Principal Component Analysis with Automobile data

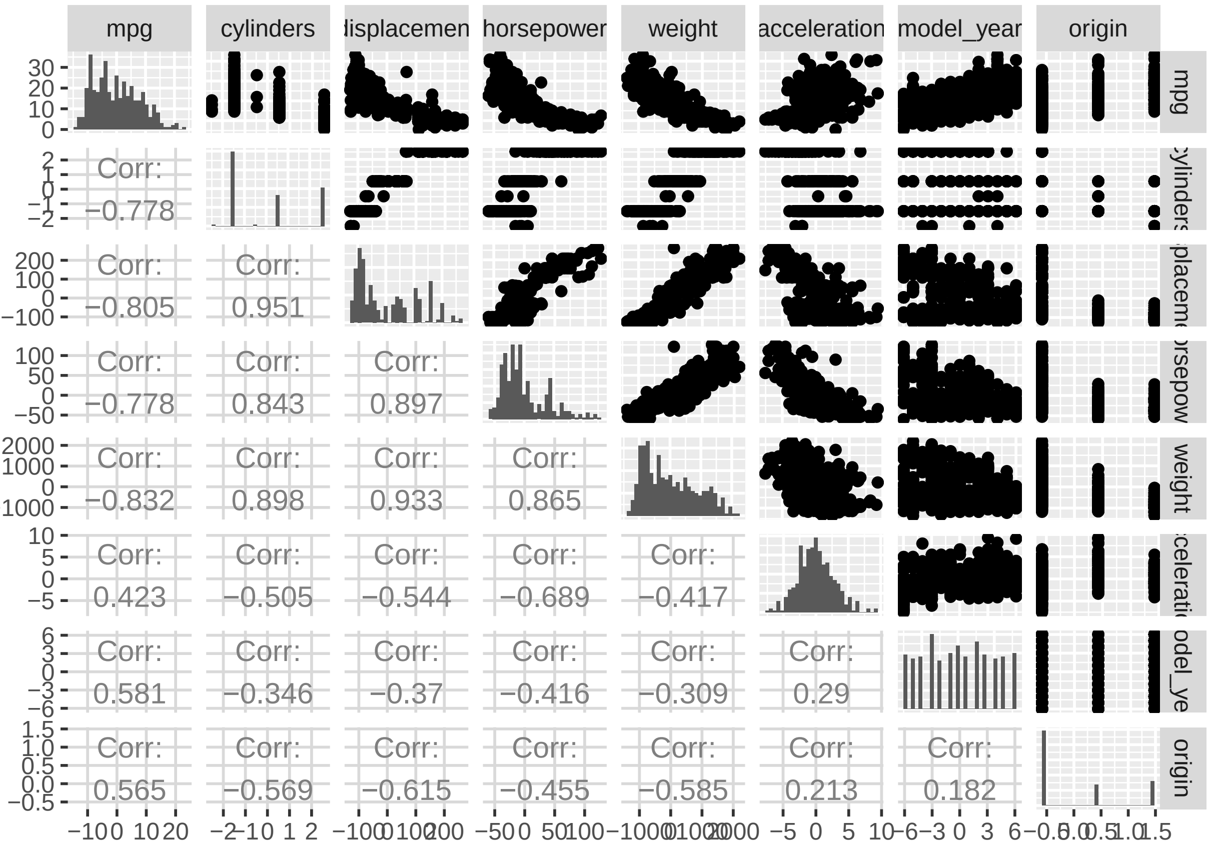
Katherine Wilkinson

1/19/2018

In this exercise you are welcome to use an existing PCA package. Here we use princomp from the GGally library. The data set used (auto-mpg.data) concernce city-cycle fuel consumption in miles per gallon (mpg) and other attributed collected for 398 vehicle instances. ## (a) Describe the Data

Describe the data and present some intial pictorial and numerical summaries, such as scatterplots, histograms, etc. Consider which variables should or should not be included in PCA on this dataset. Compare PCA on covariances and correlations.

```
##       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
```



Our original dataset has 8 variables with 398 observations. From our summary of each variable, we can see there are a few categorical variables, particularly the car name. We can see with the histogram and covariance plot (Figure 1) there there is some definite correlation between a few of our variables. Immediately, we can see that origin, cylinders and model year are categorical. There does seem to be some trends in model year, most noteable with mpg and weight. MPG has a positive correlation with model year while weight has a negative correlation with model year, suggesting that over the years, mpg for vehicles has increased and weight has decreased.

