1. Descriptive Analysis

```
>round(cor(car),1)
```

Correlation	wheel.base	length	width	height	curb.weight	engine.size	bore	stroke	compression. ratio	horsepower	peak.rpm	city.mpg	highway.mpg	price
wheel.base	1	0.9	0.8	0.6	0.8	0.6	0.6	0.3	0.2	0.4	-0.5	-0.5	-0.6	0.6
length	0.9	1	0.9	0.4	0.9	0.7	0.7	0.2	0.1	0.7	-0.5	-0.8	-0.8	0.7
width	0.8	0.9	1	0.2	0.9	0.8	0.5	0.2	0.1	0.7	-0.3	-0.7	-0.8	0.8
height	0.6	0.4	0.2	1	0.2	-0.1	0.2	0.1	0.2	-0.1	-0.2	-0.1	-0.1	0.1
curb.weight	0.8	0.9	0.9	0.2	1	0.9	0.7	0.2	0.1	0.8	-0.4	-0.8	-0.9	0.9
engine.size	0.6	0.7	0.8	-0.1	0.9	1	0.6	0.1	0	0.9	-0.4	-0.7	-0.8	0.9
bore	0.6	0.7	0.5	0.2	0.7	0.6	1	-0.1	0	0.6	-0.4	-0.6	-0.6	0.6
stroke	0.3	0.2	0.2	0.1	0.2	0.1	-0.1	1	0.1	0.1	-0.1	-0.1	-0.1	0.1
compression.ratio	0.2	0.1	0.1	0.2	0.1	0	0	0.1	1	-0.3	-0.4	0.3	0.3	-0.1
horsepower	0.4	0.7	0.7	-0.1	0.8	0.9	0.6	0.1	-0.3	1	-0.1	-0.9	-0.8	0.8
peak.rpm	-0.5	-0.5	-0.3	-0.2	-0.4	-0.4	-0.4	-0.1	-0.4	-0.1	1	0.1	0.2	-0.2
city.mpg	-0.5	-0.8	-0.7	-0.1	-0.8	-0.7	-0.6	-0.1	0.3	-0.9	0.1	1	1	-0.8
highway.mpg	-0.6	-0.8	-0.8	-0.1	-0.9	-0.8	-0.6	-0.1	0.3	-0.8	0.2	1	1	-0.8
price	0.6	0.7	0.8	0.1	0.9	0.9	0.6	0.1	-0.1	0.8	-0.2	-0.8	-0.8	1

From the correlation table, we could see that wheel base and length, wheel base and width, wheel base and curb weight, length and width, length and curb weight, length and city mpg, length and highway mpg, width and curb weight, width and engine size, width and highway mpg, width and price, curb weight and horsepower, curb weight and city mpg, curb weight and highway mpg, curb weight and price, engine size and horsepower, engine size and highway mpg, engine size and price, horsepower and city mpg, horsepower and highway mpg, horsepower and price, city mpg and price, highway mpg and price are highly correlated.

2. Simple Linear Regression

```
Coefficients:

              Estimate Std. Error t value Pr(>|t|)

(Intercept)  1.064e+02  2.921e+01   3.643 0.000914 ***

x1           4.694e-01  1.762e-01   2.664 0.011849 *

x2          -1.766e-01  1.031e-01  -1.713 0.096005 .

x3          -4.436e-01  4.323e-01  -1.026 0.312330

x4          -5.382e-01  2.671e-01  -2.015 0.052130 .

x5          -8.033e-03  3.136e-03  -2.562 0.015171 *
```

```
x6            3.916e-03  2.555e-02   0.153 0.879124
x7           -3.699e+00  2.292e+00  -1.614 0.116133
x8           -5.067e-01  1.404e+00  -0.361 0.720363
x9            5.710e-01  1.350e-01   4.229 0.000175 ***
x10          -3.650e-02  2.615e-02  -1.396 0.172070
x11          -1.551e-03  1.467e-03  -1.057 0.298284
x14           1.300e-04  1.157e-04   1.124 0.269075
```

```
x1<- car$wheel.base
x2<- car$length
x3<- car$width
x5<- car$curb.weight
x9<- car$compression.ratio
x12<- car$city.mpg
```

After getting rid of insignificant variables at .1 level.

