```r
library(tidyverse)

library(dplyr)

library(MASS)

library(ggplot2)

library(GGally)

library(DMwR)

library(car)

library(e1071)

library(caret)

library(cowplot)

library(caTools)

library(pROC)

library(ggcorrplot)

library(lattice)

library(sm)

library(Hmisc)

library(asbio)

library(MVA)

library(Hotelling)

library(Amelia)

library(grid)

library(gridExtra)

library(PerformanceAnalytics)

##Importing Data and inital analyses

#Importing csv file from a location

attr<- read.csv("C:/WD Jimit/MITA Spring 19/Ronak Parrikh/Multivariate Analysis/Dataset/HR-Employee-Attrition.csv")

attr <- as.data.frame(attr)

glimpse(attr)
```

```r
#Dimension of the dataset
dim(attr)

#View the first 5 rows of the dataset
head(attr)
summary(attr)
#Rename the Age column
colnames(attr)[1] <- "Age"
#Calculating the number of null values in each of the columns
colSums(sapply(attr,is.na))
missmap(attr,main="Missing Values VS Observed")
#Removing redundant columns
attr$EmployeeNumber<- NULL
attr$StandardHours <- NULL
attr$Over18 <- NULL
attr$EmployeeCount <- NULL
#Converting data type of categorical column
attr$Education <- factor(attr$Education)
attr$EnvironmentSatisfaction <- factor(attr$EnvironmentSatisfaction)
attr$JobInvolvement <- factor(attr$JobInvolvement)
attr$JobLevel <- factor(attr$JobLevel)
attr$JobSatisfaction <- factor(attr$JobSatisfaction)
attr$PerformanceRating <- factor(attr$PerformanceRating)
attr$RelationshipSatisfaction <- factor(attr$RelationshipSatisfaction)
attr$StockOptionLevel <- factor(attr$StockOptionLevel)
attr$WorkLifeBalance <- factor(attr$WorkLifeBalance)
#Assigning categorical and numerical variable to temporary variable
catvar<-c('BusinessTravel','Department','Education','EducationField','EnvironmentSatisfaction','Gender',
```

```
        'JobRole','JobInvolvement','JobLevel','JobSatisfaction',

'MaritalStatus','PerformanceRating','RelationshipSatisfaction','StockOptionLevel','WorkLifeBalance')
numvar<-c('Age','DailyRate','DistanceFromHome','HourlyRate',
        'MonthlyIncome','MonthlyRate','NumCompaniesWorked','PercentSalaryHike','TotalWorkingYears',
        'TrainingTimesLastYear','YearsAtCompany',
        'YearsInCurrentRole','YearsSinceLastPromotion','YearsWithCurrManager')
##Exploratory Data Analysis

#Vizualization of Attrition
attr %>%
    group_by(Attrition) %>%
    tally() %>%
    ggplot(aes(x =Attrition,y = n,fill=Attrition)) +
    geom_bar(stat = "identity") +
    theme_minimal()+
    labs(x="Attrition", y="Count of Attrition")+
    ggtitle("Attrition")+
    geom_text(aes(label = n), vjust = -0.5, position = position_dodge(0.9))

#Influence of features on Attrition
ggplot(data=attr, aes(attr$Age)) +
  geom_histogram(breaks=seq(20, 50, by=2),
              col="red",
              aes(fill=..count..))+
  labs(x="Age", y="Count")+
  scale_fill_gradient("Count", low="yellow", high="dark red")
#Checking for distributions in numerical columns
#The qqPlot show a few extreme outliers which break the assumption of 95% confidence
```

```
#normal distribution
par(mfrow = c(1,2))
hist(attr$Age,xlab='',main = 'Histogram of Age',freq = FALSE)
lines(density(attr$Age,na.rm = T))
rug(jitter(attr$Age))
qqPlot(attr$Age,main='Normal QQ plot of Age')
par(mfrow=c(1,1))

par(mfrow = c(1,2))
hist(attr$DailyRate,xlab='',main = 'Histogram of DailyRate',freq = FALSE)
lines(density(attr$DailyRate,na.rm = T))
rug(jitter(attr$DailyRate))
qqPlot(attr$DailyRate,main='Normal QQ plot of DailyRate')
par(mfrow=c(1,1))

par(mfrow = c(1,2))
hist(attr$DistanceFromHome,xlab='',main = 'Histogram of DistanceFromHome',freq = FALSE)
lines(density(attr$DistanceFromHome,na.rm = T))
rug(jitter(attr$DistanceFromHome))
qqPlot(attr$DistanceFromHome,main='Normal QQ plot of DistanceFromHome')
par(mfrow=c(1,1))

par(mfrow = c(1,2))
hist(attr$HourlyRate,xlab='',main = 'Histogram of HourlyRate',freq = FALSE)
lines(density(attr$HourlyRate,na.rm = T))
