Final Exam Answersheet

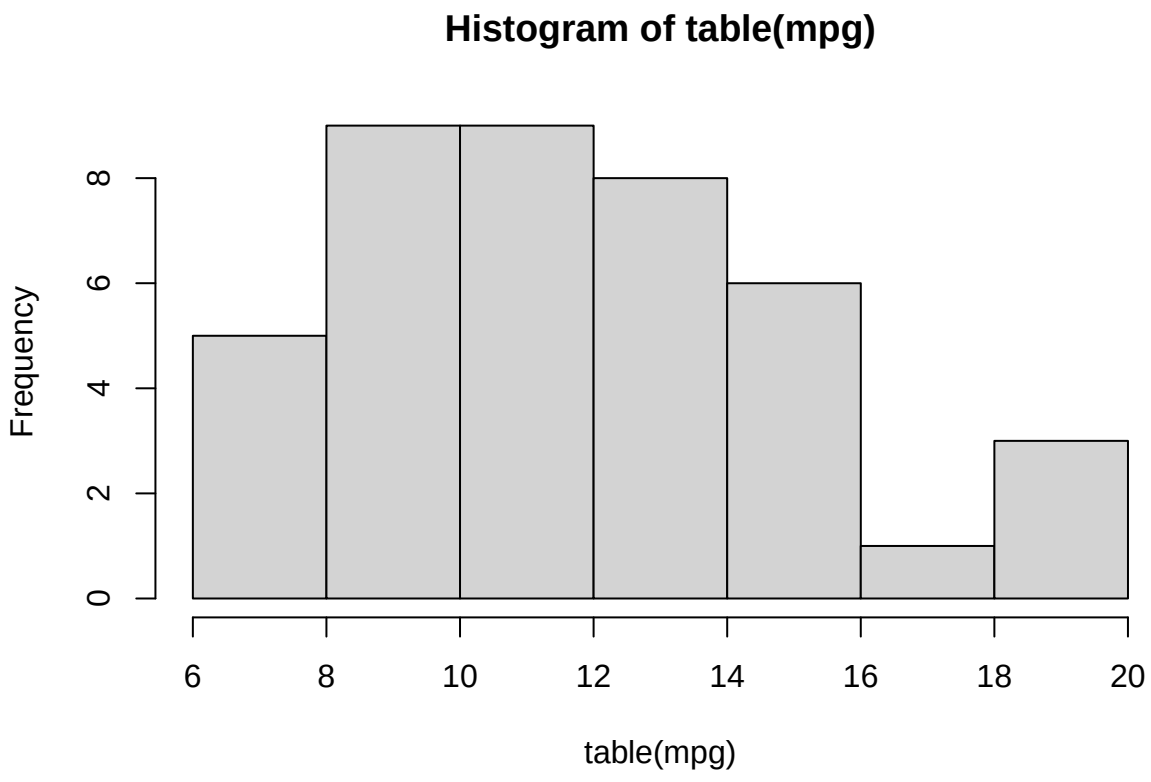
Durga Pokharel

13/03/2022

Group(B)

[Q.N.6]