Our non-categorical variables have more interesting relationships. For instance, mpg has relatively high negative correlation with both displacement and weight with -0.805 and -0.832 respectively. In contrast, horsepower is positively correlated with displacement and weight with 0.897 and 0.865 correlation respectively. Horsepower and mpg also seem to be slightly negatively correlated with a correlation of -0.778. We can also see that displacement and weight are highly positively correlated with a correlation of 0.933. Our final numeric variable, acceleration, interestingly does not have as high of correlation with any of our other variables. This can be seen easily in the scatter plots.

In order to reduce the dimensionality of our data, we are going to use Principal Component Analysis (PCA). For our PCA, we will use just our non-categorical variables, mpg, displacement, weight, horsepower, and acceleration.

Initially, we will calculate all the principa; components (PCs) using first the covariance matrix and then using the correlation matrix to view the results for unstandardized data and standardized data.

```
##       Comp.1       Comp.2       Comp.3       Comp.4       Comp.5
## 7.303222e+05 1.509427e+03 2.604651e+02 1.777346e+01 2.895191e+00
```

```
##
## Loadings:
##              Comp.1 Comp.2 Comp.3 Comp.4 Comp.5
## mpg                               0.999
## displacement -0.114 -0.946  0.303
## horsepower          -0.298 -0.949
## weight       -0.993  0.121
## acceleration                              0.996
##
##                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
```

```
##     Comp.1     Comp.2     Comp.3     Comp.4     Comp.5
## 3.92675439 0.71200563 0.22562178 0.08287545 0.05274275
```

```
##
## 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
```

```
##    Comp.1    Comp.2    Comp.3    Comp.4    Comp.5
## 1.9816040 0.8438043 0.4749966 0.2878810 0.2296579
```

The covariance matrix PCs are quite large, with our first PC at $7.3 * 10^5$. In contrast, the first PC for the correlation matrix is 3.9267. The loadings are also quite different. Since our numeric variables are measured on different scales (for instance mpg is measured differently than horsepower), we will use the PCs from our correlation matrix so our data is standardized.

(c) Percentage of variance explained

```
##     Comp.1     Comp.2     Comp.3     Comp.4     Comp.5
## 3.92675439 0.71200563 0.22562178 0.08287545 0.05274275
```

```
##              [,1]
## Comp.1 3.92675439
## Comp.2 0.71200563
## Comp.3 0.22562178
## Comp.4 0.08287545
## Comp.5 0.05274275
```

```
## [1] 0.7853509
```

