Exploratory Data Analysis performed on our dataset:

Reading the data

Hide

```
library(dplyr)
library(readr)
library(ggplot2)
library(DescTools)
library(moments)
```

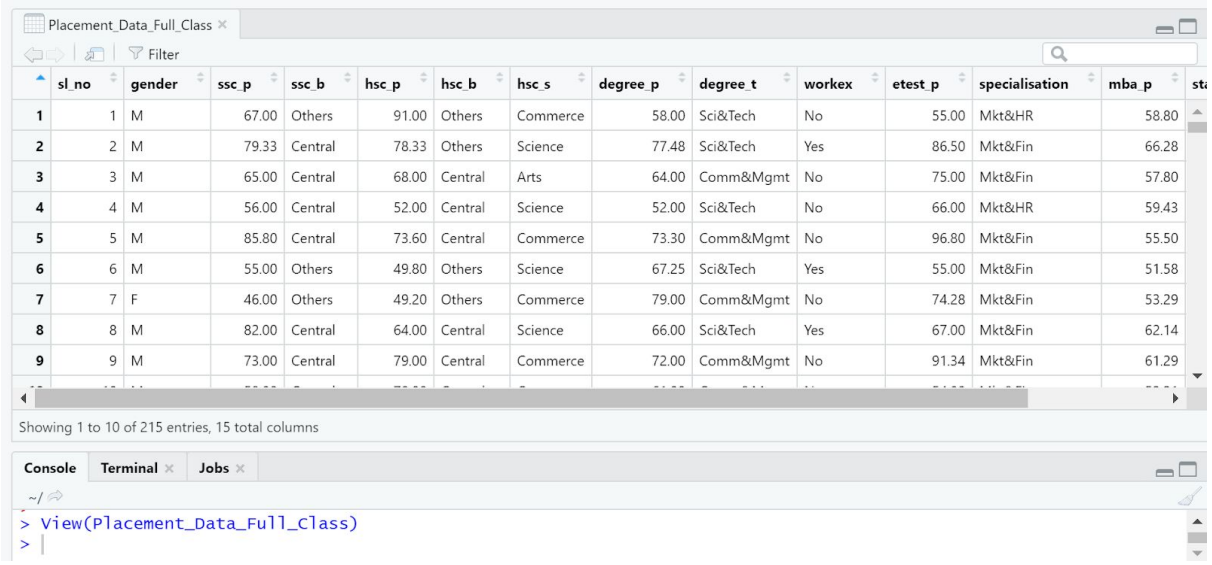
Hide

```
data<-read.csv("Placement_Data_Full_Class.csv")
head(data)
```

hsc_b <chr>	hsc_s <chr>	degree_p <dbl>	degree_t <chr>	workex <chr>	etest_p <dbl>	specialisation <chr>	mba_p <dbl>	status <chr>	salary <int>
Others	Commerce	58.00	Sci&Tech	No	55.0	Mkt&HR	58.80	Placed	270000
Others	Science	77.48	Sci&Tech	Yes	86.5	Mkt&Fin	66.28	Placed	200000
Central	Arts	64.00	Comm&Mgmt	No	75.0	Mkt&Fin	57.80	Placed	250000
Central	Science	52.00	Sci&Tech	No	66.0	Mkt&HR	59.43	Not Placed	*NA*
Central	Commerce	73.30	Comm&Mgmt	No	96.8	Mkt&Fin	55.50	Placed	425000
Others	Science	67.25	Sci&Tech	Yes	55.0	Mkt&Fin	51.58	Not Placed	*NA*

6 rows | 7-16 of 15 columns

Hide



```
summary(data)
```

```
     sl_no          gender              ssc_p          ssc_b          
 Min.   :  1.0   Length:215         Min.   :40.89   Length:215        
 1st Qu.: 54.5   Class :character   1st Qu.:60.60   Class :character  
 Median :108.0   Mode  :character   Median :67.00   Mode  :character  
 Mean   :108.0                      Mean   :67.30                     
 3rd Qu.:161.5                      3rd Qu.:75.70                     
 Max.   :215.0                      Max.   :89.40                     

     hsc_p          hsc_b              hsc_s              degree_p    
 Min.   :37.00   Length:215         Length:215         Min.   :50.00  
 1st Qu.:60.90   Class :character   Class :character   1st Qu.:61.00  
 Median :65.00   Mode  :character   Mode  :character   Median :66.00  
 Mean   :66.33                                         Mean   :66.37  
 3rd Qu.:73.00                                         3rd Qu.:72.00  
 Max.   :97.70                                         Max.   :91.00  

   degree_t            workex             etest_p     specialisation    
 Length:215         Length:215         Min.   :50.0   Length:215        
 Class :character   Class :character   1st Qu.:60.0   Class :character  
 Mode  :character   Mode  :character   Median :71.0   Mode  :character  
                                       Mean   :72.1                     
                                       3rd Qu.:83.5                     
                                       Max.   :98.0                     

     mba_p          status              salary       
 Min.   :51.21   Length:215         Min.   :200000  
 1st Qu.:57.95   Class :character   1st Qu.:240000  
 Median :62.00   Mode  :character   Median :265000  
 Mean   :62.28                      Mean   :288655  
 3rd Qu.:66.25                      3rd Qu.:300000  
 Max.   :77.89                      Max.   :940000  
                                    NA's   :67      
```

```
> glimpse(data)
Rows: 215
Columns: 15
$ sl_no          <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21…
$ gender         <chr> "M", "M", "M", "M", "M", "M", "F", "M", "M", "M", "M", "M", "F", "F", "M"…
$ ssc_p          <dbl> 67.00, 79.33, 65.00, 56.00, 85.80, 55.00, 46.00, 82.00, 73.00, 58.00, 58.…
$ ssc_b          <chr> "Others", "Central", "Central", "Central", "Central", "Others", "Others",…
$ hsc_p          <dbl> 91.00, 78.33, 68.00, 52.00, 73.60, 49.80, 49.20, 64.00, 79.00, 70.00, 61.…
$ hsc_b          <chr> "Others", "Others", "Central", "Central", "Central", "Others", "Others", …
$ hsc_s          <chr> "Commerce", "Science", "Arts", "Science", "Commerce", "Science", "Commerc…
$ degree_p       <dbl> 58.00, 77.48, 64.00, 52.00, 73.30, 67.25, 79.00, 66.00, 72.00, 61.00, 60.…
$ degree_t       <chr> "Sci&Tech", "Sci&Tech", "Comm&Mgmt", "Sci&Tech", "Comm&Mgmt", "Sci&Tech",…
$ workex         <chr> "No", "Yes", "No", "No", "No", "Yes", "No", "Yes", "No", "No", "Yes", "Ye…
$ etest_p        <dbl> 55.00, 86.50, 75.00, 66.00, 96.80, 55.00, 74.28, 67.00, 91.34, 54.00, 62.…
$ specialisation <chr> "Mkt&HR", "Mkt&Fin", "Mkt&Fin", "Mkt&HR", "Mkt&Fin", "Mkt&Fin", "Mkt&Fin"…
$ mba_p          <dbl> 58.80, 66.28, 57.80, 59.43, 55.50, 51.58, 53.29, 62.14, 61.29, 52.21, 60.…
$ status         <chr> "Placed", "Placed", "Placed", "Not Placed", "Placed", "Not Placed", "Not …
$ salary         <dbl> 270000, 200000, 250000, 0, 425000, 0, 0, 252000, 231000, 0, 260000, 25000…
>
```