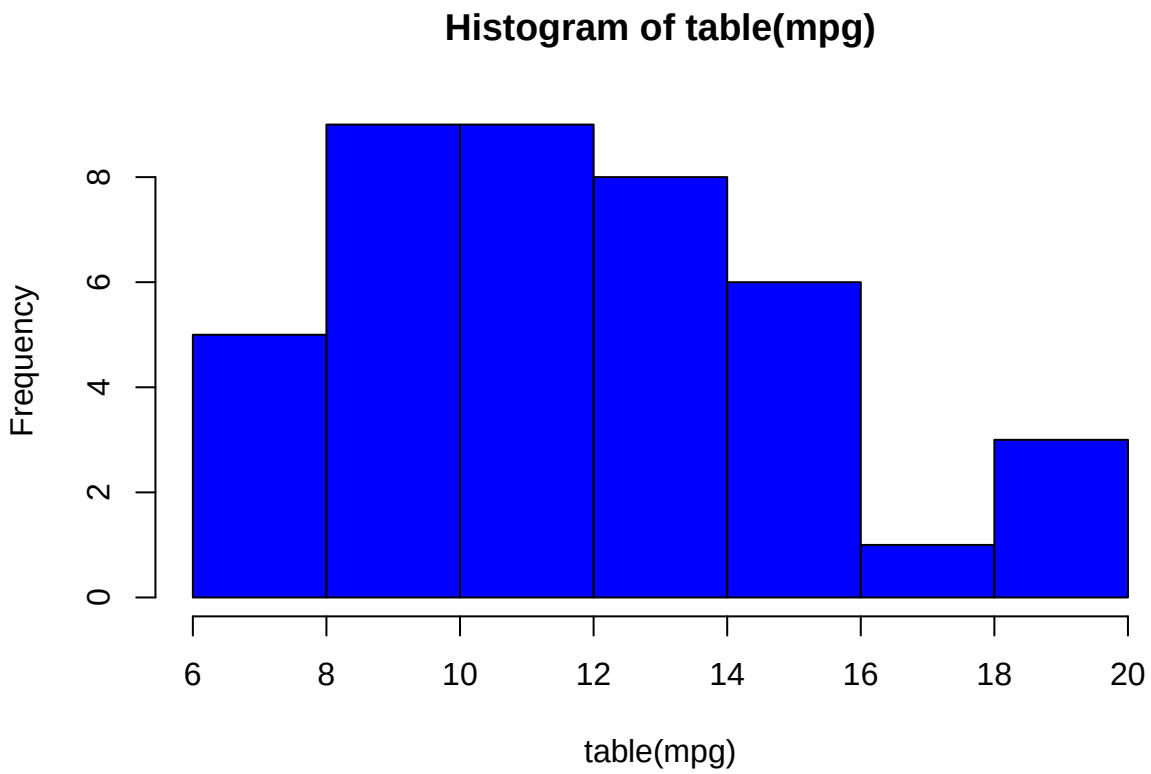
```
#a.
mpg <- c(sample(10:50,size = 500,replace = T))


#b
hist(table(mpg))
```



From histogram we see that maximum value is greater than 12 most of the value behind max value have same frequency. c.

```r
hist(table(mpg),col="blue",bin= 8)
```

```
## Warning in plot.window(xlim, ylim, "", ...): "bin" is not a graphical parameter

## Warning in title(main = main, sub = sub, xlab = xlab, ylab = ylab, ...): "bin"
## is not a graphical parameter

## Warning in axis(1, ...): "bin" is not a graphical parameter

## Warning in axis(2, ...): "bin" is not a graphical parameter
```

