hr\_miniproject.R

sumedh

Sun Jul 08 17:43:14 2018

rm(list=ls())  
gc()

## used (Mb) gc trigger (Mb) max used (Mb)  
## Ncells 501991 26.9 1129678 60.4 609162 32.6  
## Vcells 977824 7.5 8388608 64.0 1606350 12.3

setwd("D:/Great Lakes PGPDSE/Great Lakes/13 Ensemble Techniques/Mini Project")  
## Reading the data set  
hr=read.csv("HR\_Employee\_Attrition\_Data.csv",stringsAsFactors = FALSE)  
summary(hr)

## Attrition Age BusinessTravel DailyRate   
## Length:2940 Min. :18.00 Length:2940 Min. : 102.0   
## Class :character 1st Qu.:30.00 Class :character 1st Qu.: 465.0   
## Mode :character Median :36.00 Mode :character Median : 802.0   
## Mean :36.92 Mean : 802.5   
## 3rd Qu.:43.00 3rd Qu.:1157.0   
## Max. :60.00 Max. :1499.0   
## Department DistanceFromHome Education EducationField   
## Length:2940 Min. : 1.000 Min. :1.000 Length:2940   
## Class :character 1st Qu.: 2.000 1st Qu.:2.000 Class :character   
## Mode :character Median : 7.000 Median :3.000 Mode :character   
## Mean : 9.193 Mean :2.913   
## 3rd Qu.:14.000 3rd Qu.:4.000   
## Max. :29.000 Max. :5.000   
## EmployeeCount EmployeeNumber EnvironmentSatisfaction Gender   
## Min. :1 Min. : 1.0 Min. :1.000 Length:2940   
## 1st Qu.:1 1st Qu.: 735.8 1st Qu.:2.000 Class :character   
## Median :1 Median :1470.5 Median :3.000 Mode :character   
## Mean :1 Mean :1470.5 Mean :2.722   
## 3rd Qu.:1 3rd Qu.:2205.2 3rd Qu.:4.000   
## Max. :1 Max. :2940.0 Max. :4.000   
## HourlyRate JobInvolvement JobLevel JobRole   
## Min. : 30.00 Min. :1.00 Min. :1.000 Length:2940   
## 1st Qu.: 48.00 1st Qu.:2.00 1st Qu.:1.000 Class :character   
## Median : 66.00 Median :3.00 Median :2.000 Mode :character   
## Mean : 65.89 Mean :2.73 Mean :2.064   
## 3rd Qu.: 84.00 3rd Qu.:3.00 3rd Qu.:3.000   
## Max. :100.00 Max. :4.00 Max. :5.000   
## JobSatisfaction MaritalStatus MonthlyIncome MonthlyRate   
## Min. :1.000 Length:2940 Min. : 1009 Min. : 2094   
## 1st Qu.:2.000 Class :character 1st Qu.: 2911 1st Qu.: 8045   
## Median :3.000 Mode :character Median : 4919 Median :14236   
## Mean :2.729 Mean : 6503 Mean :14313   
## 3rd Qu.:4.000 3rd Qu.: 8380 3rd Qu.:20462   
## Max. :4.000 Max. :19999 Max. :26999   
## NumCompaniesWorked OverTime PercentSalaryHike PerformanceRating  
## Min. :0.000 Length:2940 Min. :11.00 Min. :3.000   
## 1st Qu.:1.000 Class :character 1st Qu.:12.00 1st Qu.:3.000   
## Median :2.000 Mode :character Median :14.00 Median :3.000   
## Mean :2.693 Mean :15.21 Mean :3.154   
## 3rd Qu.:4.000 3rd Qu.:18.00 3rd Qu.:3.000   
## Max. :9.000 Max. :25.00 Max. :4.000   
## RelationshipSatisfaction StandardHours StockOptionLevel TotalWorkingYears  
## Min. :1.000 Min. :80 Min. :0.0000 Min. : 0.00   
## 1st Qu.:2.000 1st Qu.:80 1st Qu.:0.0000 1st Qu.: 6.00   
## Median :3.000 Median :80 Median :1.0000 Median :10.00   
## Mean :2.712 Mean :80 Mean :0.7939 Mean :11.28   
## 3rd Qu.:4.000 3rd Qu.:80 3rd Qu.:1.0000 3rd Qu.:15.00   
## Max. :4.000 Max. :80 Max. :3.0000 Max. :40.00   
## TrainingTimesLastYear WorkLifeBalance YearsAtCompany YearsInCurrentRole  
## Min. :0.000 Min. :1.000 Min. : 0.000 Min. : 0.000   
## 1st Qu.:2.000 1st Qu.:2.000 1st Qu.: 3.000 1st Qu.: 2.000   
## Median :3.000 Median :3.000 Median : 5.000 Median : 3.000   
## Mean :2.799 Mean :2.761 Mean : 7.008 Mean : 4.229   
## 3rd Qu.:3.000 3rd Qu.:3.000 3rd Qu.: 9.000 3rd Qu.: 7.000   
## Max. :6.000 Max. :4.000 Max. :40.000 Max. :18.000   
## YearsSinceLastPromotion YearsWithCurrManager  
## Min. : 0.000 Min. : 0.000   
## 1st Qu.: 0.000 1st Qu.: 2.000   
## Median : 1.000 Median : 3.000   
## Mean : 2.188 Mean : 4.123   
## 3rd Qu.: 3.000 3rd Qu.: 7.000   
## Max. :15.000 Max. :17.000