Removing NA values in Salary

```
data$salary[is.na(data$salary)]<-0;
head(data)
```

hsc_b <chr>	hsc_s <chr>	degree_p <dbl>	degree_t <chr>	workex <chr>	etest_p <dbl>	specialisation <chr>	mba_p <dbl>	status <chr>	salary <dbl>
Others	Commerce	58.00	Sci&Tech	No	55.0	Mkt&HR	58.80	Placed	270000
Others	Science	77.48	Sci&Tech	Yes	86.5	Mkt&Fin	66.28	Placed	200000
Central	Arts	64.00	Comm&Mgmt	No	75.0	Mkt&Fin	57.80	Placed	250000
Central	Science	52.00	Sci&Tech	No	66.0	Mkt&HR	59.43	Not Placed	0
Central	Commerce	73.30	Comm&Mgmt	No	96.8	Mkt&Fin	55.50	Placed	425000
Others	Science	67.25	Sci&Tech	Yes	55.0	Mkt&Fin	51.58	Not Placed	0

6 rows | 7-16 of 15 columns

Some Trends:-

1. Number of male and female placed

```
male_female_job<-data%>%group_by(status,gender)%>%summarise(count=n())
```

```
`summarise()` regrouping output by 'status' (override with `.groups` argument)
```

Hide

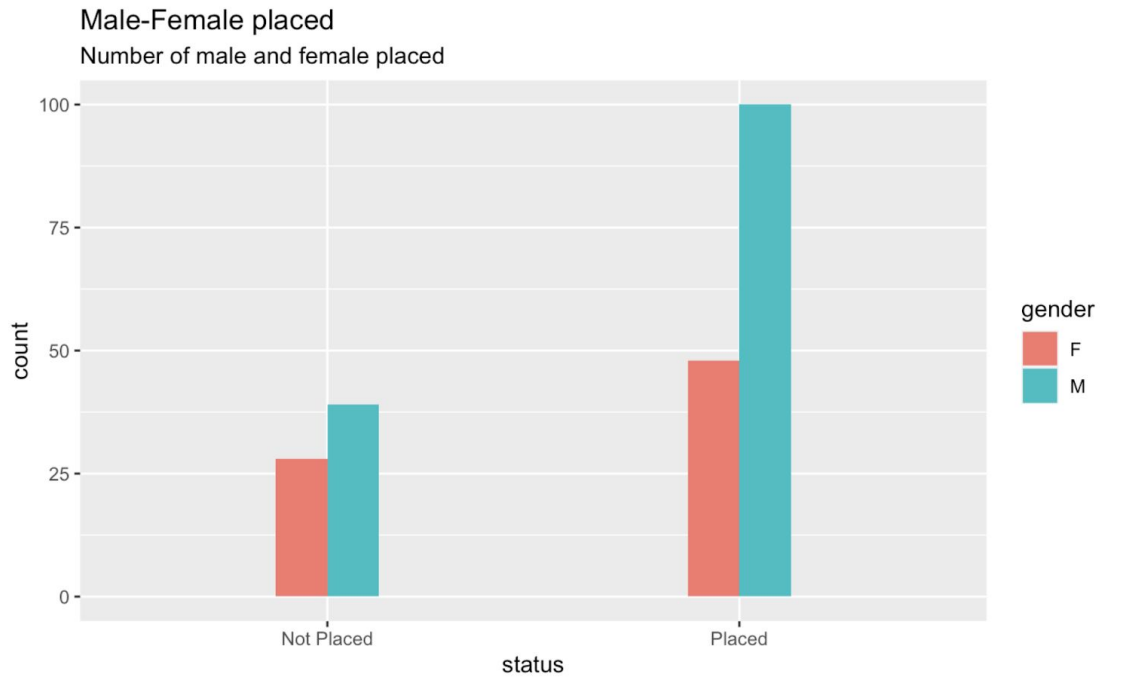
```
head(male_female_job)
```

status <chr>	gender <chr>	count <int>
Not Placed	F	28
Not Placed	M	39
Placed	F	48
Placed	M	100

