DATA WAREHOUSE AND DATA MINING

LAB ASSIGNMENT-1

NAME:HRITHIK HEM SUNDAR.B

REGNO:19MID0021

Exercise 1: Descriptive Statistics and Plots

Create an Employee Dataset in excel and store the file in .csv extension

```
   Empid Name Designation Salary Experience
1      1  abc     Manager  20000          7
2      2  def Supervisior  19000          8
3      3  ghi       Clerk  10000          6
4      4  jkl      Labour   2000          2
5      5  mno Supervisior  18000          6
6      6  pqr     Manager  25000         11
7      7  stu Supervisior  18000         10
8      8  vwx     Manager  20000          7
9      9  yza       Clerk  15000          5
10    10  bcd       Clerk  15000          5
11    11  efg     Manager  23000         10
12    12  hij       Clerk  12000          4
13    13  klm      Labour   4000          4
14    14  nop Supervisior  20000         10
15    15  qrs     Manager  20000          7
16    16  tuv      Labour   2000          2
17    17  wxy       Clerk  12000          4
18    18  zab     Manager  20000          7
19    19  cde      Labour   2000          1
20    20  fgh Supervisior  21000         10
21    21  ijk     Manager  22000          8
22    22  lmn      Labour   2000          2
23    23  opq     Manager  23000         10
24    24  rst Supervisior  20000          7
25    25  uvw      Labour   2000          2
26    26  xyz       Clerk  17000          7
27    27  zyx      Labour   3000          3
28    28  wvu Supervisior  15000          5
29    29  tsr      Labour   3000          2
30    30  qpo       Clerk  10000          3
```

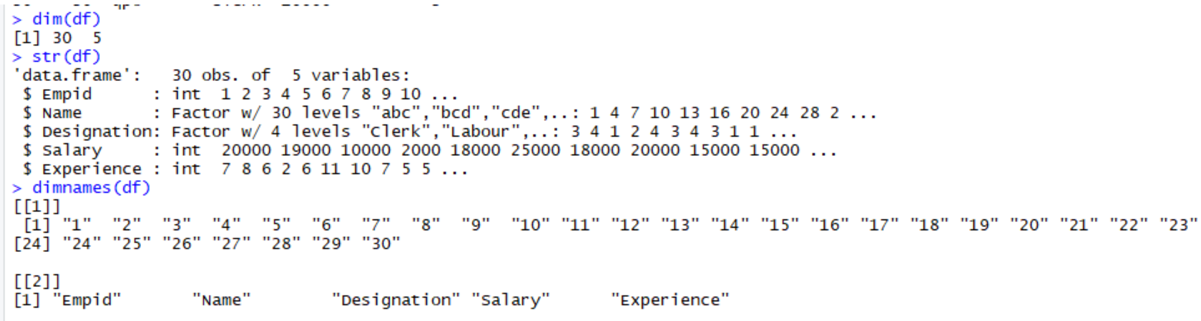
SECTION-2

1. Get the dimensions, structure, attribute name, and attribute values of the dataset

dim(df)

str(df)

dimnames(df)

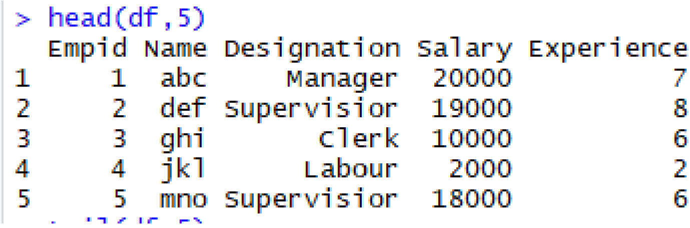


```
> dim(df)
[1] 30  5
> str(df)
'data.frame':   30 obs. of  5 variables:
 $ Empid      : int  1 2 3 4 5 6 7 8 9 10 ...
 $ Name       : Factor w/ 30 levels "abc","bcd","cde",..: 1 4 7 10 13 16 20 24 28 2 ...
 $ Designation: Factor w/ 4 levels "Clerk","Labour",..: 3 4 1 2 4 3 4 3 1 1 ...
 $ Salary     : int  20000 19000 10000 2000 18000 25000 18000 20000 15000 15000 ...
 $ Experience : int  7 8 6 2 6 11 10 7 5 5 ...
> dimnames(df)
[[1]]
 [1] "1"  "2"  "3"  "4"  "5"  "6"  "7"  "8"  "9"  "10" "11" "12" "13" "14" "15" "16" "17" "18" "19" "20" "21" "22" "23"
[24] "24" "25" "26" "27" "28" "29" "30"

[[2]]
[1] "Empid"       "Name"        "Designation" "Salary"      "Experience"
```