# Data set have 2940 observation and 34 variables is there   
str(hr)

## 'data.frame': 2940 obs. of 34 variables:  
## $ Attrition : chr "Yes" "No" "Yes" "No" ...  
## $ Age : int 41 49 37 33 27 32 59 30 38 36 ...  
## $ BusinessTravel : chr "Travel\_Rarely" "Travel\_Frequently" "Travel\_Rarely" "Travel\_Frequently" ...  
## $ DailyRate : int 1102 279 1373 1392 591 1005 1324 1358 216 1299 ...  
## $ Department : chr "Sales" "Research & Development" "Research & Development" "Research & Development" ...  
## $ DistanceFromHome : int 1 8 2 3 2 2 3 24 23 27 ...  
## $ Education : int 2 1 2 4 1 2 3 1 3 3 ...  
## $ EducationField : chr "Life Sciences" "Life Sciences" "Other" "Life Sciences" ...  
## $ EmployeeCount : int 1 1 1 1 1 1 1 1 1 1 ...  
## $ EmployeeNumber : int 1 2 3 4 5 6 7 8 9 10 ...  
## $ EnvironmentSatisfaction : int 2 3 4 4 1 4 3 4 4 3 ...  
## $ Gender : chr "Female" "Male" "Male" "Female" ...  
## $ HourlyRate : int 94 61 92 56 40 79 81 67 44 94 ...  
## $ JobInvolvement : int 3 2 2 3 3 3 4 3 2 3 ...  
## $ JobLevel : int 2 2 1 1 1 1 1 1 3 2 ...  
## $ JobRole : chr "Sales Executive" "Research Scientist" "Laboratory Technician" "Research Scientist" ...  
## $ JobSatisfaction : int 4 2 3 3 2 4 1 3 3 3 ...  
## $ MaritalStatus : chr "Single" "Married" "Single" "Married" ...  
## $ MonthlyIncome : int 5993 5130 2090 2909 3468 3068 2670 2693 9526 5237 ...  
## $ MonthlyRate : int 19479 24907 2396 23159 16632 11864 9964 13335 8787 16577 ...  
## $ NumCompaniesWorked : int 8 1 6 1 9 0 4 1 0 6 ...  
## $ OverTime : chr "Yes" "No" "Yes" "Yes" ...  
## $ PercentSalaryHike : int 11 23 15 11 12 13 20 22 21 13 ...  
## $ PerformanceRating : int 3 4 3 3 3 3 4 4 4 3 ...  
## $ RelationshipSatisfaction: int 1 4 2 3 4 3 1 2 2 2 ...  
## $ StandardHours : int 80 80 80 80 80 80 80 80 80 80 ...  
## $ StockOptionLevel : int 0 1 0 0 1 0 3 1 0 2 ...  
## $ TotalWorkingYears : int 8 10 7 8 6 8 12 1 10 17 ...  
## $ TrainingTimesLastYear : int 0 3 3 3 3 2 3 2 2 3 ...  
## $ WorkLifeBalance : int 1 3 3 3 3 2 2 3 3 2 ...  
## $ YearsAtCompany : int 6 10 0 8 2 7 1 1 9 7 ...  
## $ YearsInCurrentRole : int 4 7 0 7 2 7 0 0 7 7 ...  
## $ YearsSinceLastPromotion : int 0 1 0 3 2 3 0 0 1 7 ...  
## $ YearsWithCurrManager : int 5 7 0 0 2 6 0 0 8 7 ...

summary(hr)

