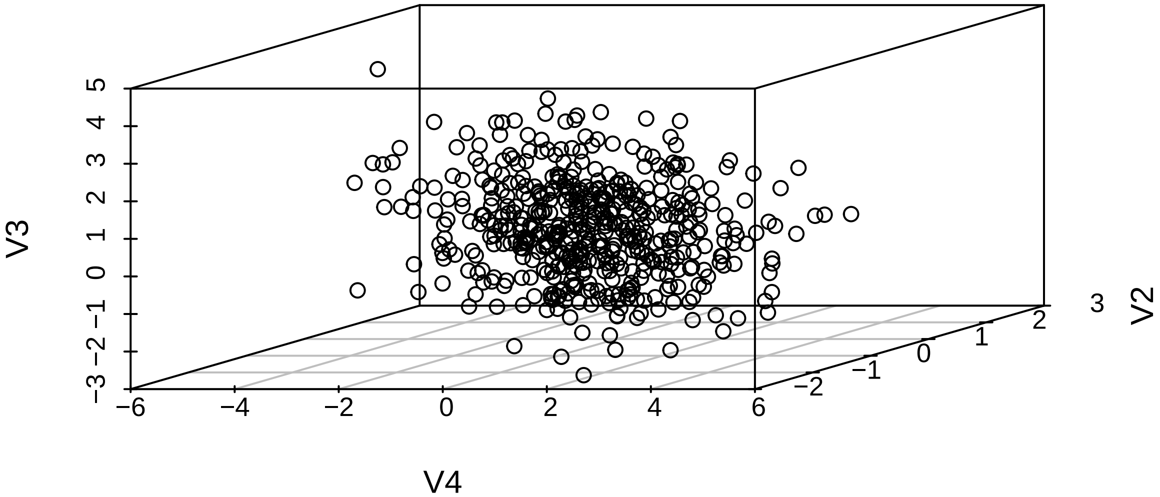
3C. Whitening

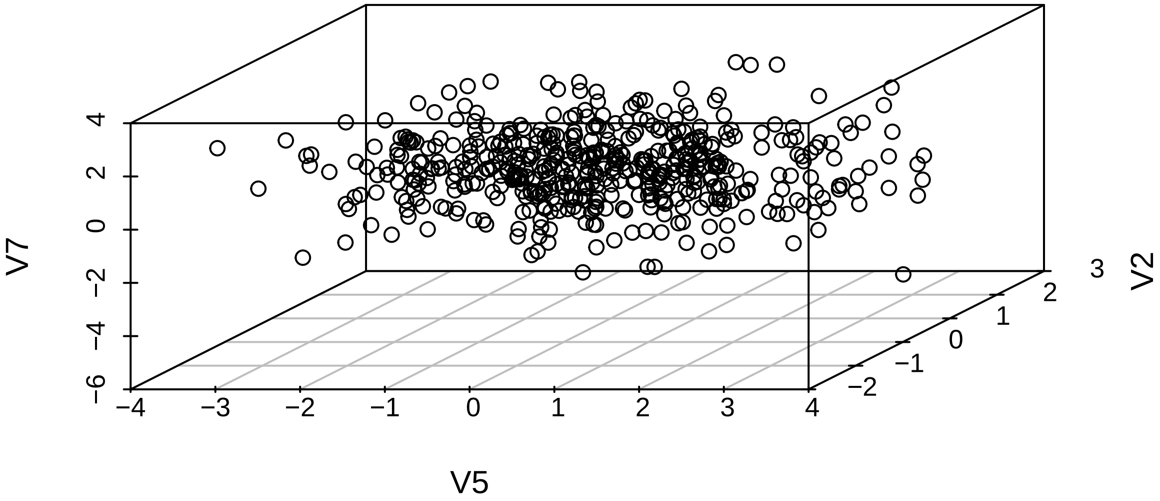
We next whitened or sphered our original data, which uncorrelated our data and created equal variance along each dimension. Our height is now centered, both of which do not provide any context. With this in mind we look back upon points 22 and 35, males we identified as being around 70 inches tall, but weighing around 250 pounds, making them outliers for our data. We now see these have the tallest height, and average weight compared to others, instead of close to average height, and much heavier weight.