2. Display

(A) First 5 Records

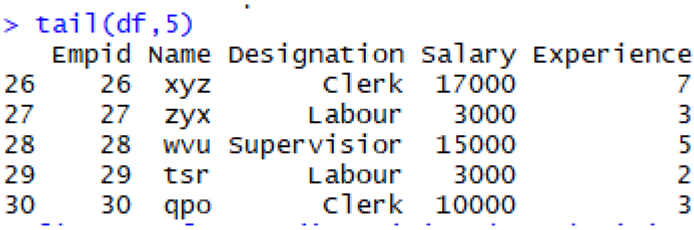
head(df,5)



```
> head(df,5)
  Empid Name Designation Salary Experience
1     1  abc     Manager  20000          7
2     2  def Supervisior  19000          8
3     3  ghi       Clerk  10000          6
4     4  jkl      Labour   2000          2
5     5  mno Supervisior  18000          6
```

(B) Last 5 Records

tail(df,5)

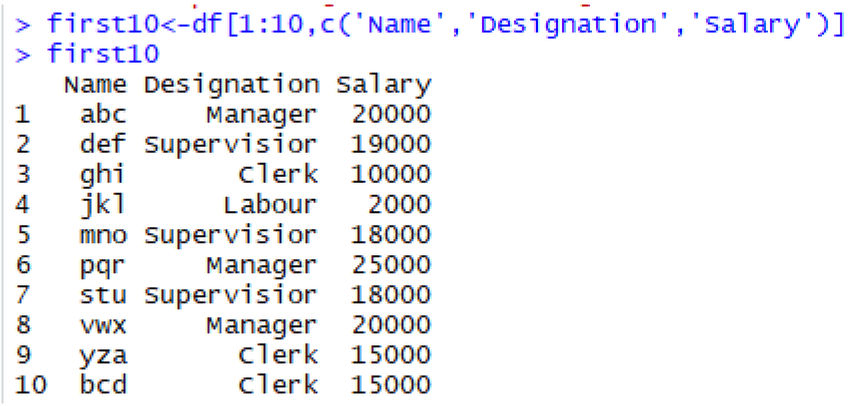


```
> tail(df,5)
   Empid Name Designation Salary Experience
26    26  xyz       Clerk  17000          7
27    27  zyx      Labour   3000          3
28    28  wvu Supervisior  15000          5
29    29  tsr      Labour   3000          2
30    30  qpo       Clerk  10000          3
```

(C) Name, Designation, Salary of First 10 records

first10<-df[1:10,c('Name','Designation','Salary')]

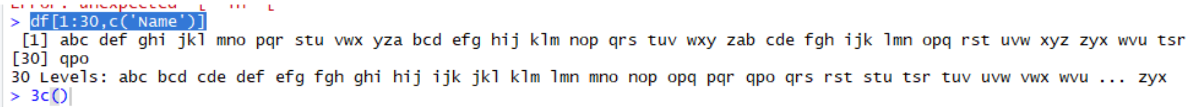
first10



```
> first10<-df[1:10,c('Name','Designation','Salary')]
> first10
   Name Designation Salary
1   abc     Manager  20000
2   def Supervisior  19000
3   ghi       Clerk  10000
4   jkl      Labour   2000
5   mno Supervisior  18000
6   pqr     Manager  25000
7   stu Supervisior  18000
8   vwx     Manager  20000
9   yza       Clerk  15000
10  bcd       Clerk  15000
```

