EXP NO : 04	DATA PREPROCESSING
DATE : 05-09-2022	

AIM:

To pre process the existing and created dataset.

DATA PREPROCESSING:

Created dataset:

Step – 1: Import the dataset

```
#Import the dataset
data1 = read.csv("DataSet1.csv")

#View the dataset
View(data1)
```

	Age	Salary	Graduate	Buys.Laptop
1	37	67000	No	Yes
2	50	83000	No	No
3	48	79000	Yes	Yes
4	44	52000	No	Yes
5	NA	58000	Yes	No
6	30	NA	Yes	Yes
7	NA	54000	No	No
8	40	48000	Yes	Yes
9	38	72000	Yes	No
10	27	61000	No	Yes
11	35	NA	No	Yes

Step – 2: Handle the missing data

```
#Handle the missing data
data1$Age = ifelse(is.na(data1$Age),ave(data1$Age,
                  FUN = function(x) mean(x,na.rm=TRUE)), data1$Age)
data1$Salary = ifelse(is.na(data1$Salary), ave(data1$Salary,
                  FUN = function(x) mean(x,na.rm=TRUE)), data1$Salary)
```

	Age	Salary	Graduate	Buys.Laptop
1	37.00000	67000.00	No	Yes
2	50.00000	83000.00	No	No
3	48.00000	79000.00	Yes	Yes
4	44.00000	52000.00	No	Yes
5	38.77778	58000.00	Yes	No
6	30.00000	63777.78	Yes	Yes
7	38.77778	54000.00	No	No
8	40.00000	48000.00	Yes	Yes
9	38.00000	72000.00	Yes	No
10	27.00000	61000.00	No	Yes
11	35.00000	63777.78	No	Yes

```
#round off the values
data1$Age = as.numeric(format(round(data1$Age,0)))
data1$Salary = as.numeric(format(round(data1$Salary,0)))
```

	Age	Salary	Graduate	Buys.Laptop
1	37	67000	No	Yes
2	50	83000	No	No
3	48	79000	Yes	Yes
4	44	52000	No	Yes
5	39	58000	Yes	No
6	30	63778	Yes	Yes
7	39	54000	No	No
8	40	48000	Yes	Yes
9	38	72000	Yes	No
10	27	61000	No	Yes
11	35	63778	No	Yes

Step – 3: Encode the categorical data

```
#Encoding categorical values
data1$Graduate = factor(data1$Graduate, levels = c("Yes","No"), labels = c(1,0))
data1$Buys.Laptop = factor(data1$Buys.Laptop, levels = c("Yes","No"), labels = c(1,0))
```

	Age	Salary	Graduate	Buys.Laptop
1	37	67000	0	1
2	50	83000	0	0
3	48	79000	1	1
4	44	52000	0	1
5	39	58000	1	0
6	30	63778	1	1
7	39	54000	0	0
8	40	48000	1	1
9	38	72000	1	0
10	27	61000	0	1
11	35	63778	0	1

Step – 4: Split the dataset into training and test sets

```
# required library for data split
library(caTools)
set.seed(123)
# returns true if observation goes to the Training set
#and false if observation goes to the test set.
split = sample.split(data1$Buys.Laptop, SplitRatio = 0.8)
#Creating the training set and test set separately
training_set = subset(data1, split == TRUE)
test_set = subset(data1, split == FALSE)
training_set
test_set
```

```
> training_set
   Age Salary Graduate Buys.Laptop
1   37  67000        0           1
2   50  83000        0           0
3   48  79000        1           1
4   44  52000        0           1
5   39  58000        1           0
6   30  63778        1           1
7   39  54000        0           0
10  27  61000        0           1
11  35  63778        0           1
```

```
> test_set
  Age Salary Graduate Buys.Laptop
8  40  48000        1           1
9  38  72000        1           0
```

Step – 5: Feature Scaling (only scaling the non-factors which are the age and the salary)

```
#Feature Scaling
training_set[, 1:2] = scale(training_set[, 1:2])
test_set[, 1:2] = scale(test_set[, 1:2])
training_set
test_set
```

```
> training_set
          Age      Salary Graduate Buys.Laptop
1  -0.2315560  0.22709209        0           1
2   1.4616972  1.75205298        0           0
3   1.2011967  1.37081276        1           1
4   0.6801957 -1.20255874        0           1
5   0.0289445 -0.63069841        1           0
6  -1.1433077 -0.07999691        1           1
7   0.0289445 -1.01193863        0           0
10 -1.5340585 -0.34476824        0           1
11 -0.4920565 -0.07999691        0           1
```

```
> test_set
        Age     Salary Graduate Buys.Laptop
8 0.7071068 -0.7071068        1           1
9 -0.7071068 0.7071068        1           0
```