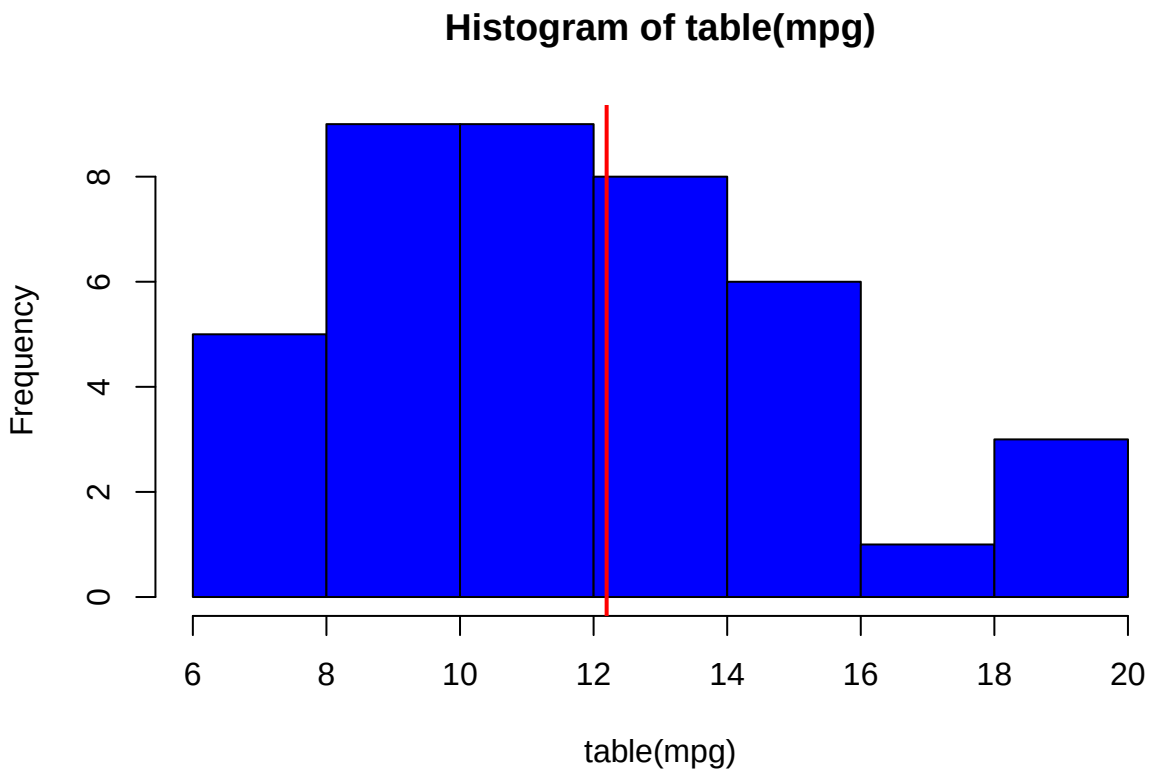


#c

```r
hist(table(mpg),col="blue",bin= 8)
```

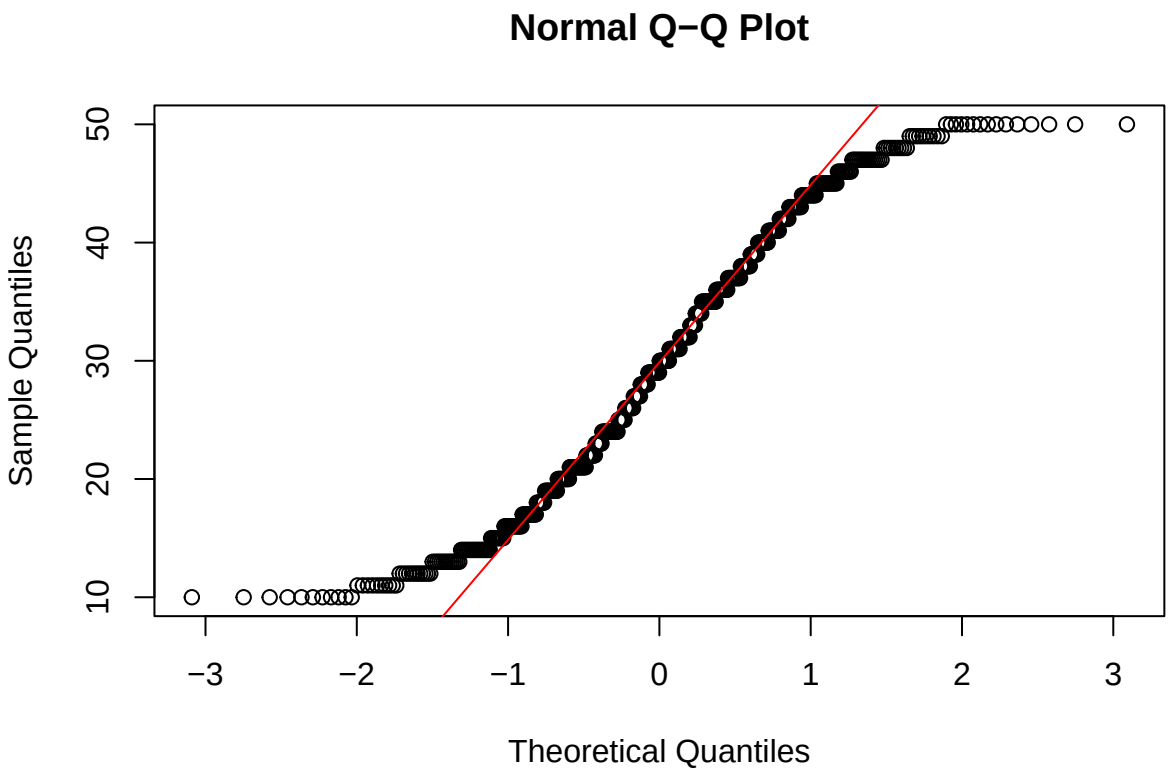
```
## Warning in plot.window(xlim, ylim, "", ...): "bin" is not a graphical parameter

## Warning in title(main = main, sub = sub, xlab = xlab, ylab = ylab, ...): "bin"
## is not a graphical parameter

## Warning in axis(1, ...): "bin" is not a graphical parameter

## Warning in axis(2, ...): "bin" is not a graphical parameter
```

```
abline(v= mean(table(mpg)),lwd=2,col="red")
```