4 rows

Hide

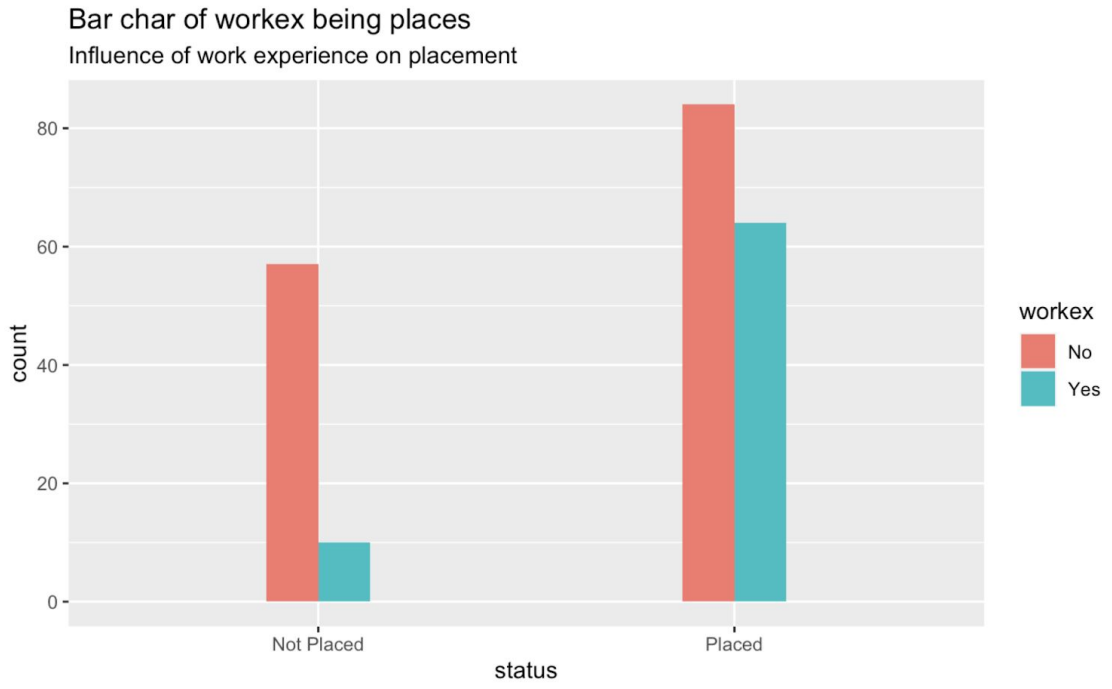
```
bar<- ggplot(data=data,aes(x=status))+geom_bar(width=0.25,aes(fill=gender),position=position_dodge())+labs(title
="Male-Female placed",subtitle="Number of male and female placed ")
bar
```



Males have higher chances of getting placed than females

2. Influence of work experience on placement

```
bar1<- ggplot(data=data,aes(x=status))+geom_bar(width=0.25,aes(fill=workex),position=position_dodge())+labs(title
="Bar char of workex being places",subtitle="Influence of work experience on placement")
bar1
```



Inference Drawn is that work expierence plays an important role in placement. People with work expierence are more likely to get placed

3. Proportion of different type of degrees who have been placed

Hide

```
degree_sci
```

status <chr>	degree_t <chr>	count <int>
Not Placed	Comm&Mgmt	43
Not Placed	Others	6
Not Placed	Sci&Tech	18
Placed	Comm&Mgmt	102
Placed	Others	5
Placed	Sci&Tech	41

6 rows

Hide

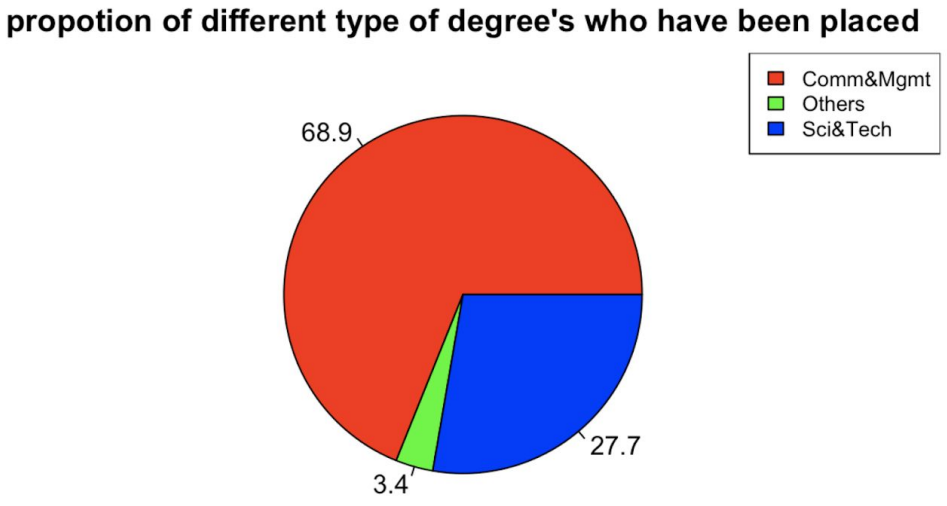
```
freq<-degree_sci%>%filter(status=="Placed")
freq
```

status <chr>	degree_t <chr>	count <int>
Placed	Comm&Mgmt	102
Placed	Others	5
Placed	Sci&Tech	41

