Exploratory Data Analysis

Importing the libraries

library(moments)  
library(dplyr)

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

library(DescTools)  
library(ggplot2)  
library(readr)

Reading the data set

data<-read.csv("Placement\_Data\_Full\_Class.csv")  
head(data)

## sl\_no gender ssc\_p ssc\_b hsc\_p hsc\_b hsc\_s degree\_p degree\_t workex  
## 1 1 M 67.00 Others 91.00 Others Commerce 58.00 Sci&Tech No  
## 2 2 M 79.33 Central 78.33 Others Science 77.48 Sci&Tech Yes  
## 3 3 M 65.00 Central 68.00 Central Arts 64.00 Comm&Mgmt No  
## 4 4 M 56.00 Central 52.00 Central Science 52.00 Sci&Tech No  
## 5 5 M 85.80 Central 73.60 Central Commerce 73.30 Comm&Mgmt No  
## 6 6 M 55.00 Others 49.80 Others Science 67.25 Sci&Tech Yes  
## etest\_p specialisation mba\_p status salary  
## 1 55.0 Mkt&HR 58.80 Placed 270000  
## 2 86.5 Mkt&Fin 66.28 Placed 200000  
## 3 75.0 Mkt&Fin 57.80 Placed 250000  
## 4 66.0 Mkt&HR 59.43 Not Placed NA  
## 5 96.8 Mkt&Fin 55.50 Placed 425000  
## 6 55.0 Mkt&Fin 51.58 Not Placed NA

Summary and Glimpse of Data

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,…  
## $ gender <chr> "M", "M", "M", "M", "M", "M", "F", "M", "M", "M", "M",…  
## $ ssc\_p <dbl> 67.00, 79.33, 65.00, 56.00, 85.80, 55.00, 46.00, 82.00…  
## $ ssc\_b <chr> "Others", "Central", "Central", "Central", "Central", …  
## $ hsc\_p <dbl> 91.00, 78.33, 68.00, 52.00, 73.60, 49.80, 49.20, 64.00…  
## $ hsc\_b <chr> "Others", "Others", "Central", "Central", "Central", "…  
## $ hsc\_s <chr> "Commerce", "Science", "Arts", "Science", "Commerce", …  
## $ degree\_p <dbl> 58.00, 77.48, 64.00, 52.00, 73.30, 67.25, 79.00, 66.00…  
## $ degree\_t <chr> "Sci&Tech", "Sci&Tech", "Comm&Mgmt", "Sci&Tech", "Comm…  
## $ workex <chr> "No", "Yes", "No", "No", "No", "Yes", "No", "Yes", "No…  
## $ etest\_p <dbl> 55.00, 86.50, 75.00, 66.00, 96.80, 55.00, 74.28, 67.00…  
## $ specialisation <chr> "Mkt&HR", "Mkt&Fin", "Mkt&Fin", "Mkt&HR", "Mkt&Fin", "…  
## $ mba\_p <dbl> 58.80, 66.28, 57.80, 59.43, 55.50, 51.58, 53.29, 62.14…  
## $ status <chr> "Placed", "Placed", "Placed", "Not Placed", "Placed", …  
## $ salary <int> 270000, 200000, 250000, NA, 425000, NA, NA, 252000, 23…

Data Cleaning

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

## sl\_no gender ssc\_p ssc\_b hsc\_p hsc\_b hsc\_s degree\_p degree\_t workex  
## 1 1 M 67.00 Others 91.00 Others Commerce 58.00 Sci&Tech No  
## 2 2 M 79.33 Central 78.33 Others Science 77.48 Sci&Tech Yes  
## 3 3 M 65.00 Central 68.00 Central Arts 64.00 Comm&Mgmt No  
## 4 4 M 56.00 Central 52.00 Central Science 52.00 Sci&Tech No  
## 5 5 M 85.80 Central 73.60 Central Commerce 73.30 Comm&Mgmt No  
## 6 6 M 55.00 Others 49.80 Others Science 67.25 Sci&Tech Yes  
## etest\_p specialisation mba\_p status salary  
## 1 55.0 Mkt&HR 58.80 Placed 270000  
## 2 86.5 Mkt&Fin 66.28 Placed 200000  
## 3 75.0 Mkt&Fin 57.80 Placed 250000  
## 4 66.0 Mkt&HR 59.43 Not Placed 0  
## 5 96.8 Mkt&Fin 55.50 Placed 425000  
## 6 55.0 Mkt&Fin 51.58 Not Placed 0