The Model for X12 can be represented as:

X12 = 106.4+ .04694 X1 - .02766 X2 - .05382 X3 – 8.033* e^3 X5+ .05710 x9

For X12, city mpg, it negatively related to car length, width, weight.

Indicating that the bigger, heavier tend to have less mile per gallon to drive in city. Also, city mpg, is positively correlated to wheel base, and compression ratio, where the greater the wheel base and compression ratio will result higher city mpg.

```
Response x13 :

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.258e+02  3.191e+01   3.941 0.000397 ***
x1           3.358e-01  1.924e-01   1.745 0.090289 .
x2          -1.001e-01  1.126e-01  -0.889 0.380607
```

```
x3          -6.686e-01  4.722e-01  -1.416 0.166138

x4          -5.225e-01  2.918e-01  -1.791 0.082500 .

x5          -1.033e-02  3.425e-03  -3.015 0.004917 **

x6          -9.928e-03  2.791e-02  -0.356 0.724339

x7          -2.487e+00  2.504e+00  -0.993 0.327809

x8           6.617e-02  1.533e+00   0.043 0.965833

x9           6.768e-01  1.475e-01   4.589 6.16e-05 ***

x10         -1.515e-02  2.856e-02  -0.531 0.599311

x11         -2.081e-03  1.603e-03  -1.299 0.203024

x14          2.016e-04  1.263e-04   1.596 0.120057

---

x1<- car$wheel.base

x4<- car$height

x5<- car$curb.weight

x9<- car$compression.ratio

x13<- car$highway.mpg
```

After getting rid of insignificant variables at .1 level.

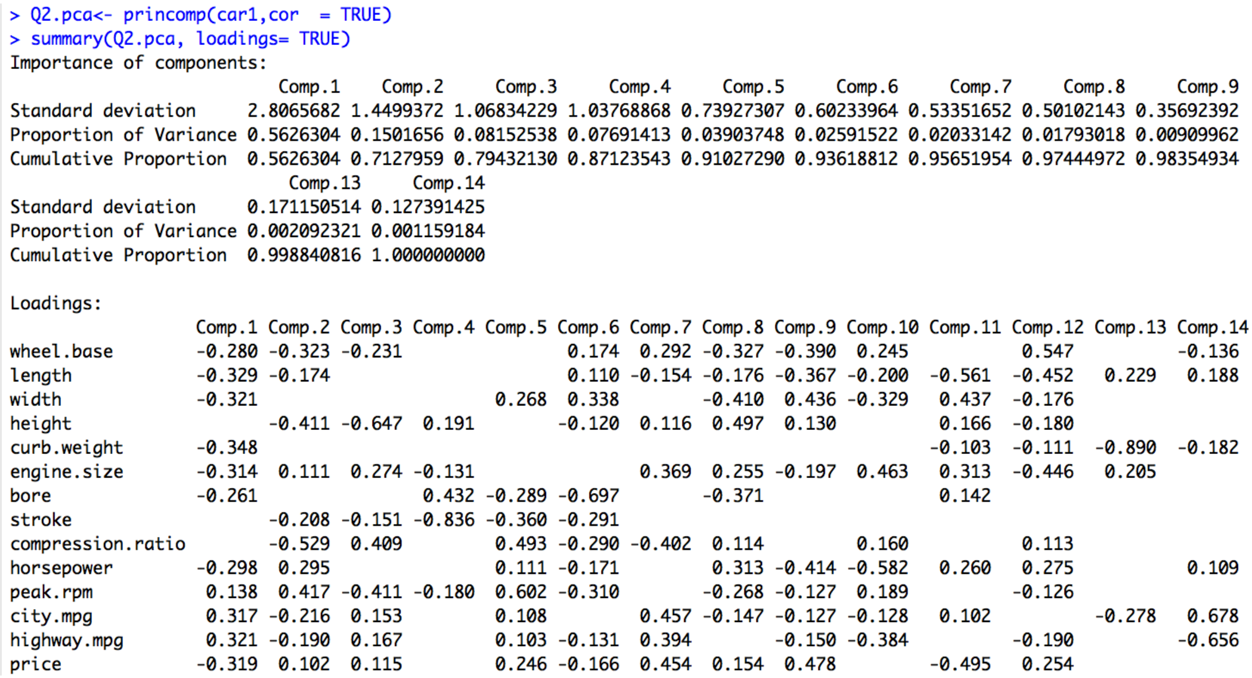
The Model for X12 can be represented as:

X13 = 125.8 + .03358 X1 - .05225 X4 - .01033 X5 + .6768 x9

The model shows that highway mpg is negatively correlated with car height, weight. Depicting that cars with greater height and weight will reduce the highway miles per gallon.

In addition, highway mpg is positively correlated with wheel base and compression ration, which a greater value in wheel base and compression ratio will increase highway mpg for cars.

3. Reducing Dimensions of object data using PCA Methodology



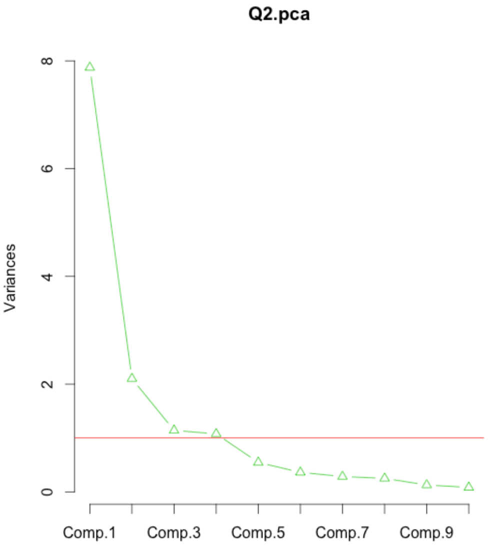
```
> Q2.pca<- princomp(car1,cor  = TRUE)
> summary(Q2.pca, loadings= TRUE)
Importance of components:
                          Comp.1    Comp.2     Comp.3     Comp.4     Comp.5     Comp.6     Comp.7     Comp.8     Comp.9
Standard deviation     2.8065682 1.4499372 1.06834229 1.03768868 0.73927307 0.60233964 0.53351652 0.50102143 0.35692392
Proportion of Variance 0.5626304 0.1501656 0.08152538 0.07691413 0.03903748 0.02591522 0.02033142 0.01793018 0.00909962
Cumulative Proportion  0.5626304 0.7127959 0.79432130 0.87123543 0.91027290 0.93618812 0.95651954 0.97444972 0.98354934
                           Comp.13     Comp.14
Standard deviation     0.171150514 0.127391425
Proportion of Variance 0.002092321 0.001159184
Cumulative Proportion  0.998840816 1.000000000

Loadings:
                  Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8 Comp.9 Comp.10 Comp.11 Comp.12 Comp.13 Comp.14
wheel.base        -0.280 -0.323 -0.231                0.174  0.292 -0.327 -0.390   0.245           0.547          -0.136
length            -0.329 -0.174                       0.110 -0.154 -0.176 -0.367  -0.200  -0.561  -0.452   0.229   0.188
width             -0.321                       0.268  0.338        -0.410  0.436  -0.329   0.437  -0.176
height                   -0.411 -0.647  0.191        -0.120  0.116  0.497  0.130           0.166  -0.180
curb.weight       -0.348                                                                  -0.103  -0.111  -0.890  -0.182
engine.size       -0.314  0.111  0.274 -0.131                0.369  0.255 -0.197   0.463   0.313  -0.446   0.205
bore              -0.261                0.432 -0.289 -0.697        -0.371                  0.142
stroke                   -0.208 -0.151 -0.836 -0.360 -0.291
compression.ratio        -0.529  0.409         0.493 -0.290 -0.402  0.114          0.160           0.113
horsepower        -0.298  0.295                0.111 -0.171         0.313 -0.414  -0.582   0.260   0.275           0.109
peak.rpm           0.138  0.417 -0.411 -0.180  0.602 -0.310        -0.268 -0.127   0.189          -0.126
city.mpg           0.317 -0.216  0.153         0.108         0.457 -0.147 -0.127  -0.128   0.102          -0.278   0.678
highway.mpg        0.321 -0.190  0.167         0.103 -0.131  0.394        -0.150  -0.384          -0.190          -0.656
price             -0.319  0.102  0.115         0.246 -0.166  0.454  0.154  0.478          -0.495   0.254
```