## Attrition Age BusinessTravel DailyRate   
## Length:2940 Min. :18.00 Length:2940 Min. : 102.0   
## Class :character 1st Qu.:30.00 Class :character 1st Qu.: 465.0   
## Mode :character Median :36.00 Mode :character Median : 802.0   
## Mean :36.92 Mean : 802.5   
## 3rd Qu.:43.00 3rd Qu.:1157.0   
## Max. :60.00 Max. :1499.0   
## Department DistanceFromHome Education EducationField   
## Length:2940 Min. : 1.000 Min. :1.000 Length:2940   
## Class :character 1st Qu.: 2.000 1st Qu.:2.000 Class :character   
## Mode :character Median : 7.000 Median :3.000 Mode :character   
## Mean : 9.193 Mean :2.913   
## 3rd Qu.:14.000 3rd Qu.:4.000   
## Max. :29.000 Max. :5.000   
## EmployeeCount EmployeeNumber EnvironmentSatisfaction Gender   
## Min. :1 Min. : 1.0 Min. :1.000 Length:2940   
## 1st Qu.:1 1st Qu.: 735.8 1st Qu.:2.000 Class :character   
## Median :1 Median :1470.5 Median :3.000 Mode :character   
## Mean :1 Mean :1470.5 Mean :2.722   
## 3rd Qu.:1 3rd Qu.:2205.2 3rd Qu.:4.000   
## Max. :1 Max. :2940.0 Max. :4.000   
## HourlyRate JobInvolvement JobLevel JobRole   
## Min. : 30.00 Min. :1.00 Min. :1.000 Length:2940   
## 1st Qu.: 48.00 1st Qu.:2.00 1st Qu.:1.000 Class :character   
## Median : 66.00 Median :3.00 Median :2.000 Mode :character   
## Mean : 65.89 Mean :2.73 Mean :2.064   
## 3rd Qu.: 84.00 3rd Qu.:3.00 3rd Qu.:3.000   
## Max. :100.00 Max. :4.00 Max. :5.000   
## JobSatisfaction MaritalStatus MonthlyIncome MonthlyRate   
## Min. :1.000 Length:2940 Min. : 1009 Min. : 2094   
## 1st Qu.:2.000 Class :character 1st Qu.: 2911 1st Qu.: 8045   
## Median :3.000 Mode :character Median : 4919 Median :14236   
## Mean :2.729 Mean : 6503 Mean :14313   
## 3rd Qu.:4.000 3rd Qu.: 8380 3rd Qu.:20462   
## Max. :4.000 Max. :19999 Max. :26999   
## NumCompaniesWorked OverTime PercentSalaryHike PerformanceRating  
## Min. :0.000 Length:2940 Min. :11.00 Min. :3.000   
## 1st Qu.:1.000 Class :character 1st Qu.:12.00 1st Qu.:3.000   
## Median :2.000 Mode :character Median :14.00 Median :3.000   
## Mean :2.693 Mean :15.21 Mean :3.154   
## 3rd Qu.:4.000 3rd Qu.:18.00 3rd Qu.:3.000   
## Max. :9.000 Max. :25.00 Max. :4.000   
## RelationshipSatisfaction StandardHours StockOptionLevel TotalWorkingYears  
## Min. :1.000 Min. :80 Min. :0.0000 Min. : 0.00   
## 1st Qu.:2.000 1st Qu.:80 1st Qu.:0.0000 1st Qu.: 6.00   
## Median :3.000 Median :80 Median :1.0000 Median :10.00   
## Mean :2.712 Mean :80 Mean :0.7939 Mean :11.28   
## 3rd Qu.:4.000 3rd Qu.:80 3rd Qu.:1.0000 3rd Qu.:15.00   
## Max. :4.000 Max. :80 Max. :3.0000 Max. :40.00   
## TrainingTimesLastYear WorkLifeBalance YearsAtCompany YearsInCurrentRole  
## Min. :0.000 Min. :1.000 Min. : 0.000 Min. : 0.000   
## 1st Qu.:2.000 1st Qu.:2.000 1st Qu.: 3.000 1st Qu.: 2.000   
## Median :3.000 Median :3.000 Median : 5.000 Median : 3.000   
## Mean :2.799 Mean :2.761 Mean : 7.008 Mean : 4.229   
## 3rd Qu.:3.000 3rd Qu.:3.000 3rd Qu.: 9.000 3rd Qu.: 7.000   
## Max. :6.000 Max. :4.000 Max. :40.000 Max. :18.000   
## YearsSinceLastPromotion YearsWithCurrManager  
## Min. : 0.000 Min. : 0.000   
## 1st Qu.: 0.000 1st Qu.: 2.000   
## Median : 1.000 Median : 3.000   
## Mean : 2.188 Mean : 4.123   
## 3rd Qu.: 3.000 3rd Qu.: 7.000   
## Max. :15.000 Max. :17.000