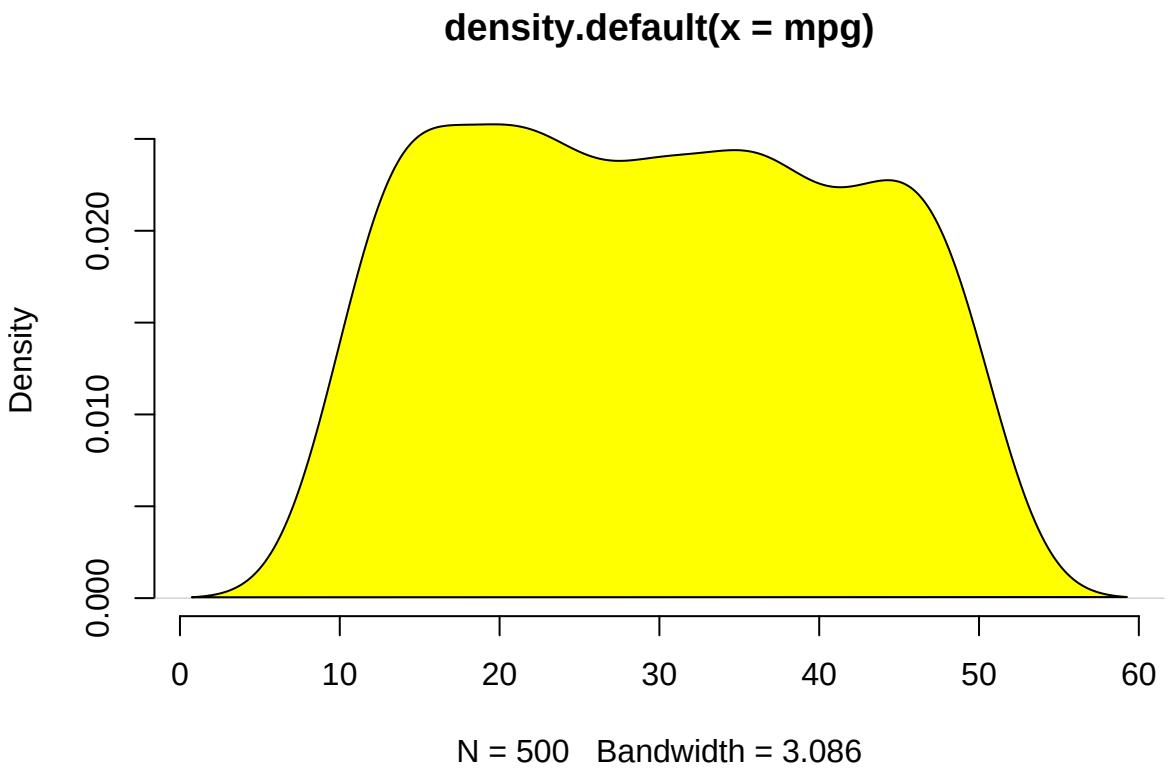
d

```
qqnorm(mpg)
qqline(mpg,col="red")
```



Figue show that our data is not normally distrubated. Actual value and observed value are far from each others. # e

```
dens <- density(mpg)
plot(dens, frame=FALSE, col= "yellow")
polygon(dens, col = "yellow")
```



Density plot also shows that our data is not normally distribuated. We could not saw bell shape curve.

[Q.N.7]

#a

```
set.seed(11)
mpg <- sample(10:50, size = 100, replace =T)
gear <- sample(3:5,size=100, replace =T)
df <- data.frame(mpg,gear)
```

b.