By using PCA, we could use 4 principal components to reduce the dimensions, as cumulatively, four principal components could represent 87% of the variance.

Based on the first principal component, it is clear that cars with smaller sizes, lighter in weights, and higher city or highway mpg could contribute to a higher PC1 component value.

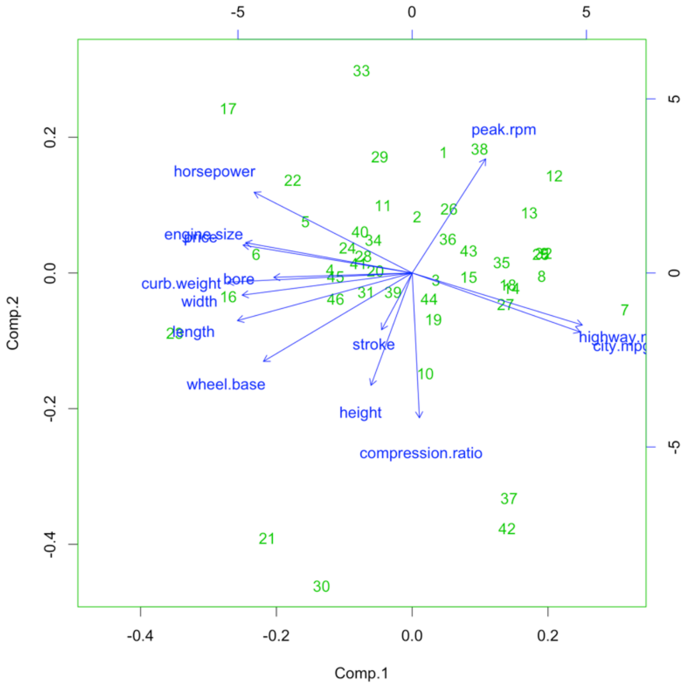
```
> screeplot(Q2.pca, type = "line")
> abline(h= 1,col= "red")
```



From the screeplot, it appears that first four principal components can reflect greater variances of the dataset.

Therefore, the dimension can be reduced by these four principal components.

```
> biplot(Q2.pca)
```



Also from the biplot, we can see that high way mpg, city mpg, peak rpm, and compression ratio is positively correlated in the first principal component. Horsepower, engine size, length width, and other variables are negatively affecting the first principal component value.

For the first principal component, it more likely reflect the car's fuel efficiency factor, as cars with higher city or highway mpg, smaller in sizes could result higher principal component values.

For example, the 7th car in the bi plot, has a higher PC1 value, which corresponding to Chevrolet, that has the highest city and highway mpg and relatively small car sizes, in the car dataset.

The second principal component more likely reflect the engine performance factor, as cars with higher horsepower, engine size, price will have a high PC2 value, similarly, these cars are negatively related the car sizes, such as length, width, and weight.