Data Normalization

normalize<-function(attribute){ return ( (attribute - min(attribute))/(max(attribute)-min(attribute))) }  
data$ssc\_p<-normalize(data$ssc\_p)  
data$hsc\_p<-normalize(data$hsc\_p)  
data$degree\_p<-normalize(data$degree\_p)  
data$etest\_p<-normalize(data$etest\_p)  
data$mba\_p<-normalize(data$mba\_p)  
data$salary<-normalize(data$salary)  
head(data)

## sl\_no gender ssc\_p ssc\_b hsc\_p hsc\_b hsc\_s degree\_p  
## 1 1 M 0.5382395 Others 0.8896211 Others Commerce 0.19512195  
## 2 2 M 0.7924139 Central 0.6808896 Others Science 0.67024390  
## 3 3 M 0.4970109 Central 0.5107084 Central Arts 0.34146341  
## 4 4 M 0.3114822 Central 0.2471170 Central Science 0.04878049  
## 5 5 M 0.9257885 Central 0.6029654 Central Commerce 0.56829268  
## 6 6 M 0.2908679 Others 0.2108731 Others Science 0.42073171  
## degree\_t workex etest\_p specialisation mba\_p status salary  
## 1 Sci&Tech No 0.1041667 Mkt&HR 0.28448276 Placed 0.2872340  
## 2 Sci&Tech Yes 0.7604167 Mkt&Fin 0.56484258 Placed 0.2127660  
## 3 Comm&Mgmt No 0.5208333 Mkt&Fin 0.24700150 Placed 0.2659574  
## 4 Sci&Tech No 0.3333333 Mkt&HR 0.30809595 Not Placed 0.0000000  
## 5 Comm&Mgmt No 0.9750000 Mkt&Fin 0.16079460 Placed 0.4521277  
## 6 Sci&Tech Yes 0.1041667 Mkt&Fin 0.01386807 Not Placed 0.0000000

Exploratory Data Analysis

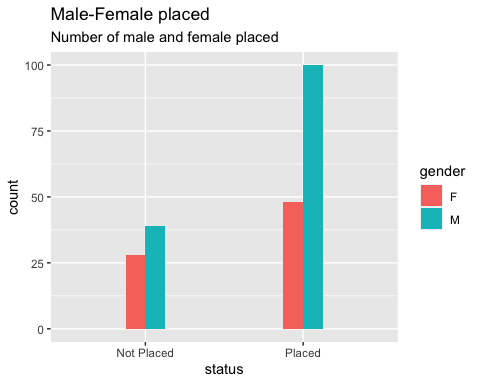
male\_female\_job<-data%>%group\_by(status,gender)%>%summarize(count=n())

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

head(male\_female\_job)

## # A tibble: 4 x 3  
## # Groups: status [2]  
## status gender count  
## <chr> <chr> <int>  
## 1 Not Placed F 28  
## 2 Not Placed M 39  
## 3 Placed F 48  
## 4 Placed M 100

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



bar1<-ggplot(data=data,aes(x=status))+geom\_bar(width=0.25,aes(fill=workex),position=position\_dodge())+labs(title="Bar chart of workex being places",subtitle="Influence of work experience on placement")  
bar1



degree\_sci<-data%>%group\_by(status,degree\_t)%>%summarize(count=n())

