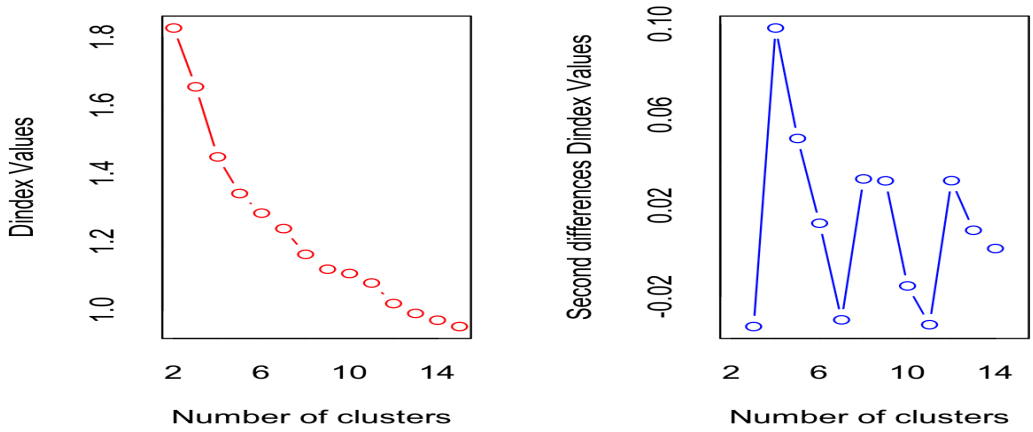
K-means Clustering
Car Spec Data

<R>

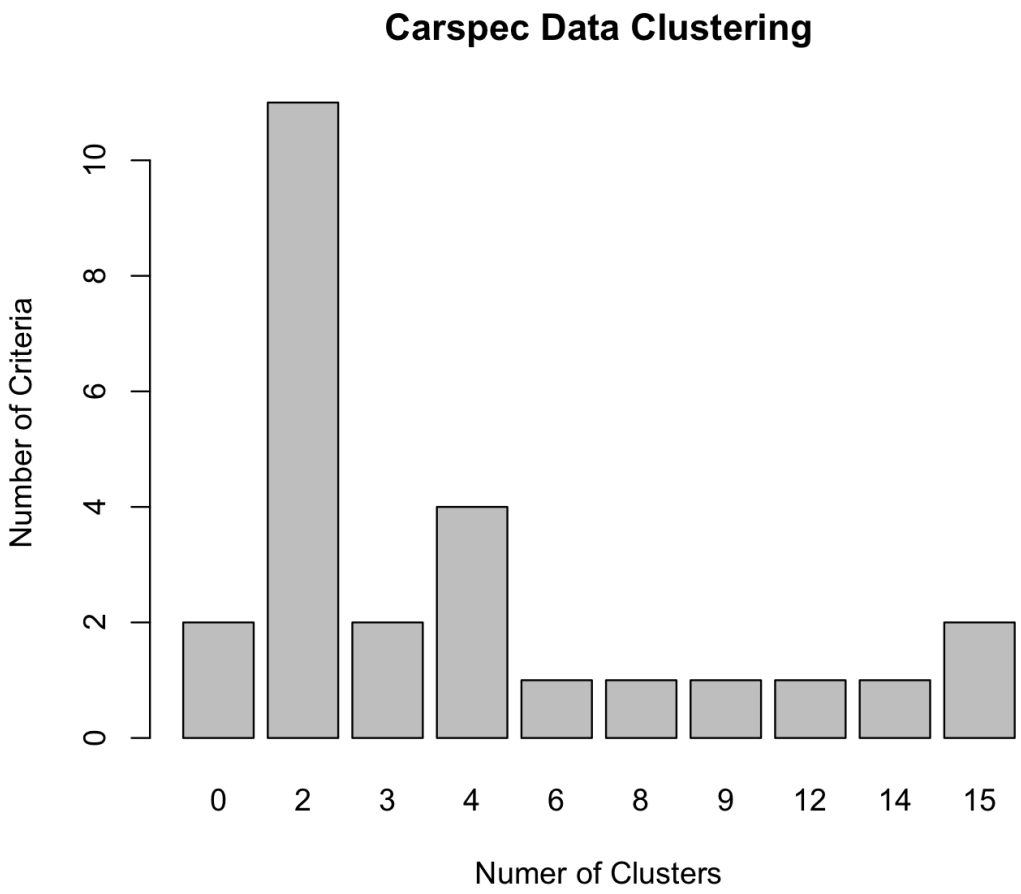
```
> car <- read.csv("/Users/ichung-gi/Documents/spdier2/Clustering/Auto.csv",
header=TRUE);      Read Auto.csv data
> car$horsepower <- as.numeric(car$horsepower) Make type from factor to
numeric, because scale() function needs numeric type
> data <- scale(car[-1])   Make Gaussian distribution and nomalization
> library(NbClust)
> cluster <- NbClust(data, min.nc=2, max.nc=15, method="kmeans")  Find
best k value
```




```
*** : The Hubert index is a graphical method of determining the number of clusters.
                In the plot of Hubert index, we seek a significant knee that corresponds to a
                significant increase of the value of the measure i.e the significant peak in Hubert
                index second differences plot.

*** : The D index is a graphical method of determining the number of clusters.
                In the plot of D index, we seek a significant knee (the significant peak in Dindex
                second differences plot) that corresponds to a significant increase of the value of
                the measure.

*******************************************************************
* Among all indices:
* 11 proposed 2 as the best number of clusters
* 2 proposed 3 as the best number of clusters
* 4 proposed 4 as the best number of clusters
* 1 proposed 6 as the best number of clusters
* 1 proposed 8 as the best number of clusters
* 1 proposed 9 as the best number of clusters
* 1 proposed 12 as the best number of clusters
* 1 proposed 14 as the best number of clusters
* 2 proposed 15 as the best number of clusters

                   ***** Conclusion *****

* According to the majority rule, the best number of clusters is  2
```

The best number of clusters is 2!!