3 rows

Hide

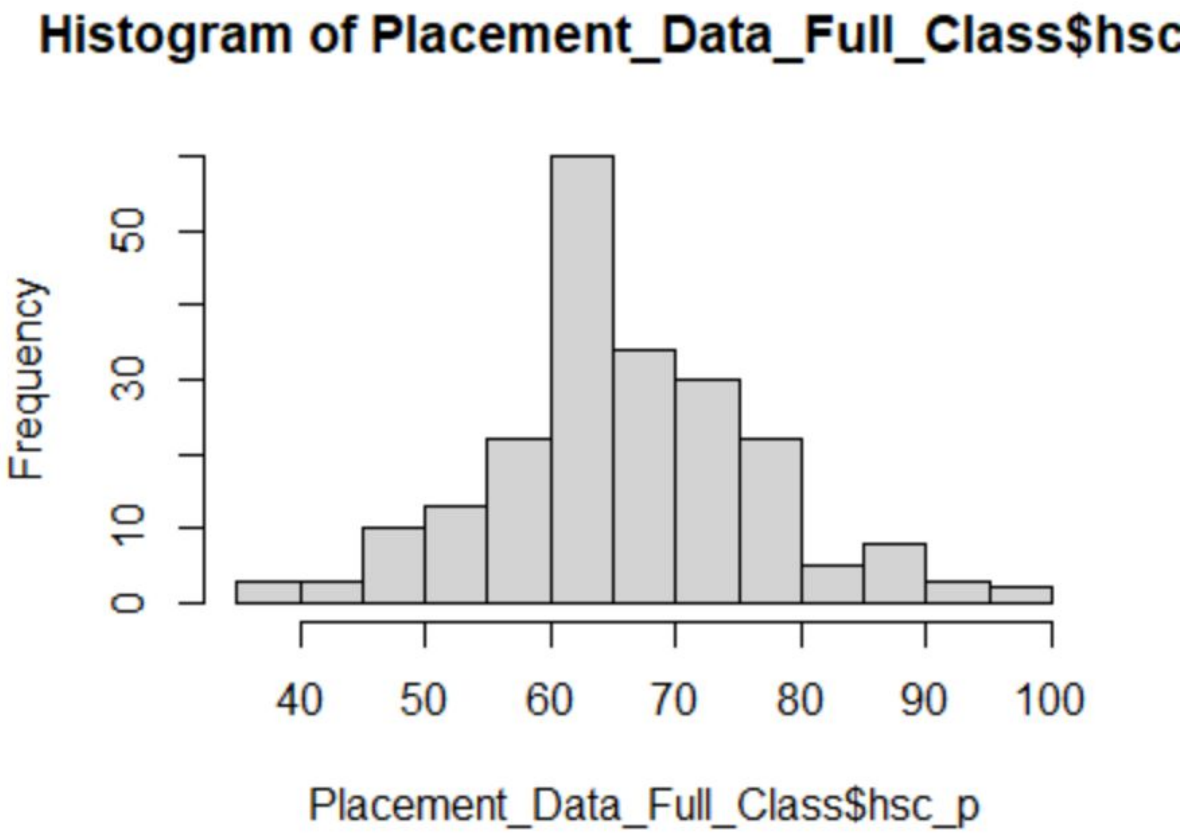
```
piepercent<- round(100*freq$count/sum(freq$count), 1)
pie(freq$count,labels=piepercent,main="propotion of different type of degree's who have been placed",col = rainbo
w(length(freq$count)))
legend("topright",freq$degree_t , cex = 0.8,fill = rainbow(length(freq$count)))
```



Inference: In this data set we see that Comm&Mgmt has more frequency of being placed

4.Histogram for High School Percentage

```
> hist(Placement_Data_Full_Class$hsc_p)
>
```

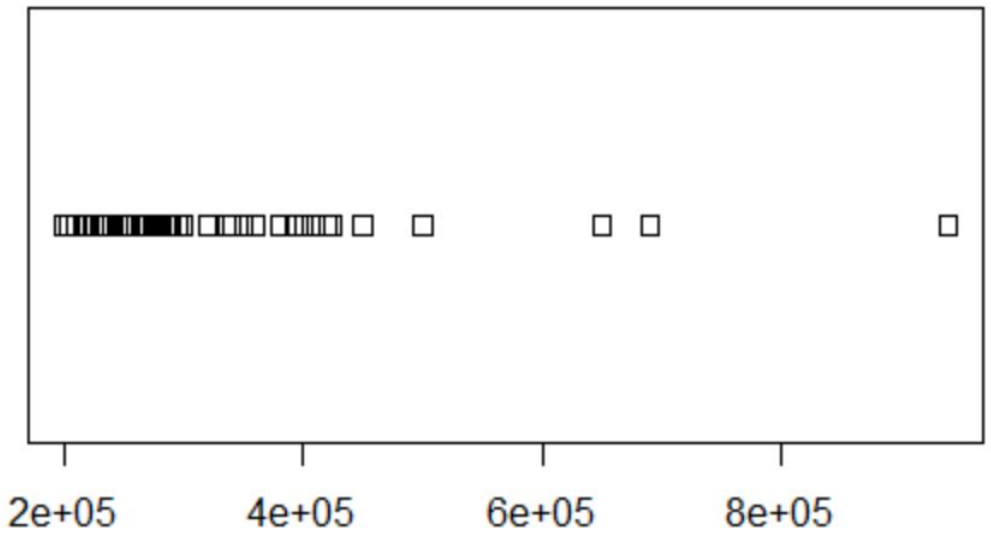


5.Correlation between salary and percentage

```
> sal<-Placement_Data_Full_Class$salary
> sscp<-Placement_Data_Full_Class$ssc_p
> m<-cor.test(sal,sscp,method = "pearson")
> m

        Pearson's product-moment correlation

data:  sal and sscp
t = 0.42716, df = 146, p-value = 0.6699
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.1267359  0.1955594
sample estimates:
       cor
0.03533034
```