rug(jitter(attr$HourlyRate))
qqPlot(attr$HourlyRate,main='Normal QQ plot of HourlyRate')
par(mfrow=c(1,1))
```

```
par(mfrow = c(1,2))
hist(attr$MonthlyIncome,xlab='',main = 'Histogram of Monthly Income',freq = FALSE)
lines(density(attr$MonthlyIncome,na.rm = T))
rug(jitter(attr$MonthlyIncome))
qqPlot(attr$MonthlyIncome,main='Normal QQ plot of Monthly Income')
par(mfrow=c(1,1))


par(mfrow = c(1,2))
hist(attr$NumCompaniesWorked,xlab='',main = 'Histogram of NumCompaniesWorked',freq = FALSE)
lines(density(attr$NumCompaniesWorked,na.rm = T))
rug(jitter(attr$NumCompaniesWorked))
qqPlot(attr$NumCompaniesWorked,main='Normal QQ plot of NumCompaniesWorked')
par(mfrow=c(1,1))

par(mfrow = c(1,2))
hist(attr$PercentSalaryHike,xlab='',main = 'Histogram of PercentSalaryHike',freq = FALSE)
lines(density(attr$PercentSalaryHike,na.rm = T))
rug(jitter(attr$PercentSalaryHike))
qqPlot(attr$PercentSalaryHike,main='Normal QQ plot of PercentSalaryHike')
par(mfrow=c(1,1))

par(mfrow = c(1,2))
hist(attr$TrainingTimesLastYear,xlab='',main = 'Histogram of TrainingTimesLastYear',freq = FALSE)
lines(density(attr$TrainingTimesLastYear,na.rm = T))
rug(jitter(attr$TrainingTimesLastYear))
qqPlot(attr$TrainingTimesLastYear,main='Normal QQ plot of TrainingTimesLastYear')
par(mfrow=c(1,1))
```

```
par(mfrow = c(1,2))
hist(attr$YearsAtCompany,xlab='',main = 'Histogram of YearsAtCompany',freq = FALSE)
lines(density(attr$YearsAtCompany,na.rm = T))
rug(jitter(attr$YearsAtCompany))
qqPlot(attr$YearsAtCompany,main='Normal QQ plot of YearsAtCompany')
par(mfrow=c(1,1))

par(mfrow = c(1,2))
hist(attr$YearsInCurrentRole,xlab='',main = 'Histogram of YearsInCurrentRole',freq = FALSE)
lines(density(attr$YearsInCurrentRole,na.rm = T))
rug(jitter(attr$YearsInCurrentRole))
qqPlot(attr$YearsInCurrentRole,main='Normal QQ plot of YearsInCurrentRole')
par(mfrow=c(1,1))

par(mfrow = c(1,2))
hist(attr$YearsSinceLastPromotion,xlab='',main = 'Histogram of YearsSinceLastPromotion',freq = FALSE)
lines(density(attr$YearsSinceLastPromotion,na.rm = T))
rug(jitter(attr$YearsSinceLastPromotion))
qqPlot(attr$YearsSinceLastPromotion,main='Normal QQ plot of YearsSinceLastPromotion')
par(mfrow=c(1,1))

par(mfrow = c(1,2))
hist(attr$YearsWithCurrManager,xlab='',main = 'Histogram of YearsWithCurrManager',freq = FALSE)
lines(density(attr$YearsWithCurrManager,na.rm = T))
rug(jitter(attr$YearsWithCurrManager))
qqPlot(attr$YearsWithCurrManager,main='Normal QQ plot of YearsWithCurrManager')
par(mfrow=c(1,1))

#Boxplot distributions for our numeric columns
```

```
#The dashed line shows the mean and the dark center line shows the median
#Difference between these two lines depict the deviation from the central limit theorem
#Boxplot distributions for  Age
boxplot(attr$Age, ylab = "Age")
rug(jitter(attr$Age), side = 2)
abline(h = mean(attr$Age, na.rm = T), lty = 2)
#Plotting the Age  with 3 lines for mean, median and mean+std
plot(attr$Age, xlab = "")
abline(h = mean(attr$Age, na.rm = T), lty = 1)
abline(h = mean(attr$Age, na.rm = T) + sd(attr$Age, na.rm = T),lty = 2)
abline(h = median(attr$Age, na.rm = T), lty = 3)
identify(attr$Age)

#Boxplot distributions for Daily rate
boxplot(attr$DailyRate, ylab = "DailyRate",outline = TRUE)
rug(jitter(attr$DailyRate), side = 2)
abline(h = mean(attr$DailyRate, na.rm = T), lty = 2)
#Plotting the DailyRate  with 3 lines for mean, median and mean+std
plot(attr$DailyRate, xlab = "")
abline(h = mean(attr$DailyRate, na.rm = T), lty = 1)
abline(h = mean(attr$DailyRate, na.rm = T) + sd(attr$DailyRate, na.rm = T),lty = 2)
abline(h = median(attr$DailyRate, na.rm = T), lty = 3)
identify(attr$DailyRate)

#Boxplot distributions for Distance from home
boxplot(attr$DistanceFromHome, ylab = "DistanceFromHome",outline = TRUE)
rug(jitter(attr$DistanceFromHome), side = 2)
abline(h = mean(attr$DistanceFromHome, na.rm = T), lty = 2)
#Plotting the Distance from home  with 3 lines for mean, median and mean+std
```

