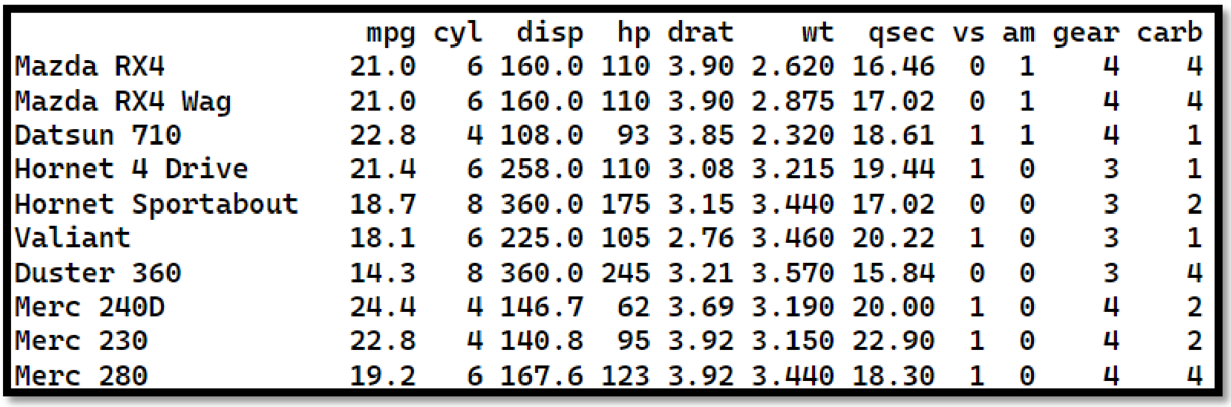
DECISION TREE CONSTRUCTION USING MTCARS DATASET

1. Loading Data set

data=mtcars

print(data)



```
                   mpg cyl  disp  hp drat    wt  qsec vs am gear carb
Mazda RX4         21.0   6 160.0 110 3.90 2.620 16.46  0  1    4    4
Mazda RX4 Wag     21.0   6 160.0 110 3.90 2.875 17.02  0  1    4    4
Datsun 710        22.8   4 108.0  93 3.85 2.320 18.61  1  1    4    1
Hornet 4 Drive    21.4   6 258.0 110 3.08 3.215 19.44  1  0    3    1
Hornet Sportabout 18.7   8 360.0 175 3.15 3.440 17.02  0  0    3    2
Valiant           18.1   6 225.0 105 2.76 3.460 20.22  1  0    3    1
Duster 360        14.3   8 360.0 245 3.21 3.570 15.84  0  0    3    4
Merc 240D         24.4   4 146.7  62 3.69 3.190 20.00  1  0    4    2
Merc 230          22.8   4 140.8  95 3.92 3.150 22.90  1  0    4    2
Merc 280          19.2   6 167.6 123 3.92 3.440 18.30  1  0    4    4
```