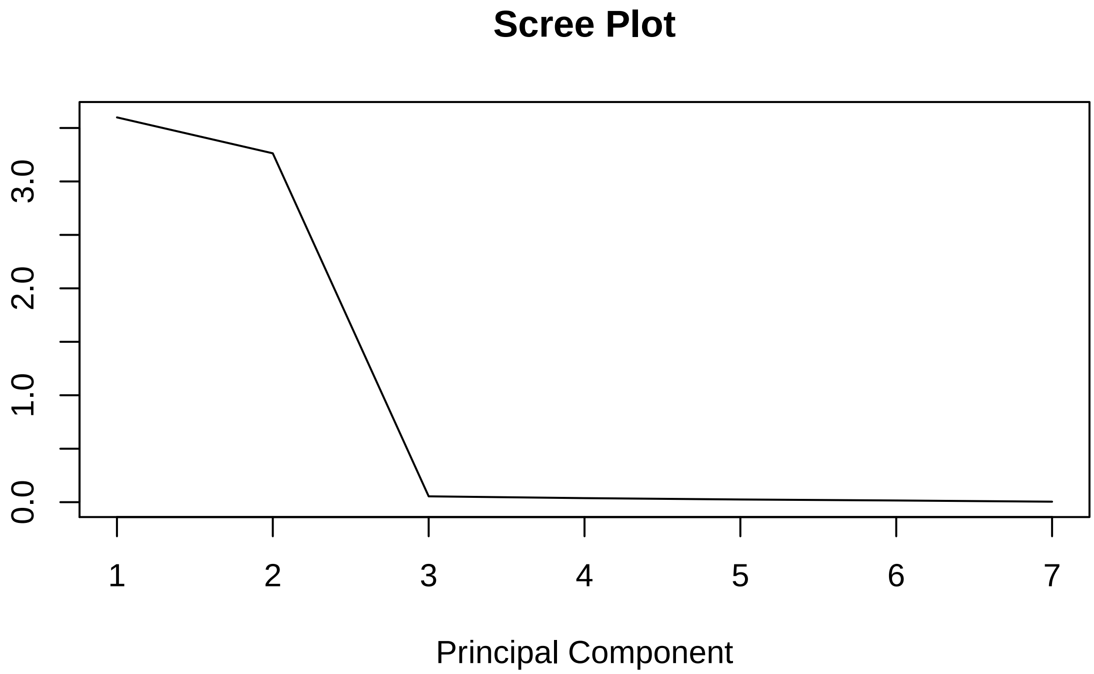


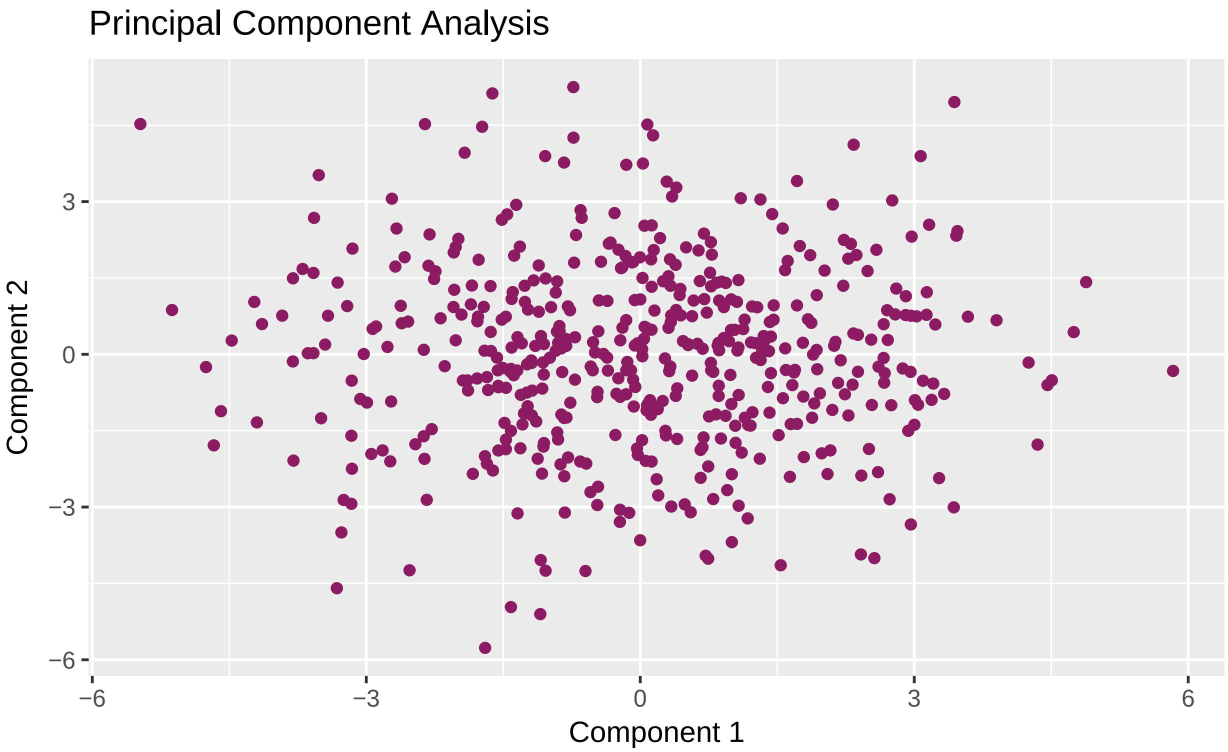
4. PCA by hand

Before using Principal Component Analysis we examine our dataset of unknown origin by randomly plotting 3 of the 7 variables twice. These visualizations show that our data could be reduced dimensionally. In order to do this we center and scale our data, then compute the correlation matrix. We center these data to ensure that our first principal component is fit around the origin and correctly determines the direction of most variation, which can be an issue if these data are not centered. Next we compute the eigenvectors and values, with the idea to maximize the variation explained by our different principal components. We compute a scree plot to determine how many principal components to use to recover most of the variation explained by our original dataset, and decide upon 2. This is because these two dimensions cover 98% of variation. Choosing two principal components also makes it easier for interpretation, although in this case we do not have any additional information about what these variables mean. Next, we transform our original dataset by multiplying it by the 2 eigenvectors (decided upon by those corresponding with the highest eigenvalues). Finally, we calculate the percentage variation explained by our two principal components, 98%.