plot(attr$DistanceFromHome, xlab = "")

abline(h = mean(attr$DistanceFromHome, na.rm = T), lty = 1)

abline(h = mean(attr$DistanceFromHome, na.rm = T) + sd(attr$DistanceFromHome, na.rm = T),lty = 2)

abline(h = median(attr$DistanceFromHome, na.rm = T), lty = 3)

identify(attr$DistanceFromHome)

#Boxplot distributions for Monthly Income

boxplot(attr$MonthlyIncome, ylab = "Monthly Income")

rug(jitter(attr$MonthlyIncome), side = 2)

abline(h = mean(attr$MonthlyIncome, na.rm = T), lty = 2)

#Plotting the Monthly Income and Age with 3 lines for mean, median and mean+std

plot(attr$MonthlyIncome, xlab = "")

abline(h = mean(attr$MonthlyIncome, na.rm = T), lty = 1)

abline(h = mean(attr$MonthlyIncome, na.rm = T) + sd(attr$MonthlyIncome, na.rm = T),lty = 2)

abline(h = median(attr$MonthlyIncome, na.rm = T), lty = 3)

identify(attr$MonthlyIncome)

#Boxplot distributions for NumCompaniesWorked

boxplot(attr$NumCompaniesWorked, ylab = "NumCompaniesWorked")

rug(jitter(attr$NumCompaniesWorked), side = 2)

abline(h = mean(attr$NumCompaniesWorked, na.rm = T), lty = 2)

#Plotting the NumCompaniesWorked with 3 lines for mean, median and mean+std

plot(attr$NumCompaniesWorked, xlab = "")

abline(h = mean(attr$NumCompaniesWorked, na.rm = T), lty = 1)

abline(h = mean(attr$NumCompaniesWorked, na.rm = T) + sd(attr$NumCompaniesWorked, na.rm = T),lty = 2)

abline(h = median(attr$NumCompaniesWorked, na.rm = T), lty = 3)