The graph shows k values.



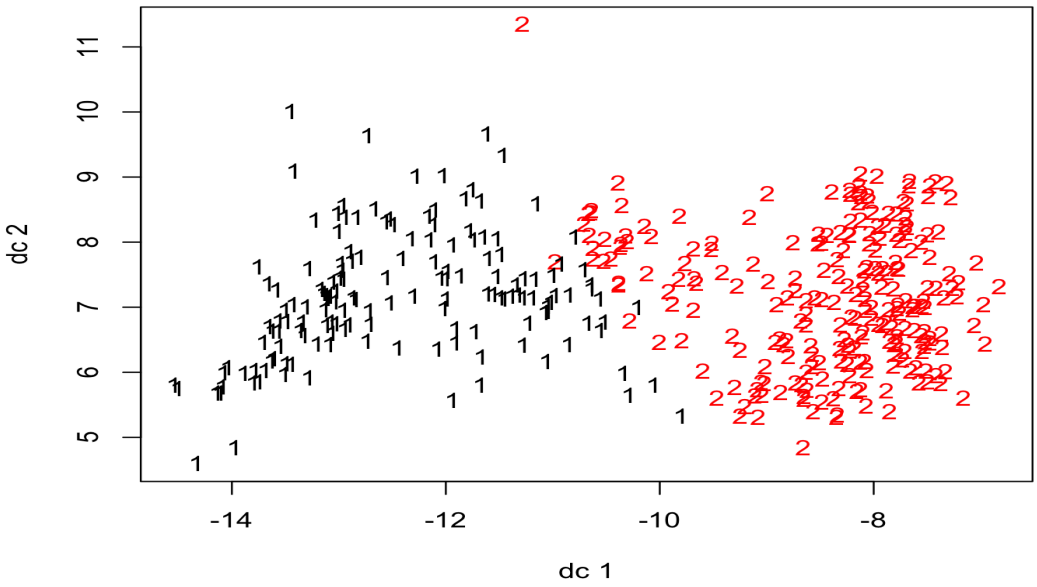
Summary of normalization car spec data

```
> summary(data)
   cylinders        displacement       horsepower         weight        acceleration          year        
 Min.   :-1.4448   Min.   :-1.2027   Min.   :-1.6916   Min.   :-1.6007   Min.   :-2.74752   Min.   :-1.624649  
 1st Qu.:-0.8571   1st Qu.:-0.8578   1st Qu.:-0.8545   1st Qu.:-0.8813   1st Qu.:-0.63843   1st Qu.:-0.811642  
 Median :-0.8571   Median :-0.4554   Median : 0.3176   Median :-0.2008   Median :-0.02024   Median : 0.001365  
 Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.00000   Mean   : 0.000000  
 3rd Qu.: 1.4937   3rd Qu.: 0.6559   3rd Qu.: 0.9203   3rd Qu.: 0.7533   3rd Qu.: 0.56158   3rd Qu.: 0.814372  
 Max.   : 1.4937   Max.   : 2.5050   Max.   : 1.4226   Max.   : 2.5589   Max.   : 3.36158   Max.   : 1.627379  
     origin       
 Min.   :-0.7156  
 1st Qu.:-0.7156  
 Median :-0.7156  
 Mean   : 0.0000  
 3rd Qu.: 0.5304  
 Max.   : 1.7765  
```

Summary of k-means clustering (shows cluster means and clustering vector each elements)