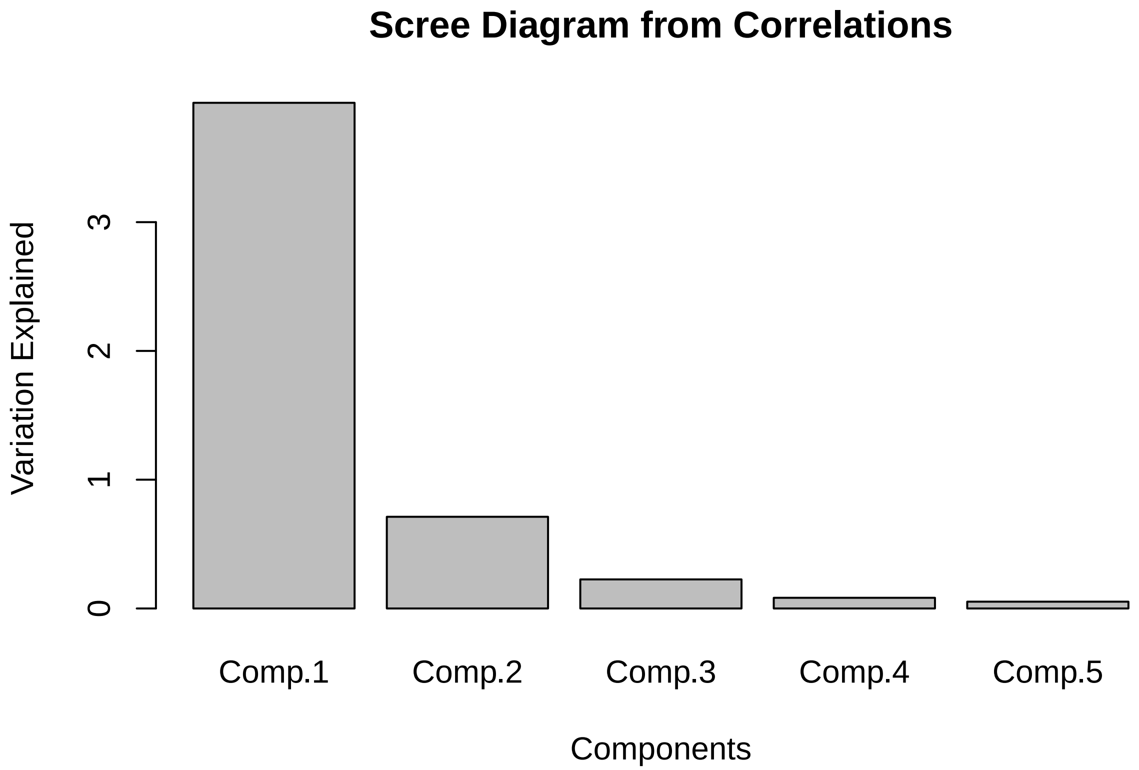
```
## [1] 0.927752
```

```
## [1] 0.7853509
```

```
## [1] 0.1424011
```

```
## [1] 0.04512436
```

```
## [1] 0.01657509
## [1] 0.01054855
```



Initially there are 5 PCs, which summed together, explain all the variance. Each PC explains a percentage of the total variance, as seen in the table below.

% PC 1	% PC 2	% PC 3	% PC 4	% PC 5
78.54%	14.24%	4.51%	1.66%	1.05%

From here, we can see that with just the first 2 PCs, 92.775% of the variance is explained. We can also see from the scree diagram that there is a significant jump in variance explained after the 3rd PC. Thus, moving forward, we will use just the first two principal components.