To perform goodnessof fit we need to check normality and is independant variables and depandent variables variance is equal we can do it in following way.

```
with(df,shapiro.test(mpg[gear==3]))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  mpg[gear == 3]
## W = 0.91092, p-value = 0.00354
```

Here p value is less than 0.05 hence it does not follow normal distrbution.

```
with(df,shapiro.test(mpg[gear==4]))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  mpg[gear == 4]
## W = 0.93061, p-value = 0.05096
```

Here P value is greater than 0.05 hence it follow normal distribution.

```
with(df,shapiro.test(mpg[gear==5]))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  mpg[gear == 5]
## W = 0.94923, p-value = 0.1748
```

Here P value also greater than 0.05 hence it follows normal distribution.

In our data all the three categories of gear did not follow the normal distibution hence it is not normal.

b.

To check variance we let's do it. We need to used levene test insted of variance test because we have 3 categories.

```
library(car)
```

```
## Loading required package: carData
```

```
leveneTest(mpg~as.factor(gear),data=df)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value Pr(>F)
## group  2  0.7238 0.4875
##       97
```

Here p value is greater than 0.05 hence variance between depandent variable mpg and independent variabls gear have equal.

c

```
summary(aov(mpg~gear,data=df))
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## gear          1     16   15.65   0.107  0.744
## Residuals    98  14360  146.53
```

Here p value is greater than 0.05 hence h0 accepted. That means means across category is same hence we do not need to do post hoic test. #e. No, can not used this test for data to used it in to our data dependent variable should always follow normal distibution but above data did not satisfy the normality test.

[Q.N.8]

#a.

```
set.seed(11)
mpg <- sample(10:50,size= 200, replace = T)
am <- sample(0:1,size= 200, replace = T)
wt <- sample(1:10,size=200,replace = T)
hp <- sample(125:400,size=200, replace=T)
df <-data.frame(mpg,am,wt,hp)
```

b

```
set.seed(11)
ind <- sample(2,nrow(df), replace = T, prob = c(0.7,0.3))
train_data <- df[ind==1,]
test_data <- df[ind==2,]
```

#c I fit the linear model using mpg as dependent variable.

```
model <- lm(mpg~wt,data= train_data)
```

#d

```
library(caret)
```

```
## Warning: package 'caret' was built under R version 4.1.2
```

```
## Loading required package: ggplot2
```

```
##
## Attaching package: 'ggplot2'
```

```
## The following object is masked _by_ '.GlobalEnv':
##
##     mpg
```

```
## Loading required package: lattice
```

```
set.seed(11)
ind <- sample(2,nrow(df), replace = T, prob = c(0.7,0.3))
train_data <- df[ind==1,]
test_data <- df[ind==2,]
pred <- predict(model,test_data)
R2 <- R2(pred, test_data$mpg)
R2
```

```
## [1] 0.0540566
```

```
RMSE <- RMSE(pred, test_data$mpg)
RMSE
```

```
## [1] 14.1642
```

Here cofficent of determination is only 0.026. That is only 2.6% of variablity explain by independent variable. To do BLUE test it should have R2 greater than 50% independent variable and dependent variable must be normal and value of a and b are significant. Anova must be valid. It did not satisfied condition of blue test.

#e.

```
pred <- predict(model,test_data)
pred
```

```
##        6        9       13       14       15       35       46       50
## 28.55184 31.64755 32.88583 29.17098 29.17098 32.26669 29.17098 29.17098
##       58       62       65       75       85       89       92       95
## 32.26669 29.79012 30.40926 31.02840 32.26669 30.40926 32.26669 27.93270
##       97      101      108      111      114      117      123      128
## 29.17098 32.26669 32.88583 31.64755 32.88583 27.31355 29.79012 29.17098
##      129      131      138      139      140      149      152      154
## 28.55184 30.40926 31.64755 30.40926 28.55184 29.79012 31.64755 31.02840
##      166      169      172      174      178      181      185      190
## 31.02840 29.79012 29.79012 29.79012 30.40926 32.26669 32.26669 31.64755
##      193      194      196      197      200
## 31.64755 29.79012 28.55184 30.40926 32.88583
```

[Q.N.9]

