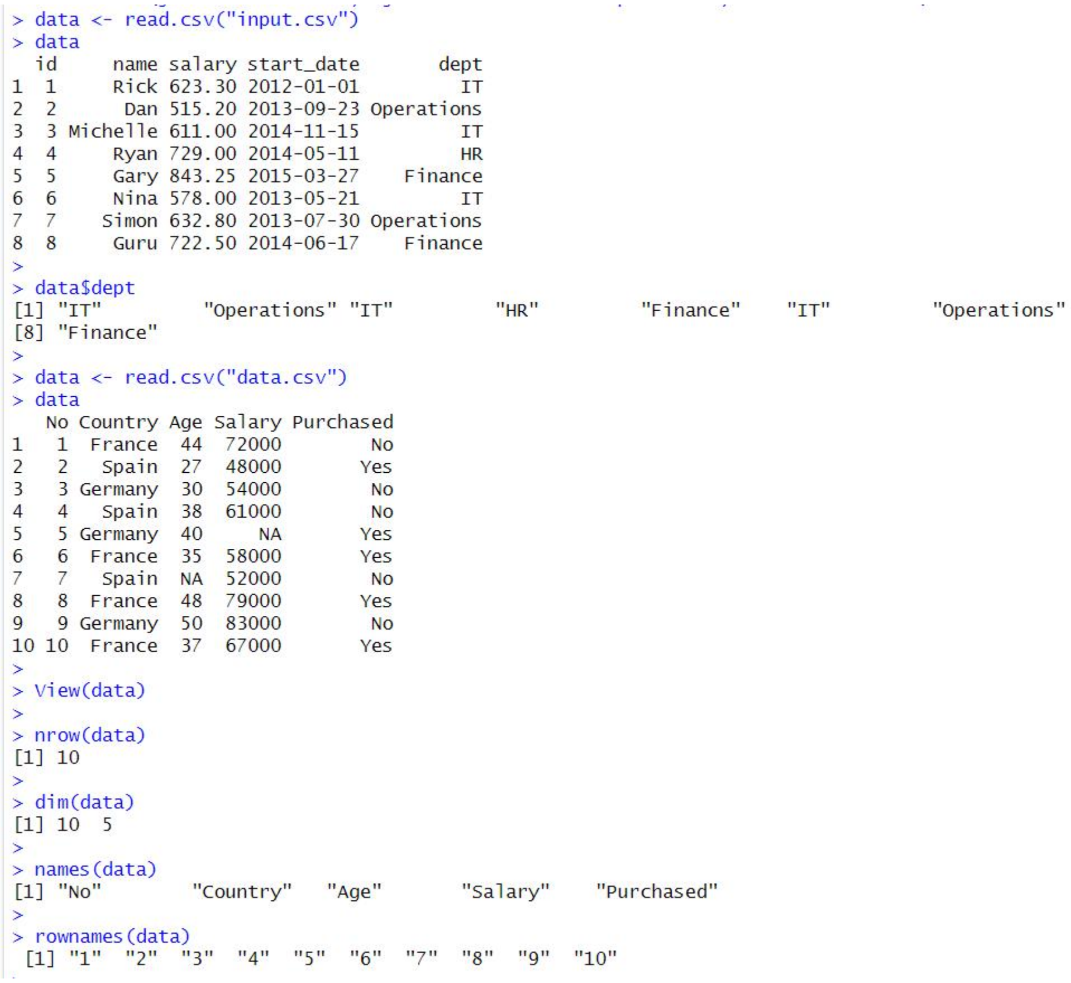
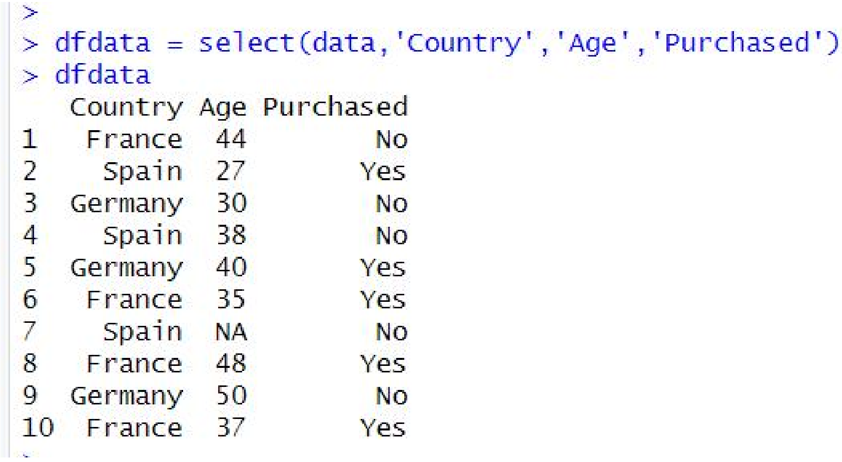
Practical No.7