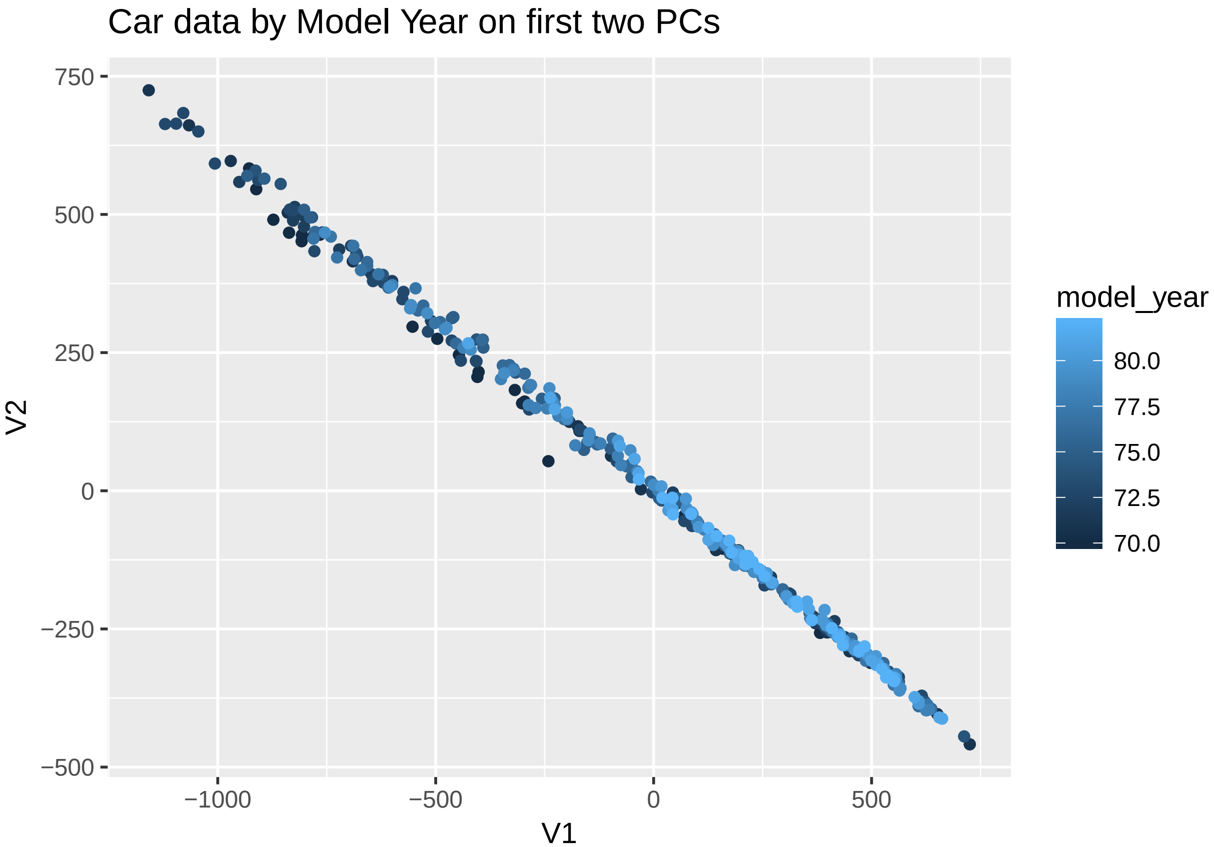
Factor Loadings

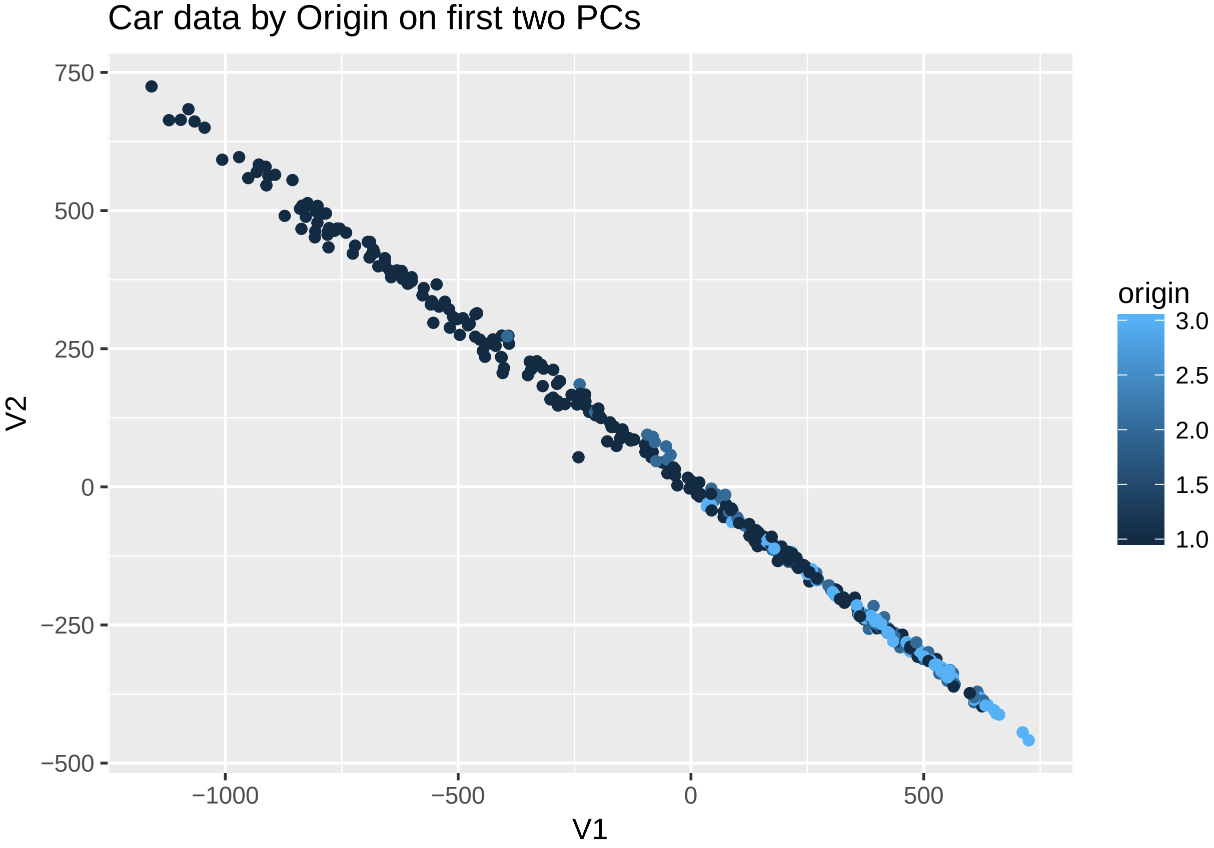
```
##
## Loadings:
##              Comp.1 Comp.2 Comp.3 Comp.4 Comp.5
## mpg           0.444 -0.304  0.839
## displacement -0.483  0.135  0.371 -0.476  0.620
## horsepower   -0.484 -0.124  0.206  0.826  0.160
## weight       -0.471  0.326  0.305 -0.159 -0.744
## acceleration  0.335  0.876  0.150  0.257  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
```