identify(attr$NumCompaniesWorked)

```
#Boxplot distributions for  PercentSalaryHike

boxplot(attr$PercentSalaryHike, ylab = "PercentSalaryHike")

rug(jitter(attr$PercentSalaryHike), side = 2)

abline(h = mean(attr$PercentSalaryHike, na.rm = T), lty = 2)

#Plotting the PercentSalaryHike  with 3 lines for mean, median and mean+std

plot(attr$PercentSalaryHike, xlab = "")

abline(h = mean(attr$PercentSalaryHike, na.rm = T), lty = 1)

abline(h = mean(attr$PercentSalaryHike, na.rm = T) + sd(attr$PercentSalaryHike, na.rm = T),lty = 2)

abline(h = median(attr$PercentSalaryHike, na.rm = T), lty = 3)

identify(attr$PercentSalaryHike)




#Boxplot distributions for  TotalWorkingYears

boxplot(attr$TotalWorkingYears, ylab = "TotalWorkingYears")

rug(jitter(attr$TotalWorkingYears), side = 2)

abline(h = mean(attr$TotalWorkingYears, na.rm = T), lty = 2)

#Plotting the TotalWorkingYears  with 3 lines for mean, median and mean+std

plot(attr$TotalWorkingYears, xlab = "")

abline(h = mean(attr$TotalWorkingYears, na.rm = T), lty = 1)

abline(h = mean(attr$TotalWorkingYears, na.rm = T) + sd(attr$TotalWorkingYears, na.rm = T),lty = 2)

abline(h = median(attr$TotalWorkingYears, na.rm = T), lty = 3)

identify(attr$TotalWorkingYears)


#Boxplot distributions for  TrainingTimesLastYear

boxplot(attr$TrainingTimesLastYear, ylab = "TrainingTimesLastYear")

rug(jitter(attr$TrainingTimesLastYear), side = 2)

abline(h = mean(attr$TrainingTimesLastYear, na.rm = T), lty = 2)

#Plotting the TrainingTimesLastYear  with 3 lines for mean, median and mean+std
```

```
plot(attr$TrainingTimesLastYear, xlab = "")

abline(h = mean(attr$TrainingTimesLastYear, na.rm = T), lty = 1)

abline(h = mean(attr$TrainingTimesLastYear, na.rm = T) + sd(attr$TrainingTimesLastYear, na.rm = T),lty = 2)

abline(h = median(attr$TrainingTimesLastYear, na.rm = T), lty = 3)

identify(attr$TrainingTimesLastYear)


#Boxplot distributions for  YearsAtCompany

boxplot(attr$YearsAtCompany, ylab = "YearsAtCompany")

rug(jitter(attr$YearsAtCompany), side = 2)

abline(h = mean(attr$YearsAtCompany, na.rm = T), lty = 2)

#Plotting the Years at Company  with 3 lines for mean, median and mean+std

plot(attr$YearsAtCompany, xlab = "")

abline(h = mean(attr$YearsAtCompany, na.rm = T), lty = 1)

abline(h = mean(attr$YearsAtCompany, na.rm = T) + sd(attr$YearsAtCompany, na.rm = T),lty = 2)

abline(h = median(attr$YearsAtCompany, na.rm = T), lty = 3)

identify(attr$YearsAtCompany)


#Boxplot distributions for  YearsInCurrentRole

boxplot(attr$YearsInCurrentRole, ylab = "YearsInCurrentRole")

rug(jitter(attr$YearsInCurrentRole), side = 2)

abline(h = mean(attr$YearsInCurrentRole, na.rm = T), lty = 2)

#Plotting the YearsInCurrentRole  with 3 lines for mean, median and mean+std

plot(attr$YearsInCurrentRole, xlab = "")

abline(h = mean(attr$YearsInCurrentRole, na.rm = T), lty = 1)

abline(h = mean(attr$YearsInCurrentRole, na.rm = T) + sd(attr$YearsInCurrentRole, na.rm = T),lty = 2)

abline(h = median(attr$YearsInCurrentRole, na.rm = T), lty = 3)

identify(attr$YearsInCurrentRole)
```

```
#Boxplot distributions for  YearsSinceLastPromotion

boxplot(attr$YearsSinceLastPromotion, ylab = "YearsSinceLastPromotion")

rug(jitter(attr$YearsSinceLastPromotion), side = 2)

abline(h = mean(attr$YearsSinceLastPromotion, na.rm = T), lty = 2)

#Plotting the YearsSinceLastPromotion  with 3 lines for mean, median and mean+std

plot(attr$YearsSinceLastPromotion, xlab = "")

abline(h = mean(attr$YearsSinceLastPromotion, na.rm = T), lty = 1)

abline(h = mean(attr$YearsSinceLastPromotion, na.rm = T) + sd(attr$YearsSinceLastPromotion, na.rm =
T),lty = 2)

abline(h = median(attr$YearsSinceLastPromotion, na.rm = T), lty = 3)

identify(attr$YearsSinceLastPromotion)


#Boxplot distributions for  YearsWithCurrManager

boxplot(attr$YearsWithCurrManager, ylab = "YearsWithCurrManager")

rug(jitter(attr$YearsWithCurrManager), side = 2)

abline(h = mean(attr$YearsWithCurrManager, na.rm = T), lty = 2)

#Boxplot distributions for  YearsWithCurrManager

plot(attr$YearsWithCurrManager, xlab = "")

abline(h = mean(attr$YearsWithCurrManager, na.rm = T), lty = 1)

abline(h = mean(attr$YearsWithCurrManager, na.rm = T) + sd(attr$YearsWithCurrManager, na.rm =