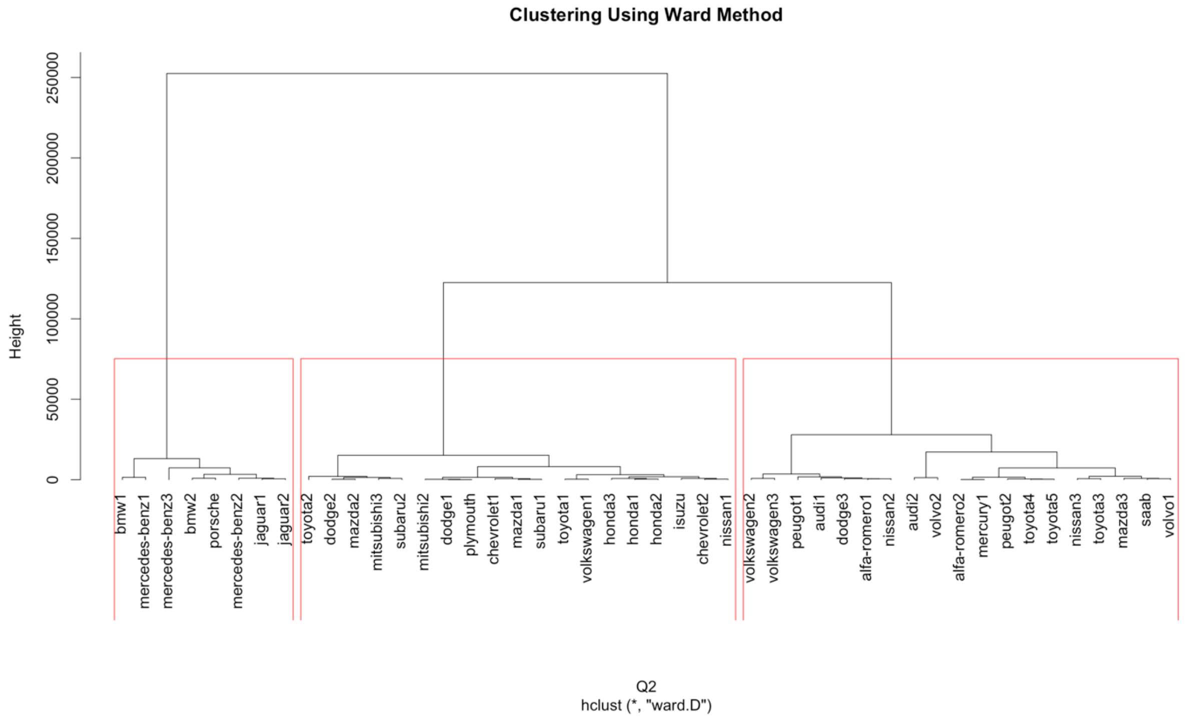
For example, the 33rd car, which has the highest PC2 value, happens to be a Porsche, which has a higher price, greater horsepower, and engine size compare to other cars.

Data Visualization

4. Data Visualization

```
hc444<-hclust(Q2, "ward")
```

```
plot(hc444,hang=-1,labels = car[,2], main = "Clustering Using Ward
Method");re4<-rect.hclust(hc444,k=3,border="red")
```



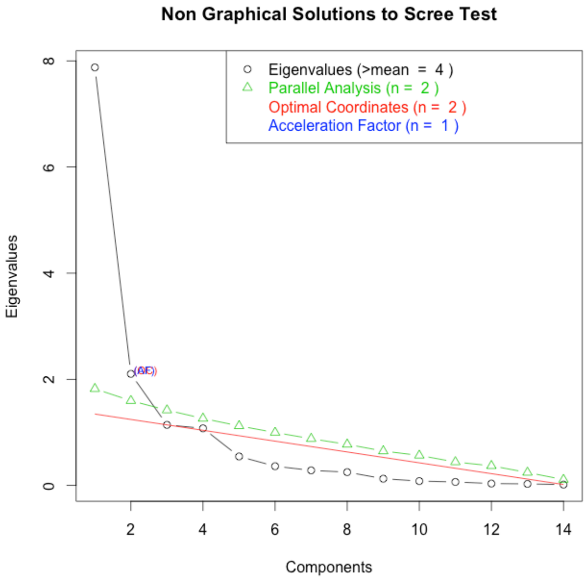
By using ward method, we can divide 46 cars into 3 groups. Based on the clustering, the first group of cars seem to have higher prices, and greater horsepower, and less fuel efficient.