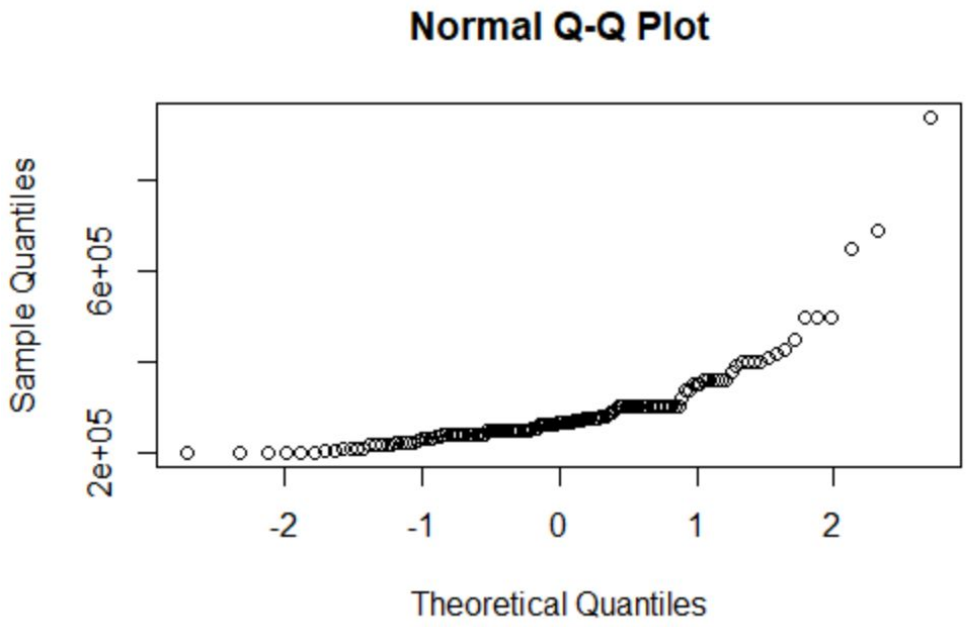
6.Strip plot for salary

```
> stripchart(Placement_Data_Full_Class$salary)
> qqnorm(Placement_Data_Full_Class$salary)
>
```



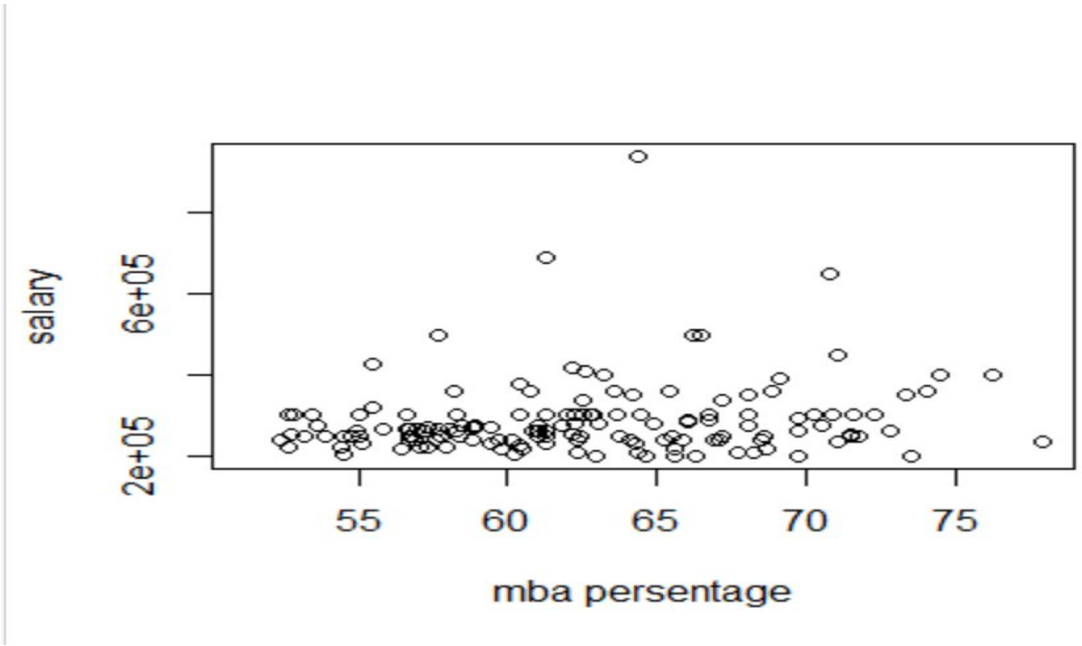
7.QQ plot for salary

```
> stripchart(Placement_Data_Full_Class$salary)
> qqnorm(Placement_Data_Full_Class$salary)
>
```



