Mtcars analysis

Wojciech Kolasa

2025-07-23

The data comes from R's built-in dataset called `mtcars`. Each observation consists of 11 features describing various parameters of selected car models. The aim of this project is to perform both supervised and unsupervised classification and to describe each variable.

Analysis Plan:

1. Presentation of summary statistics including histograms
2. Univariate analysis of outliers and multimodality
3. Correlation analysis and dimensionality reduction
4. Clustering analysis
5. Supervised classification using random forests

```
cat('Basic statistical measures','\n')
```

```
## Basic statistical measures
```

```
cat('\n')
```

```
apply(mtcars[,], 2 ,summary)
```

```
##              mpg    cyl     disp       hp     drat      wt     qsec     vs
## Min.    10.40000 4.0000  71.1000  52.0000 2.760000 1.51300 14.50000 0.0000
## 1st Qu. 15.42500 4.0000 120.8250  96.5000 3.080000 2.58125 16.89250 0.0000
## Median  19.20000 6.0000 196.3000 123.0000 3.695000 3.32500 17.71000 0.0000
## Mean    20.09062 6.1875 230.7219 146.6875 3.596563 3.21725 17.84875 0.4375
## 3rd Qu. 22.80000 8.0000 326.0000 180.0000 3.920000 3.61000 18.90000 1.0000
## Max.    33.90000 8.0000 472.0000 335.0000 4.930000 5.42400 22.90000 1.0000
##              am   gear   carb
## Min.    0.00000 3.0000 1.0000
## 1st Qu. 0.00000 3.0000 2.0000
## Median  0.00000 4.0000 2.0000
## Mean    0.40625 3.6875 2.8125
## 3rd Qu. 1.00000 4.0000 4.0000
## Max.    1.00000 5.0000 8.0000
```

```
cat('\n')
```

```
cat('Standard deviation','\n')
```

```
## Standard deviation
```

```
cat('\n')
```

```
apply(mtcars[,] , 2 , sd)
```

```
##         mpg         cyl        disp          hp        drat          wt
##   6.0269481   1.7859216 123.9386938  68.5628685   0.5346787   0.9784574
##        qsec          vs          am        gear        carb
##   1.7869432   0.5040161   0.4989909   0.7378041   1.6152000
```

```
cat('\n')
```

```
cat('Skewness','\n')
```

```
## Skewness
```

```
cat('\n')
```

```
apply(mtcars[,] , 2 ,skewness)
```

```
##        mpg        cyl       disp         hp       drat         wt       qsec
##  0.6404399 -0.1831287  0.4002724  0.7614356  0.2788734  0.4437855  0.3870456
##         vs         am       gear       carb
##  0.2519763  0.3817709  0.5546495  1.1021304
```

```
cat('\n')
```

```
cat('Kurtosis','\n')
```

```
## Kurtosis
```

```
cat('\n')
```

```
apply(mtcars[,] , 2 ,kurtosis)
```

```
##      mpg      cyl     disp       hp     drat       wt     qsec       vs
## 2.799467 1.319032 1.910317 3.052233 2.435116 3.172471 3.553753 1.063492
##       am     gear     carb
## 1.145749 2.056790 4.536121
```

Summary and Conclusions:

The variables `disp` and `hp` have the highest minimum values and also the highest mean and median. They also have high standard deviation values, indicating low concentration around the mean. The rest of the variables show low standard deviation, suggesting high concentration around the mean. All variables have kurtosis greater than 0, indicating leptokurtic distributions. The variable `carb` has the highest kurtosis and the strongest presence of extreme values.

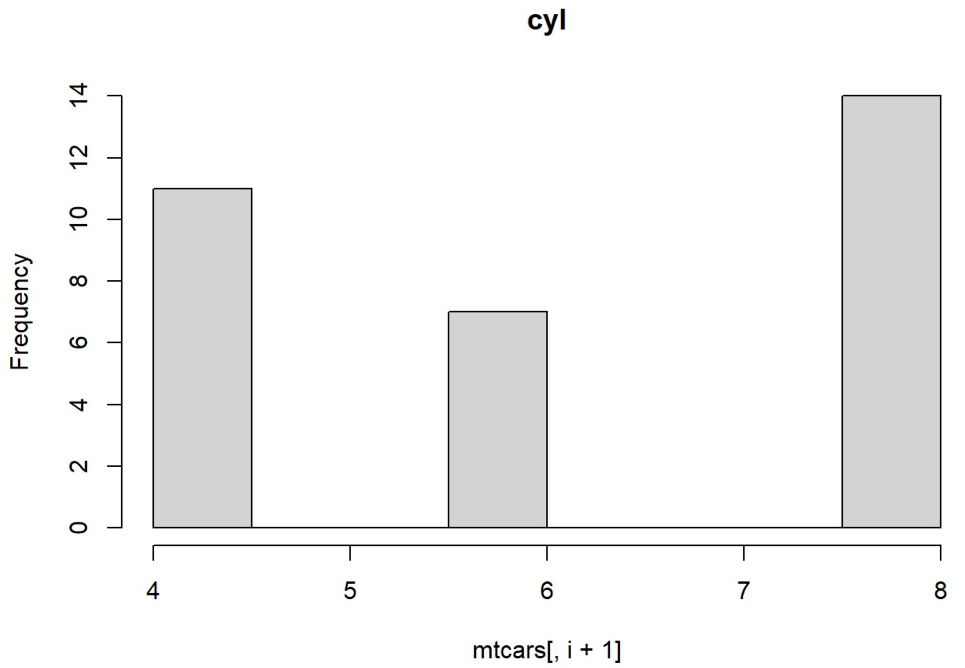
```
for(i in 1:11) print(grubbs.test(mtcars[,i]))
```