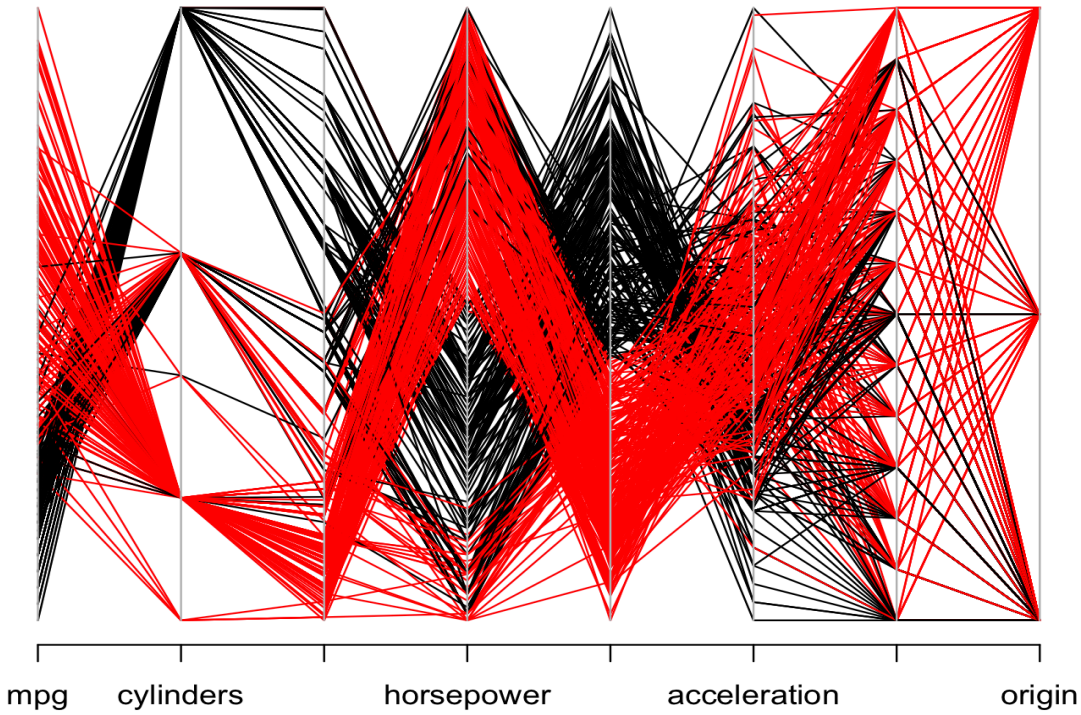
```
> fit.km <- kmeans(car, 2)
>
> fit.km
K-means clustering with 2 clusters of sizes 156, 241

Cluster means:
       mpg cylinders displacement horsepower   weight acceleration     year   origin
1 16.45385  7.237179     301.6538   32.43590 3879.532     14.54167 74.76282 1.064103
2 28.08714  4.307054     123.5456   63.86722 2381.689     16.21203 76.79253 1.904564

Clustering vector:
  [1] 1 1 1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 2 1 1 1 1 2 2 2 2 1 1 1 1 1 1 1 1 1 1 2 2 1 1 2 2 2 2 2 2 2 2
 [59] 2 2 2 2 1 1 1 1 1 1 1 1 2 1 1 1 1 2 2 2 2 2 2 2 2 2 2 1 1 1 1 1 1 1 1 1 1 1 1 1 2 1 2 2 2 2 1 1 1 1 2 2 2 2 2 2 2 2 1
[117] 1 2 2 2 2 2 1 2 2 1 2 2 2 1 2 2 2 2 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 2 2 1 1 1 1 1 1 1 1 1 1 1 2 1 1 2 2 2 2 2 2 2
[175] 2 2 1 2 2 2 2 2 2 2 2 2 2 1 1 1 1 1 1 2 2 2 2 2 2 1 1 1 1 2 2 2 2 1 1 1 2 1 1 1 1 1 1 2 2 2 2 2 2 1 1 1 1 1 1 1 1 1 1
[233] 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 1 1 1 1 1 2 2 2 1 1 1 2 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 1 2 1 2 2 1 2 2 1 1 1 1 1 1 1
[291] 1 1 1 2 2 2 2 2 1 1 1 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
[349] 2 2 2 2 2 2 2 2 2 2 2 2 1 1 2 2 1 1 2 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2

Within cluster sum of squares by cluster:
[1] 40366278 33460846
 (between_SS / total_SS =  74.5 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss" "betweenss"    "size"
[8] "iter"         "ifault"
```

>library(fpc)
>plotcluster(data.train, fit.km$cluster)



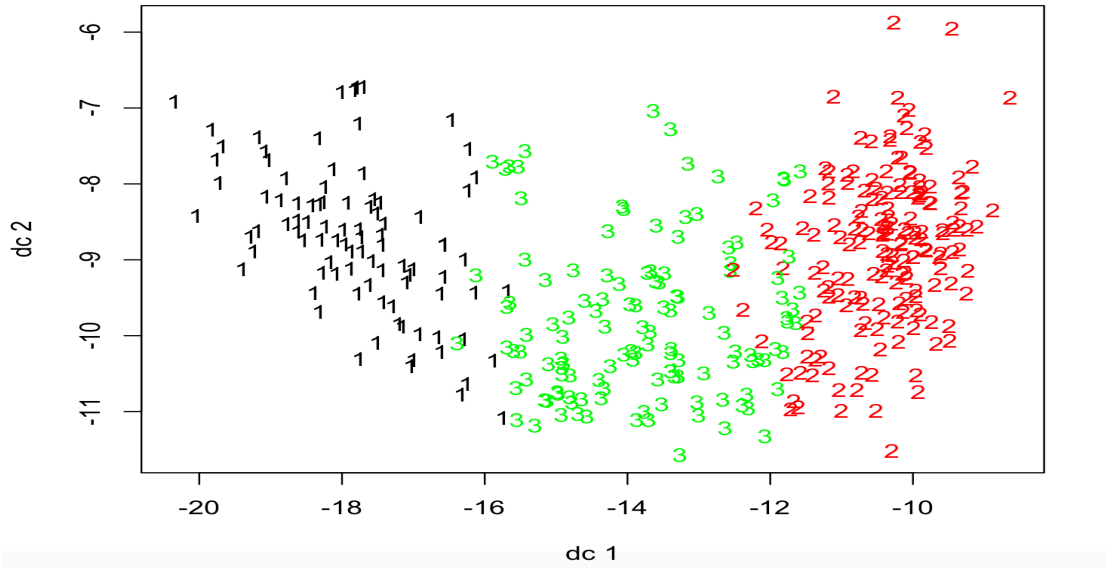
dc1 and dc2 mean PCA components which are most two important components