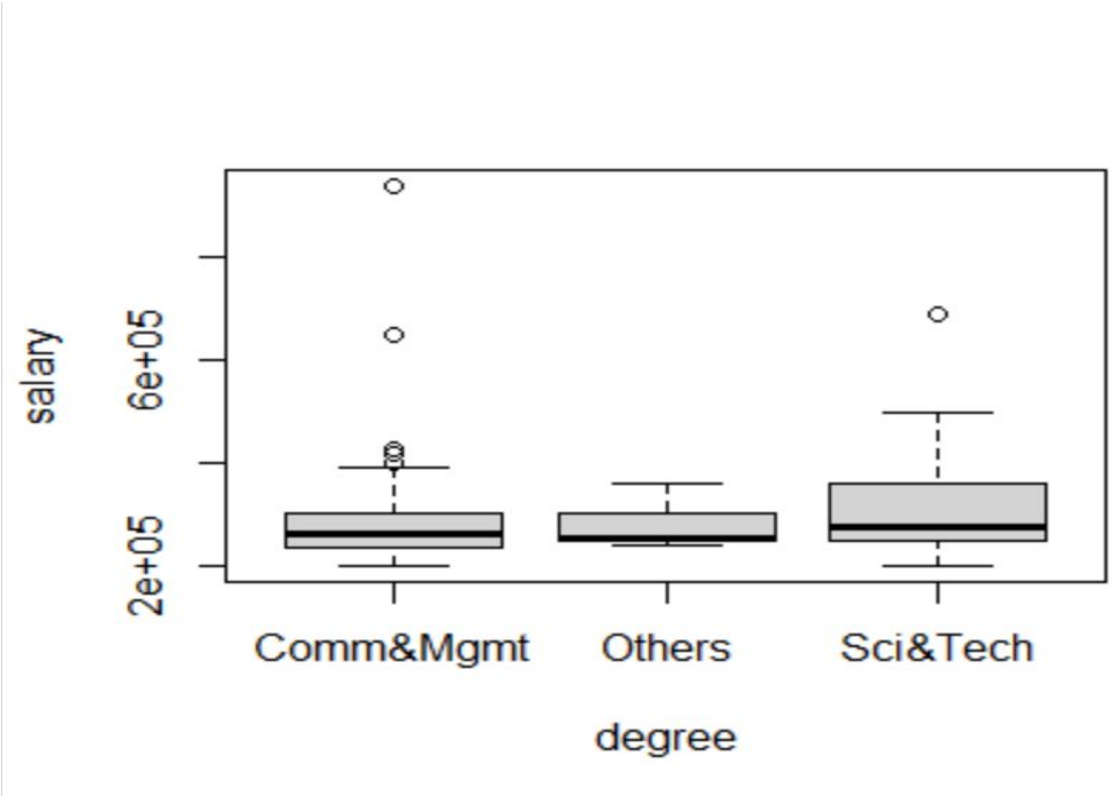
8.Scatter plot for salary and MBA percentage

```
> sal<-Placement_Data_Full_Class$salary
> mbap<-Placement_Data_Full_Class$mba_p
> plot(mbap,sal,xlab="mba persentage",ylab="salary")
>
```

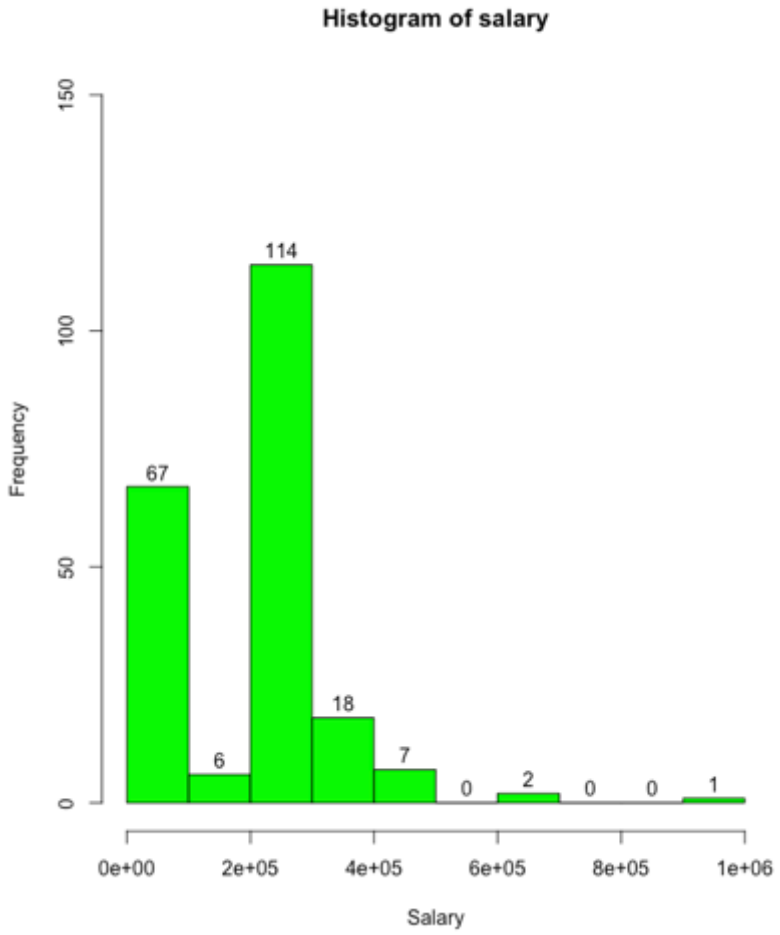


9.Box Plot for Salary and Degree Type

```
> sal<-Placement_Data_Full_Class$salary
> degree<-Placement_Data_Full_Class$degree_t
> boxplot(sal~degree,xlab="degree",ylab="salary")
>
```



