DM Practical 1

--- By Jebin

Exam Ready Cheatsheet (Refer at your own Risk)

Question 1 : The dataset ToyotaCorolla.csv contains data on used cars on sale during the late summer of 2004 in the Netherlands. It has 1436 records containing details on attributes, including Price, Age, Kilometers, HP, and other specifications.

Ans :

a) **Show the different subsets of the whole dataset**

To read dataset :

tc <- read.csv("D:/Dataset/ToyottaCorolla.csv",header = TRUE)

Command What It Does

dim(tc)	Shows rows and columns
names(tc)	Shows column names
head(tc)	Displays first 6 rows
View(tc)	Opens the dataset for viewing
tc[1:10, 1]	Selects first 10 rows of 1st column
tc[1:10, 4]	Selects first 10 rows of 4th column
tc[1:10,]	Selects first 10 rows with all columns
tc[5, 1:2]	Selects 5th row, first 2 columns
tc[5, 6]	Selects 5th row, 6th column
tc[5, c(1:2, 4, 8:10)]	Selects 5th row with specific columns
tc$FuelType[1:20]	Selects first 20 values from FuelType
length(tc$FuelType)	Counts values in FuelType column

b) Find mean of and summary statistics for the dataset

mean(tc$Price)	Calculates the mean of Price
mean(tc$FuelType)	❗ Error, because FuelType is not numeric
summary(tc$Price)	Gives summary statistics for Price
summary(tc)	Provides summary of all columns

c) The dataset has two categorical attributes, Fuel Type and Metallic. Convert these to binary variables so that categorical data is transformed into dummies.

You have two **categorical attributes** in your dataset:

1. **FuelType**: Likely a factor column (e.g., "Diesel", "Petrol", etc.)
2. **Metallic**: Another categorical column (e.g., "Yes", "No", etc.)

Code :

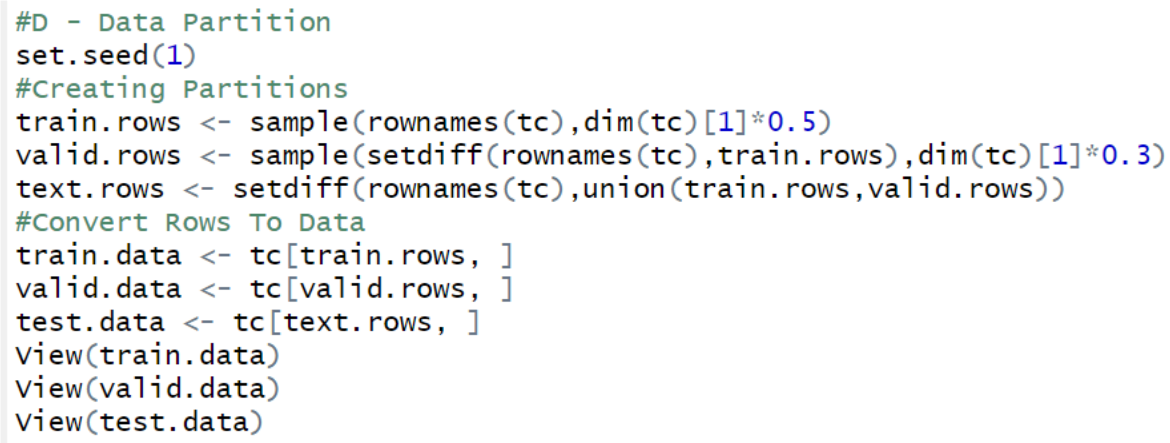
```
convert <- model.matrix(~0 + FuelType + Automatic, data = tc)

convert <- as.data.frame(convert)

View(convert)
```

d) Prepare the dataset for data mining techniques of supervised learning by creating partitions in R. Select all the variables and use default values for the random seed and partitioning percentages for training (50%), validation (30%), and test (20%) sets. Describe the roles that these partitions will play in modeling.

Ans :

```
#D - Data Partition

set.seed(1)

#Creating Partitions

train.rows <- sample(rownames(tc),dim(tc)[1]*0.5)

valid.rows <- sample(setdiff(rownames(tc),train.rows),dim(tc)[1]*0.3)

text.rows <- setdiff(rownames(tc),union(train.rows,valid.rows))

#Convert Rows To Data

train.data <- tc[train.rows, ]

valid.data <- tc[valid.rows, ]
```

```
test.data <- tc[text.rows, ]

View(train.data)

View(valid.data)

View(test.data)
```



```
#D - Data Partition
set.seed(1)
#Creating Partitions
train.rows <- sample(rownames(tc),dim(tc)[1]*0.5)
valid.rows <- sample(setdiff(rownames(tc),train.rows),dim(tc)[1]*0.3)
text.rows <- setdiff(rownames(tc),union(train.rows,valid.rows))
#Convert Rows To Data
train.data <- tc[train.rows, ]
valid.data <- tc[valid.rows, ]
test.data <- tc[text.rows, ]
View(train.data)
View(valid.data)
View(test.data)
```

e) Explore the data using the data visualization capabilities of R. Which of the pairs among the variables seem to be correlated?

```
install.packages("ggplot2")

library("ggplot2")

tcsub <- tc[1:50, ]

#Barplot

ggplot(data=tcsub,mapping=aes(x=FuelType)) + geom_bar()

#or

ggplot(tc,mapping=aes(x=FuelType)) + geom_bar()

ggplot(tcsub,aes(x=FuelType)) + geom_bar(fill="blue",color="black") +

 labs(x="FuelType",y="Frequency",title="Purchases by  Fuel type")
```

Now Scatter Plot

```
#Scattr Plot

ggplot(tc,aes(x=Age, y= Price))+ geom_point(color="red",size=3)
ggplot(tc,aes(x=Age,y=Price,color=FuelType)) +geom_point(size=3)
```

```
ggplot(tc,aes(x=Age,y=Price,color=FuelType)) +geom_point(size=3)+
 geom_smooth(method='lm')
```

```
ggplot(tc,aes(x=Age,y=Price,color=FuelType)) +geom_point(size=3)+
 geom_smooth(method='lm')+scale_color_manual(values=c("red","blue","yellow"))
```

Solution:

Dataset used: ToyotaCorolla.csv

Description : Price : Price offered in Euro

Age: Age in years

KM: Accumulated kilometres

FuelType: Fuel type(Petrol,Diesel,CNG) categorical data

HP: Horse power (Unit of measurement of power)

MetColor: Metallic color (Yes-1,No-1)

Automatic: Automatic(Yes-1,No-0)

Doors: No. Of Doors

Weight: Weight in kilograms