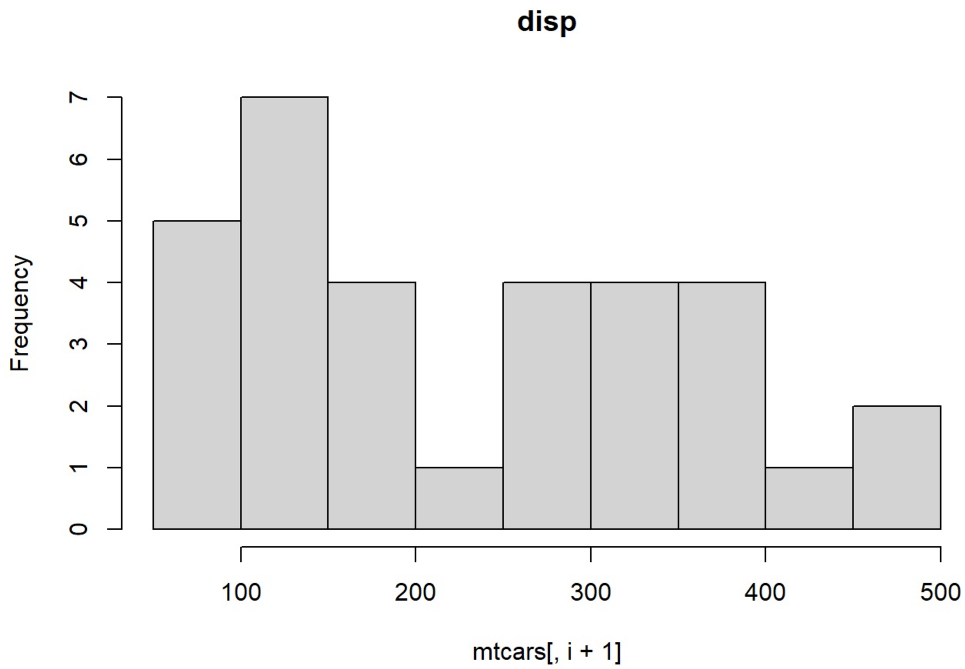
```
##
##  Grubbs test for one outlier
##
## data:  mtcars[, i]
## G = 2.29127, U = 0.82518, p-value = 0.276
## alternative hypothesis: highest value 33.9 is an outlier
##
##
##  Grubbs test for one outlier
##
## data:  mtcars[, i]
## G = 1.22486, U = 0.95004, p-value = 1
## alternative hypothesis: lowest value 4 is an outlier
##
##
##  Grubbs test for one outlier
##
## data:  mtcars[, i]
## G = 1.9468, U = 0.8738, p-value = 0.7363
## alternative hypothesis: highest value 472 is an outlier
##
##
##  Grubbs test for one outlier
##
## data:  mtcars[, i]
## G = 2.74657, U = 0.74881, p-value = 0.05564
## alternative hypothesis: highest value 335 is an outlier
##
##
##  Grubbs test for one outlier
##
## data:  mtcars[, i]
## G = 2.4939, U = 0.7929, p-value = 0.1419
## alternative hypothesis: highest value 4.93 is an outlier
##
##
##  Grubbs test for one outlier
##
## data:  mtcars[, i]
## G = 2.25534, U = 0.83063, p-value = 0.3083
## alternative hypothesis: highest value 5.424 is an outlier
##
##
##  Grubbs test for one outlier
##
## data:  mtcars[, i]
## G = 2.82675, U = 0.73393, p-value = 0.04021
## alternative hypothesis: highest value 22.9 is an outlier
##
##
##  Grubbs test for one outlier
##
## data:  mtcars[, i]
## G = 1.11604, U = 0.95853, p-value = 1
## alternative hypothesis: highest value 1 is an outlier
##
##
##  Grubbs test for one outlier
##
## data:  mtcars[, i]
## G = 1.18990, U = 0.95285, p-value = 1
## alternative hypothesis: highest value 1 is an outlier
##
##
##  Grubbs test for one outlier
##
## data:  mtcars[, i]
## G = 1.77893, U = 0.89462, p-value = 1
## alternative hypothesis: highest value 5 is an outlier
##
##
##  Grubbs test for one outlier
##
## data:  mtcars[, i]
## G = 3.21168, U = 0.65653, p-value = 0.006787
## alternative hypothesis: highest value 8 is an outlier
```

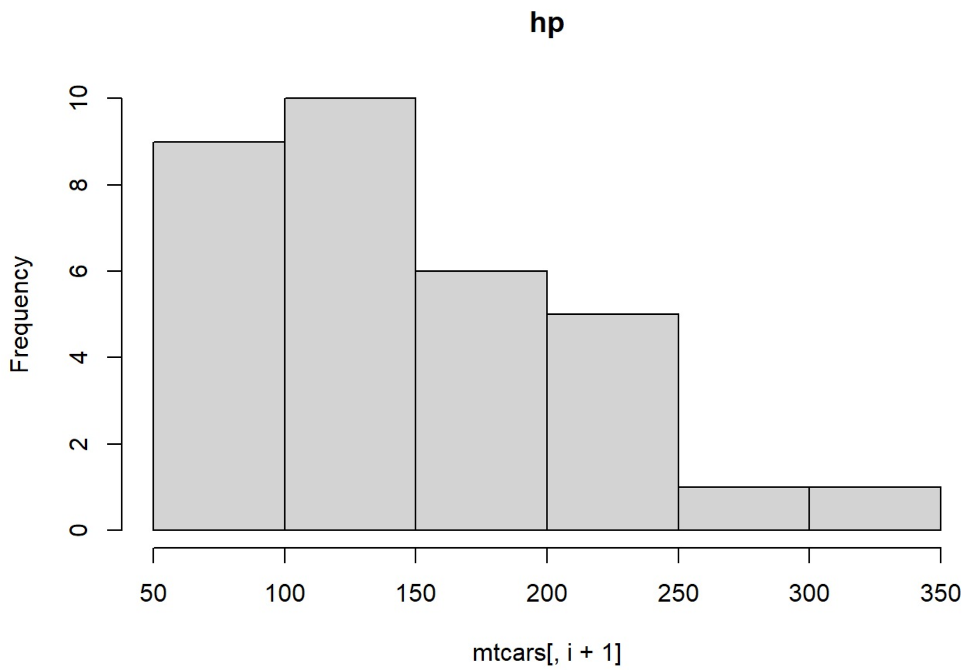
Conclusions:

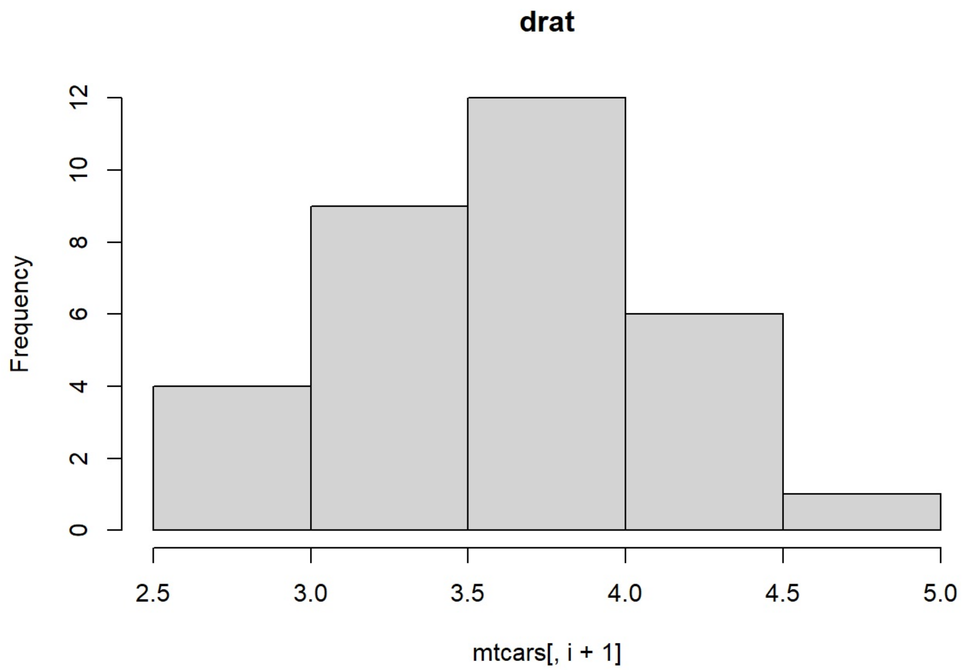
For the variables `qsec` and `carb` , the Grubbs test p-value is less than 0.05, indicating the presence of outliers that must be considered during clustering analysis.