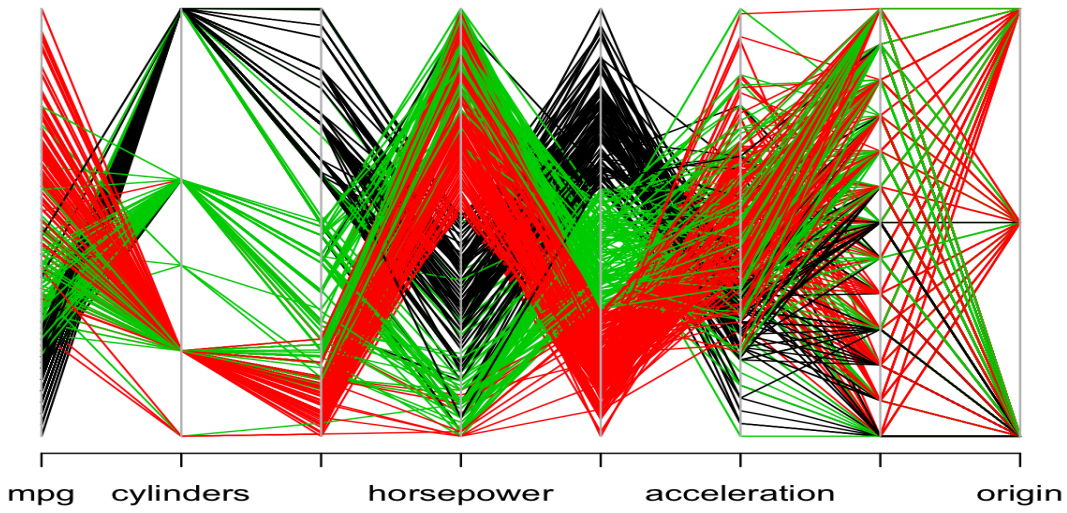
library(MASS)
parcoord(data.train, fit.km$cluster)



K-means clustering car spec data (k is 3)

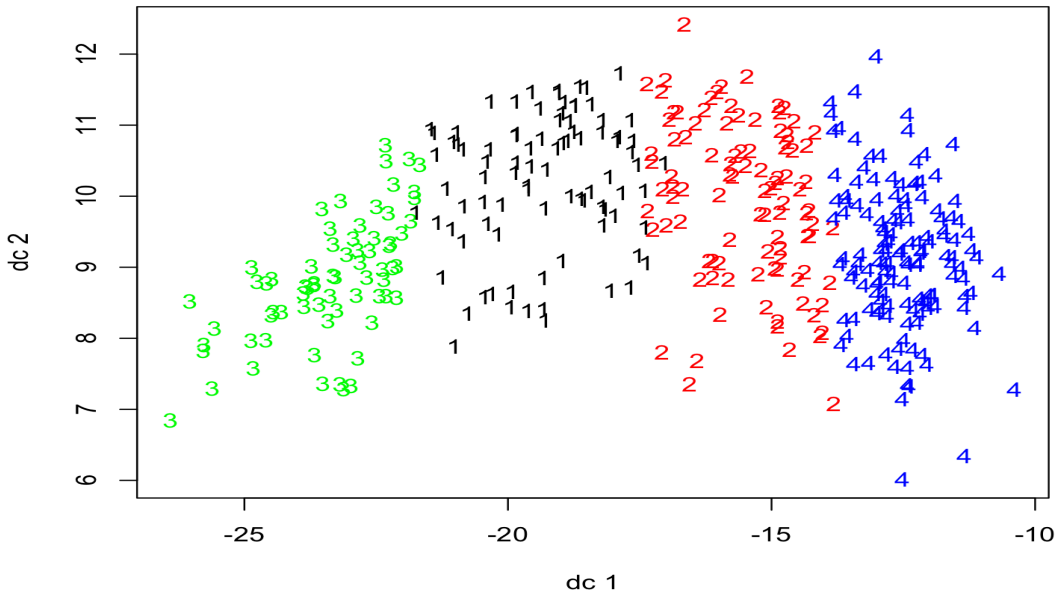
```
> fit.km <- kmeans(car, 3)
>
> fit.km
K-means clustering with 3 clusters of sizes 93, 176, 128

Cluster means:
       mpg cylinders displacement horsepower   weight acceleration     year   origin
1 14.61505  7.870968     343.3441   30.64516 4218.032     13.39462 73.96774 1.010753
2 29.82330  4.039773     106.0653   67.72159 2202.216     16.39148 76.70455 2.062500
3 21.31016  5.656250     204.9531   44.39844 3119.742     15.97656 76.49219 1.312500

Clustering vector:
  [1] 3 1 3 3 3 1 1 1 1 3 3 1 3 2 3 3 2 2 2 3 2 2 2 2 1 1 1 2 2 2 2 2 3 3 3 1 1 1 1 1 1 1 3 2 3 3 2 2 2 2 2 2 2 2
 [59] 2 2 2 2 1 1 1 1 1 1 1 1 1 2 1 1 1 1 3 2 3 3 2 2 2 2 2 2 1 1 1 1 1 1 1 1 1 1 1 1 1 3 3 3 3 2 1 1 1 1 3 2 2 2 2 2 2 2 1
[117] 1 2 2 2 3 3 2 3 1 3 3 3 3 3 2 2 2 2 1 3 3 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 2 2 3 3 3 3 1 1 1 1 1 1 1 1 3 3 3 2 2 3 2 3 2 2
[175] 3 2 3 3 3 3 3 2 2 2 2 2 2 1 1 1 1 3 3 3 3 2 2 2 2 2 3 3 3 3 2 2 2 2 2 3 1 3 3 1 1 1 1 1 2 2 2 2 2 1 1 1 1 3 3 3 3 1 1 1
[233] 1 2 3 2 3 3 2 2 2 2 3 3 2 2 2 2 2 2 3 1 3 3 3 3 3 3 3 3 3 3 3 3 3 3 1 2 2 2 2 2 3 3 2 3 3 3 3 3 2 2 3 3 3 3 3 1 1 1 1 1
[291] 1 3 1 2 2 2 3 3 1 3 3 2 2 2 2 3 2 3 2 2 2 2 2 2 3 3 3 3 2 3 2 2 2 2 3 2 2 2 3 2 2 2 3 2 2 2 3 2 2 2 2 2 2 2 2 2 2 2
[349] 2 2 2 2 2 2 2 2 2 2 2 3 3 3 3 3 1 3 3 2 2 2 2 2 3 3 2 2 2 2 2 2 2 2 2 2 2 2 2 3 3 2 3 3 2 3 3 2 2 2 3

Within cluster sum of squares by cluster:
[1] 11899432 10879486 11586711
 (between_SS / total_SS =  88.1 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss" "betweenss"    "size"
[8] "iter"         "ifault"
> library(fpc)
```