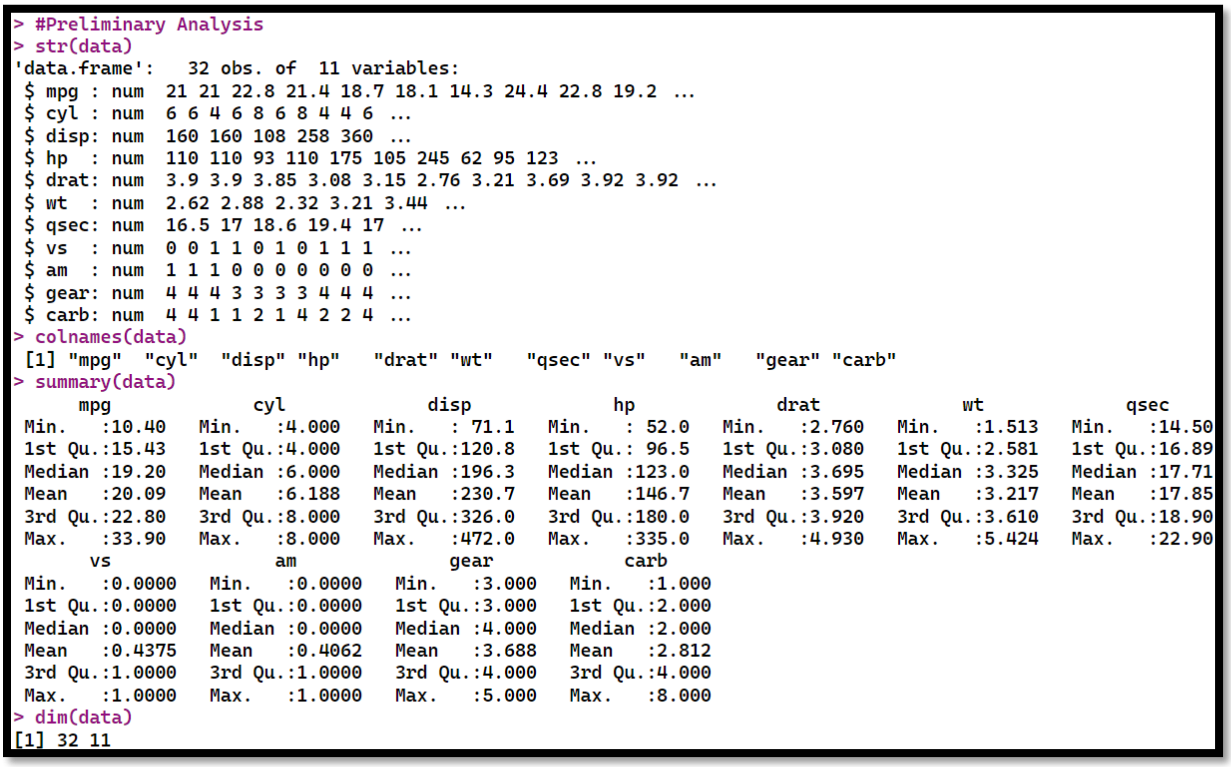
2. Preliminary Analysis

str(data)

colnames(data)

summary(data)

dim(data)

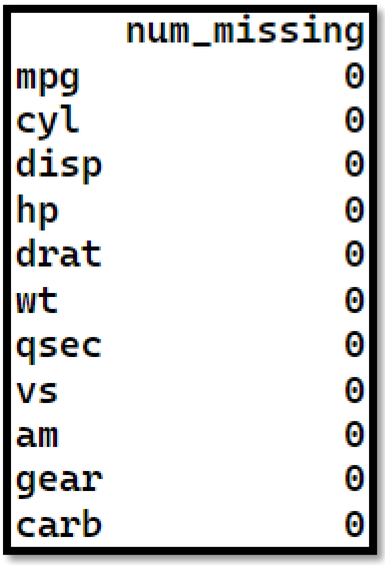


```
> #Preliminary Analysis
> str(data)
'data.frame':	32 obs. of  11 variables:
 $ mpg : num  21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
 $ cyl : num  6 6 4 6 8 6 8 4 4 6 ...
 $ disp: num  160 160 108 258 360 ...
 $ hp  : num  110 110 93 110 175 105 245 62 95 123 ...
 $ drat: num  3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3.92 3.92 ...
 $ wt  : num  2.62 2.88 2.32 3.21 3.44 ...
 $ qsec: num  16.5 17 18.6 19.4 17 ...
 $ vs  : num  0 0 1 1 0 1 0 1 1 1 ...
 $ am  : num  1 1 1 0 0 0 0 0 0 0 ...
 $ gear: num  4 4 4 3 3 3 3 4 4 4 ...
 $ carb: num  4 4 1 1 2 1 4 2 2 4 ...
> colnames(data)
 [1] "mpg"  "cyl"  "disp" "hp"   "drat" "wt"   "qsec" "vs"   "am"   "gear" "carb"
> summary(data)
      mpg             cyl             disp             hp             drat             wt             qsec      
 Min.   :10.40   Min.   :4.000   Min.   : 71.1   Min.   : 52.0   Min.   :2.760   Min.   :1.513   Min.   :14.50
 1st Qu.:15.43   1st Qu.:4.000   1st Qu.:120.8   1st Qu.: 96.5   1st Qu.:3.080   1st Qu.:2.581   1st Qu.:16.89
 Median :19.20   Median :6.000   Median :196.3   Median :123.0   Median :3.695   Median :3.325   Median :17.71
 Mean   :20.09   Mean   :6.188   Mean   :230.7   Mean   :146.7   Mean   :3.597   Mean   :3.217   Mean   :17.85
 3rd Qu.:22.80   3rd Qu.:8.000   3rd Qu.:326.0   3rd Qu.:180.0   3rd Qu.:3.920   3rd Qu.:3.610   3rd Qu.:18.90
 Max.   :33.90   Max.   :8.000   Max.   :472.0   Max.   :335.0   Max.   :4.930   Max.   :5.424   Max.   :22.90
       vs               am              gear            carb      
 Min.   :0.0000   Min.   :0.0000   Min.   :3.000   Min.   :1.000
 1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:3.000   1st Qu.:2.000
 Median :0.0000   Median :0.0000   Median :4.000   Median :2.000
 Mean   :0.4375   Mean   :0.4062   Mean   :3.688   Mean   :2.812
 3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:4.000   3rd Qu.:4.000
 Max.   :1.0000   Max.   :1.0000   Max.   :5.000   Max.   :8.000
> dim(data)
[1] 32 11
```