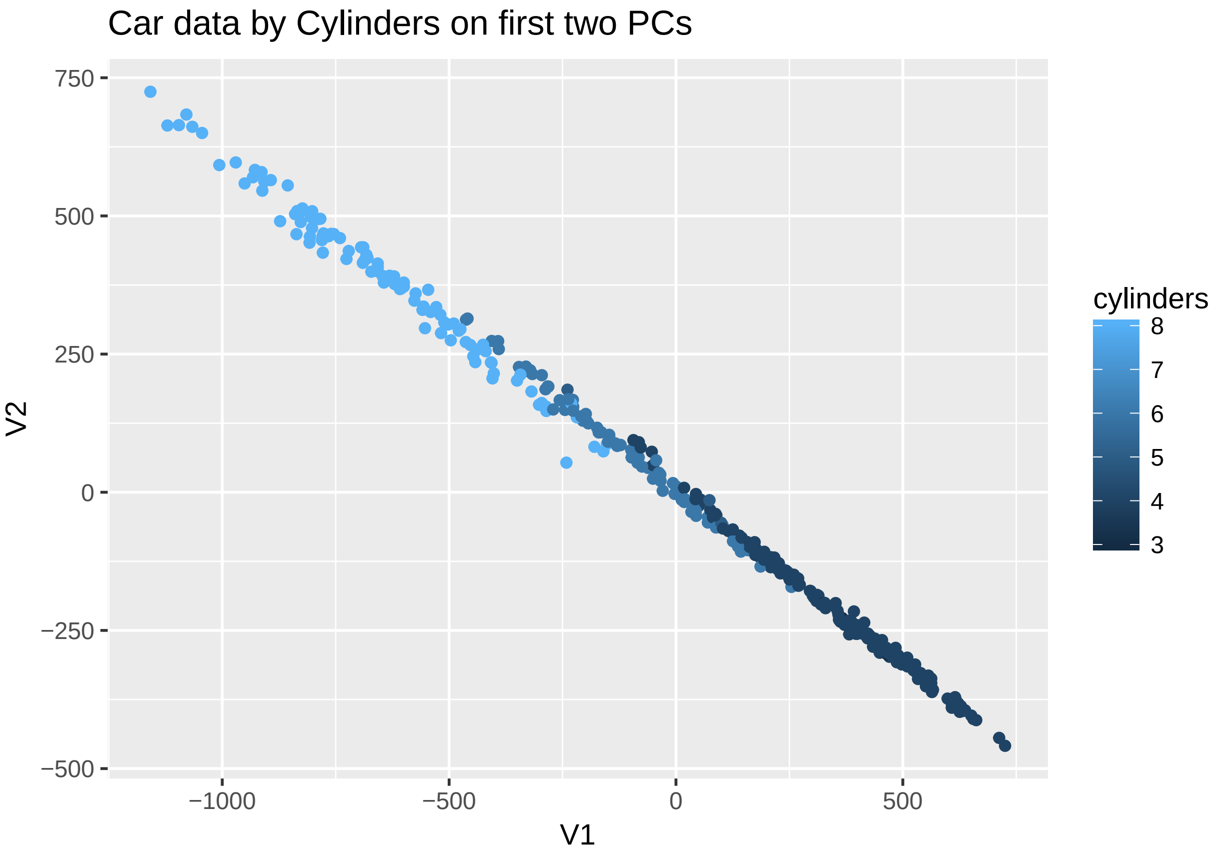
The first PC is not super highly correlated with any of the variables, as seen by the factor loadings above. The first PC is negatively correlated with displacement, weight, and horsepower, thus it increases as these variables decrease. Alternatively, as mpg and acceleration increase, the first component also increases.