(D) Name of all records

```
df[1:30,c('Name')]
```

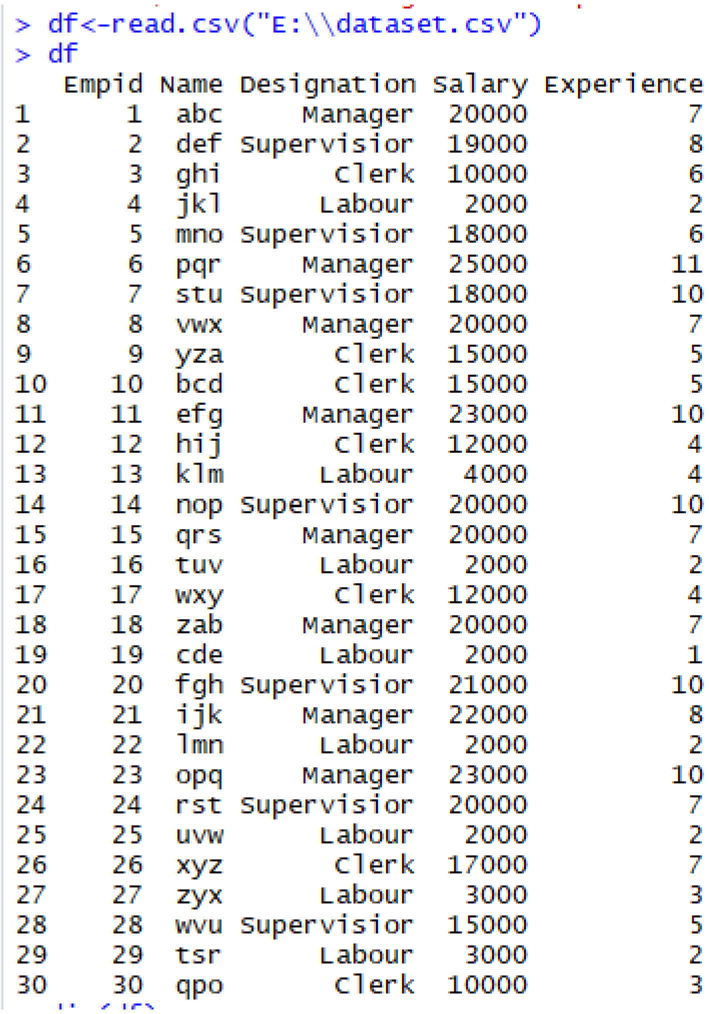


```
> df[1:30,c('Name')]
 [1] abc def ghi jkl mno pqr stu vwx yza bcd efg hij klm nop qrs tuv wxy zab cde fgh ijk lmn opq rst uvw xyz zyx wvu tsr
[30] qpo
30 Levels: abc bcd cde def efg fgh ghi hij ijk jkl klm lmn mno nop opq pqr qpo qrs rst stu tsr tuv uvw vwx wvu ... zyx
> 3c()
```

(E) All records

df<-read.csv("E:\\dataset.csv")

df

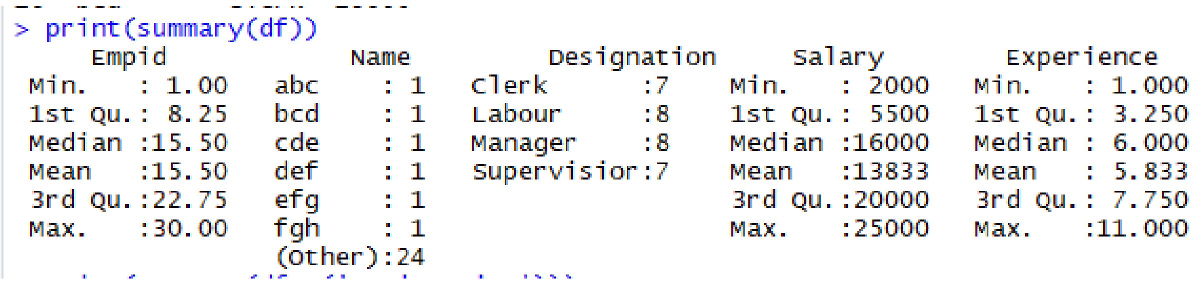


```
> df<-read.csv("E:\\dataset.csv")
> df
   Empid Name Designation Salary Experience
1      1  abc     Manager  20000          7
2      2  def Supervisior  19000          8
3      3  ghi       Clerk  10000          6
4      4  jkl      Labour   2000          2
5      5  mno Supervisior  18000          6
6      6  pqr     Manager  25000         11
7      7  stu Supervisior  18000         10
8      8  vwx     Manager  20000          7
9      9  yza       Clerk  15000          5
10    10  bcd       Clerk  15000          5
11    11  efg     Manager  23000         10
12    12  hij       Clerk  12000          4
13    13  klm      Labour   4000          4
14    14  nop Supervisior  20000         10
15    15  qrs     Manager  20000          7
16    16  tuv      Labour   2000          2
17    17  wxy       Clerk  12000          4
18    18  zab     Manager  20000          7
19    19  cde      Labour   2000          1
20    20  fgh Supervisior  21000         10
21    21  ijk     Manager  22000          8
22    22  lmn      Labour   2000          2
23    23  opq     Manager  23000         10
24    24  rst Supervisior  20000          7
25    25  uvw      Labour   2000          2
26    26  xyz       Clerk  17000          7
27    27  zyx      Labour   3000          3
28    28  wvu Supervisior  15000          5
29    29  tsr      Labour   3000          2
30    30  qpo       Clerk  10000          3
```