The second group contain most of the Japanese cars such as Honda, Toyota, and Mitsubishi compared to the third group, and these cars seem to be more fuel efficient, and less expensive than group1.

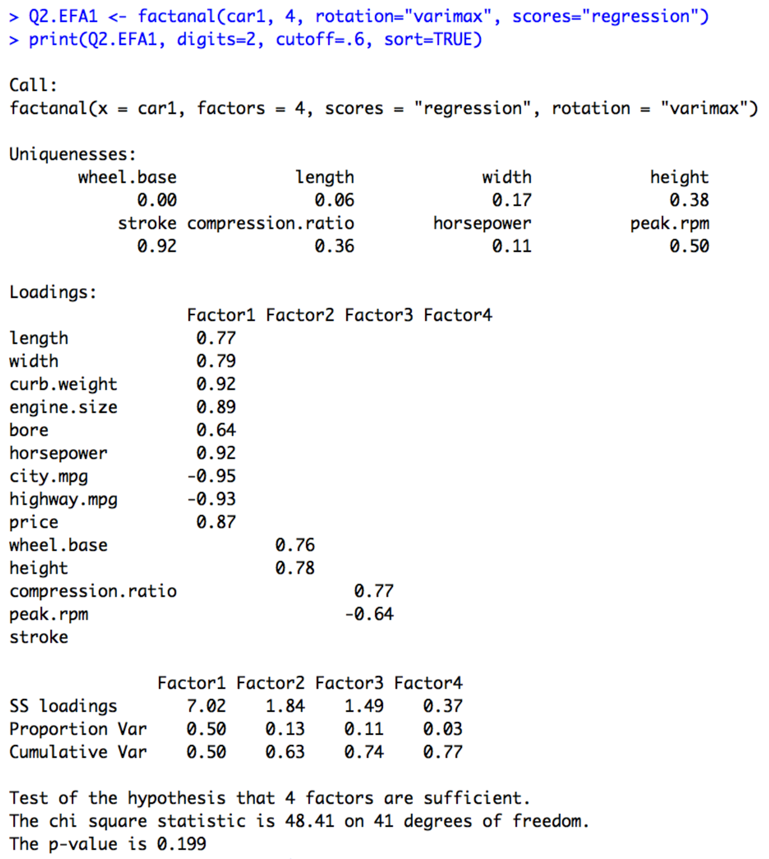
The third group contain most of European made cars such as Volkswagen, Audi, and Volvo. Generally speaking, these cars have engine performance between luxury cars in group 1 and less expensive cars in group 2. The price is moderate, not extremely expensive.