The second componenet however, is relatively highly correlated with acceleration at -0.876. When we look beyond the first two compoenent, we can see that the third component is highly correlated with mpg at 0.839, but is not very correlated with the other variables. The last two PCs are not correlated at all with mpg and only have higher correlation with a single variable each (horsepower and weight respectively).

Plot data projected on first two PCs







The plots of the data projected on the first two PCs do not appear to show any major outliers, although there is one point right about in the middle of the graph that may be a slight outlier. There also does appear to be some categorical attributes, particularly by cylinder and (slightly) origin. That is, the data is sepatated and grouped by number of cylinders in the third graph.

Bootstrapping

Compute boostrap confidence intervals for the percentage of variance explained by the first 2 PCs.

Bootstrapping our data, we get many different samples of PCs and thus many different percentage of variance explained by the first 2 PCs that we retained for this dataset. From this we calculate the following 95% confidence interval for the percentage of variance explained by the first and the second principal component.

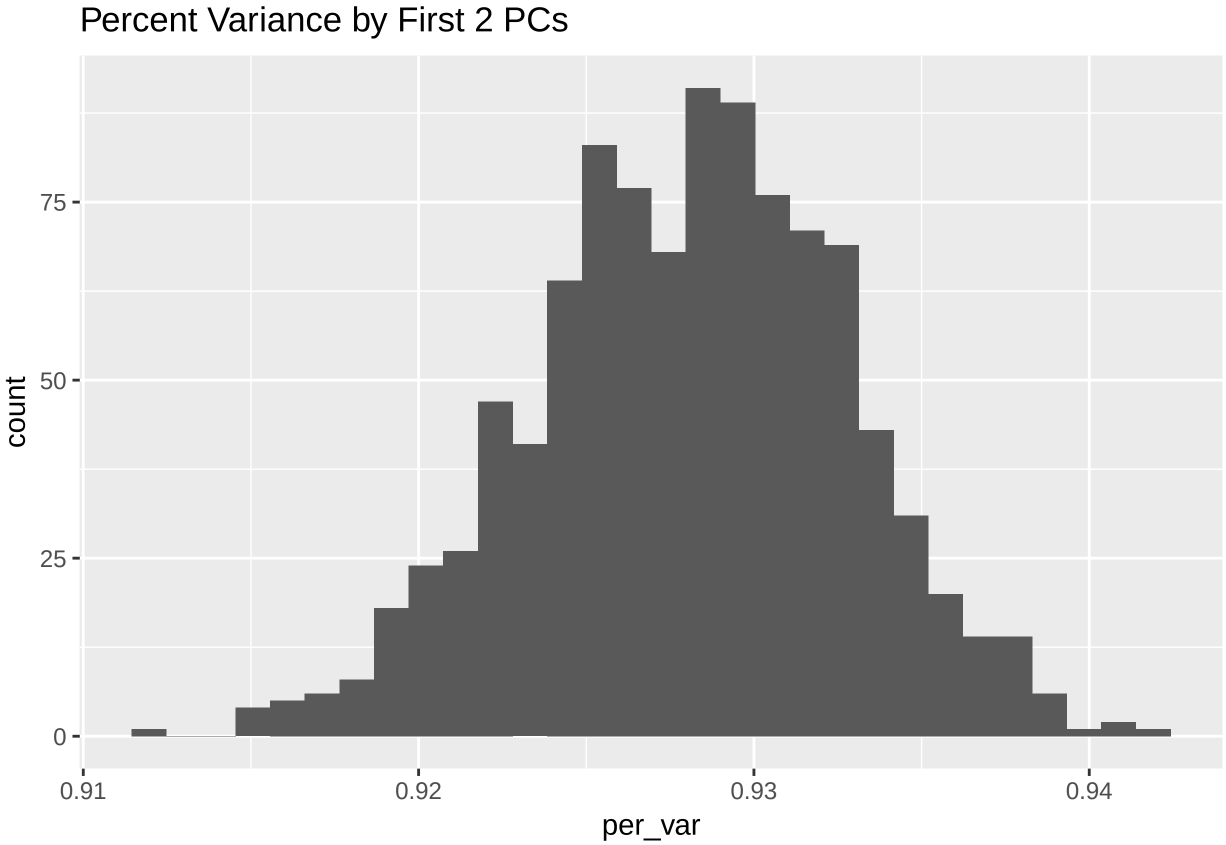
CI	% from PC1	% from PC2
2.5%	77.801%	13.647%
97.5%	80.778%	16.127%