3. Display the following statistical measures of the dataset

a) mean, median, 3 quartile distribution of the variables

print(summary(df))



```
> print(summary(df))
     Empid              Name          Designation       Salary          Experience
 Min.   : 1.00   abc      : 1   Clerk      :7   Min.   : 2000   Min.   : 1.000
 1st Qu.: 8.25   bcd      : 1   Labour     :8   1st Qu.: 5500   1st Qu.: 3.250
 Median :15.50   cde      : 1   Manager    :8   Median :16000   Median : 6.000
 Mean   :15.50   def      : 1   Supervisior:7   Mean   :13833   Mean   : 5.833
 3rd Qu.:22.75   efg      : 1                   3rd Qu.:20000   3rd Qu.: 7.750
 Max.   :30.00   fgh      : 1                   Max.   :25000   Max.   :11.000
                 (Other):24
```

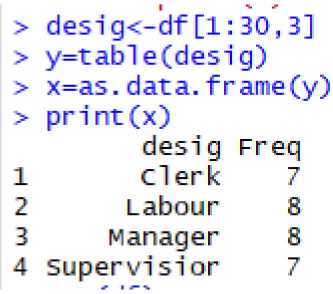
b) Frequency of designation

desig<-df[1:30,3]

y=table(desig)

x=as.data.frame(y)

print(x)



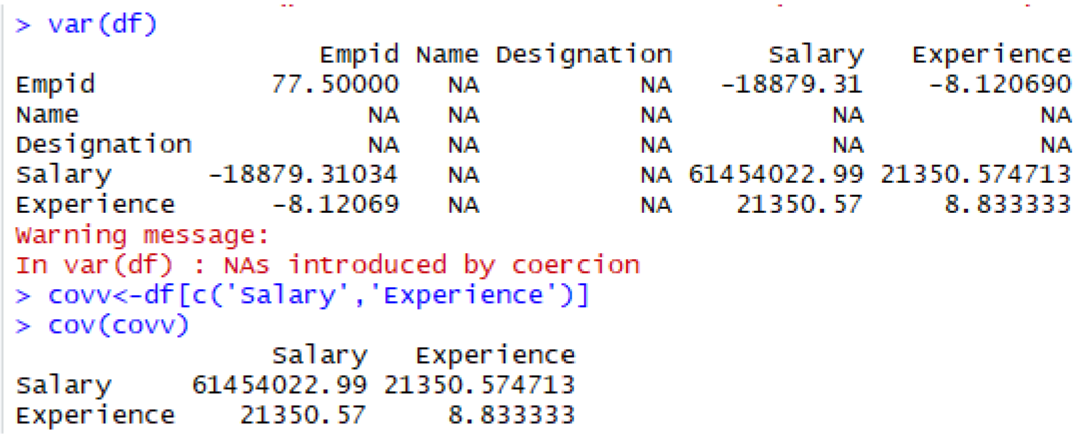
```
> desig<-df[1:30,3]
> y=table(desig)
> x=as.data.frame(y)
> print(x)
        desig Freq
1       Clerk    7
2      Labour    8
3     Manager    8
4 Supervisior    7
```

c) Variance and Covariance

var(df)

covv<-df[c('Salary','Experience')]

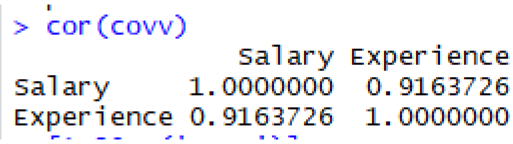
cov(covv)



```
> var(df)
                 Empid Name Designation       Salary   Experience
Empid         77.50000   NA          NA    -18879.31    -8.120690
Name                NA   NA          NA           NA           NA
Designation         NA   NA          NA           NA           NA
Salary      -18879.31034   NA          NA  61454022.99 21350.574713
Experience     -8.12069   NA          NA     21350.57     8.833333
Warning message:
In var(df) : NAs introduced by coercion
> covv<-df[c('Salary','Experience')]
> cov(covv)
                Salary   Experience
Salary     61454022.99 21350.574713
Experience    21350.57     8.833333
```

d) Correlation of salary to experience

cor(covv)



```
> cor(covv)
              Salary Experience
Salary     1.0000000  0.9163726
Experience 0.9163726  1.0000000
```