5. Factor Analysis Using Exploratory Factor Analysis (EFA)

```
> plotnScree(Q2.nS)
```



From the scree plot, there seems to have 4 factors that has an eigenvalue greater than mean. Based on optimal coordinates, we might choose 2 factors in total, but as the third and fourth factors have eigenvalues higher than the mean, I will still use four factors in EFA.



```
> Q2.EFA1 <- factanal(car1, 4, rotation="varimax", scores="regression")
> print(Q2.EFA1, digits=2, cutoff=.6, sort=TRUE)

Call:
factanal(x = car1, factors = 4, scores = "regression", rotation = "varimax")

Uniquenesses:
       wheel.base            length             width            height        curb.weight       engine.size              bore
             0.00              0.06              0.17              0.38              0.02              0.07              0.51
           stroke compression.ratio        horsepower          peak.rpm           city.mpg       highway.mpg             price
             0.92              0.36              0.11              0.50              0.00              0.04              0.14

Loadings:
                  Factor1 Factor2 Factor3 Factor4
length             0.77
width              0.79
curb.weight        0.92
engine.size        0.89
bore               0.64
horsepower         0.92
city.mpg          -0.95
highway.mpg       -0.93
price              0.87
wheel.base                 0.76
height                     0.78
compression.ratio                  0.77
peak.rpm                          -0.64
stroke

               Factor1 Factor2 Factor3 Factor4
SS loadings       7.02    1.84    1.49    0.37
Proportion Var    0.50    0.13    0.11    0.03
Cumulative Var    0.50    0.63    0.74    0.77

Test of the hypothesis that 4 factors are sufficient.
The chi square statistic is 48.41 on 41 degrees of freedom.
The p-value is 0.199
```

By using Exploratory Factor Analysis, we could come up with four factors, as the fourth factor does not contribute anything the the factorial analysis, I will only choose three factors.

From the loadings of the factors, the factor 1 reflect the car fuel efficiency, for which cars that are more fuel efficient will have lower factor value in factor 1.

The factor 2 reflect the height factor of the cars, as cars with greater height, greater wheel base will result higher value in factor 2.

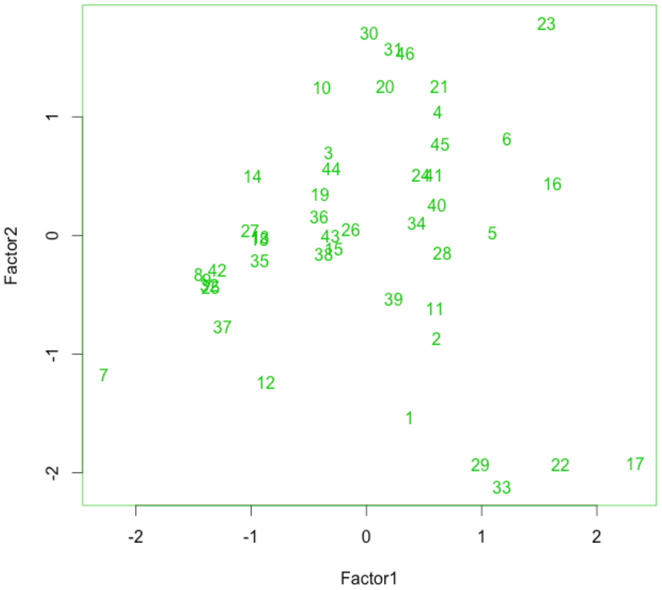
The factor 3 reflect the car engine performance. A higher compression ratio with lower engine peak rpm indicates a more efficient and powerful engine, therefore, the factor 3 value will be higher as well.

Accordingly, the car dataset can be divided into three latent variables, fuel efficiency, height, and engine performance.

```
> plotnScree(Q2.nS)

> Q2.EFA.scores <- Q2.EFA1$scores[,1:2]

> plot(Q2.EFA.scores,type="n")   set up plot
> text(Q2.EFA1$scores[,1],Q2.EFA1$scores[,2])   add variable names
```



As we already come up with three factors in the previous question. Therefore, we are able to better classify the data in the car dataset.

The dimensions have been reduced. In the score plot, I only used the two factors' scores in explaining the data.