Existing dataset:

Step – 1: Import the dataset

```
#Import dataset
data2 = read.csv("DataSet2.csv")

#View the dataset
View(data2)
```

	age	workclass	occupation	race	sex	hours_per_week	income.greater.than.50K
1	39	State-gov	Adm-clerical	White	Male	40	No
2	50	Self-emp-not-inc	Exec-managerial	White	Male	13	No
3	NA	Private	Handlers-cleaners	White	Male	40	No
4	53	Private	Handlers-cleaners	Black	Male	NA	No
5	28	Private	Prof-specialty	Black	Female	40	No
6	NA	Private	Exec-managerial	White	Female	40	No
7	49	Private	Other-service	Black	Female	16	No
8	52	Self-emp-not-inc	Exec-managerial	White	Male	45	Yes
9	NA	Private	Prof-specialty	White	Female	50	Yes
10	42	Private	Exec-managerial	White	Male	40	Yes
11	37	Private	Exec-managerial	Black	Male	80	Yes
12	30	State-gov	Prof-specialty	Asian-Pac-Islander	Male	40	Yes
13	NA	Private	Adm-clerical	White	Female	30	No
14	32	Private	Sales	Black	Male	50	No
15	40	Private	Craft-repair	Asian-Pac-Islander	Male	NA	Yes

Step – 2: Handle the missing data

```
#Handle the missing data
data2$age = ifelse(is.na(data2$age),ave(data2$age,
                   FUN = function(x) mean(x,na.rm=TRUE)), data2$age)
data2$hours_per_week = ifelse(is.na(data2$hours_per_week), ave(data2$hours_per_week,
                   FUN = function(x) mean(x,na.rm=TRUE)), data2$hours_per_week)
```

	age	workclass	occupation	race	sex	hours_per_week	income.greater.than.50K
1	39.00000	State-gov	Adm-clerical	White	Male	40.00000	No
2	50.00000	Self-emp-not-inc	Exec-managerial	White	Male	13.00000	No
3	41.09091	Private	Handlers-cleaners	White	Male	40.00000	No
4	53.00000	Private	Handlers-cleaners	Black	Male	40.30769	No
5	28.00000	Private	Prof-specialty	Black	Female	40.00000	No
6	41.09091	Private	Exec-managerial	White	Female	40.00000	No
7	49.00000	Private	Other-service	Black	Female	16.00000	No
8	52.00000	Self-emp-not-inc	Exec-managerial	White	Male	45.00000	Yes
9	41.09091	Private	Prof-specialty	White	Female	50.00000	Yes
10	42.00000	Private	Exec-managerial	White	Male	40.00000	Yes
11	37.00000	Private	Exec-managerial	Black	Male	80.00000	Yes
12	30.00000	State-gov	Prof-specialty	Asian-Pac-Islander	Male	40.00000	Yes
13	41.09091	Private	Adm-clerical	White	Female	30.00000	No
14	32.00000	Private	Sales	Black	Male	50.00000	No
15	40.00000	Private	Craft-repair	Asian-Pac-Islander	Male	40.30769	Yes

```
#round off the values
data2$age = as.numeric(format(round(data2$age,0)))
data2$hours_per_week = as.numeric(format(round(data2$hours_per_week,0)))
```

	age	workclass	occupation	race	sex	hours_per_week	income.greater.than.50K
1	39	State-gov	Adm-clerical	White	Male	40	No
2	50	Self-emp-not-inc	Exec-managerial	White	Male	13	No
3	41	Private	Handlers-cleaners	White	Male	40	No
4	53	Private	Handlers-cleaners	Black	Male	40	No
5	28	Private	Prof-specialty	Black	Female	40	No
6	41	Private	Exec-managerial	White	Female	40	No
7	49	Private	Other-service	Black	Female	16	No
8	52	Self-emp-not-inc	Exec-managerial	White	Male	45	Yes
9	41	Private	Prof-specialty	White	Female	50	Yes
10	42	Private	Exec-managerial	White	Male	40	Yes
11	37	Private	Exec-managerial	Black	Male	80	Yes
12	30	State-gov	Prof-specialty	Asian-Pac-Islander	Male	40	Yes
13	41	Private	Adm-clerical	White	Female	30	No
14	32	Private	Sales	Black	Male	50	No
15	40	Private	Craft-repair	Asian-Pac-Islander	Male	40	Yes