3. Checking the missing value

df=data.frame(num_missing=colSums(is.na(data)))

print(df)



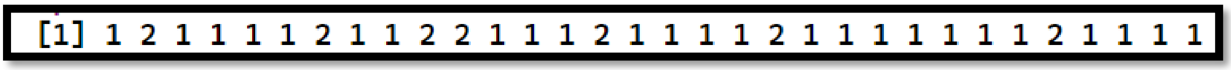
```
     num_missing
mpg            0
cyl            0
disp           0
hp             0
drat           0
wt             0
qsec           0
vs             0
am             0
gear           0
carb           0
```

4. Partitioning of data set into training and testing data

set.seed(555)

ind=sample(2,nrow(data),replace=TRUE,prob=c(0.8,0.2))

print(ind)



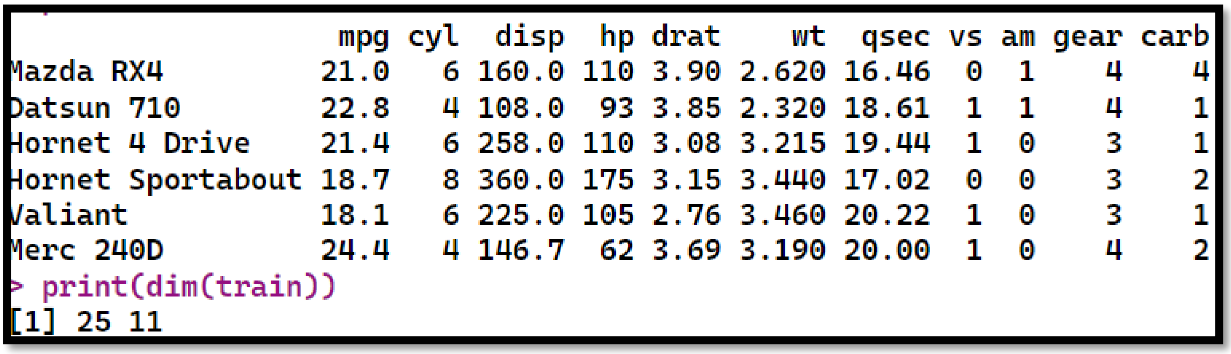
```
 [1] 1 2 1 1 1 1 2 1 1 2 2 1 1 1 2 1 1 1 1 2 1 1 1 1 1 1 1 2 1 1 1 1
```

5. Creation of Training data set

train=data[ind==1,]

print(head(train))

print(dim(train))



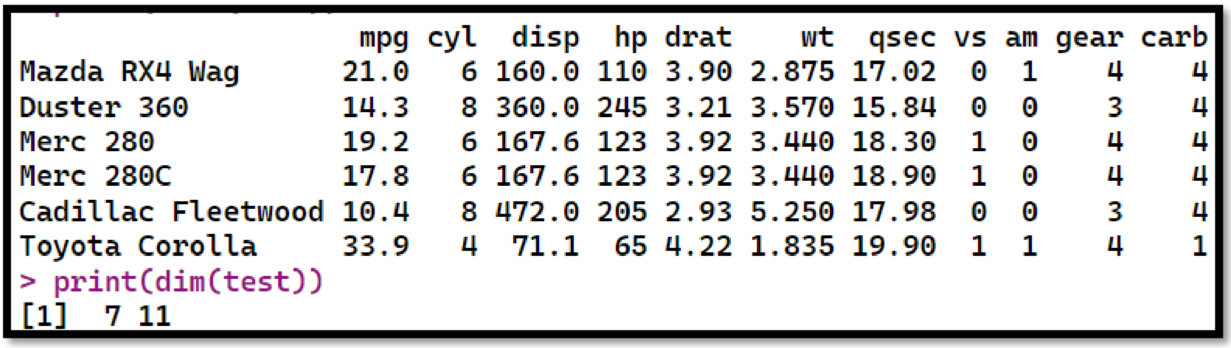
```
                   mpg cyl  disp  hp drat    wt  qsec vs am gear carb
Mazda RX4         21.0   6 160.0 110 3.90 2.620 16.46  0  1    4    4
Datsun 710        22.8   4 108.0  93 3.85 2.320 18.61  1  1    4    1
Hornet 4 Drive    21.4   6 258.0 110 3.08 3.215 19.44  1  0    3    1
Hornet Sportabout 18.7   8 360.0 175 3.15 3.440 17.02  0  0    3    2
Valiant           18.1   6 225.0 105 2.76 3.460 20.22  1  0    3    1
Merc 240D         24.4   4 146.7  62 3.69 3.190 20.00  1  0    4    2
> print(dim(train))
[1] 25 11
```