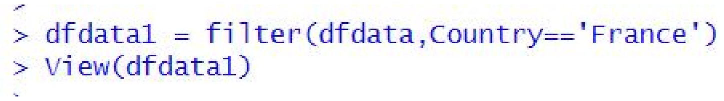
Title: - Introduction to R Graphics and Data Preprocessing
Aim: - To perform data preprocessing using R programming.
Lab Objectives: -
Students will understand following R programming concepts:
I. Importing dataset
II. Handling the Missing Data
III. Encoding Categorical Data
IV. Splitting the Dataset into the Training and Test sets
V. Feature Scaling



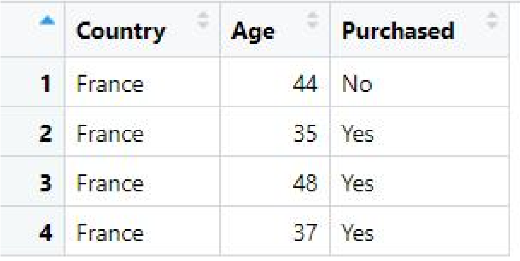
```
> data <- read.csv("input.csv")
> data
  id     name salary start_date       dept
1  1     Rick 623.30 2012-01-01         IT
2  2      Dan 515.20 2013-09-23 Operations
3  3 Michelle 611.00 2014-11-15         IT
4  4     Ryan 729.00 2014-05-11         HR
5  5     Gary 843.25 2015-03-27    Finance
6  6     Nina 578.00 2013-05-21         IT
7  7    Simon 632.80 2013-07-30 Operations
8  8     Guru 722.50 2014-06-17    Finance
>
> data$dept
[1] "IT"         "Operations" "IT"         "HR"         "Finance"    "IT"         "Operations"
[8] "Finance"
>
> data <- read.csv("data.csv")
> data
   No Country Age Salary Purchased
1   1  France  44  72000        No
2   2   Spain  27  48000       Yes
3   3 Germany  30  54000        No
4   4   Spain  38  61000        No
5   5 Germany  40     NA       Yes
6   6  France  35  58000       Yes
7   7   Spain  NA  52000        No
8   8  France  48  79000       Yes
9   9 Germany  50  83000        No
10 10  France  37  67000       Yes
>
> View(data)
>
> nrow(data)
[1] 10
>
> dim(data)
[1] 10  5
>
> names(data)
[1] "No"        "Country"   "Age"       "Salary"    "Purchased"
>
> rownames(data)
 [1] "1"  "2"  "3"  "4"  "5"  "6"  "7"  "8"  "9"  "10"
```



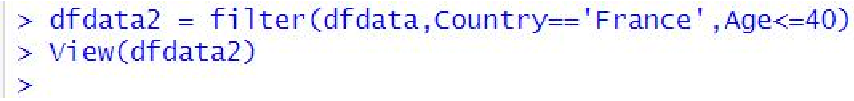
```
>
> dfdata = select(data,'Country','Age','Purchased')
> dfdata
   Country Age Purchased
1   France  44        No
2    Spain  27       Yes
3  Germany  30        No
4    Spain  38        No
5  Germany  40       Yes
6   France  35       Yes
7    Spain  NA        No
8   France  48       Yes
9  Germany  50        No
10  France  37       Yes
```



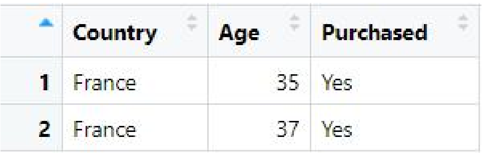
```
> dfdata1 = filter(dfdata,Country=='France')
> View(dfdata1)
```