We can also calculate a confidence interval for the sum of the percentage of variance explained by the first 2 PCs:

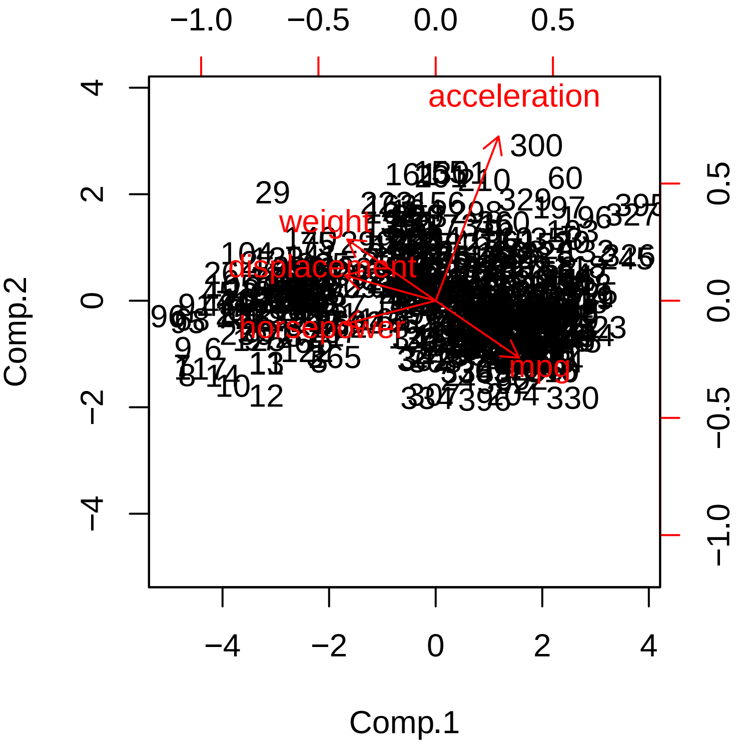
2.5%	97.5%
91.87%	93.76%

This follows our results we calculated above where we got 92.775% variance explained by the first two PCs.

We can also see this in our histogram of the bootstrap samples of the percent variance explained by the first two PCs, where most of our points fall within the confidence interval.



Biplot



Finally, we can look at a PCA biplot and see how each of our data points align with our variable directions. We can see that all of our data points are fairly clustered along the weight and mpg, which look to be nearly perfectly negatively correlated. However, there is a clump of data points that have less horsepower. We can also see clearly that acceleration is not very correlated with any of the other variables.