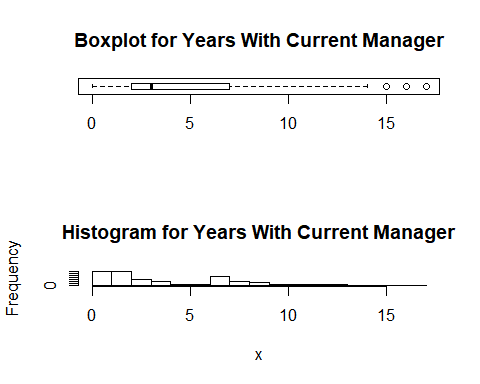
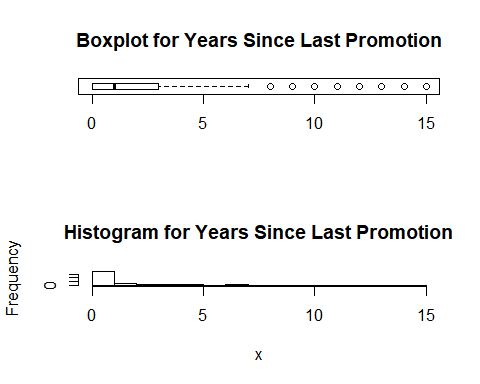
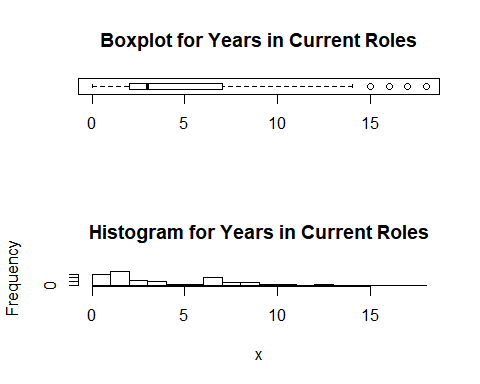
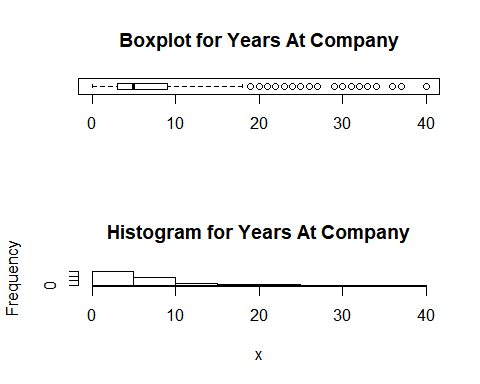
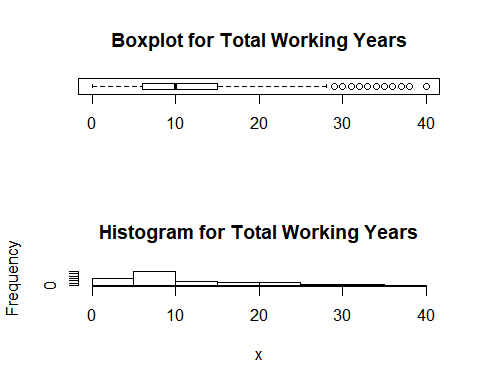
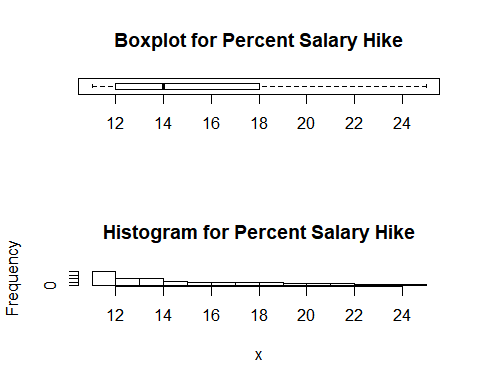
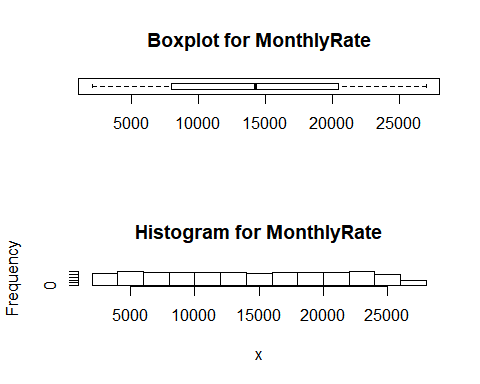
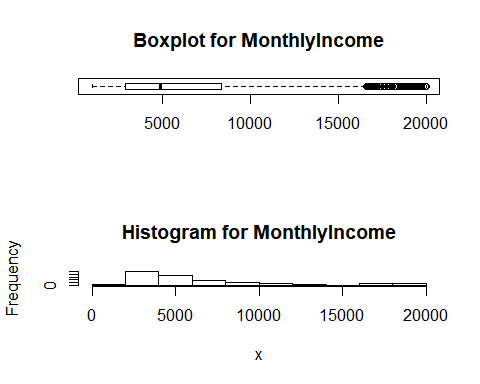
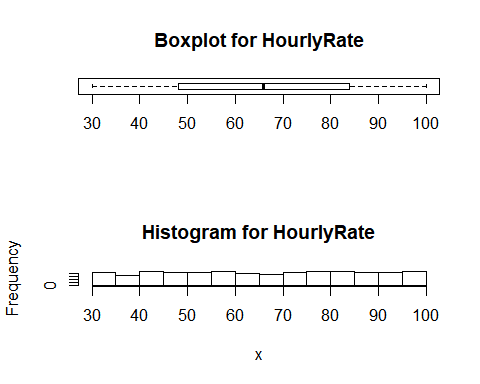
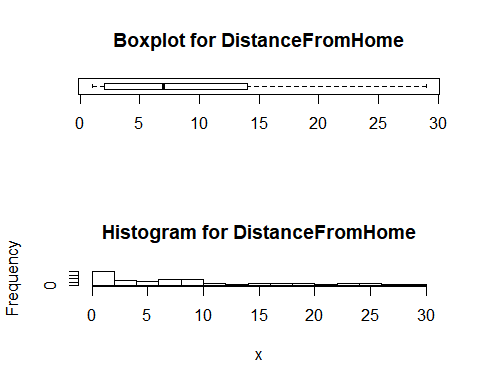
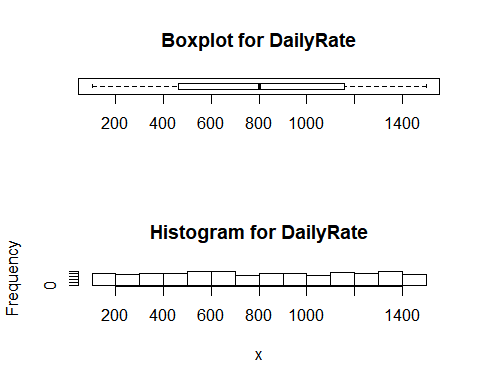
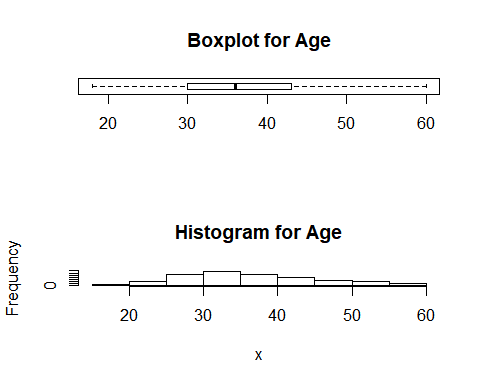
# There is no missing values in the dataset  
# Changing value in attrition 0 and 1 format  
hr$Attrition[hr$Attrition=="No"]=0  
hr$Attrition[hr$Attrition=="Yes"]=1  
# Changing value in BusinessTravel 1,2 and 3  
hr$BusinessTravel[hr$BusinessTravel=="Non-Travel"]=1  
hr$BusinessTravel[hr$BusinessTravel=="Travel\_Frequently"]=2  
hr$BusinessTravel[hr$BusinessTravel=="Travel\_Rarely"]=3  
# Changing value in Department 1, 2 and 3  
hr$Department[hr$Department=="Human Resources"]=1  
hr$Department[hr$Department=="Research & Development"]=2  
hr$Department[hr$Department=="Sales"]=3  
# Changing value of EducationField into the 1 to 4 format  
hr$EducationField[hr$EducationField=="Human Resources"]=1  
hr$EducationField[hr$EducationField=="Life Sciences"]=2  
hr$EducationField[hr$EducationField=="Markerting"]=3  
hr$EducationField[hr$EducationField=="Medical"]=4  
hr$EducationField[hr$EducationField=="Other"]=5  
hr$EducationField[hr$EducationField=="Technical Degree"]=6  
# Changing value of Gender into 1 and 2 format  
hr$Gender[hr$Gender=="Female"]=1  
hr$Gender[hr$Gender=="Male"]=2  
  