Scree Plot
0.0
1.0
2.0
3.0
1
2
3
4
5
6
7
Principal Component


Principal Component Analysis
Component 2
Component 1
3
0
−3
−6
−6
−3
0
3
6

```
## [1] 0.9803036
```

5A. Description

We conduct a principal components analysis examining 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 relationship more closely for a handful of variables. As expected mpg and weight have a negative relationship, so heavier vehicles have worse gas mileage than lighter vehicles, on average. The relationship is not strictly linear, and appears to be more curved, closer to a negative quadratic relationship. Of course there are some vehicles that perform much better than their weight class peers (around 3,000 lbs), for example, the 1982 Oldsmobile Cutlass Ciera and the 1981 Volvo, both diesel engines, compared to their non-diesel peers.

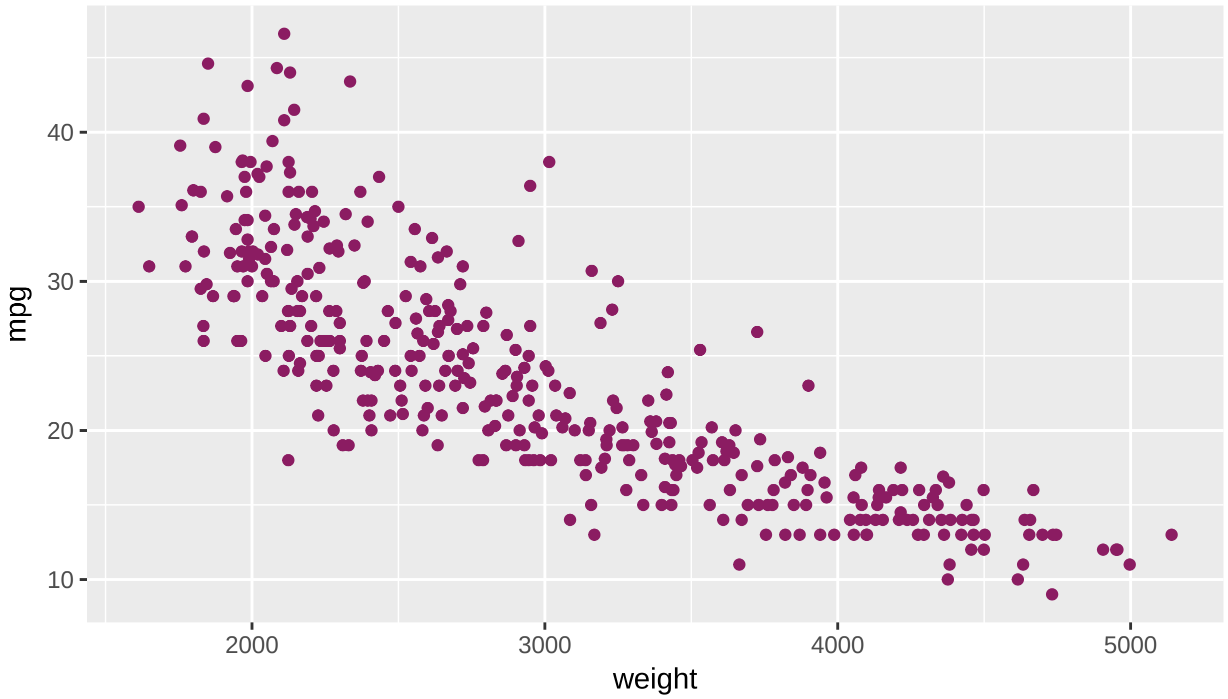
Examining the relationship between horsepower and mpg we notice again a negative, somewhat curved/quadratic relationship. Typically as horsepower increases the gas mileage decreases, which we would expect. Horsepower roughly translate to the amount of energy exerted, and the more vehicles exert, the worse fuel efficieny they have.

We also examine the relationship betweeen weight and acceleration, which appears to be slightly negative, but has a lot of variation. On average, heavier cars have slightly less acceleration, but the amount of variation is likely also due to the number of engine cylinders as well. So a car with the same weight, but very different acceleration is likely related to a difference in engine composition as well.