#a.

```
set.seed(11)
mpg <-sample(10:50, size = 300, replace = T)
am <- sample(0:1,size = 300, replace = T)
wt <- sample(0:10,size=300, replace = T)
hp <- sample(125:400,size=300, replace = T)
df <- data.frame(mpg, am,wt,hp)
```

#b

```
set.seed(11)
ind <- sample(2, nrow(df), replace = T, prob=c(0.8,0.2))
train <- df[ind==1,]
test <- df[ind==2,]
```

#c

```
log.mod <- train(am~., data= train,methods= "glm", family= "binomial")
```

```
## note: only 2 unique complexity parameters in default grid. Truncating the grid to 2 .
```

#d.

```
library(caret)
pred <- predict(log.mod,test)
pred
```

```
##         6         9        13        14        35        46        50        58
## 0.8472333 0.5133000 0.3990000 0.3117667 0.4992667 0.4528667 0.7349333 0.8238000
##        75        85        89        95       114       117       123       128
## 0.4422333 0.3347000 0.8037333 0.4662333 0.5459333 0.5360333 0.6554000 0.5749667
##       131       140       149       166       169       172       178       181
## 0.7058000 0.5657000 0.5396333 0.8154667 0.2060333 0.7373667 0.5678000 0.7854667
##       185       190       193       194       197       206       212       216
## 0.2373333 0.3558000 0.7216000 0.5095000 0.3125333 0.7349000 0.6200333 0.6839333
##       217       219       220       221       222       223       238       241
## 0.5887667 0.2684667 0.6646333 0.1663667 0.6090667 0.6136333 0.7820333 0.2792667
##       244       245       248       251       253       257       276       282
## 0.8303333 0.2994667 0.7486667 0.8038667 0.6020333 0.5332667 0.4181000 0.4120667
##       295
## 0.5088333
```

#e .

```
library(caret)

#confusionMatrix(pred, data= test$am)
```

e numbers question code could not run in my R studio so, inorder to knit it I comment it In above chunck.

[Q.N.10]

#a.

```
data <- mtcars
head(data)
```

```
##                    mpg cyl disp  hp drat    wt  qsec vs am gear carb
## Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
## Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
## Datsun 710        22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
## Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  1  0    3    1
## Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0    3    2
## Valiant           18.1   6  225 105 2.76 3.460 20.22  1  0    3    1
```

Using head function we can get top 6 row of data frame. In above result we can see the top 6 row together with all columns of mtcars data frame.

```
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 ...
```

Function `str` check the data type of variables present in the table. Here we can see all the data type of variable is numerical.

#b

```
pca1 <- prcomp(data)
pca1
```

```
## Standard deviations (1, .., p=11):
##  [1] 136.5330479  38.1480776   3.0710166   1.3066508   0.9064862   0.6635411
##  [7]   0.3085791   0.2859604   0.2506973   0.2106519   0.1984238
##
## Rotation (n x k) = (11 x 11):
##               PC1          PC2          PC3          PC4         PC5
## mpg  -0.038118199  0.009184847  0.982070847  0.047634784 -0.08832843
## cyl   0.012035150 -0.003372487 -0.063483942 -0.227991962  0.23872590
## disp  0.899568146  0.435372320  0.031442656 -0.005086826 -0.01073597
## hp    0.434784387 -0.899307303  0.025093049  0.035715638  0.01655194
## drat -0.002660077 -0.003900205  0.039724928 -0.057129357 -0.13332765
## wt    0.006239405  0.004861023 -0.084910258  0.127962867 -0.24354296
## qsec -0.006671270  0.025011743 -0.071670457  0.886472188 -0.21416101
## vs   -0.002729474  0.002198425  0.004203328  0.177123945 -0.01688851
## am   -0.001962644 -0.005793760  0.054806391 -0.135658793 -0.06270200
## gear -0.002604768 -0.011272462  0.048524372 -0.129913811 -0.27616440
## carb  0.005766010 -0.027779208 -0.102897231 -0.268931427 -0.85520810
##               PC6          PC7           PC8          PC9         PC10
## mpg  -0.143790084 -0.039239174  2.271040e-02 -0.002790139  0.030630361
```

```
## cyl  -0.793818050  0.425011021 -1.890403e-01  0.042677206  0.131718534
## disp  0.007424138  0.000582398 -5.841464e-04  0.003532713 -0.005399132
## hp    0.001653685 -0.002212538  4.748087e-06 -0.003734085  0.001862554
## drat  0.227229260  0.034847411 -9.385817e-01 -0.014131110  0.184102094
## wt   -0.127142296 -0.186558915  1.561907e-01 -0.390600261  0.829886844
## qsec -0.189564973  0.254844548 -1.028515e-01 -0.095914479 -0.204240658
## vs    0.102619063 -0.080788938 -2.132903e-03  0.684043835  0.303060724
## am    0.205217266  0.200858874 -2.273255e-02 -0.572372433 -0.162808201
## gear  0.334971103  0.801625551  2.174878e-01  0.156118559  0.203540645
## carb -0.283788381 -0.165474186  3.972219e-03  0.127583043 -0.239954748
##               PC11
## mpg  -0.0158569365
## cyl   0.1454453628
## disp  0.0009420262
## hp   -0.0021526102
## drat -0.0973818815
## wt   -0.0198581635
## qsec  0.0110677880
## vs    0.6256900918
## am    0.7331658036
## gear -0.1909325849
## carb  0.0557957968
```