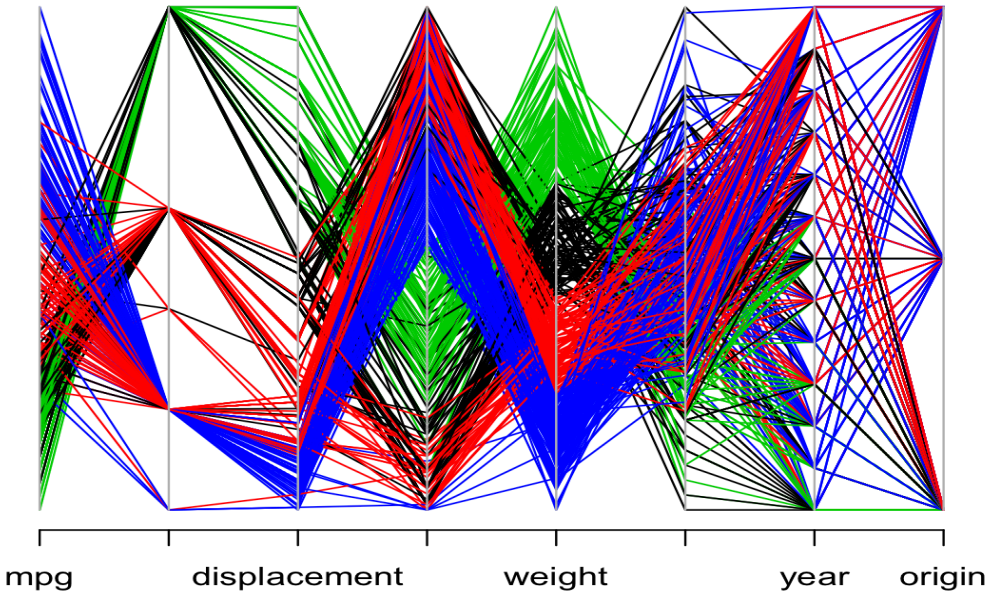




K-means clustering car spec data (k is 4)

```
> fit.km <- kmeans(car, 4)
>
> fit.km
K-means clustering with 4 clusters of sizes 89, 99, 69, 140

Cluster means:
       mpg cylinders displacement horsepower   weight acceleration     year   origin
1 18.35281  6.640449    259.96629   32.08989 3484.483     15.67528 75.58427 1.112360
2 24.56970  4.717172    155.14141   59.81818 2752.535     15.71919 77.18182 1.555556
3 13.99130  8.000000    356.53623   32.71014 4366.594     13.04203 73.60870 1.000000
4 30.74714  3.978571     98.11071   67.26429 2109.136     16.60286 76.59286 2.164286

Clustering vector:
  [1] 1 1 1 1 1 3 3 3 3 1 1 1 1 1 4 2 2 2 4 4 2 4 4 4 2 3 3 3 3 4 4 4 4 2 1 1 1 1 3 3 3 3 3 3 3 2 4 1 1 4 4 4 4 4 4 4 4 4
 [59] 4 4 4 4 3 3 3 3 3 1 3 3 3 3 4 1 3 3 3 2 2 2 4 4 4 2 4 4 3 1 3 3 1 3 3 3 3 3 3 3 1 1 1 2 2 2 4 3 3 3 3 2 4 4 4 4 4 2 4 3
[117] 3 4 4 2 2 1 2 2 1 2 2 2 1 4 2 4 2 1 1 1 3 3 3 3 3 4 4 4 4 4 4 4 4 2 4 4 1 1 1 1 3 3 3 3 1 1 1 1 2 1 1 4 2 2 2 2 4 2
[175] 2 4 1 2 2 2 2 4 2 4 2 4 4 3 3 3 3 1 1 2 2 4 4 4 4 1 1 1 1 4 4 4 4 2 1 3 1 2 1 3 3 1 1 4 4 4 4 4 4 1 3 3 3 3 1 1 1 1 3 3 3
[233] 3 4 2 4 2 4 4 4 4 4 2 2 2 4 4 4 4 4 1 1 1 1 1 2 2 1 1 1 2 1 1 1 1 1 3 4 2 4 4 2 2 2 2 4 2 1 2 1 4 4 1 2 2 2 1 1 1 1 3 1 3
[291] 3 1 3 4 4 4 4 2 1 1 1 1 4 4 4 4 4 2 2 2 2 4 4 4 4 2 2 2 2 1 4 2 2 2 4 4 2 4 4 4 4 2 1 4 4 4 4 2 4 2 2 4 2 2 2 2 2 4 4 4 4 4 4