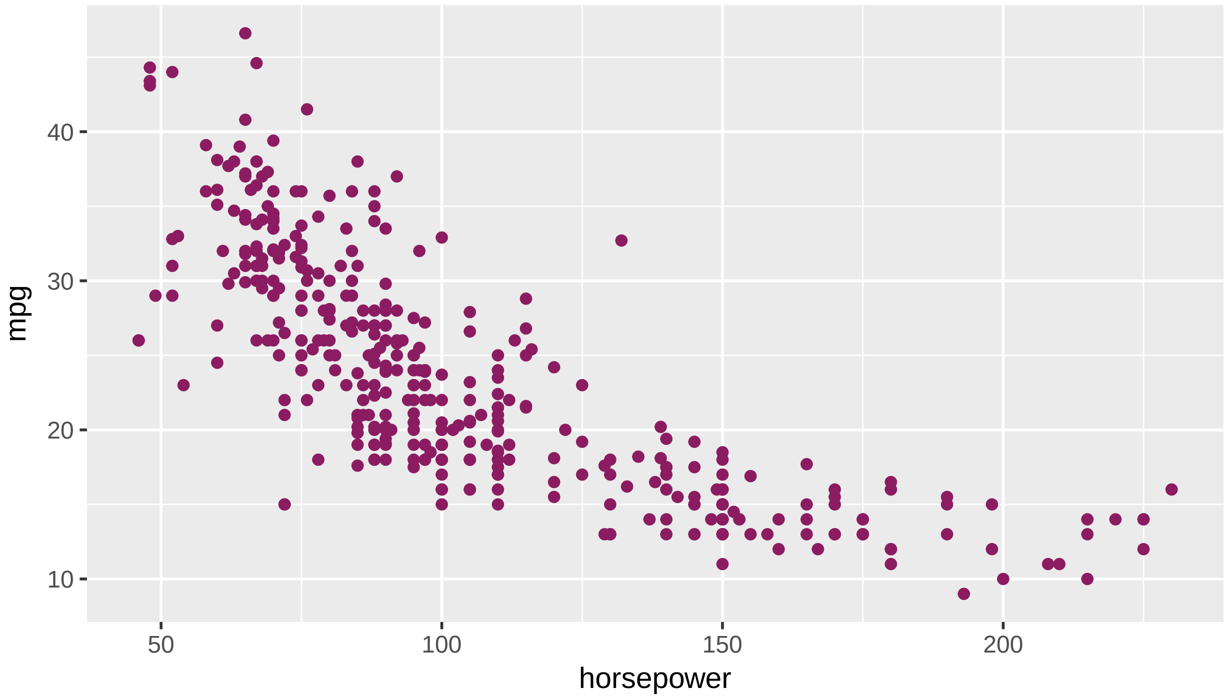
Lastly we examine our data by region and notice the vehicles from the United States have more horsepower and are heavier than those from Western Europe and Japan. Again, this seems reasonable because the USA produces many more trucks and large vehicles than the more compact vehicles produced in Western Europe and Japan.

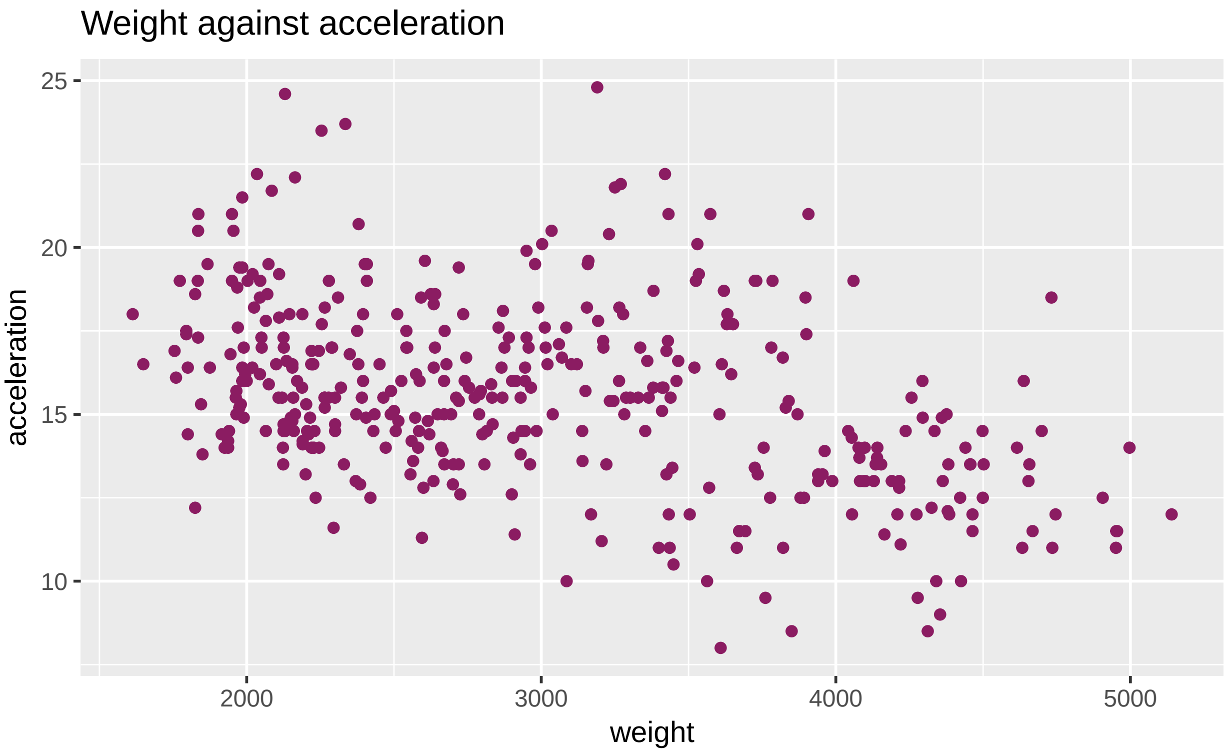
```
##       mpg          cylinders      displacement     horsepower   
##  Min.   : 9.00   Min.   :3.000   Min.   : 68.0   Min.   : 46.0  
##  1st Qu.:17.50   1st Qu.:4.000   1st Qu.:104.2   1st Qu.: 75.0  
##  Median :23.00   Median :4.000   Median :148.5   Median : 93.5  
##  Mean   :23.51   Mean   :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  
```