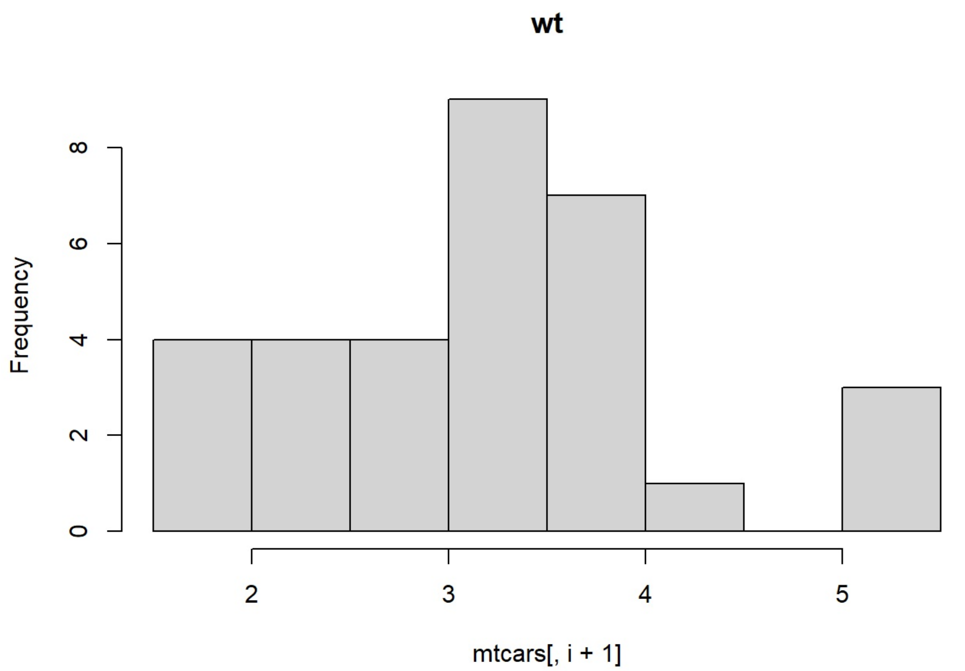
```
for(i in 1:10) hist(mtcars[,i+1], main=colnames(mtcars)[i+1])
```

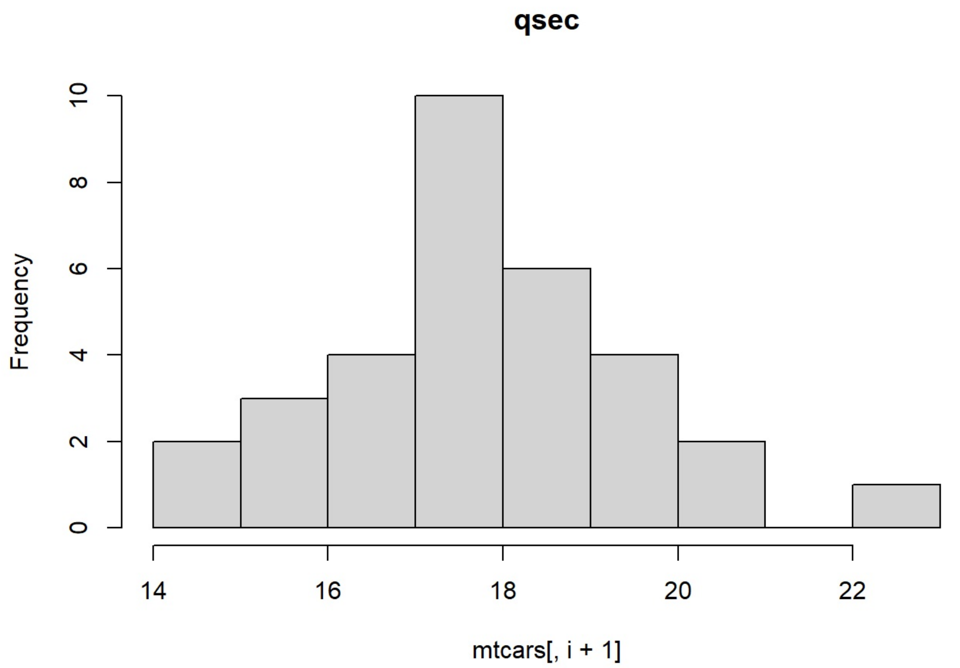

cyl
Frequency
0
2
4
6
8
10
12
14
4
5
6
7
8
mtcars[, i + 1]

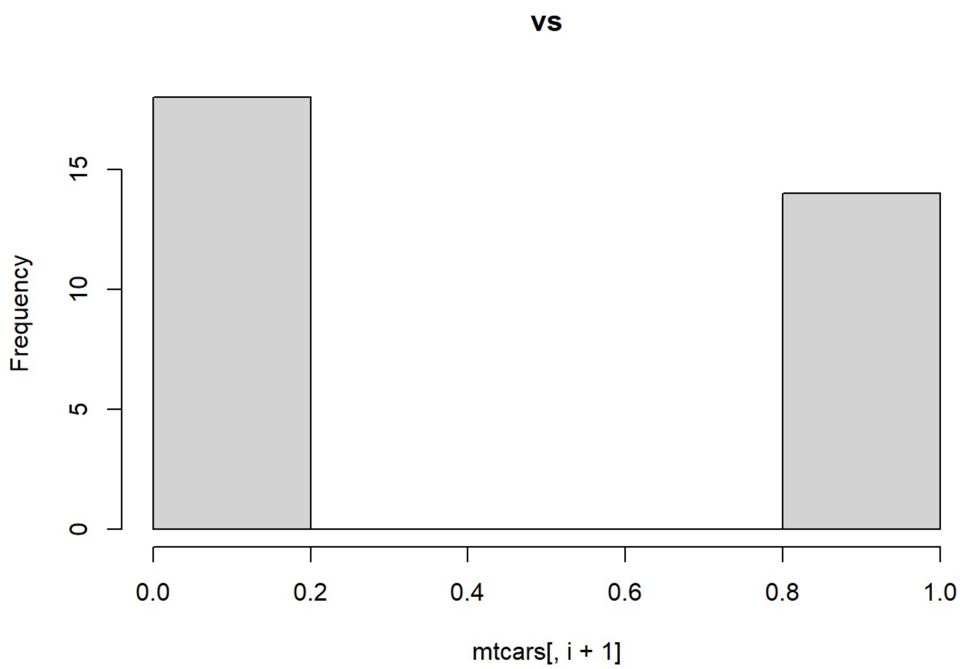

disp
Frequency
0
1
2
3
4
5
6
7
100
200
300
400
500
mtcars[, i + 1]