10. Histogram for salary



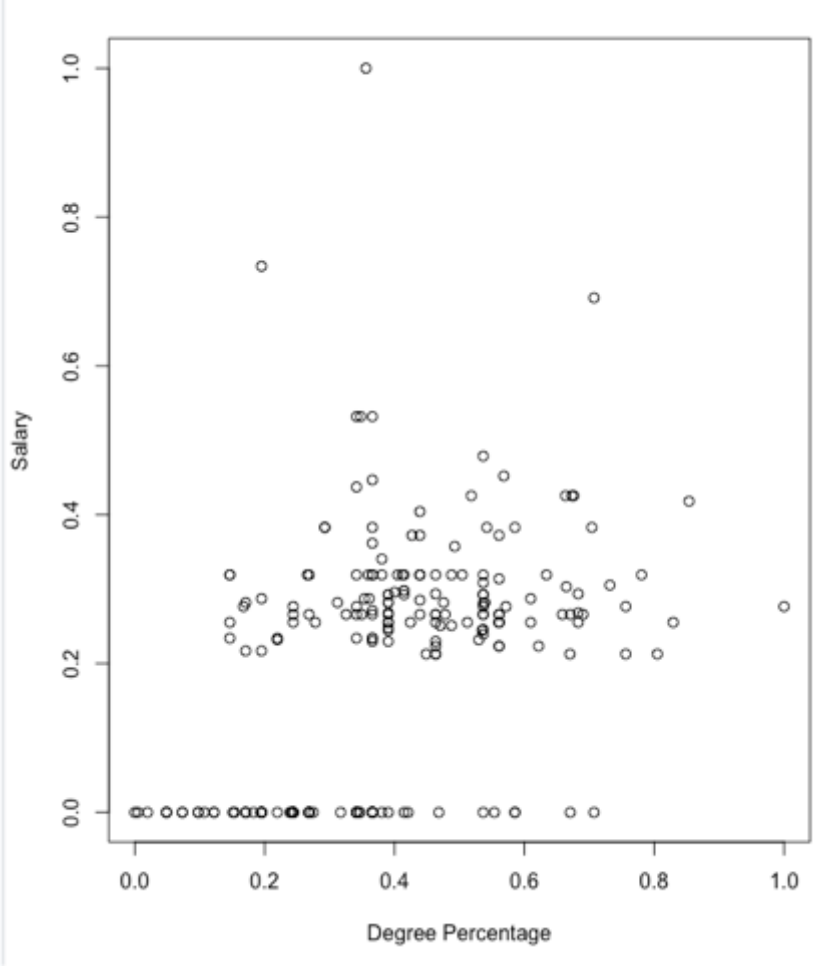
11. Skewness and Kurtosis of Salary attribute

```
> library(moments)
> salary<-data$salary
> skewness(salary)
[1] 0.4435234
> kurtosis(salary)
[1] 4.589852
>
```

Since the kurtosis value is greater than 3 , it is leptokurtic

12. Scatter plot between Salary and Degree Percentage

```
> normalize <- function(x) {
+     return ((x - min(x)) / (max(x) - min(x)))
+ }
> data$salary<-normalize(data$salary)
> data$degree_p<-normalize(data$degree_p)
> plot(data$degree_p,data$salary,xlab="Degree Percentage",ylab="Salary")
>
```



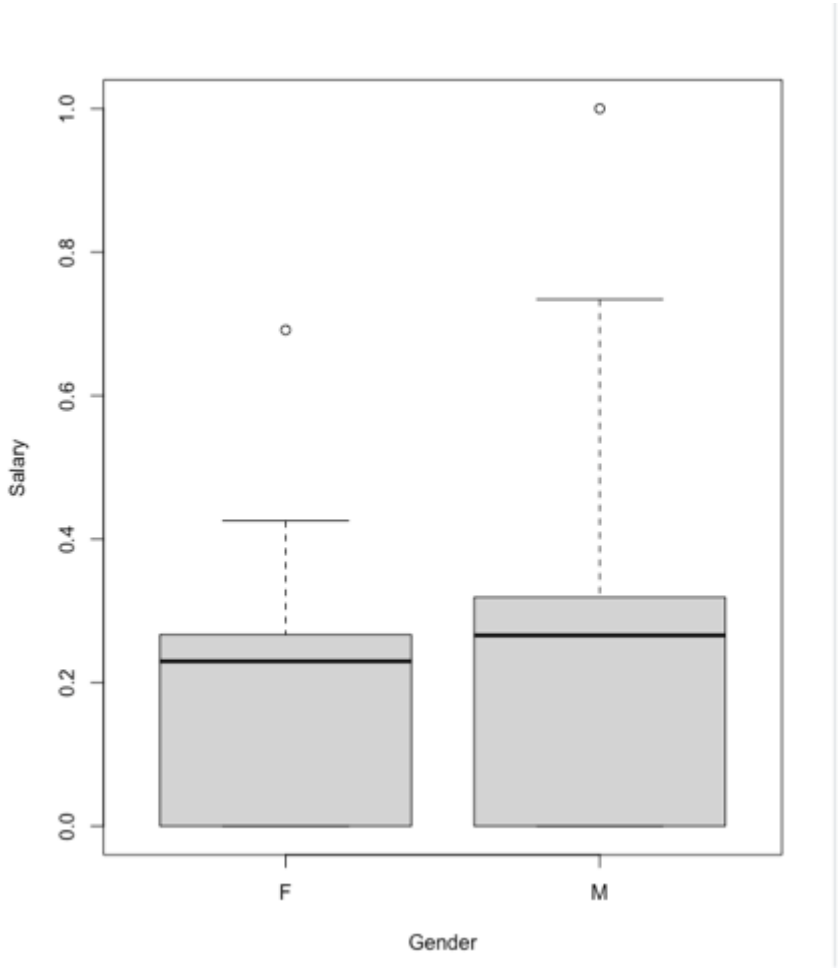
13. The correlation between degree percentage and salary offered

```
> salary<-data$salary
> pdegree<-data$degree_p
> c<-cor.test(salary,pdegree,method="pearson")
> c

        Pearson's product-moment correlation

data:  salary and pdegree
t = 6.5292, df = 213, p-value = 4.767e-10
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.2904370 0.5140841
sample estimates:
      cor
0.4083708

>
```

From the scatterplot and the r value ,we can infer that the degree percentage and salary have a positive relationship