Weight against MPG



horsepower against MPG





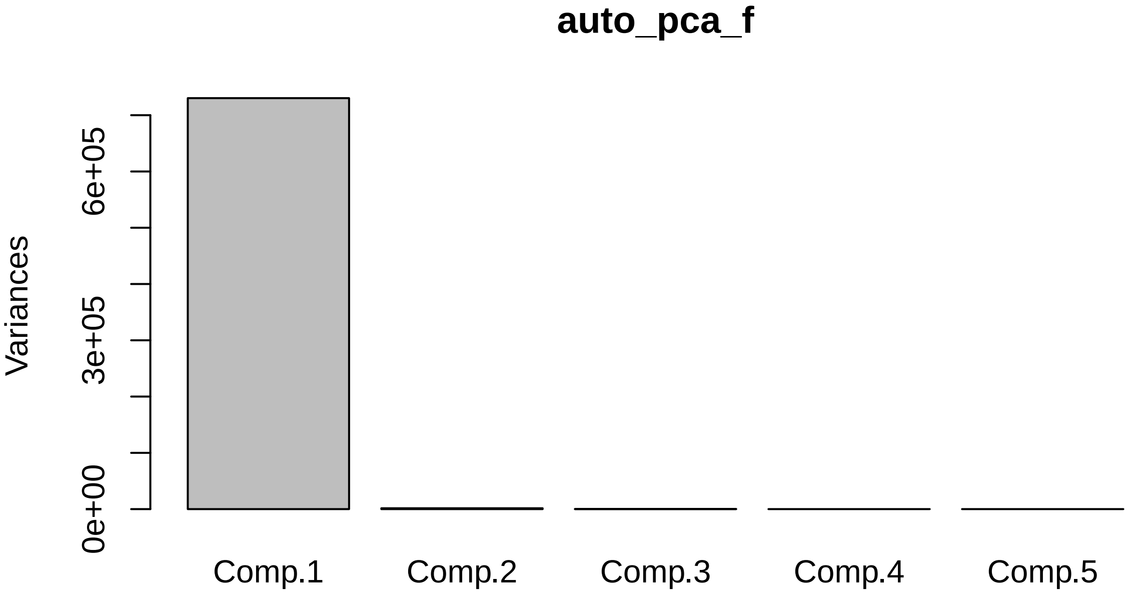
```
##   origin freq
## 1      1  249
## 2      2   70
## 3      3   79
```

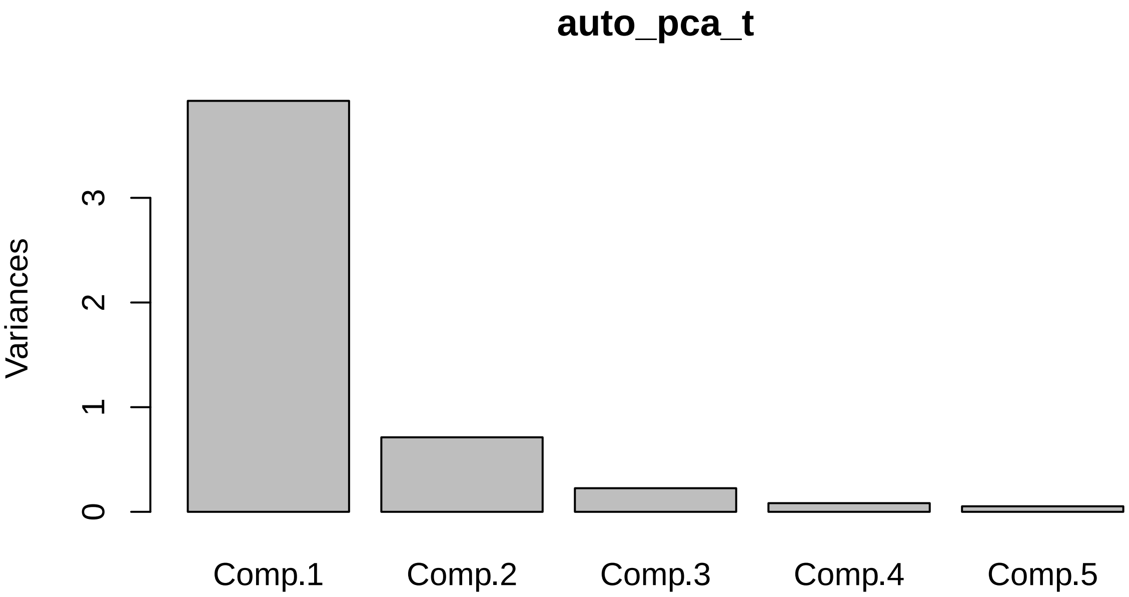
```
##   cylinders freq
## 1         3    4
## 2         4  204
## 3         5    3
## 4         6   84
## 5         8  103
```