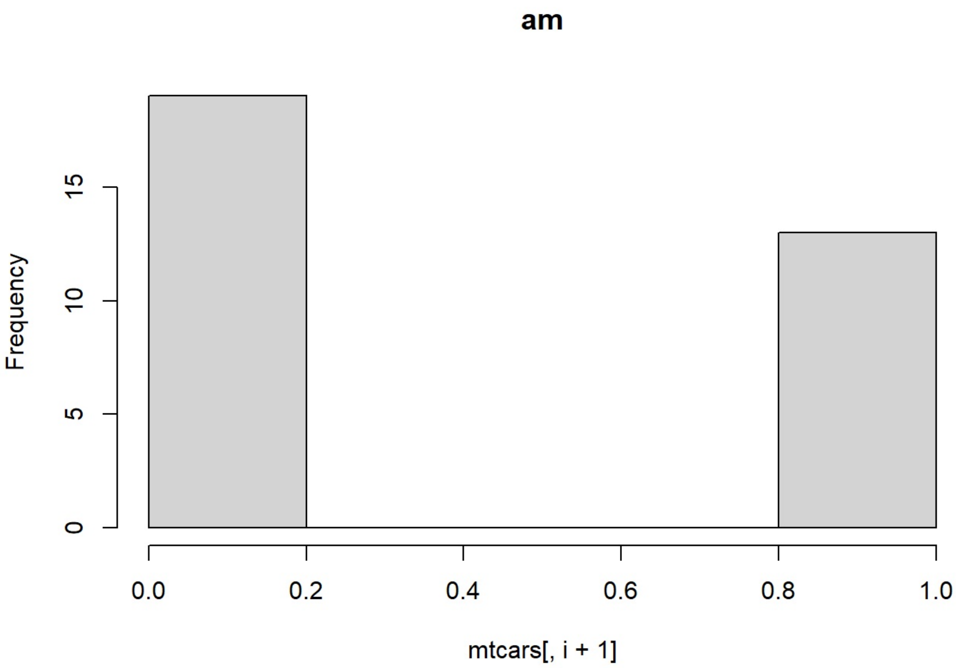

hp
Frequency
0
2
4
6
8
10
50
100
150
200
250
300
350
mtcars[, i + 1]

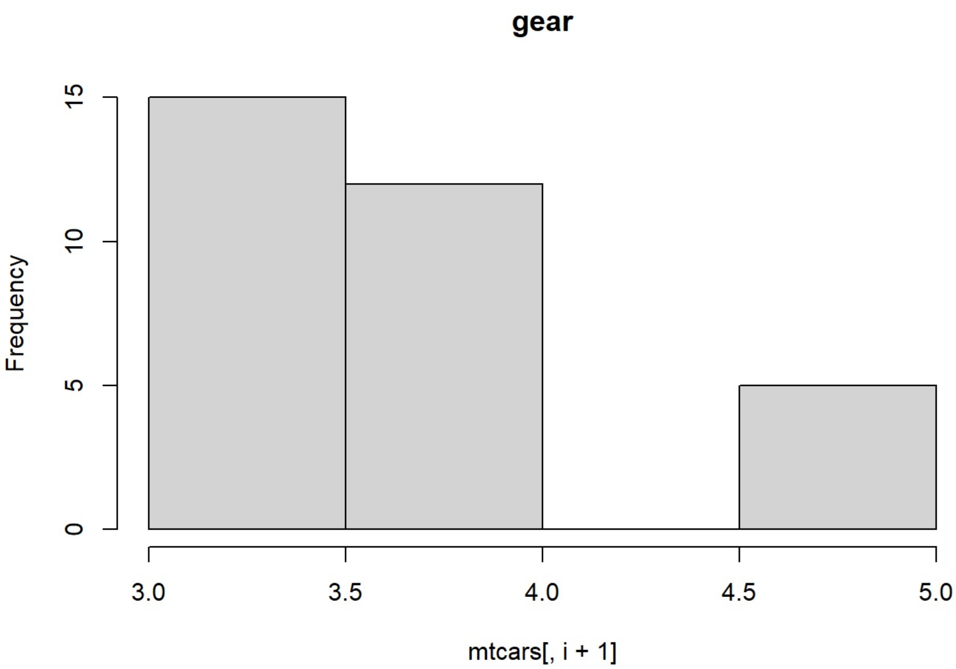

drat
Frequency
0
2
4
6
8
10
12
2.5
3.0
3.5
4.0
4.5
5.0
mtcars[, i + 1]

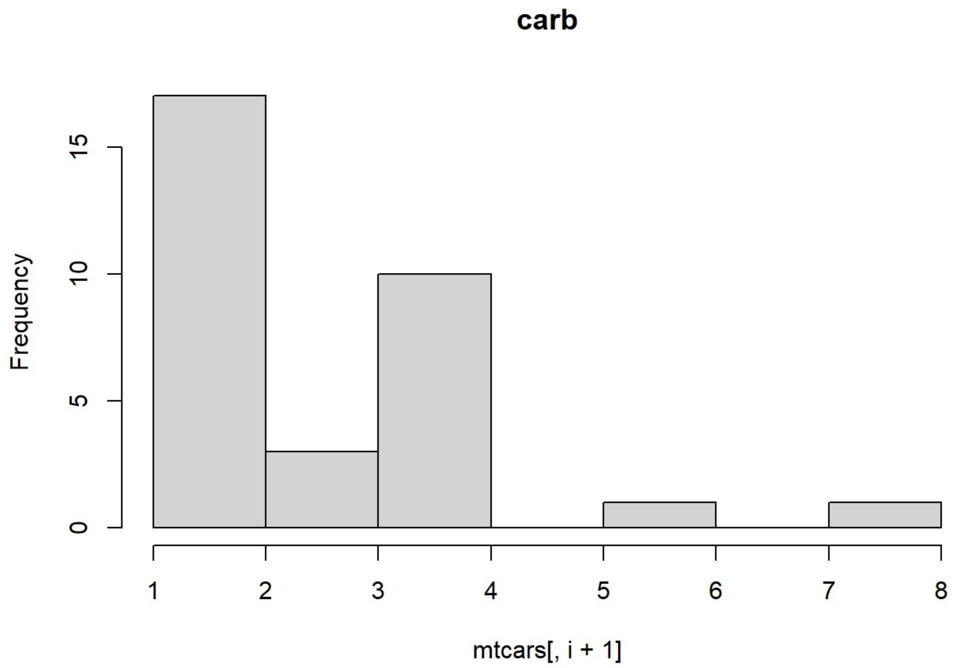

wt
Frequency
0
2
4
6
8
2
3
4
5
mtcars[, i + 1]


qsec
Frequency
0
2
4
6
8
10
14
16
18
20
22
mtcars[, i + 1]


vs
Frequency
0
5
10
15
0.0
0.2
0.4
0.6
0.8
1.0
mtcars[, i + 1]


am
Frequency
0
5
10
15
0.0
0.2
0.4
0.6
0.8
1.0
mtcars[, i + 1]


gear
Frequency
0
5
10
15
3.0
3.5
4.0
4.5
5.0
mtcars[, i + 1]