#Changing the format of JobRole into 1,2,3 and 4 format  
hr$JobRole[hr$JobRole=="Healthcare Representative"]=1  
hr$JobRole[hr$JobRole=="Human Resources"]=2  
hr$JobRole[hr$JobRole=="Laboratory Technician"]=3  
hr$JobRole[hr$JobRole=="Manager"]=4  
hr$JobRole[hr$JobRole=="Manufacturing Director"]=5  
hr$JobRole[hr$JobRole=="Research Director"]=6  
hr$JobRole[hr$JobRole=="Research Scientist"]=7  
hr$JobRole[hr$JobRole=="Sales Executive"]=8  
hr$JobRole[hr$JobRole=="Sales Representative"]=9  
# Changing the format of MaritalStatus 1,2 and 3  
hr$MaritalStatus[hr$MaritalStatus=="Divorced"]=1  
hr$MaritalStatus[hr$MaritalStatus=="Married"]=2  
hr$MaritalStatus[hr$MaritalStatus=="Single"]=3  
# Changing the format of OverTIme into 0 and 1 format  
hr$OverTime[hr$OverTime=="No"]=0  
hr$OverTime[hr$OverTime=="Yes"]=1  
head(hr,2)

## Attrition Age BusinessTravel DailyRate Department DistanceFromHome  
## 1 1 41 3 1102 3 1  
## 2 0 49 2 279 2 8  
## Education EducationField EmployeeCount EmployeeNumber  
## 1 2 2 1 1  
## 2 1 2 1 2  
## EnvironmentSatisfaction Gender HourlyRate JobInvolvement JobLevel  
## 1 2 1 94 3 2  
## 2 3 2 61 2 2  
## JobRole JobSatisfaction MaritalStatus MonthlyIncome MonthlyRate  
## 1 8 4 3 5993 19479  
## 2 7 2 2 5130 24907  
## NumCompaniesWorked OverTime PercentSalaryHike PerformanceRating  
## 1 8 1 11 3  
## 2 1 0 23 4  
## RelationshipSatisfaction StandardHours StockOptionLevel  
## 1 1 80 0  
## 2 4 80 1  
## TotalWorkingYears TrainingTimesLastYear WorkLifeBalance YearsAtCompany  
## 1 8 0 1 6  
## 2 10 3 3 10  
## YearsInCurrentRole YearsSinceLastPromotion YearsWithCurrManager  
## 1 4 0 5  
## 2 7 1 7

# EmployeeCount have to removed beacuse all is having 1 value  
# EmployeeNumber have to be removed because it is not impacting the result  
# StandardHours have to be removed beacuse in all rows having the same value.  
View(hr)  
hr=hr[ , -c(9,10,26)]  
View(hr)  
dim(hr)

## [1] 2940 31

d=c(2,4,6,11,17,18,21,25,28,29,30,31)  
colname=c("Age","DailyRate","DistanceFromHome","HourlyRate","MonthlyIncome","MonthlyRate","Percent Salary Hike","Total Working Years","Years At Company","Years in Current Roles","Years Since Last Promotion","Years With Current Manager")  
  
hrf=hr[,c(2,4,6,11,17,18,21,25,28,29,30,31)]  
pl = colnames(hrf)  
colname=c("Age","DailyRate","DistanceFromHome","HourlyRate","MonthlyIncome","MonthlyRate","Percent Salary Hike","Total Working Years","Years At Company","Years in Current Roles","Years Since Last Promotion","Years With Current Manager")  
par(mfrow=c(2,1))  
for (i in 1:length(pl)) {  
   
 x <- hrf[,i]  
 boxplot(x,horizontal = T,main = paste("Boxplot for", colname[i]))  
 hist(x,main = paste("Histogram for", colname[i]))  
}



# From the Box plot and the Histogram for many attributes some is not showing outliers and some is showing outliers due to do the mistach between the no. of experiece , Salary , position and many other factors are also associated with outliers which is valid .  
# So we can't remove the outliers from the dataset as it hold true.