Step – 3: Encode the categorical data

```
#Encoding categorical values
data2$income.greater.than.50K = factor(data2$income.greater.than.50K, levels = c("Yes","No"), labels = c(1,0))
```

	age	workclass	occupation	race	sex	hours_per_week	income.greater.than.50K
1	39	State-gov	Adm-clerical	White	Male	40	0
2	50	Self-emp-not-inc	Exec-managerial	White	Male	13	0
3	41	Private	Handlers-cleaners	White	Male	40	0
4	53	Private	Handlers-cleaners	Black	Male	40	0
5	28	Private	Prof-specialty	Black	Female	40	0
6	41	Private	Exec-managerial	White	Female	40	0
7	49	Private	Other-service	Black	Female	16	0
8	52	Self-emp-not-inc	Exec-managerial	White	Male	45	1
9	41	Private	Prof-specialty	White	Female	50	1
10	42	Private	Exec-managerial	White	Male	40	1
11	37	Private	Exec-managerial	Black	Male	80	1
12	30	State-gov	Prof-specialty	Asian-Pac-Islander	Male	40	1
13	41	Private	Adm-clerical	White	Female	30	0
14	32	Private	Sales	Black	Male	50	0
15	40	Private	Craft-repair	Asian-Pac-Islander	Male	40	1

Step – 4: Split the dataset into training and test sets

```
# required library for data split
library(caTools)
set.seed(123)
# returns true if observation goes to the Training set
#and false if observation goes to the test set.
split = sample.split(data2$income.greater.than.50K, SplitRatio = 0.8)
#Creating the training set and test set separately
training_set = subset(data2, split == TRUE)
test_set = subset(data2, split == FALSE)
training_set
test_set
```

```
> training_set
   age         workclass         occupation                race     sex hours_per_week income.greater.than.50K
1   39         State-gov       Adm-clerical               White    Male            40                       0
2   50  Self-emp-not-inc    Exec-managerial               White    Male            13                       0
3   41           Private  Handlers-cleaners               White    Male            40                       0
4   53           Private  Handlers-cleaners               Black    Male            40                       0
6   41           Private    Exec-managerial               White  Female            40                       0
7   49           Private      Other-service               Black  Female            16                       0
8   52  Self-emp-not-inc    Exec-managerial               White    Male            45                       1
10  42           Private    Exec-managerial               White    Male            40                       1
11  37           Private    Exec-managerial               Black    Male            80                       1
12  30         State-gov     Prof-specialty  Asian-Pac-Islander    Male            40                       1
14  32           Private              Sales               Black    Male            50                       0
15  40           Private       Craft-repair  Asian-Pac-Islander    Male            40                       1

> test_set
   age workclass      occupation   race     sex hours_per_week income.greater.than.50K
5   28   Private  Prof-specialty  Black  Female            40                       0
9   41   Private  Prof-specialty  White  Female            50                       1
13  41   Private    Adm-clerical  White  Female            30                       0
```

Step – 5: Feature Scaling (only scaling the non-factors which are the age and hours per week)

```
#Feature Scaling
training_set[, c(1,6)] = scale(training_set[, c(1,6)])
test_set[, c(1,6)] = scale(test_set[, c(1,6)])
training_set
test_set
```

```
> training_set
         age         workclass         occupation                 race     sex hours_per_week income.greater.than.50K
1  -0.4225923         State-gov       Adm-clerical                White    Male    -0.02008859                      0
2   1.0453600 Self-emp-not-inc     Exec-managerial                White    Male    -1.64726398                      0
3  -0.1556919           Private   Handlers-cleaners                White    Male    -0.02008859                      0
4   1.4457106           Private   Handlers-cleaners                Black    Male    -0.02008859                      0
6  -0.1556919           Private     Exec-managerial                White  Female    -0.02008859                      0
7   0.9119098           Private       Other-service                Black  Female    -1.46646671                      0
8   1.3122604 Self-emp-not-inc     Exec-managerial                White    Male     0.28124019                      1
10 -0.0222417           Private     Exec-managerial                White    Male    -0.02008859                      1
11 -0.6894927           Private     Exec-managerial                Black    Male     2.39054163                      1
12 -1.6236442         State-gov      Prof-specialty Asian-Pac-Islander    Male    -0.02008859                      1
14 -1.3567438           Private               Sales                Black    Male     0.58256897                      0
15 -0.2891421           Private       Craft-repair Asian-Pac-Islander    Male    -0.02008859                      1
```

```
> test_set
          age workclass      occupation   race     sex hours_per_week income.greater.than.50K
5  -1.1547005   Private  Prof-specialty  Black  Female              0                       0
9   0.5773503   Private  Prof-specialty  White  Female              1                       1
13  0.5773503   Private    Adm-clerical  White  Female             -1                       0
```

RESULT:

Thus, the data pre-processing steps has been implemented.