[349] 4 4 4 4 4 4 4 4 4 4 2 2 1 1 2 2 1 1 2 1 2 2 4 2 2 2 2 2 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 2 2 2 2 2 4 2 2 4 4 2 2

Within cluster sum of squares by cluster:
[1] 5246583 3370035 5732953 4713795
 (between_SS / total_SS =  93.4 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss" "betweenss"    "size"
[8] "iter"         "ifault"
> library(fpc)
```





Wine data K-means clustering

Wine data

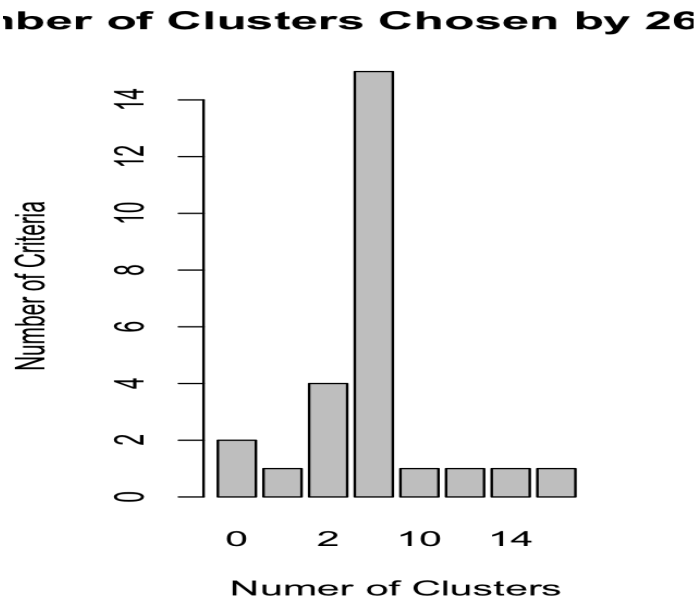
```
> data(wine, package="rattle")
> head(wine)
  Type Alcohol Malic  Ash Alcalinity Magnesium Phenols Flavanoids Nonflavanoids Proanthocyanins Color  Hue Dilution
1    1   14.23  1.71 2.43       15.6       127    2.80       3.06          0.28            2.29  5.64 1.04     3.92
2    1   13.20  1.78 2.14       11.2       100    2.65       2.76          0.26            1.28  4.38 1.05     3.40
3    1   13.16  2.36 2.67       18.6       101    2.80       3.24          0.30            2.81  5.68 1.03     3.17
4    1   14.37  1.95 2.50       16.8       113    3.85       3.49          0.24            2.18  7.80 0.86     3.45
5    1   13.24  2.59 2.87       21.0       118    2.80       2.69          0.39            1.82  4.32 1.04     2.93
6    1   14.20  1.76 2.45       15.2       112    3.27       3.39          0.34            1.97  6.75 1.05     2.85
  Proline
1    1065
2    1050
3    1185
4    1480
5     735
6    1450
```