Conclusions:

Histogram analysis suggests that normality should be tested for `hp` and `drat`, and multimodality for `disp` and `wt`.

Normality Tests:

```
shapiro.test(mtcars$hp)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  mtcars$hp
## W = 0.93342, p-value = 0.04881
```

```
shapiro.test(mtcars$drat)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  mtcars$drat
## W = 0.94588, p-value = 0.1101
```

Conclusions:

The hypothesis of normality for variables `hp` and `drat` is rejected.

Multimodality Analysis:

```
silverman.test(mtcars$disp,k=1)
```

```
## Silvermantest: Testing the null hypothesis that the number of modes is <=  1
## The resulting p-value is  0.1781782
```

```
silverman.test(mtcars$wt,k=1)
```

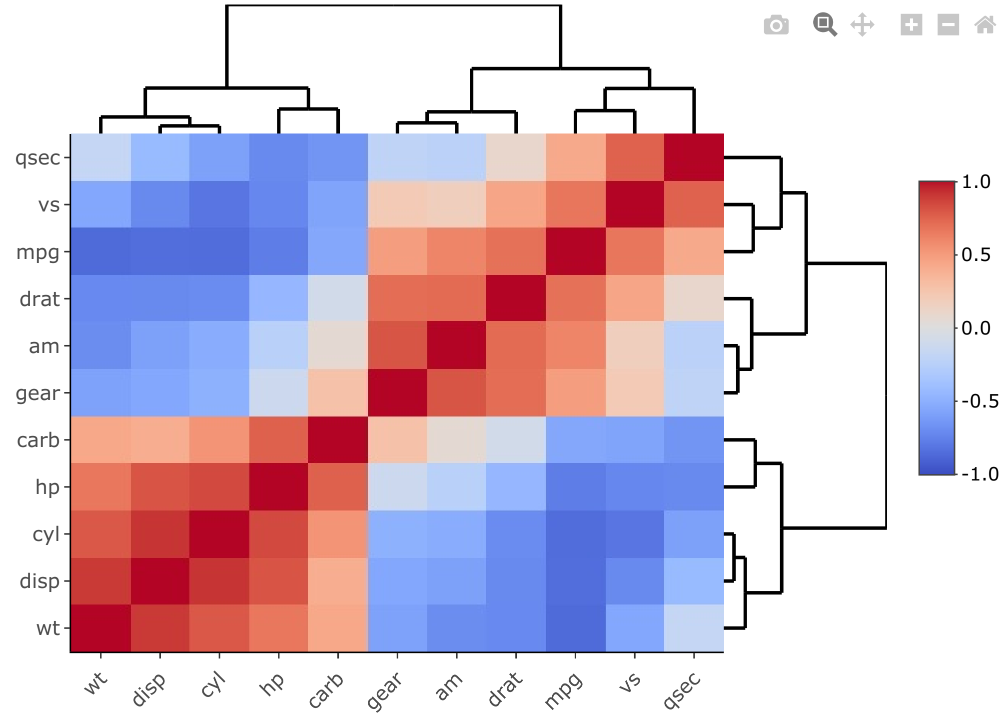
```
## Silvermantest: Testing the null hypothesis that the number of modes is <=  1
## The resulting p-value is  0.08608609
```

Conclusions:

The distributions of `disp` and `wt` are not multimodal, as the significance level exceeds 5%.

Correlation Analysis using Pearson's coefficient

```
heatmaply_cor(cor(mtcars[,],method='pearson'))
```

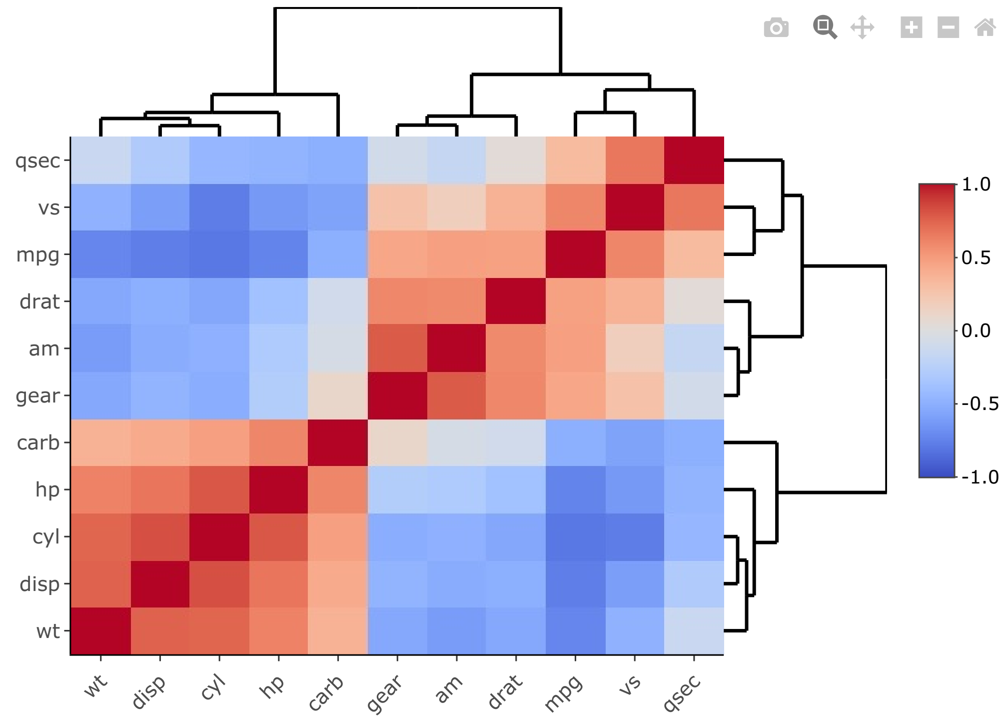


Conclusions:

Positive correlations are observed within the groups: (`wt` , `disp` , `cyl` , `hp`), (`gear` , `am` , `drat`), and (`vs` , `qsec`). A weaker positive correlation is present between `vs` and `carb` .