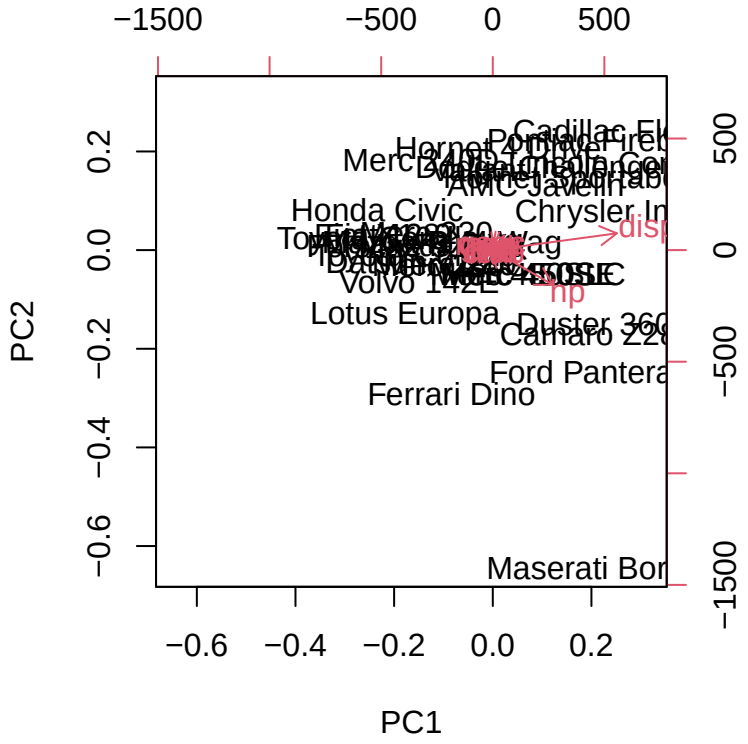
Here, 11 components of pca are seen. There are respectively from PC1 to PC11.

#d

```
biplot(pca1, lavels= rownames(data))
```

```
## Warning in plot.window(...): "lavels" is not a graphical parameter

## Warning in plot.xy(xy, type, ...): "lavels" is not a graphical parameter

## Warning in axis(side = side, at = at, labels = labels, ...): "lavels" is not a
## graphical parameter

## Warning in axis(side = side, at = at, labels = labels, ...): "lavels" is not a
## graphical parameter

## Warning in box(...): "lavels" is not a graphical parameter

## Warning in title(...): "lavels" is not a graphical parameter

## Warning in text.default(x, xlabs, cex = cex[1L], col = col[1L], ...): "lavels"
## is not a graphical parameter

## Warning in plot.window(...): "lavels" is not a graphical parameter

## Warning in plot.xy(xy, type, ...): "lavels" is not a graphical parameter

## Warning in title(...): "lavels" is not a graphical parameter
```

```
## Warning in axis(3, col = col[2L], ...): "lavels" is not a graphical parameter

## Warning in axis(4, col = col[2L], ...): "lavels" is not a graphical parameter

## Warning in text.default(y, labels = ylabs, cex = cex[2L], col = col[2L], :
## "lavels" is not a graphical parameter
```

