Du.Minghao.HW10.R

duter

2021-02-01

############ Question 2 ############  
  
# (a) Write the R-code to create a matrix in which the first column consists of  
# x variable taking integer values from 1 to 6 and the second column taking 6 random  
# values from a standard normal distribution. Provide the matrix you generated.  
  
mtrx <- matrix(c((1:6), rnorm(6)), ncol = 2, byrow = F)  
mtrx

## [,1] [,2]  
## [1,] 1 2.383235196  
## [2,] 2 0.003378615  
## [3,] 3 -0.520355738  
## [4,] 4 -0.052373043  
## [5,] 5 0.737788234  
## [6,] 6 0.899108450

# (b) What is the key difference between a matrix and a data frame in R?  
# Provide an example to demonstrate the difference. (1 point)  
  
setClass('Person', representation(name = 'character', age = 'numeric'))  
john <- list(list(new(  
 'Person', name = 'John Smith', age = 42  
)))  
  
mtx <- matrix(c(1:9), ncol = 3, byrow = T)  
mtx

## [,1] [,2] [,3]  
## [1,] 1 2 3  
## [2,] 4 5 6  
## [3,] 7 8 9

# mtx[4, ] <- c(10, 11, 12) # Error in `[<-`(`\*tmp\*`, 4, , value = c(10, 11, 12)) : subscript out of bounds  
mtx[1, 1] <- john # Matrix structure compromised: variable mtx is now a list object  
class(mtx)

## [1] "list"

df <- as.data.frame(matrix(c(1:9), ncol = 3, byrow = T))  
df

## V1 V2 V3  
## 1 1 2 3  
## 2 4 5 6  
## 3 7 8 9

df[4, ] <- c(10, 11, 12) # Dataframe allows appending new row to existing collection  
df[1, 1] <- john # Dataframe allows storing heterogeneous data  
df

## V1 V2 V3  
## 1 <S4 class 'Person' [package ".GlobalEnv"] with 2 slots> 2 3  
## 2 4 5 6  
## 3 7 8 9  
## 4 10 11 12

############ Question 3 ############  
  
# Download the Auto data set, from the course Blackboard page to answer questions (a) - (e).  
# Make sure that the missing values have been removed from the data.  
  
auto <- read.csv('Auto.csv',  
 stringsAsFactors = T,  
 na.strings = '?')  
nrow(auto)

## [1] 397

auto <- na.omit(auto)  
rownames(auto) <- NULL  
nrow(auto)

## [1] 392

head(auto)

## mpg cylinders displacement horsepower weight acceleration year origin  
## 1 18 8 307 130 3504 12.0 70 1  
## 2 15 8 350 165 3693 11.5 70 1  
## 3 18 8 318 150 3436 11.0 70 1  
## 4 16 8 304 150 3433 12.0 70 1  
## 5 17 8 302 140 3449 10.5 70 1  
## 6 15 8 429 198 4341 10.0 70 1  
## name  
## 1 chevrolet chevelle malibu  
## 2 buick skylark 320  
## 3 plymouth satellite  
## 4 amc rebel sst  
## 5 ford torino  
## 6 ford galaxie 500

summary(auto)

## mpg cylinders displacement horsepower weight   
## Min. : 9.00 Min. :3.000 Min. : 68.0 Min. : 46.0 Min. :1613   
## 1st Qu.:17.00 1st Qu.:4.000 1st Qu.:105.0 1st Qu.: 75.0 1st Qu.:2225   
## Median :22.75 Median :4.000 Median :151.0 Median : 93.5 Median :2804   
## Mean :23.45 Mean :5.472 Mean :194.4 Mean :104.5 Mean :2978   
## 3rd Qu.:29.00 3rd Qu.:8.000 3rd Qu.:275.8 3rd Qu.:126.0 3rd Qu.:3615   
## Max. :46.60 Max. :8.000 Max. :455.0 Max. :230.0 Max. :5140   
##   
## acceleration year origin name   
## Min. : 8.00 Min. :70.00 Min. :1.000 amc matador : 5   
## 1st Qu.:13.78 1st Qu.:73.00 1st Qu.:1.000 ford pinto : 5   
## Median :15.50 Median :76.00 Median :1.000 toyota corolla : 5   
## Mean :15.54 Mean :75.98 Mean :1.577 amc gremlin : 4   
## 3rd Qu.:17.02 3rd Qu.:79.00 3rd Qu.:2.000 amc hornet : 4   
## Max. :24.80 Max. :82.00 Max. :3.000 chevrolet chevette: 4   
## (Other) :365

# (b) What is the range and median of each quantitative predictor?  
# (Hint: For range, use the range() function.)  
  
quant <- sapply(auto[, 1:7], function(c) c(range(c), median(c)))  
row.names(quant) <- c('MIN', 'MAX', 'MID')  
quant

## mpg cylinders displacement horsepower weight acceleration year  
## MIN 9.00 3 68 46.0 1613.0 8.0 70  
## MAX 46.60 8 455 230.0 5140.0 24.8 82  
## MID 22.75 4 151 93.5 2803.5 15.5 76