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

```
##          Comp.1    Comp.2      Comp.3     Comp.4      Comp.5
## [1,]  -536.4594 -50.72157  10.8562082 -0.9172728 -1.90062244
## [2,]  -730.3736 -79.06730  -8.9625698 -1.1586047  0.38311509
## [3,]  -470.9978 -75.34599  -5.0521831 -0.3447421 -0.89622572
## [4,]  -466.4329 -62.46467  -9.3084497 -2.4381965 -0.03862074
## [5,]  -481.6914 -55.69105  -0.4563547 -1.7861272 -2.43922934
## [6,] -1383.9297 -85.38258 -14.6642740  4.1519564  0.47457211
```

```
##         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
```

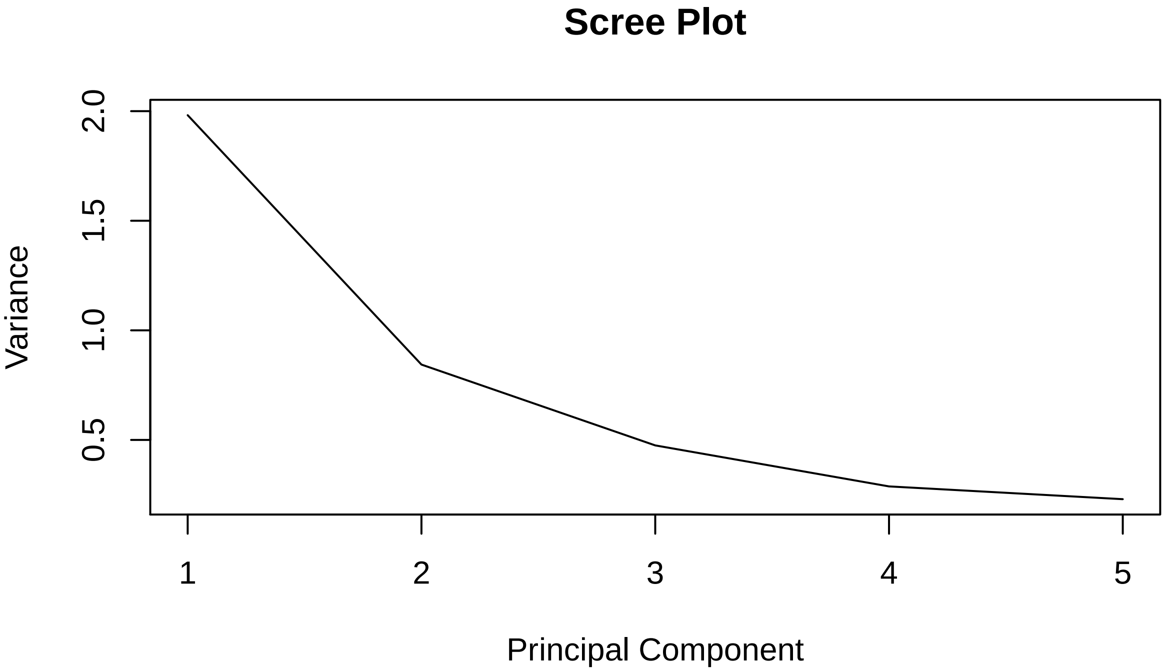




5B/C. Initial PCA / variable retention

Next we conduct our principal component analysis on mpg, displacement, horsepower, weight, and acceleration. We center our data then compute the correlation and covariance matrix, comparing our results. We notice the correlation matrix loads across multiple principal components (accounting for scaling) while the covariance matrix loads almost all of the weight onto the first principal component (not accounting for scaling). As a result we choose our correlation matrix, and after examining our screen plot decide 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
## 0.7853509 0.9277520 0.9728764 0.9894514 1.0000000
```



5D. Factor Loadings

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/weighting of our different variables. The first factor weighs 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
```