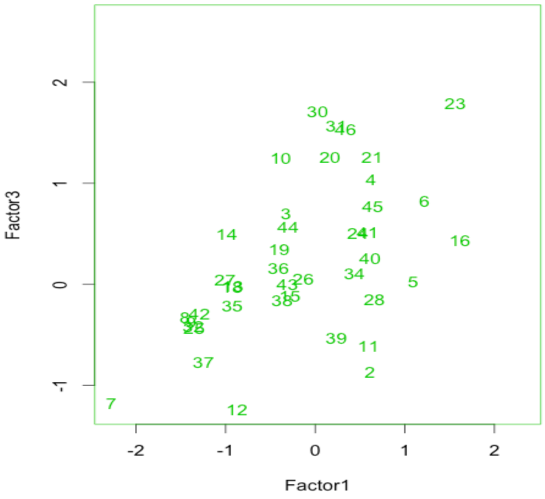
As the factor 1 reflects the fuel efficiency, therefore, the lowest score in factor 1 will indicate the most fuel efficiency car in the dataset, which is the 7th car, the Chevrolet. This is same result we can get from the PCA analysis for the PC1. In addition, those 33rd, 22nd, 17th cars (Porsche, Mercedes, and jaguar2) who has the higher factor 1 scores, will indicate the least fuel efficiency and correspondently have higher prices.

For the Factor 2, the 23rd car Mercedes, and 30th car peugot1 who have a relatively higher factor 2 scores, will indicates these two cars seem to higher than other cars.

```
> Q2.EFA.scores1 <- Q2.EFA1$scores[,c(1,3)]

> plot(Q2.EFA.scores1,type="n")   set up plot

> text(Q2.EFA1$scores[,1],Q2.EFA1$scores[,2])
```

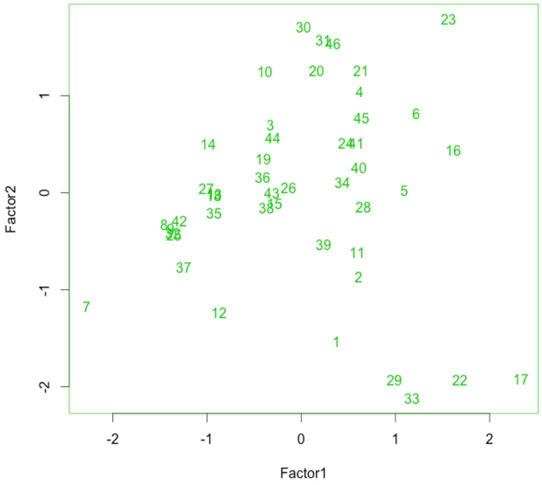


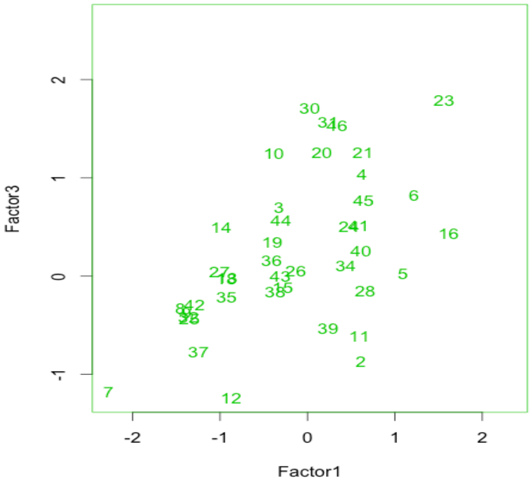
When plot factor scores of factor 1 and factor 3, we will have a better understanding of car fuel efficiency and engine performance.

The 23rd car Mercedes, and 30th car peugot1 have the highest score in factor 3, and therefore these two cars have the best engine performance.

Additionally, the 23rd car Mercedes, also has a relatively higher score in factor, therefore, given the fact that it has a powerful engine, it is a car that is not fuel efficient and expensive.

Conclusion





Based on the EFA, I noticed that most of the cars manufactured are moderately fuel efficient, and have an average car engine.

From the factors between factor 1 and factor 2, the cars that are least fuel efficient and have higher prices, will have smaller car sizes, probably because smaller car sizes could lead to faster car acceleration, and therefore less fuel efficient.