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

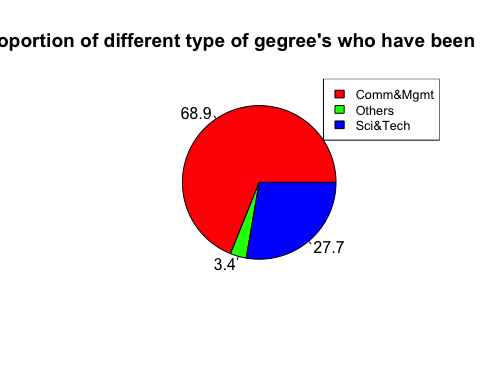
degree\_sci

## # A tibble: 6 x 3  
## # Groups: status [2]  
## status degree\_t count  
## <chr> <chr> <int>  
## 1 Not Placed Comm&Mgmt 43  
## 2 Not Placed Others 6  
## 3 Not Placed Sci&Tech 18  
## 4 Placed Comm&Mgmt 102  
## 5 Placed Others 5  
## 6 Placed Sci&Tech 41

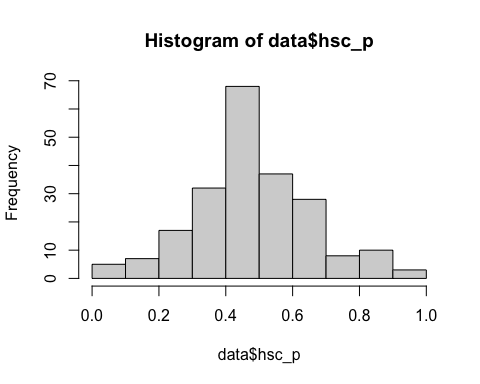
freq<-degree\_sci%>%filter(status=="Placed")  
freq

## # A tibble: 3 x 3  
## # Groups: status [1]  
## status degree\_t count  
## <chr> <chr> <int>  
## 1 Placed Comm&Mgmt 102  
## 2 Placed Others 5  
## 3 Placed Sci&Tech 41

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



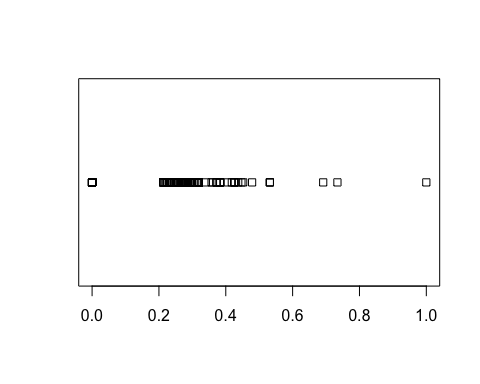
hist(data$hsc\_p)



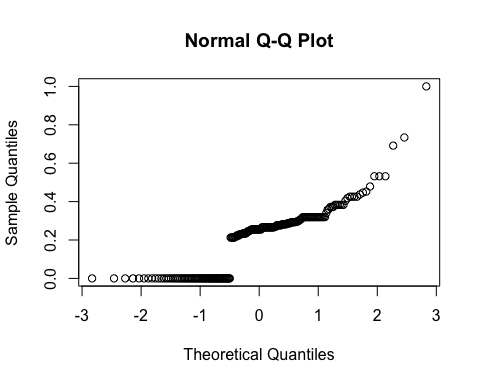
sal<-data$salary  
sscp<-data$ssc\_p  
m<-cor.test(sal,sscp,ethod="pearson")  
m

##   
## Pearson's product-moment correlation  
##   
## data: sal and sscp  
## t = 9.317, df = 213, p-value < 2.2e-16  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## 0.4356523 0.6267672  
## sample estimates:  
## cor   
## 0.5380897

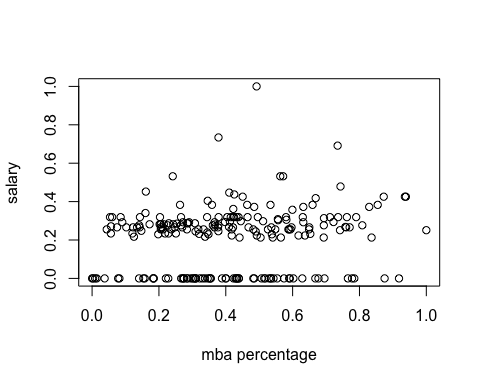
stripchart(data$salary)