Correlation Analysis using Kendall's coefficient

```
heatmaply_cor(cor(mtcars[,],method='kendall'))
```



Conclusions:

The variables are less correlated in the same areas as with Pearson's method.

PCA

```
prcomp(mtcars[,-1])->pca.mtcars
summary(pca.mtcars)
```

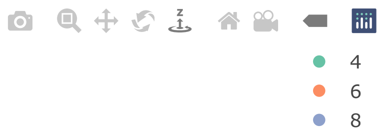
```
## Importance of components:
##                             PC1      PC2     PC3     PC4     PC5     PC6    PC7
## Standard deviation     136.4339 38.14648 1.31542 0.96553 0.76789 0.32620 0.2892
## Proportion of Variance   0.9273  0.07249 0.00009 0.00005 0.00003 0.00001 0.0000
## Cumulative Proportion    0.9273  0.99982 0.99991 0.99995 0.99998 0.99999 1.0000
##                           PC8   PC9   PC10
## Standard deviation     0.2508 0.222 0.1999
## Proportion of Variance 0.0000 0.000 0.0000
## Cumulative Proportion  1.0000 1.000 1.0000
```

```
pca.mtcars$rotation[,1]
```

```
##          cyl         disp           hp         drat           wt         qsec
##  0.012042615  0.900235270  0.435074057 -0.002661394  0.006242550 -0.006676533
##           vs           am         gear         carb
## -0.002731293 -0.001963245 -0.002606103  0.005767541
```

```
df.pca=data.frame(pc1=pca.mtcars$x[,1],pc2=pca.mtcars$x[,2],pc3=pca.mtcars$x[,3],kl=as.factor(mtcars$cyl))
plot_ly(df.pca,x=~pc1,y=~pc2,z=~pc3,color = ~kl,type='scatter3d')
```

```
## No scatter3d mode specifed:
##   Setting the mode to markers
##   Read more about this attribute -> https://plotly.com/r/reference/#scatter-mode
```



WebGL is not supported by your browser - visit https://get.webgl.org for more info

Conclusions:

The first principal component explains 92.73% of the total variance. The variable `am` has the strongest influence on PC1. Clear cluster separation is visible along the PC1 axis.

Unsupervised classification using k-means clustering

```
kmeans(mtcars[,-1], centers=3)->km.mtcars.3
table(km.mtcars.3$cluster,mtcars$cyl)
```

```
##     
##       4  6  8
##   1 11  5  0
##   2  0  2  5
##   3  0  0  9
```

```
df.km<-data.frame(x=pca.mtcars$x[,1],y=pca.mtcars$x[,2],z=pca.mtcars$x[,3],type=as.factor(km.mtcars.3$cluster))
plot_ly(df.km,x=~x,y=~y, z=~z,color=~type,type ='scatter3d')
```

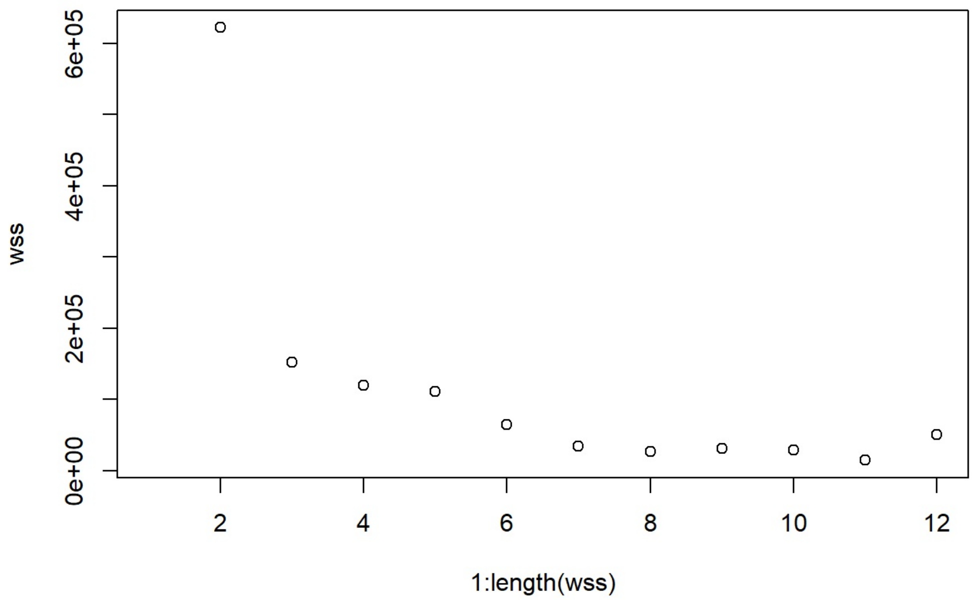
```
## No scatter3d mode specifed:
##   Setting the mode to markers
##   Read more about this attribute -> https://plotly.com/r/reference/#scatter-mode
```

Conclusions:

A clear division into 3 clusters is visible, with one of the clusters being the most dispersed.

Determining the optimal number of clusters

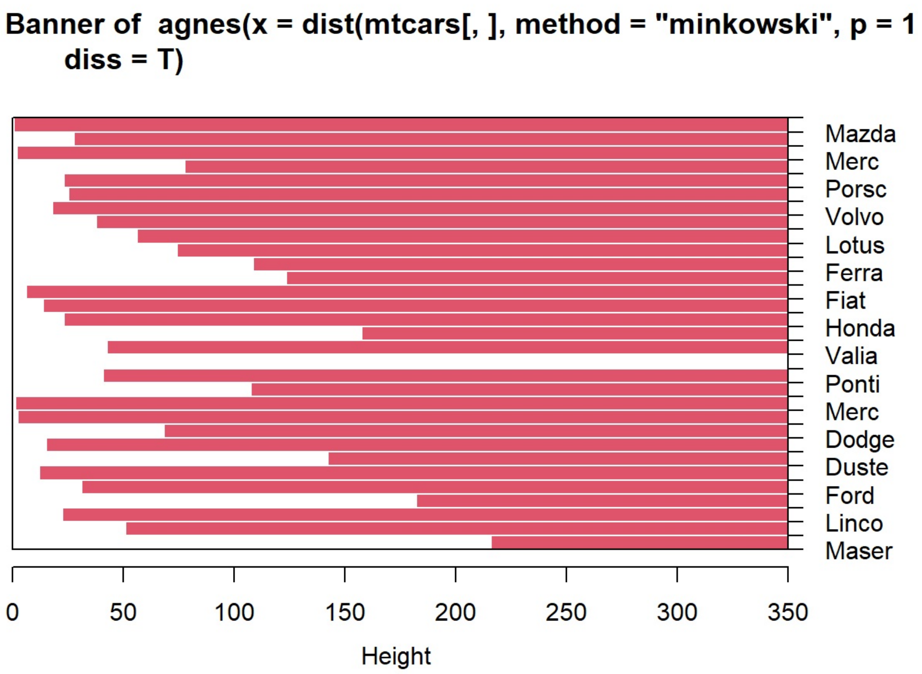
```
wss<-NA
for(i in 1:11) wss<-c(wss,kmeans(mtcars[,-1], centers=i)$tot.withinss)
plot(1:length(wss),wss)
```