6. Creation of Testing data set

test=data[ind==2,]

print(head(test))

print(dim(test))



```
                   mpg cyl  disp  hp drat    wt  qsec vs am gear carb
Mazda RX4 Wag     21.0   6 160.0 110 3.90 2.875 17.02  0  1    4    4
Duster 360        14.3   8 360.0 245 3.21 3.570 15.84  0  0    3    4
Merc 280          19.2   6 167.6 123 3.92 3.440 18.30  1  0    4    4
Merc 280C         17.8   6 167.6 123 3.92 3.440 18.90  1  0    4    4
Cadillac Fleetwood 10.4  8 472.0 205 2.93 5.250 17.98  0  0    3    4
Toyota Corolla    33.9   4  71.1  65 4.22 1.835 19.90  1  1    4    1
> print(dim(test))
[1]  7 11
```

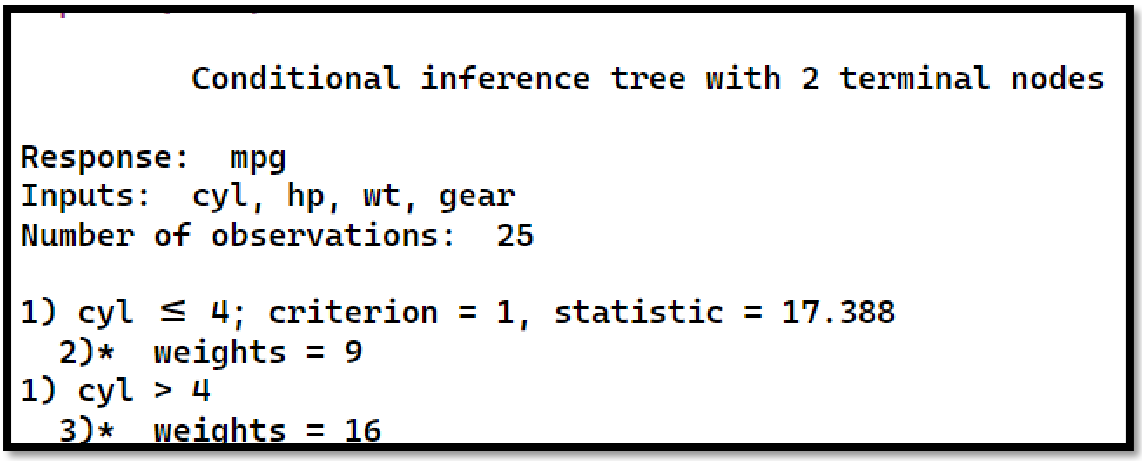
7.Creation of Decision tree

library(party)