T),lty = 2)

abline(h = median(attr$YearsWithCurrManager, na.rm = T), lty = 3)

identify(attr$YearsWithCurrManager)


#Chi Plot for inspecting the independence

chi.plot(attr$MonthlyIncome,attr$Age)


#Plotting joint boxplots for various categories wrt numerical column Monthly Income

bwplot(attr$Department ~ attr$Age, data=attr, ylab='Department',xlab='Age')
```

```
bwplot(attr$Gender ~ attr$Age, data=attr, ylab='Gender',xlab='Age')

bwplot(attr$EducationField ~ attr$Age, data=attr, ylab='EducationField',xlab='Age')

bwplot(attr$JobRole ~ attr$Age, data=attr, ylab='JobRole',xlab='Age')

bwplot(attr$MaritalStatus ~ attr$MonthlyIncome, data=attr, ylab='MaritalStatus',xlab='Age')

bwplot(attr$BusinessTravel ~ attr$Age, data=attr, ylab='BusinessTravel',xlab='Age')

#Plotting stripplots for various categories wrt numerical column TotalCharges

bwplot(attr$Department ~ attr$Age, data=attr,panel=panel.bpplot,
    probs=seq(.01,.49,by=.01), datadensity=TRUE, ylab='Department',xlab='Age')

bwplot(attr$Gender ~ attr$Age, data=attr,panel=panel.bpplot,
    probs=seq(.01,.49,by=.01), datadensity=TRUE, ylab='Gender',xlab='Age')

bwplot(attr$EducationField ~ attr$Age, data=attr,panel=panel.bpplot,
    probs=seq(.01,.49,by=.01), datadensity=TRUE, ylab='EducationField',xlab='Age')

bwplot(attr$JobRole ~ attr$Age, data=attr,panel=panel.bpplot,
    probs=seq(.01,.49,by=.01), datadensity=TRUE, ylab='JobRole',xlab='Age')

bwplot(attr$MartialStatus ~ attr$Age, data=attr,panel=panel.bpplot,
    probs=seq(.01,.49,by=.01), datadensity=TRUE, ylab='MartialStatus',xlab='Age')

bwplot(attr$BusinessTravel ~ attr$Age, data=attr,panel=panel.bpplot,
    probs=seq(.01,.49,by=.01), datadensity=TRUE, ylab='BusinessTravel',xlab='Age')


data<-attr[,c('Age','DailyRate','DistanceFromHome','HourlyRate',

'MonthlyIncome','MonthlyRate','NumCompaniesWorked','PercentSalaryHike','TotalWorkingYears',
        'TrainingTimesLastYear','YearsAtCompany',
        'YearsInCurrentRole','YearsSinceLastPromotion','YearsWithCurrManager')]

chart.Correlation(data,histogram = TRUE,pch=19)

#--------------------------------------------------------------------------------

##Creating Temporary Variables

#--------------------------------------------------------------------------------
```

```
#Converting double/int columns to numeric
numeric_col <- c("Age","DailyRate","DistanceFromHome","HourlyRate",

"MonthlyIncome","MonthlyRate","NumCompaniesWorked","PercentSalaryHike","TotalWorkingYears",
        "TrainingTimesLastYear","YearsAtCompany",
        "YearsInCurrentRole","YearsSinceLastPromotion","YearsWithCurrManager")
attr[numeric_col] <- sapply(attr[numeric_col], as.numeric)

#Take out the numeric columns from categorical columns and storing them as a seperate dataframe
attr_i <- attr[,c("Age","DailyRate","DistanceFromHome","HourlyRate",

"MonthlyIncome","MonthlyRate","NumCompaniesWorked","PercentSalaryHike","TotalWorkingYears",
         "TrainingTimesLastYear","YearsAtCompany",
         "YearsInCurrentRole","YearsSinceLastPromotion","YearsWithCurrManager")]
attr_i <- data.frame(scale(attr_i))

#Creating temporary variables for the categorical data
attr_c <- attr[,-c(2,3,5,8,10,11,12,13,14,15,19,21,22,23)]
temporary<- data.frame(sapply(attr_c,function(x) data.frame(model.matrix(~x-1,data =attr_c))[,-1]))
head(temporary)
View(attr)

#Combining the temporary and the numeric columns and create the final dataset
attr_final <- cbind(attr_i,temporary)
head(attr_final)
glimpse(attr_final)



# solve the error "Figure margins too large"
```

```
par("mar")
par(mar=c(1,1,1,1))
graphics.off()
dev.off()
##Matrix Plots, Covariance and Corelations Plots
#ScatterPlot matrix
pairs(attr_final[,1:5],pch=".",cex=1.5)

#CorrelationMatrix
cormatrix <- round(cor(attr_final),4)
cormatrix
#Heatmap for correlation matrix
#Negative correlations are shown in blue and positive in red
col<- colorRampPalette(c("blue", "white", "red"))(20)
heatmap(cormatrix, col=col, symm=TRUE)

##Test of Significance

#T-Test
#Null Hypothesis - The two means are equal
#Alternate Hypothesis - Difference in the two means is not zero
#pvalue >= 0.05, accept null hypothesis
#Or
#else accept the alternate hypothesis
```