# (c) What is the mean and standard deviation of each quantitative predictor?  
  
quant <- sapply(auto[, 1:7], function(c) c(mean(c), sd(c)))  
row.names(quant) <- c('MEAN', 'STDDEV')  
quant

## mpg cylinders displacement horsepower weight acceleration  
## MEAN 23.445918 5.471939 194.412 104.46939 2977.5842 15.541327  
## STDDEV 7.805007 1.705783 104.644 38.49116 849.4026 2.758864  
## year  
## MEAN 75.979592  
## STDDEV 3.683737

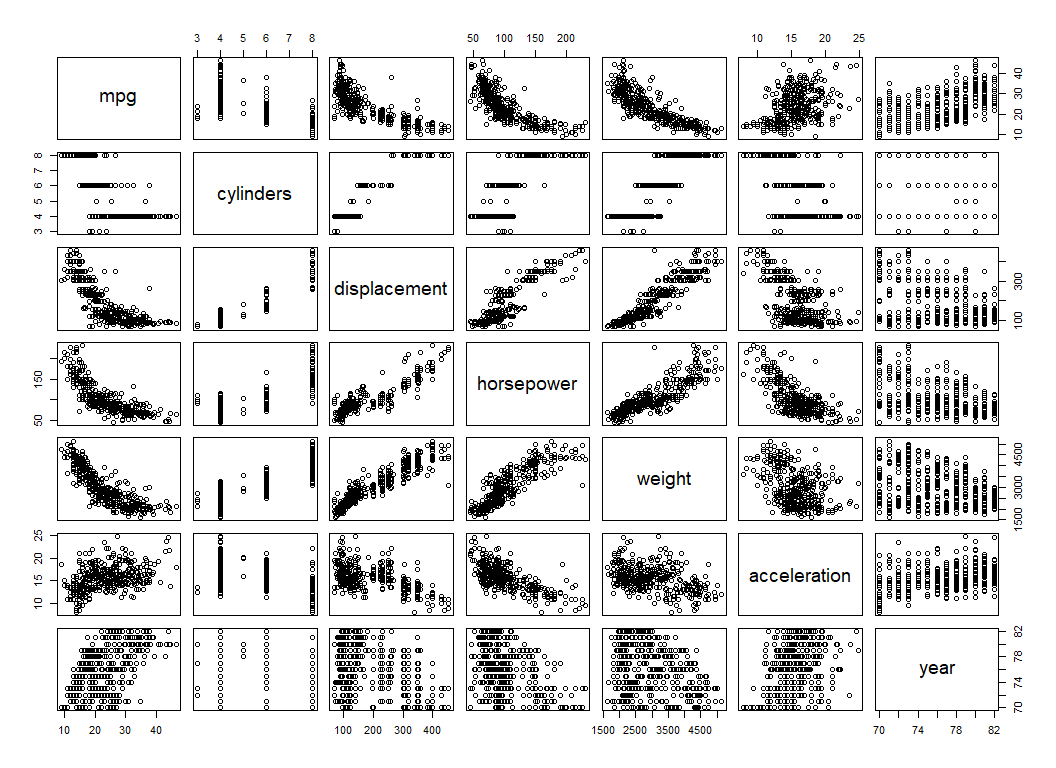
# (d) Remove the 25th through 115th observations. What is the range, median, mean,  
# and standard deviation of each predictor in the subset of the data that remains?  
  
quant <- sapply(auto[-(25:115), 1:7], function(c) c(range(c), median(c), mean(c), sd(c)))  
row.names(quant) <- c('MIN', 'MAX', 'MID', 'MEAN', 'STDDEV')  
quant

## mpg cylinders displacement horsepower weight acceleration  
## MIN 11.000000 3.000000 68.00000 46.00000 1649.0000 8.000000  
## MAX 46.600000 8.000000 455.00000 230.00000 4699.0000 24.800000  
## MID 24.500000 4.000000 140.00000 90.00000 2720.0000 15.600000  
## MEAN 25.015947 5.272425 180.46844 98.63455 2868.7708 15.743189  
## STDDEV 7.637553 1.610032 95.90959 34.35597 755.6678 2.760072  
## year  
## MIN 70.00000  
## MAX 82.00000  
## MID 77.00000  
## MEAN 77.20266  
## STDDEV 3.31091

# (e) Suppose that we wish to predict gas mileage (mpg) on the basis of other variables.  
# Using the full data set which variables do you believe will be useful in predicting mpg?  
# Explain your answer using plots and correlation coefficients of the data.  
  
res <- sapply(auto[, 2:7], function(c) cor(auto$mpg, c))  
res

## cylinders displacement horsepower weight acceleration year   
## -0.7776175 -0.8051269 -0.7784268 -0.8322442 0.4233285 0.5805410

full <- plot(auto[, 1:7])



plt <-  
 barplot(res,  
 names.arg = names(res),  
 ylim = c(-1, 1),  
 ylab = 'Correlation Coefficients')  
text(  
 plt,  
 y = round(unname(res), 4),  
 label = round(unname(res), 4),  
 pos = 3,  
 col = "red"  
)