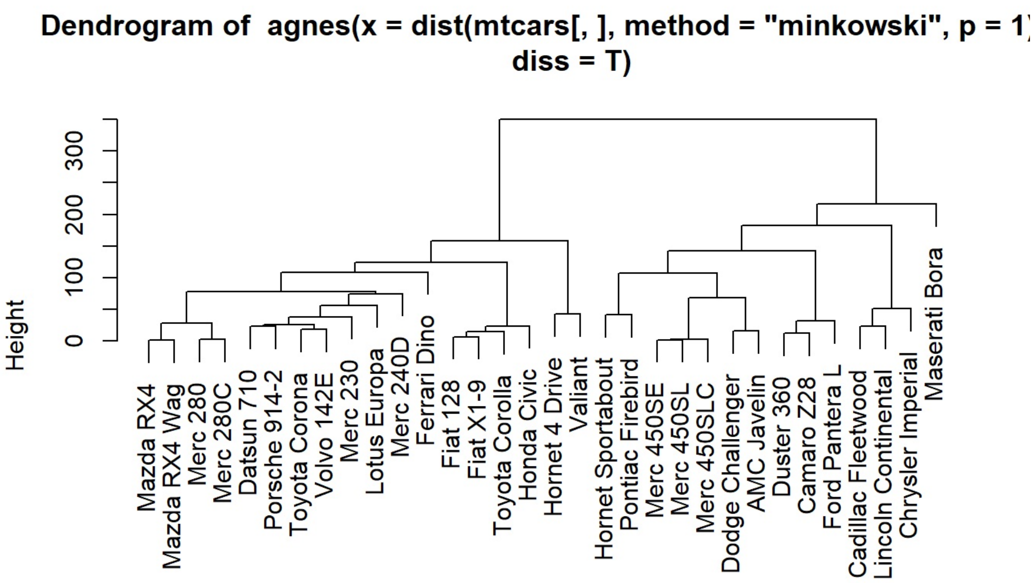
Conclusion:

Using the elbow method, it can be concluded that 3 or 4 clusters are most optimal.

```
plot(agnes(dist(mtcars[,],method='minkowski',p=1),diss=T))
```

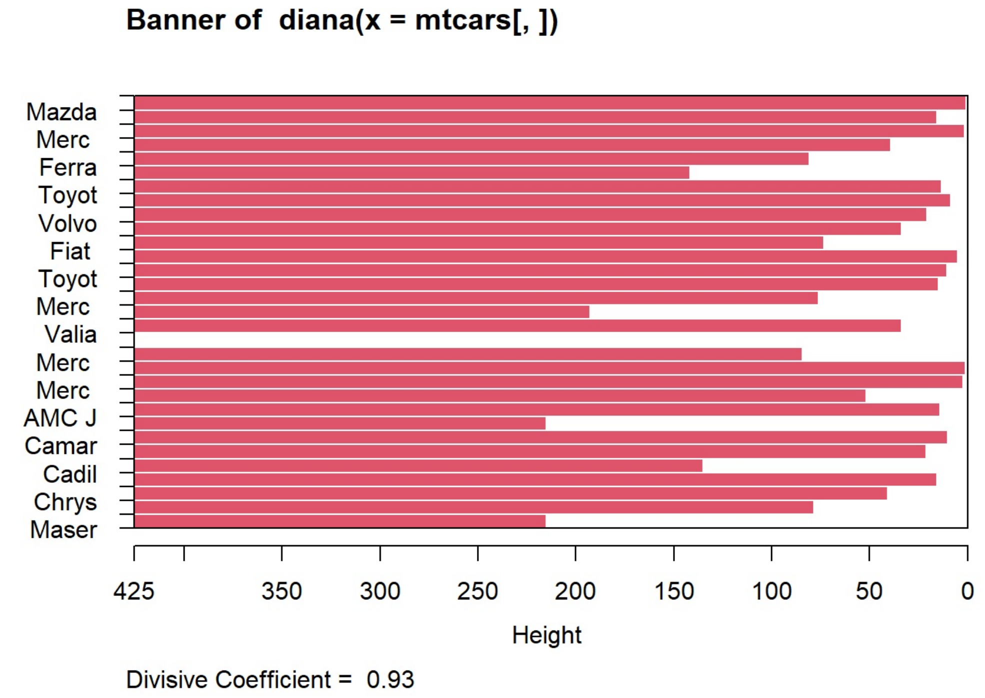


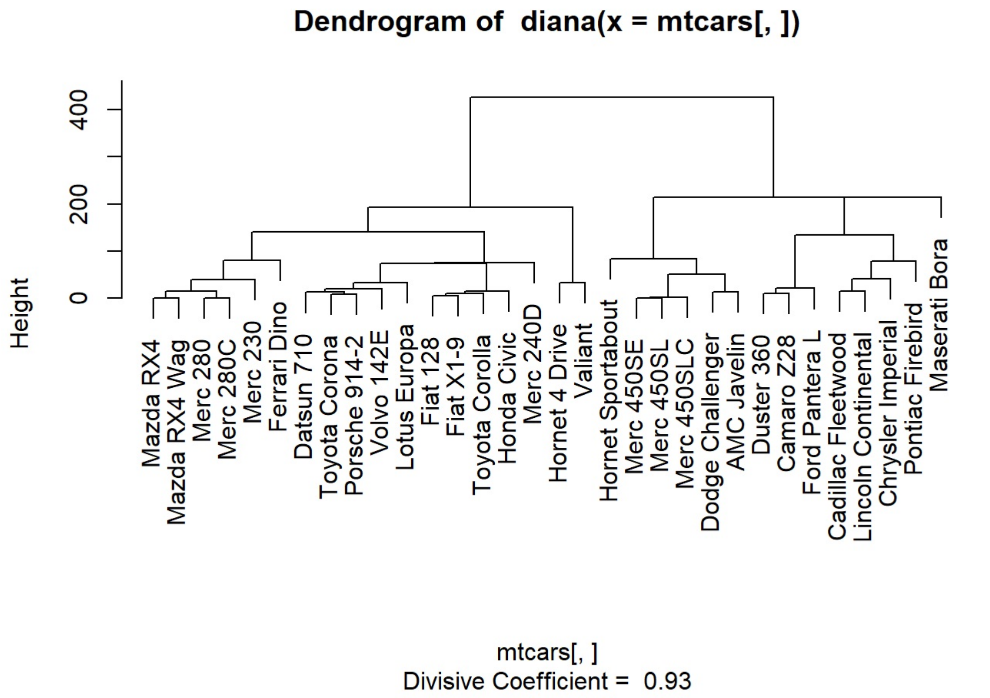
Agglomerative Coefficient = 0.91



dist(mtcars[,], method = "minkowski", p = 1)
Agglomerative Coefficient = 0.91

```
plot(diana(mtcars[,]))
```





Conclusions:

Both methods yield similar results. The coefficients differ by only 0.02.