#Univariate mean comparison using t test

#Monthly Income and Attrition

```
with(data=attr,t.test(attr$MonthlyIncome[attr$Attrition=="Yes"],attr$MonthlyIncome[attr$Attrition=="No"],var.equal=TRUE))
```

#HourlyRate and Attrition

```
with(data=attr,t.test(attr$HourlyRate[attr$Attrition=="Yes"],attr$HourlyRate[attr$Attrition=="No"],var.equal=TRUE))
```

#Daily Rate and Attrition

```
with(data=attr,t.test(attr$DailyRate[attr$Attrition=="Yes"],attr$DailyRate[attr$Attrition=="No"],var.equal=TRUE))
```

#Age and Attrition

```
with(data=attr,t.test(attr$Age[attr$Attrition=="Yes"],attr$Age[attr$Attrition=="No"],var.equal=TRUE))
```

#DistanceFromHome and Attrition

```
with(data = attr,t.test(attr$DistanceFromHome[attr$Attrition=="Yes"],attr$Age[attr$Attrition=="No"],var.equal = TRUE))
```

#Monthly Income and Gender

```
with(data = attr,t.test(attr$MonthlyIncome[attr$Gender=="Male"],attr$MonthlyIncome[attr$Gender=="Female"],var.equal = TRUE))
```

#DistanceFromHome and Gender

```
with(data = attr,t.test(attr$DistanceFromHome[attr$Gender=="Male"],attr$DistanceFromHome[attr$Gender=="Female"],var.equal = TRUE))
```

```r
#Multivariate mean comparison using Hotelling t test

#Monthly Income and gender
t2testgender <- hotelling.test(attr$MonthlyIncome + attr$DistanceFromHome ~ attr$Gender, data=attr)
cat("T2 statistic =",t2testgender$stat[[1]],"\n")
print(t2testgender)

#Monthly Income and Attrition
t2testattr <- hotelling.test(attr$MonthlyIncome + attr$DistanceFromHome ~ attr$Attrition, data=attr)
cat("T2 statistic =",t2testattr$stat[[1]],"\n")
print(t2testattr)
```