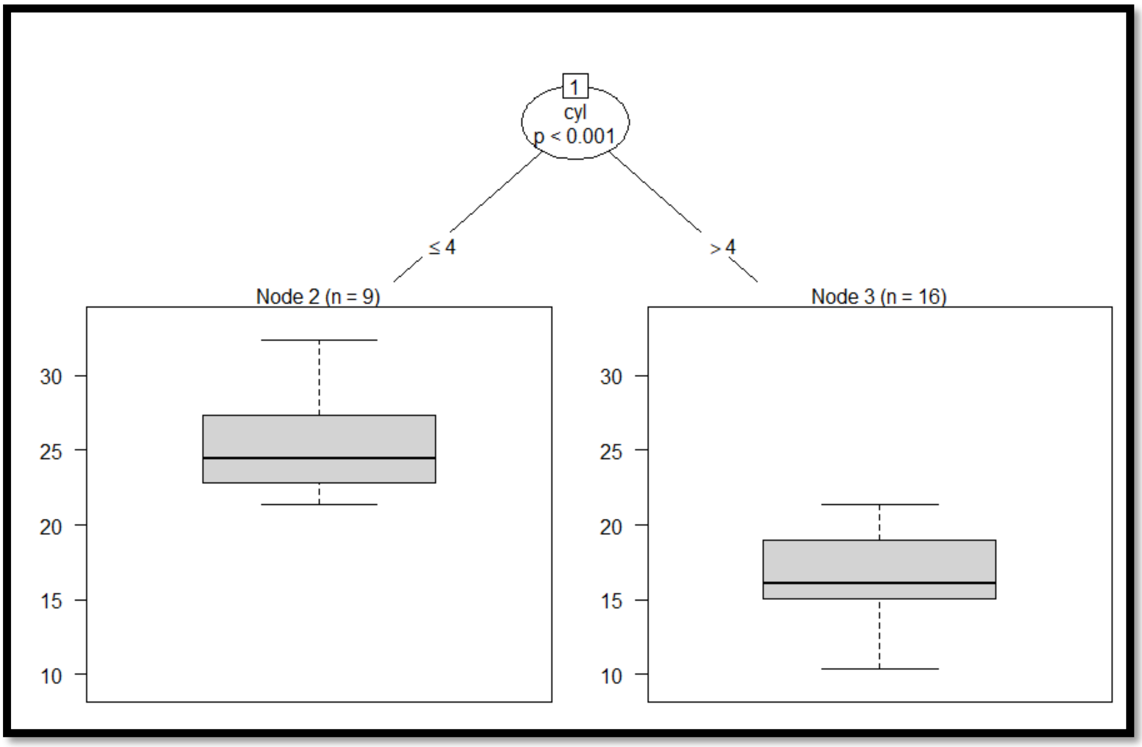
tree=ctree(mpg~cyl+hp+wt+gear,train)

print(tree)

plot(tree)

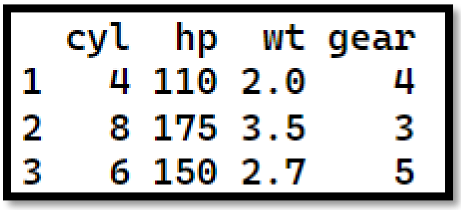


```
         Conditional inference tree with 2 terminal nodes

Response:  mpg
Inputs:  cyl, hp, wt, gear
Number of observations:  25

1) cyl ≤ 4; criterion = 1, statistic = 17.388
  2)*  weights = 9
1) cyl > 4
  3)*  weights = 16
```



8. Create new data for testing the model

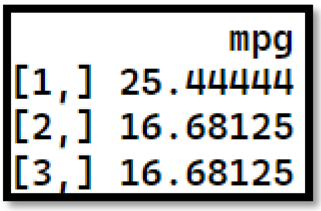
```
new_data=data.frame(
  cyl=c(4,8,6),
  hp=c(110,175,150),
  wt=c(2,3.5,2.7),
  gear=c(4,3,5)
)
print(new_data)
```



```
  cyl  hp  wt gear
1   4 110 2.0    4
2   8 175 3.5    3
3   6 150 2.7    5
```

9. Model prediction on new data

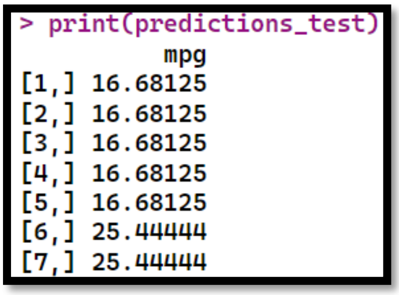
```
predictions_newdata=predict(tree,newdata=new_data,type="response")
print(predictions_newdata)
```



```
          mpg
[1,] 25.44444
[2,] 16.68125
[3,] 16.68125
```

10. Finding prediction using test data

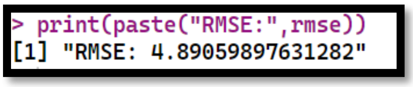
```
predictions_test=predict(tree,newdata=test,type="response")
print(predictions_test)
```



```
> print(predictions_test)
          mpg
[1,] 16.68125
[2,] 16.68125
[3,] 16.68125
[4,] 16.68125
[5,] 16.68125
[6,] 25.44444
[7,] 25.44444
```

11. Finding Root mean square error

```
rmse=sqrt(mean((predictions_test-test$mpg)^2))
print(paste("RMSE:",rmse))
```



```
> print(paste("RMSE:",rmse))
[1] "RMSE: 4.89059897631282"
```