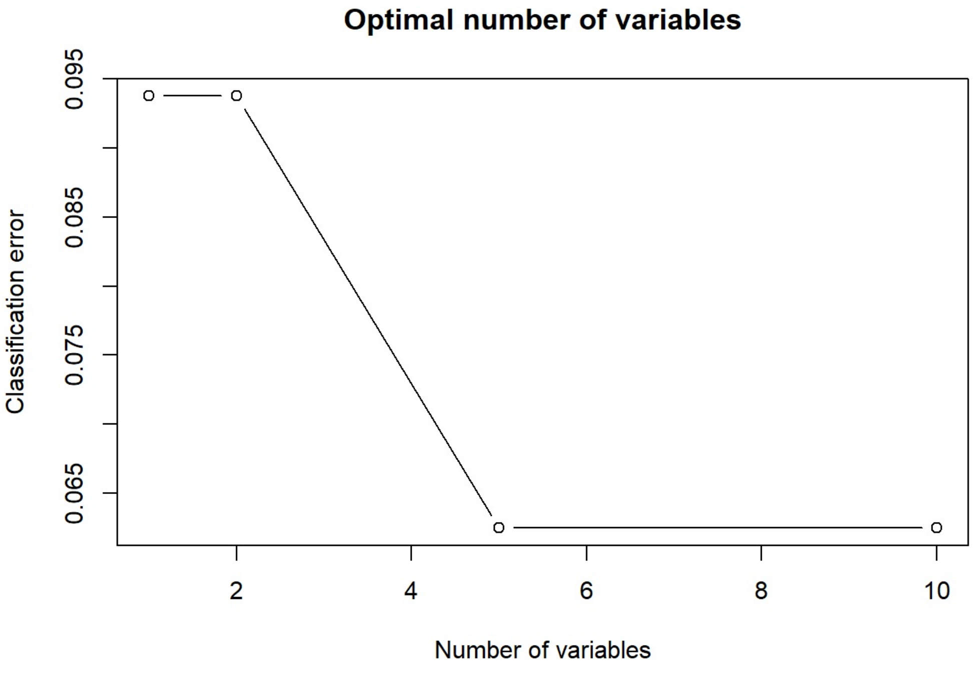
Supervised Classification

We use the random forest method.

```
rfcv(mtcars[,-2],as.factor(mtcars$cyl))->rf.ic
rf.ic$error.cv
```

```
##      10       5       2       1
## 0.06250 0.06250 0.09375 0.09375
```

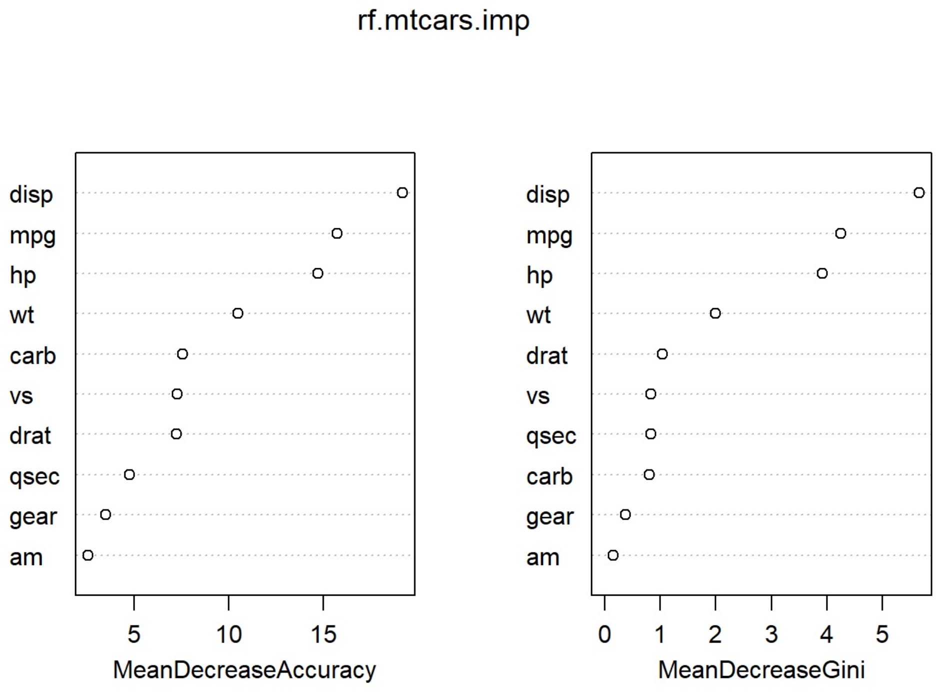
```
plot(rf.ic$n.var, rf.ic$error.cv, type="b",
     xlab="Number of variables", ylab="Classification error",
     main="Optimal number of variables")
```



Conclusion:

The smallest cross-validation error occurs when the number of variables is between 2 and 5.

```
randomForest(mtcars[,-2],as.factor(mtcars$cyl),importance = T)->rf.mtcars.imp
varImpPlot(rf.mtcars.imp)
```



Conclusions:

According to the accuracy drop criterion, the most important variables are `disp` , `mpg` , `hp` , and `wt` . This number of variables is acceptable based on previous results.

```
randomForest(mtcars[,c(3,1,4,6)],as.factor(mtcars$cyl))->rf.mtcars.sel
rf.mtcars.sel$confusion
```

```
##    4 6  8 class.error
## 4 10 1  0  0.09090909
## 6  1 5  1  0.28571429
## 8  0 0 14  0.00000000
```

Conclusion:

Using these variables, the maximum classification error is 28%.

Final Conclusions

The variables `disp` and `hp` have high mean and standard deviation values. The Grubbs test indicated outliers in `qsec` and `carb`, which should be considered in cluster analysis. Histogram analysis helped identify variables for normality and multimodality testing. Normality tests showed that `hp` and `drat` do not follow a normal distribution, while Silverman's test did not confirm multimodality for `disp` and `wt`. PCA analysis showed that the first principal component (PC1) explains 92.73% of the variance, mainly influenced by `am`. K-means clustering and the elbow method indicated 3 or 4 optimal clusters. Random forests are not an optimal classification model in this case, as the maximum classification error is 28%.