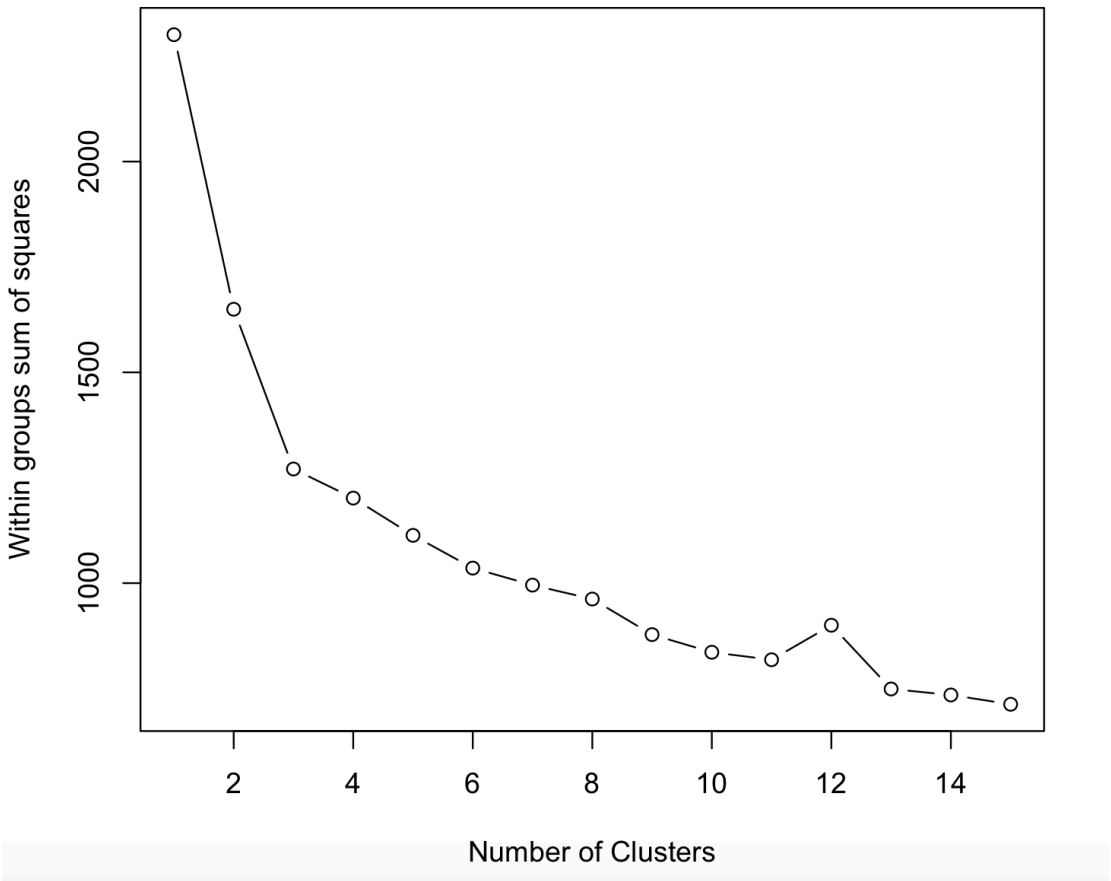
Summary of normalization wine data

```
> data.train <- scale(wine[-1])
> summary(data.train)
    Alcohol             Malic              Ash             Alcalinity          Magnesium          Phenols       
 Min.   :-2.42739   Min.   :-1.4290   Min.   :-3.66881   Min.   :-2.663505   Min.   :-2.0824   Min.   :-2.10132  
 1st Qu.:-0.78603   1st Qu.:-0.6569   1st Qu.:-0.57051   1st Qu.:-0.687199   1st Qu.:-0.8221   1st Qu.:-0.88298  
 Median : 0.06083   Median :-0.4219   Median :-0.02375   Median : 0.001514   Median :-0.1219   Median : 0.09569  
 Mean   : 0.00000   Mean   : 0.0000   Mean   : 0.00000   Mean   : 0.000000   Mean   : 0.0000   Mean   : 0.00000  
 3rd Qu.: 0.83378   3rd Qu.: 0.6679   3rd Qu.: 0.69615   3rd Qu.: 0.600395   3rd Qu.: 0.5082   3rd Qu.: 0.80672  
 Max.   : 2.25341   Max.   : 3.1004   Max.   : 3.14745   Max.   : 3.145637   Max.   : 4.3591   Max.   : 2.53237  
   Flavanoids       Nonflavanoids     Proanthocyanins       Color              Hue              Dilution       
 Min.   :-1.6912   Min.   :-1.8630   Min.   :-2.06321   Min.   :-1.6297   Min.   :-2.08884   Min.   :-1.8897  
 1st Qu.:-0.8252   1st Qu.:-0.7381   1st Qu.:-0.59560   1st Qu.:-0.7929   1st Qu.:-0.76540   1st Qu.:-0.9496  
 Median : 0.1059   Median :-0.1756   Median :-0.06272   Median :-0.1588   Median : 0.03303   Median : 0.2371  
 Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.00000   Mean   : 0.0000   Mean   : 0.00000   Mean   : 0.0000  
 3rd Qu.: 0.8467   3rd Qu.: 0.6078   3rd Qu.: 0.62741   3rd Qu.: 0.4926   3rd Qu.: 0.71116   3rd Qu.: 0.7864  
 Max.   : 3.0542   Max.   : 2.3956   Max.   : 3.47527   Max.   : 3.4258   Max.   : 3.29241   Max.   : 1.9554  
    Proline       
 Min.   :-1.4890  
 1st Qu.:-0.7824  
 Median :-0.2331  
 Mean   : 0.0000  
 3rd Qu.: 0.7561  
 Max.   : 2.9631  
```

The best number of clusters is 3

```
> nc <- NbClust(data.train,
+               min.nc=2, max.nc=15,
+               method="kmeans")
*** : The Hubert index is a graphical method of determining the number of clusters.
                In the plot of Hubert index, we seek a significant knee that corresponds to a
                significant increase of the value of the measure i.e the significant peak in Hubert
                index second differences plot.

*** : The D index is a graphical method of determining the number of clusters.
                In the plot of D index, we seek a significant knee (the significant peak in Dindex
                second differences plot) that corresponds to a significant increase of the value of
                the measure.

*******************************************************************
* Among all indices:
* 4 proposed 2 as the best number of clusters
* 15 proposed 3 as the best number of clusters
* 1 proposed 10 as the best number of clusters
* 1 proposed 12 as the best number of clusters
* 1 proposed 14 as the best number of clusters
* 1 proposed 15 as the best number of clusters

                   ***** Conclusion *****

* According to the majority rule, the best number of clusters is  3


*******************************************************************
> barplot(table(nc$Best.n[1,]),
+         xlab="Numer of Clusters",
+         ylab="Number of Criteria",
+         main="Number of Clusters Chosen by 26 Criteria")
```


