APPLIED DATASCIENCE LAB-3

Name :- E Hemasundar Yadav

Roll :- AP19110010481

Sec :- CSE-H

#importing the data set.

data <- read.csv('canadafuel.csv', sep = ",")

head(data,10)

sample Output :-

Year MAKE MODEL VEHICLE.CLASS ENGINE.SIZE CYLINDERS TRANSMISSION

1 2000 ACURA 1.6EL COMPACT 1.6 4 A4

2 2000 ACURA 1.6EL COMPACT 1.6 4 M5

3 2000 ACURA 3.2TL MID-SIZE 3.2 6 AS5

4 2000 ACURA 3.5RL MID-SIZE 3.5 6 A4

5 2000 ACURA INTEGRA SUBCOMPACT 1.8 4 A4

6 2000 ACURA INTEGRA SUBCOMPACT 1.8 4 M5

7 2000 ACURA INTEGRA GSR/TYPE R SUBCOMPACT 1.8 4 M5

8 2000 ACURA NSX SUBCOMPACT 3.0 6 AS4

9 2000 ACURA NSX SUBCOMPACT 3.2 6 M6

10 2000 AUDI A4 COMPACT 1.8 4 A5

FUEL FUEL.CONSUMPTION CO

1 X 10.5 216

2 X 9.8 205

3 Z 13.7 265

4 Z 15.0 301

5 X 11.4 230

6 X 10.6 219

7 Z 10.8 223

8 Z 15.2 306

9 Z 15.4 306

10 Z 12.9 251

Shape of the dataset

Dim(data)

Sample output

639 10

#the type of the data column

sapply(data,class)

sample output :-

Year MAKE MODEL VEHICLE.CLASS

"integer" "character" "character" "character"

ENGINE.SIZE CYLINDERS TRANSMISSION FUEL

"numeric" "integer" "character" "character"

FUEL.CONSUMPTION CO

"numeric" "integer"

# gives the summary of the columns in the dataset

summary(data)

sample ouput :-

Year MAKE MODEL

Min. :2000 Length:639 Length:639

1st Qu.:2000 Class :character Class :character

Median :2000 Mode :character Mode :character

Mean :2000

3rd Qu.:2000

Max. :2000

VEHICLE.CLASS ENGINE.SIZE CYLINDERS

Length:639 Min. :1.000 Min. : 3.000

Class :character 1st Qu.:2.200 1st Qu.: 4.000

Mode :character Median :3.000 Median : 6.000

Mean :3.266 Mean : 5.806

3rd Qu.:4.300 3rd Qu.: 6.000

Max. :8.000 Max. :12.000

TRANSMISSION FUEL FUEL.CONSUMPTION

Length:639 Length:639 Min. : 4.90

Class :character Class :character 1st Qu.:12.50

Mode :character Mode :character Median :14.40

Mean :14.71

3rd Qu.:16.60

Max. :30.20

CO

Min. :104.0

1st Qu.:253.0

Median :288.0

Mean :296.8

3rd Qu.:343.0

Max. :582.0

# Pre-processing

# returns total null values in dataset

sum(is.na(data))

sample output :-

0

# returns null values in dataset column wise

apply(data,2,function(x) sum(is.na(x)))

sample output :-

Year MAKE MODEL

0 0 0

VEHICLE.CLASS ENGINE.SIZE CYLINDERS

0 0 0

TRANSMISSION FUEL FUEL.CONSUMPTION

0 0 0

CO

0

# checking the correlation between to atributes

correlation <-cor(data[['ENGINE.SIZE']],data[['FUEL.CONSUMPTION']])

correlation

0.8581696

# plotting the data

# Data visualization

# plot the graph

model\_data <- lm(data$FUEL.CONSUMPTION~data$ENGINE.SIZE)

summary(model\_data)

print(model\_data)

plot(FUEL.CONSUMPTION~ENGINE.SIZE, data=data)

abline(model\_data,col="red")