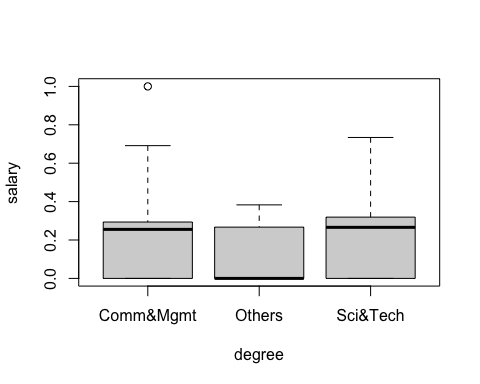
qqnorm(data$salary)



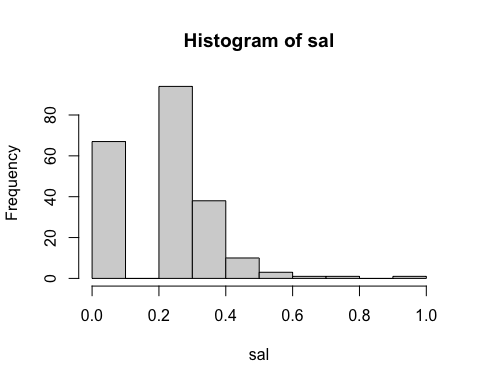
mbap<-data$mba\_p  
sal<-data$salary  
plot(mbap,sal,xlab="mba percentage",ylab="salary")



degree<-data$degree\_t  
boxplot(sal~degree,xlab="degree",ylab="salary")



hist(sal)



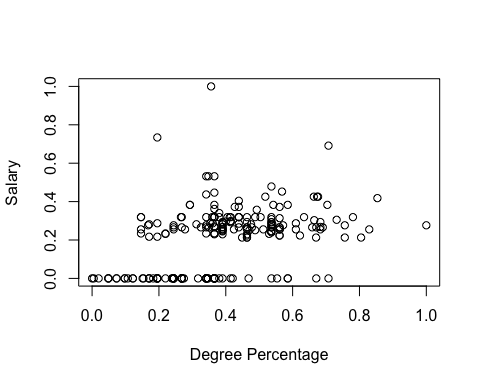
skewness(sal)

## [1] 0.4435234

kurtosis(sal)

## [1] 4.589852

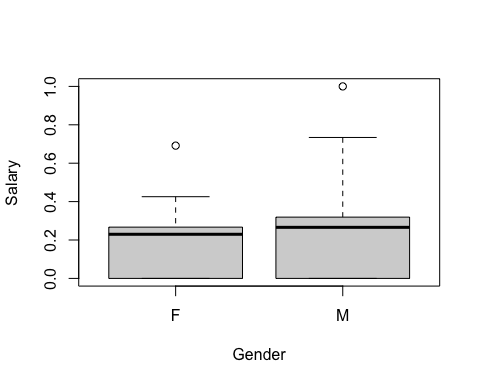
plot(data$degree\_p,data$salary,xlab="Degree Percentage", ylab="Salary")



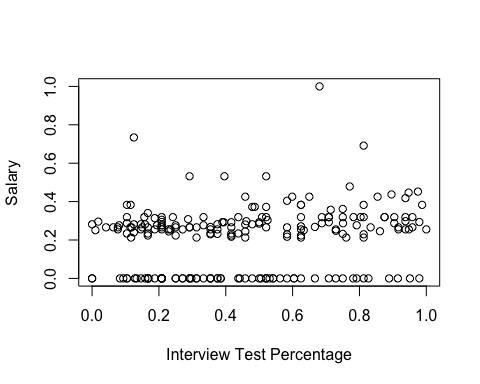
c<-cor.test(data$salary,data$degree\_p,method="pearson")  
c

##   
## Pearson's product-moment correlation  
##   
## data: data$salary and data$degree\_p  
## 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

boxplot(data$salary~data$gender,xlab="Gender",ylab="Salary")



salary<-data$salary  
etest<-data$etest\_p  
plot(etest,salary,xlab="Interview Test Percentage",ylab="Salary")



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

specialisation<-data$specialisation  
stable<-table(specialisation)  
names(stable)[which(stable==max(stable))]

## [1] "Mkt&Fin"