4. Draw the following

A) Pie chart on designation

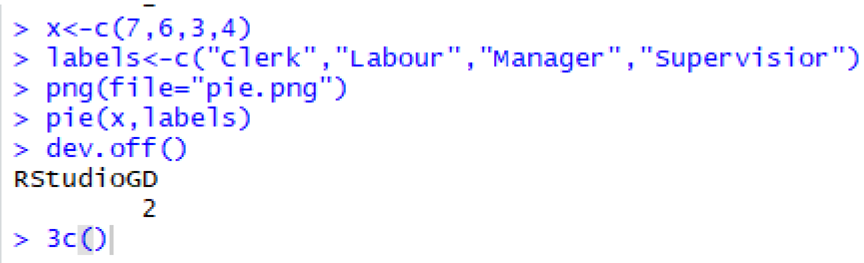
x<-c(7,6,3,4)

labels<-c("Clerk","Labour","Manager","Supervisior")

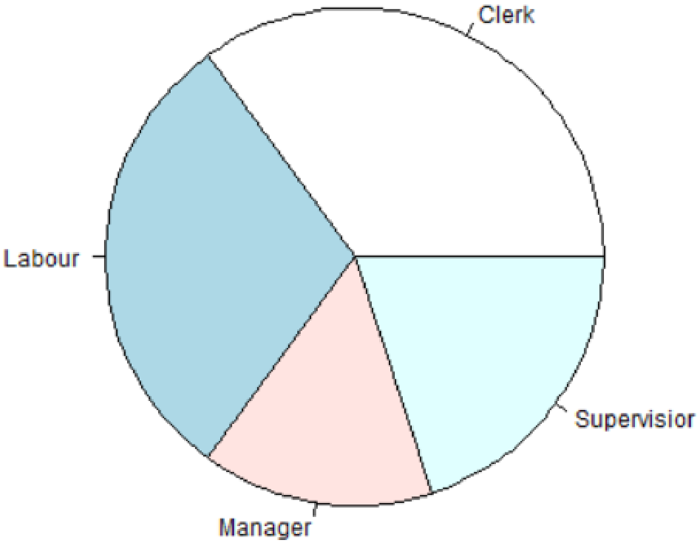
png(file="pie.png")

pie(x,labels)

dev.off()



```
> x<-c(7,6,3,4)
> labels<-c("Clerk","Labour","Manager","Supervisior")
> png(file="pie.png")
> pie(x,labels)
> dev.off()
RStudioGD
        2
> 3c()
```



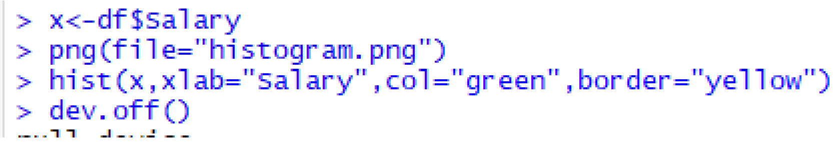
B) Histogram of Salary

x<-df$Salary

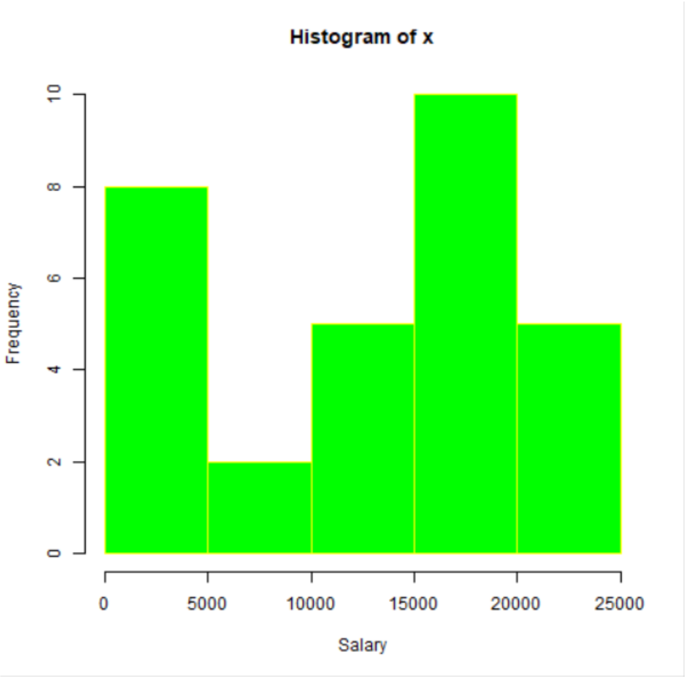
png(file="histogram.png")