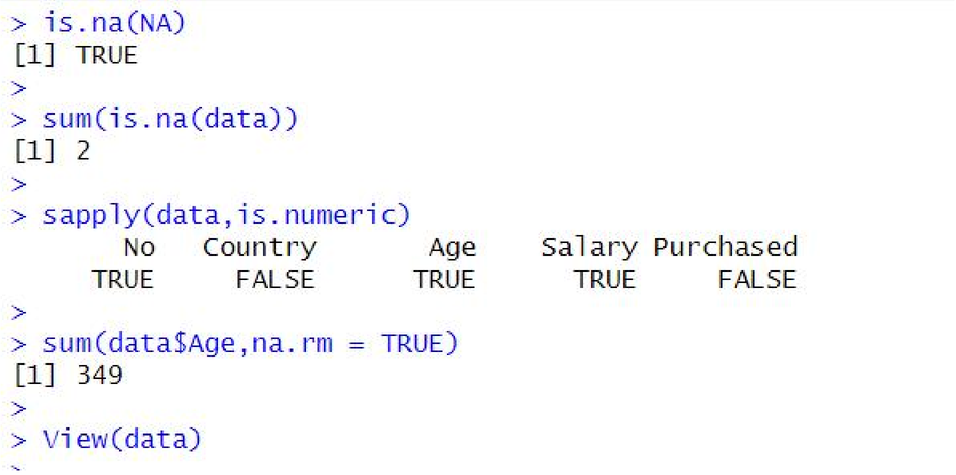
	Country	Age	Purchased
1	France	44	No
2	France	35	Yes
3	France	48	Yes
4	France	37	Yes



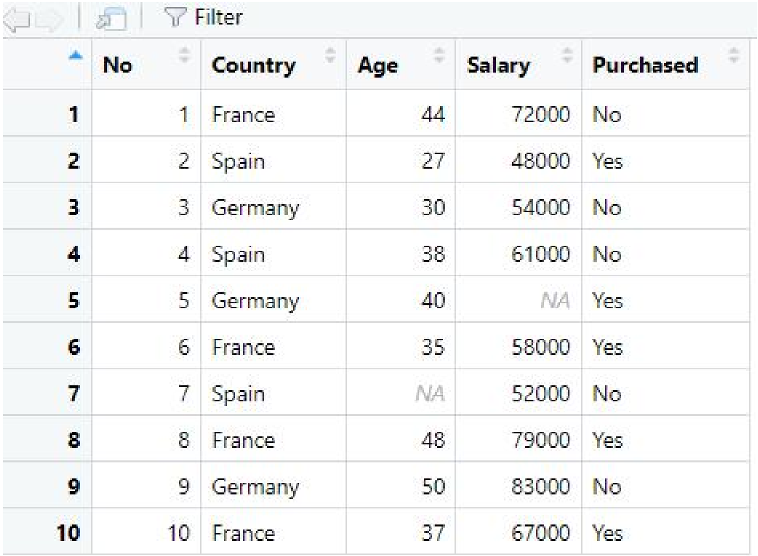
```
> dfdata2 = filter(dfdata,Country=='France',Age<=40)
> View(dfdata2)
>
```



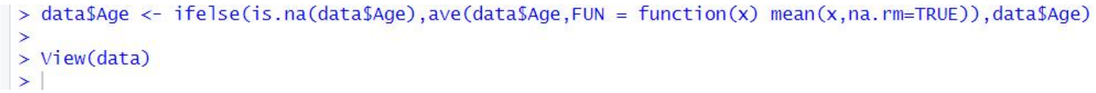
	Country	Age	Purchased
1	France	35	Yes
2	France	37	Yes



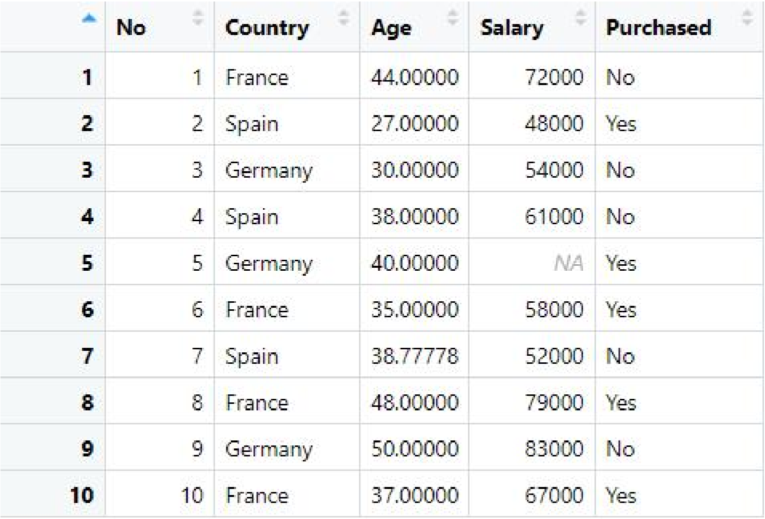
```
> is.na(NA)
[1] TRUE
>
> sum(is.na(data))
[1] 2
>
> sapply(data,is.numeric)
       No   Country       Age    Salary Purchased 
     TRUE     FALSE      TRUE      TRUE     FALSE 
>
> sum(data$Age,na.rm = TRUE)
[1] 349
>
> View(data)
```