14. Variation in the salary for male and female

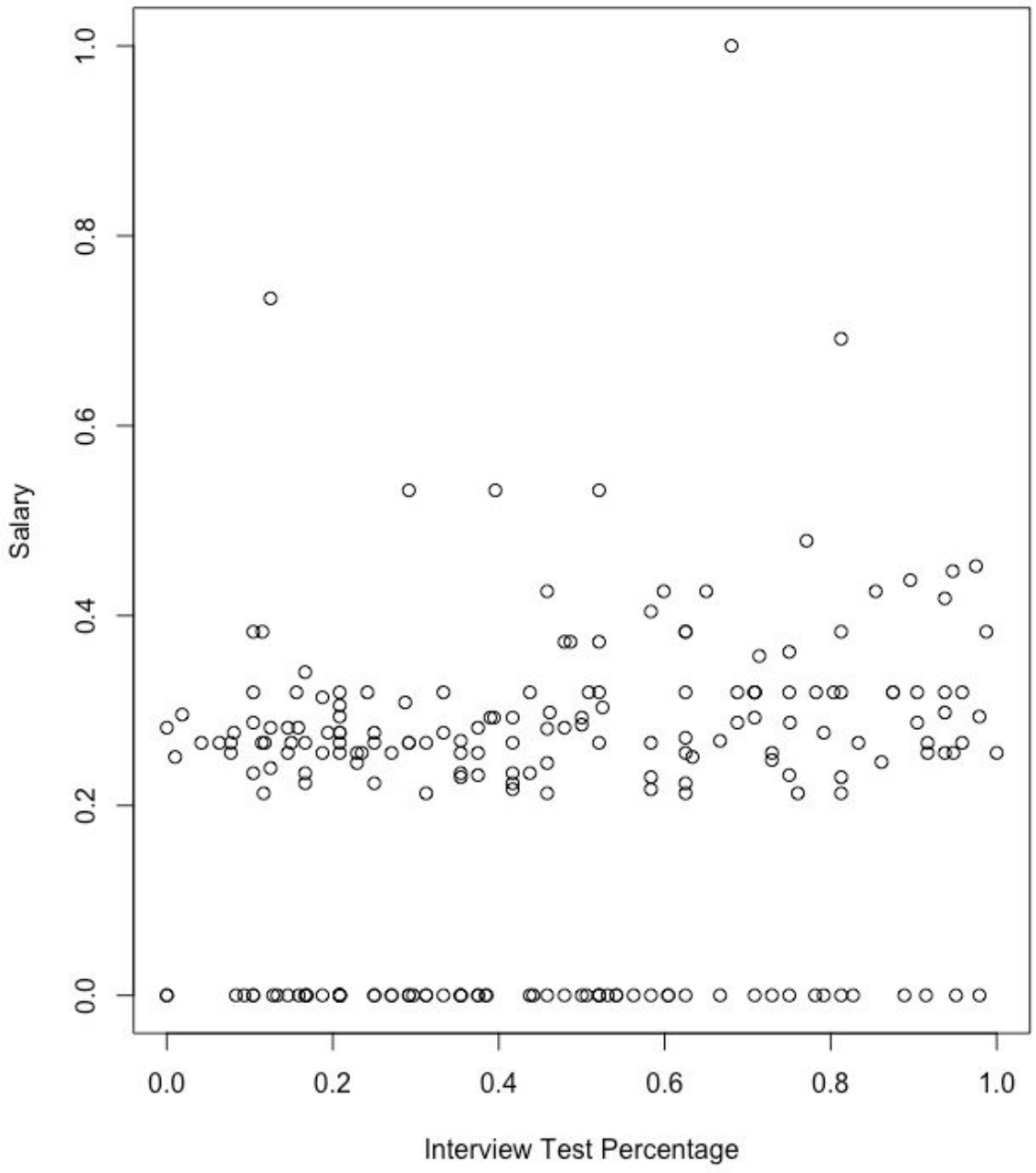
```
> salary<-data$salary
> gender<-data$gender
> boxplot(salary~gender,xlab="Gender",ylab="Salary")
>
```



From the above plot it is evident that there are not many outliers

15. Scatterplot between Interview test percentage versus the salary offered for each student

```
> salary<-data$salary
> etest<-data$etest_p
> plot(etest,salary,xlab="Interview Test Percentage",ylab="Salary")
>
```


Salary
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0
Interview Test Percentage

16. Correlation between interview test percentage and salary offered

```
> salary<-data$salary
> etest<-data$etest_p
> c<-cor.test(salary,etest,method="pearson")
> c

        Pearson's product-moment correlation

data:  salary and etest
t = 2.778, df = 213, p-value = 0.005958
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.0545488 0.3129612
sample estimates:
      cor
0.1869877

>
```

From the scatterplot and r value, we can see a positive relation between interview test percentage and salary . But the magnitude of r is less .

17. Which specialization is most common ?

```
> specialisation<-data$specialisation
> stable<-table(specialisation)
> names(stable)[which(stable==max(stable))]
[1] "Mkt&Fin"
>
```

The mode of the specialization attribute is found out.
Thus Marketing and Finance is the most common specialization .

Certain initial insights gained after performing Exploratory Data Analysis :

1. From the graph plotted , it can be observed that males have a higher chance of getting placed though it is only one of the factors.
2. In the early stages , looks like people with work experience have higher chance of being placed as against freshers.
3. In this particular dataset, Comm&Mngt have higher frequency of getting placed , though this could be due to the data collection method but as fas as the initial Exploratory Data Analysis is concerned , this is the trend observed.
4. Visualising the distribution of the salary attribute, the skewness and kurtosis value is found out . The kurtosis value is greater than 3 . Thus it is leptokurtic.
5. A scatter plot is plotted between salary and percentage , and the correlation coefficient value is found to be 0.4083 indicating a positive relation.
6. A boxplot is plotted to understand the variation in salary for male and female. From the plot, it is evident there are hardly any outliers .
7. The mode of the specialization attribute is found out and it can be seen that Marketing and Finance is the most common specialization.
8. There is a positive correlation between salary and secondary school percentage.
9. The scatterplot between MBA percentage and salary show no patterns.
10. There are no outliers in other degree types but there are outliers present in comm&mgmt and in sci&tech.