ıber of Clusters Chosen by 26
Number of Criteria
0
2
4
6
8
10
12
14
0
2
10
14
Numer of Clusters

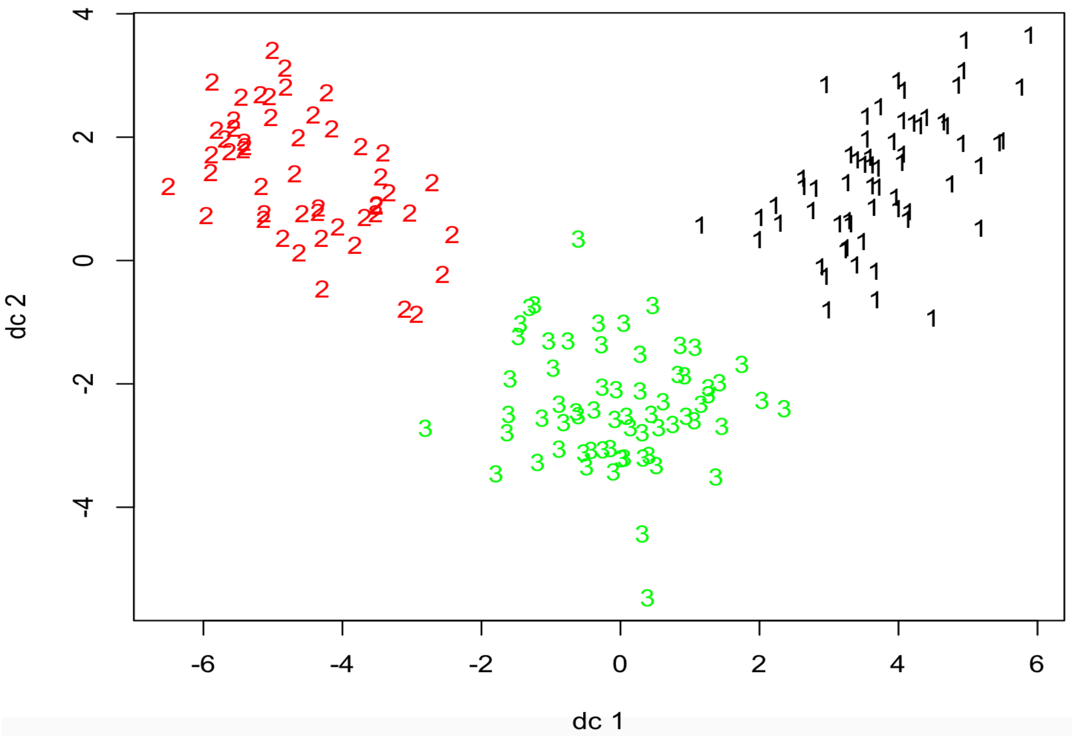

Within groups sum of squares
1000
1500
2000
2
4
6
8
10
12
14
Number of Clusters

The result of k-means clustering

```
> fit.km <- kmeans(data.train, 3)
> fit.km
K-means clustering with 3 clusters of sizes 62, 51, 65

Cluster means:
     Alcohol      Malic        Ash Alcalinity   Magnesium     Phenols Flavanoids Nonflavanoids Proanthocyanins
1  0.8328826 -0.3029551  0.3636801 -0.6084749  0.57596208  0.88274724 0.97506900   -0.56050853      0.57865427
2  0.1644436  0.8690954  0.1863726  0.5228924 -0.07526047 -0.97657548 -1.21182921    0.72402116     -0.77751312
3 -0.9234669 -0.3929331 -0.4931257  0.1701220 -0.49032869 -0.07576891 0.02075402   -0.03343924      0.05810161
       Color        Hue   Dilution    Proline
1  0.1705823  0.4726504  0.7770551  1.1220202
2  0.9388902 -1.1615122 -1.2887761 -0.4059428
3 -0.8993770  0.4605046  0.2700025 -0.7517257

Clustering vector:
  [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 [59] 1 3 3 2 3 3 3 3 3 3 3 3 3 3 3 1 3 3 3 3 3 3 3 3 3 2 3 3 3 3 3 3 3 3 3 3 3 3 1 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
[117] 3 3 2 3 3 1 3 3 3 3 3 3 3 3 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
[175] 2 2 2 2

Within cluster sum of squares by cluster:
[1] 385.6983 326.3537 558.6971
 (between_SS / total_SS =  44.8 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss" "betweenss"    "size"
[8] "iter"         "ifault"
> |
```

```
> library(fpc)
> plotcluster(data.train, fit.km$cluster)
```

Express plot cluster



```
> library(MASS)
> parcoord(data.train, fit.km$cluster)
```

The result of parallel coordination plot