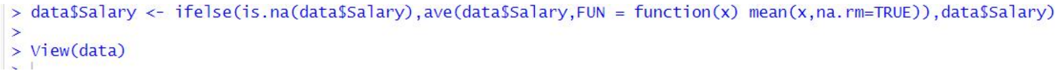
	No	Country	Age	Salary	Purchased
1	1	France	44	72000	No
2	2	Spain	27	48000	Yes
3	3	Germany	30	54000	No
4	4	Spain	38	61000	No
5	5	Germany	40	NA	Yes
6	6	France	35	58000	Yes
7	7	Spain	NA	52000	No
8	8	France	48	79000	Yes
9	9	Germany	50	83000	No
10	10	France	37	67000	Yes



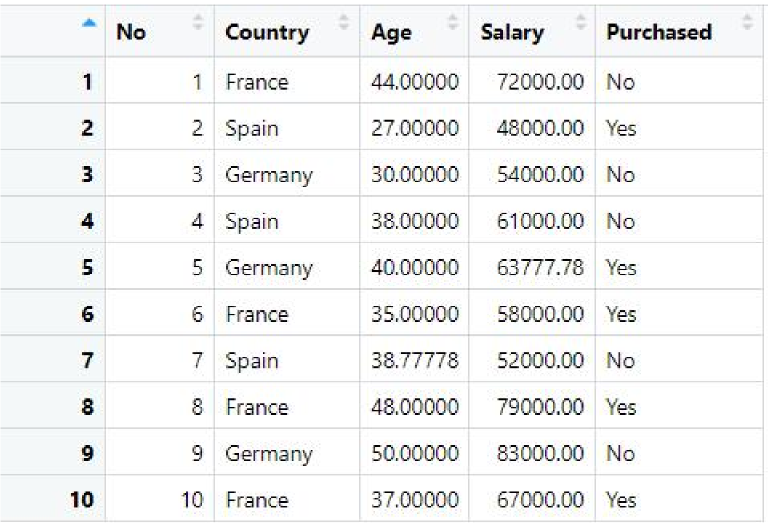
```
> data$Age <- ifelse(is.na(data$Age),ave(data$Age,FUN = function(x) mean(x,na.rm=TRUE)),data$Age)
>
> View(data)
>
```



	No	Country	Age	Salary	Purchased
1	1	France	44.00000	72000	No
2	2	Spain	27.00000	48000	Yes
3	3	Germany	30.00000	54000	No
4	4	Spain	38.00000	61000	No
5	5	Germany	40.00000	*NA*	Yes
6	6	France	35.00000	58000	Yes
7	7	Spain	38.77778	52000	No
8	8	France	48.00000	79000	Yes
9	9	Germany	50.00000	83000	No
10	10	France	37.00000	67000	Yes



```
> data$Salary <- ifelse(is.na(data$Salary),ave(data$Salary,FUN = function(x) mean(x,na.rm=TRUE)),data$Salary)
>
> View(data)
```



	No	Country	Age	Salary	Purchased
1	1	France	44.00000	72000.00	No
2	2	Spain	27.00000	48000.00	Yes
3	3	Germany	30.00000	54000.00	No
4	4	Spain	38.00000	61000.00	No
5	5	Germany	40.00000	63777.78	Yes
6	6	France	35.00000	58000.00	Yes
7	7	Spain	38.77778	52000.00	No
8	8	France	48.00000	79000.00	Yes
9	9	Germany	50.00000	83000.00	No
10	10	France	37.00000	67000.00	Yes