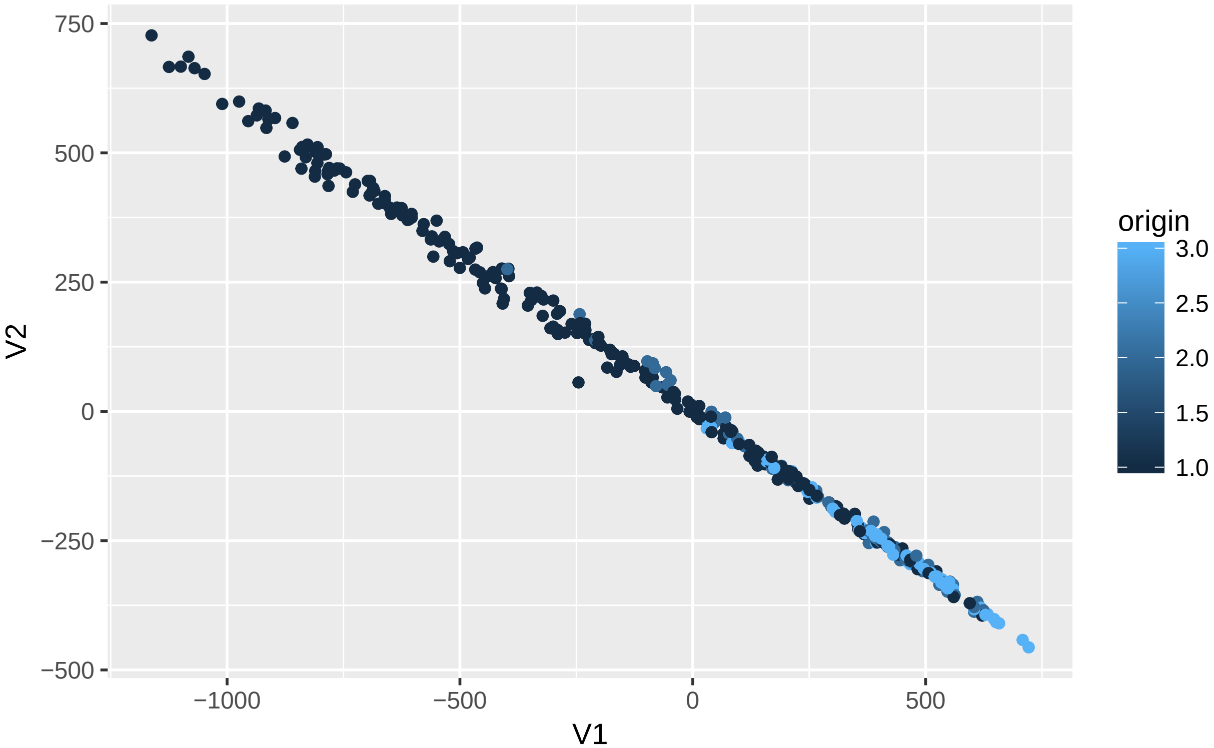
5E. Projections

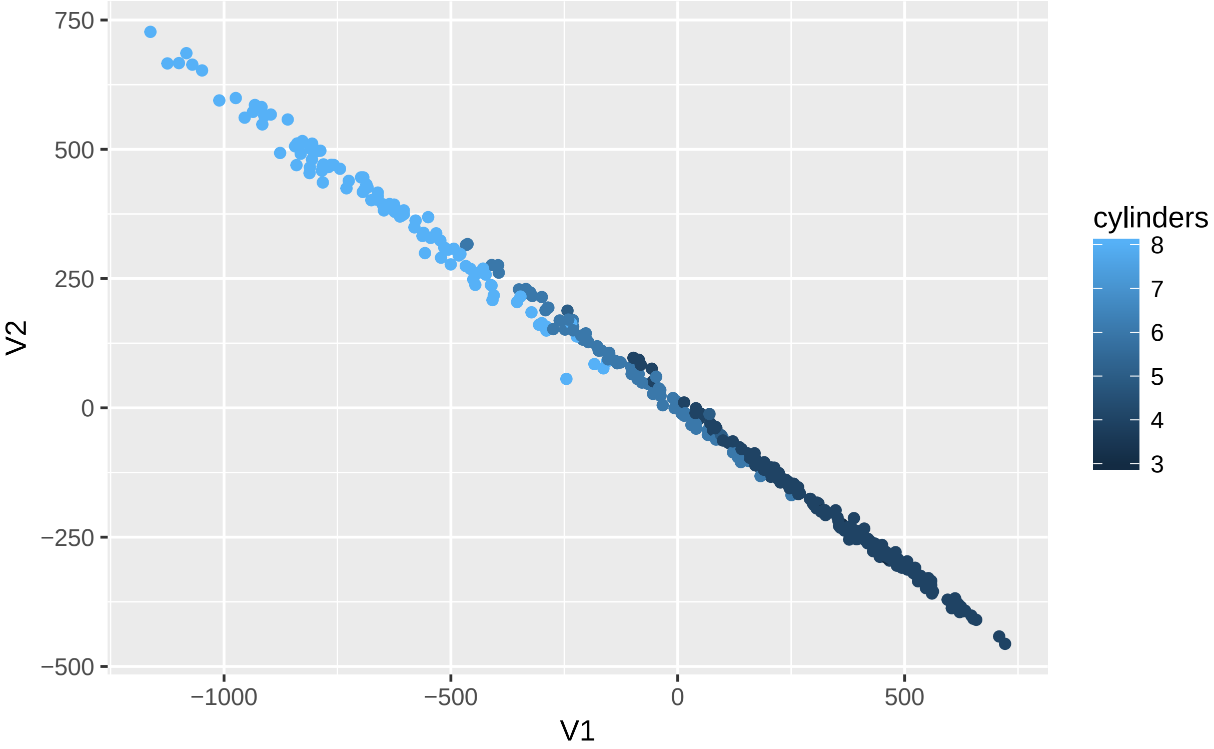
We next project our data onto our first two principal components. We notice our data are tight bound, without much variation, and a strong negative relationship between component 2 and 1, so as component 1 increases, component 2 decreases. Based on what we noticed about factor loadings we could conjecture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