hist(x,xlab="Salary",col="green",border="yellow")

```
dev.off()
```

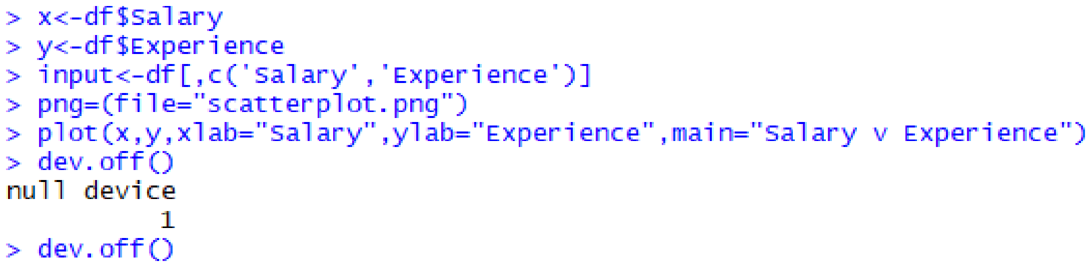


```
> x<-df$Salary
> png(file="histogram.png")
> hist(x,xlab="Salary",col="green",border="yellow")
> dev.off()
null device
```

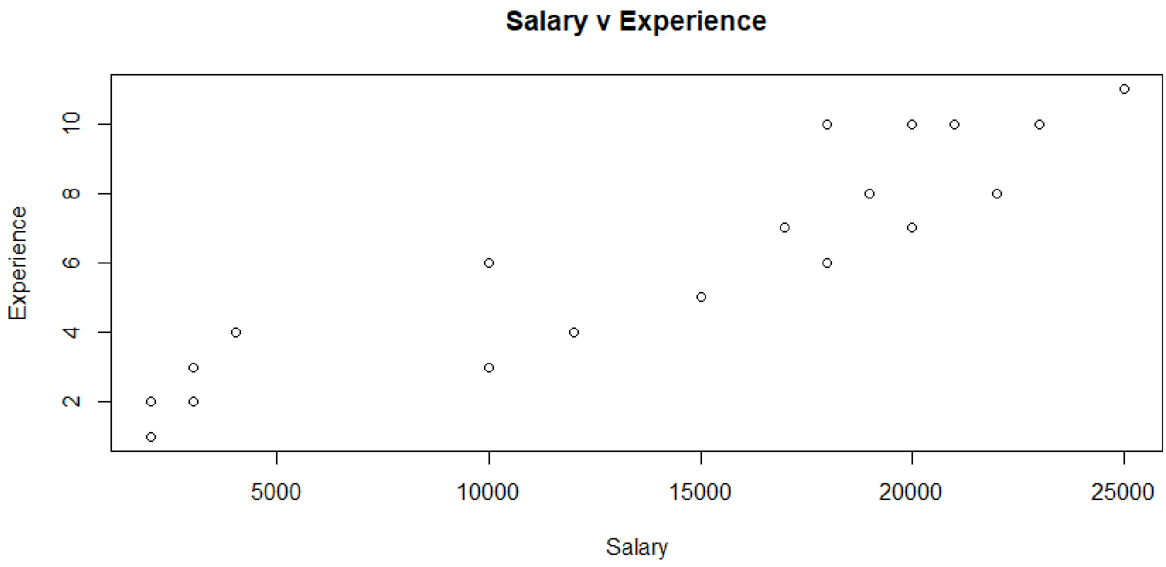


C) Scatter plot of Salary to Experience

```
x<-df$Salary

y<-df$Experience

input<-df[,c('Salary','Experience')]

png=(file="scatterplot.png")

plot(x,y,xlab="Salary",ylab="Experience",main="Salary v Experience")

dev.off()
```



```
> x<-df$Salary
> y<-df$Experience
> input<-df[,c('Salary','Experience')]
> png=(file="scatterplot.png")
> plot(x,y,xlab="Salary",ylab="Experience",main="Salary v Experience")
> dev.off()
null device
          1
> dev.off()
```


Salary v Experience
Experience
2
4
6
8
10
5000
10000
15000
20000
25000
Salary