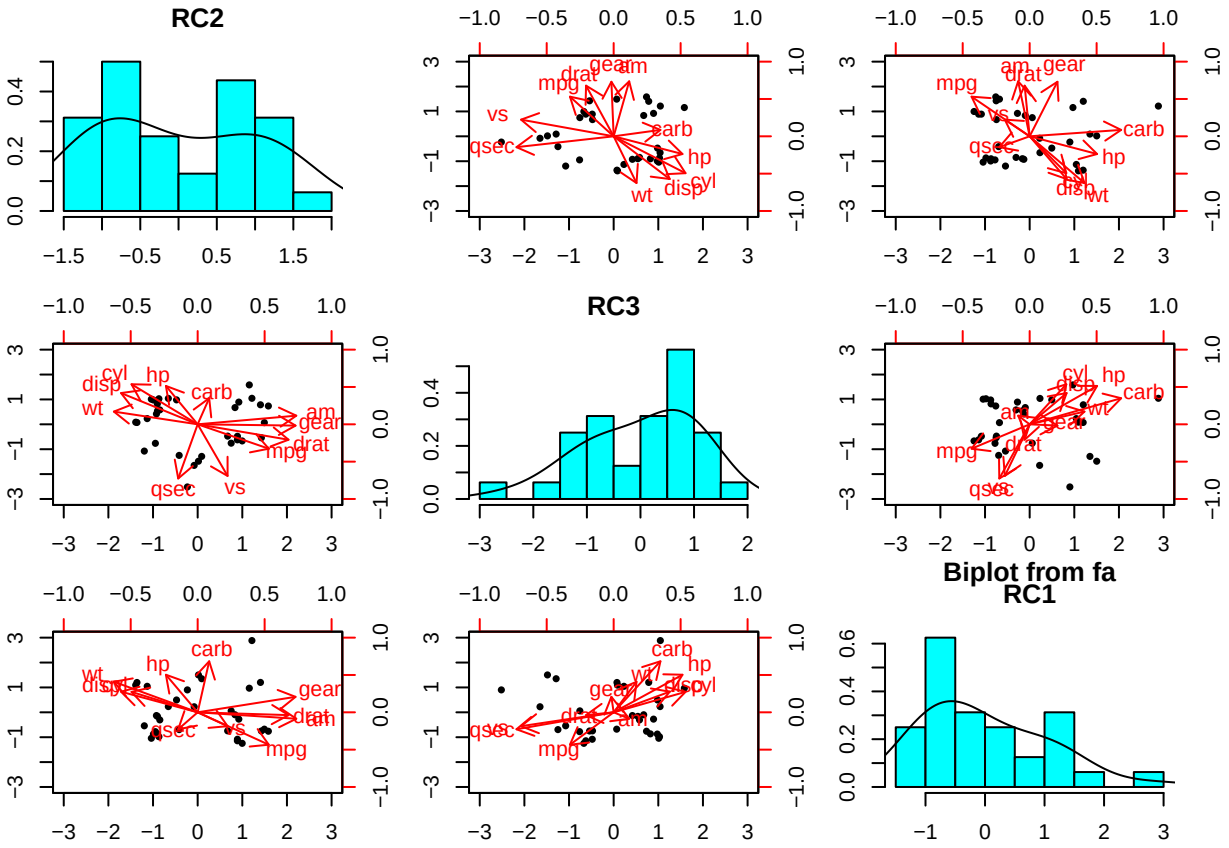


#e

```
library(psych)
```

```
## Warning: package 'psych' was built under R version 4.1.2

##
## Attaching package: 'psych'

## The following objects are masked from 'package:ggplot2':
##
##     %+%, alpha

## The following object is masked from 'package:car':
##
##     logit
```

```
fa1 <- psych::principal(data, nfactors = 3, rotate = "varimax")
biplot(fa1, labels = rownames(fa1))
```



Here, we can see biplot RC2 in the top and RC3 in to second row and RC1 in last row. PCA obtained from `VARIMAX` in not true pca.