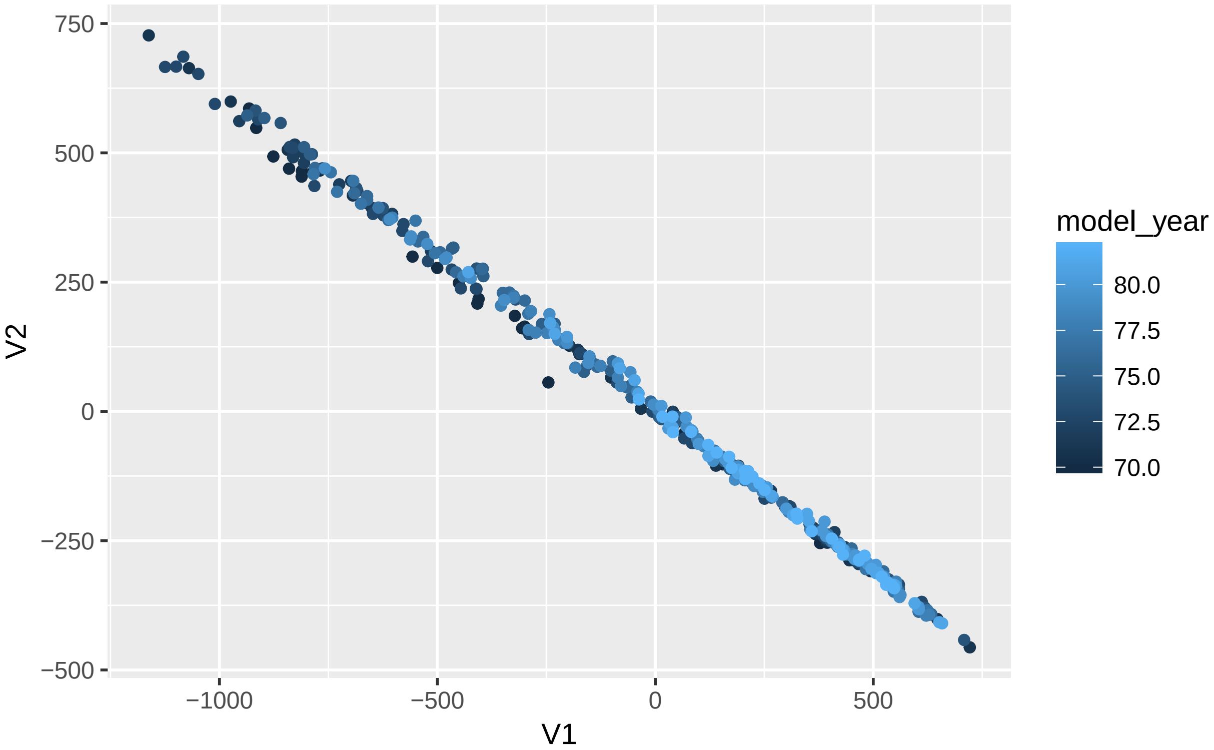
In particular we examine cylinder, model year, and origin. Overall it seems data are more distinguishable according to cylinder than the other categorical variables. For engine cylinders we see that cylinders map almost perfectly onto our principal components, with the 8 cylinder engines mapping to a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


V2
750
500
250
0
−250
−500
−1000
−500
0
500
V1
origin
3.0
2.5
2.0
1.5
1.0


V2
750
500
250
0
−250
−500
−1000
−500
0
500
V1
cylinders
8
7
6
5
4
3



5F. Bootstrapping

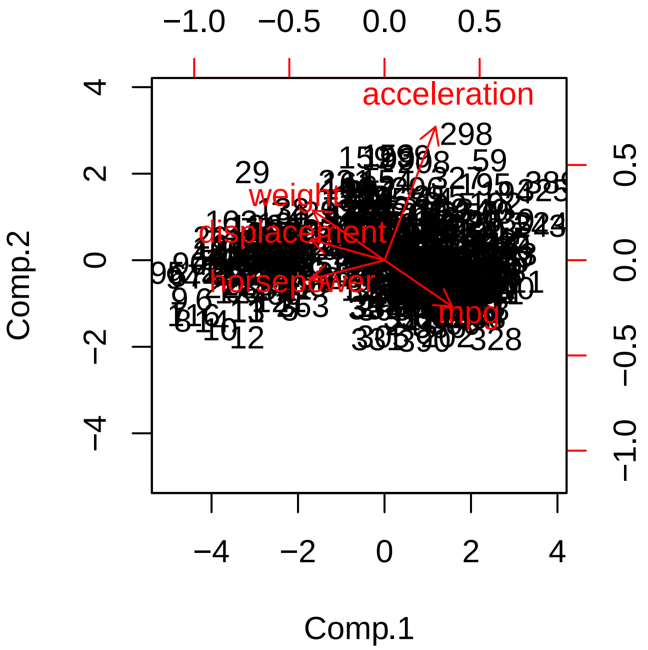
Next we create bootstrapped confidence intervals to understand the percentage of variation explained by the first two principal components. We estimate with 95% confidence that our first principal component value (eigenvalue) is between 3.82 and 4.04. For our second principal component we expect it to be between .629 and .805.

```
##     2.5%    97.5% 
## 3.813883 4.036532
```

```
##      2.5%     97.5% 
## 0.6333011 0.8095232
```

5G. Biplot

Finally we make a biplot 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.



−1.0
−0.5
0.0
0.5
acceleration
weight
displacement
horsepower
mpg
298
59
29
12
328
Comp.2
Comp.1
4
2
0
−2
−4
0.5
0.0
−0.5
−1.0