Employee Attrition

Association Rules using arules package

```
#importing Libraries
library(readr)
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(tidyverse)
```

```
## ── Attaching packages ─────────────────────────────────────────────────────── tidyverse 1.2.1 ──
```

```
## ✔ ggplot2 3.2.1     ✔ purrr   0.3.2
## ✔ tibble  2.1.3     ✔ stringr 1.4.0
## ✔ tidyr   1.0.0     ✔ forcats 0.4.0
```

```
## ── Conflicts ────────────────────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
```

```
library(tidyr)
library(imputeTS)
```

```
## Registered S3 method overwritten by 'xts':
##   method     from
##   as.zoo.xts zoo
```

```
## Registered S3 method overwritten by 'quantmod':
##   method            from
##   as.zoo.data.frame zoo
```

```
## Registered S3 methods overwritten by 'forecast':
##   method             from
##   fitted.fracdiff    fracdiff
##   residuals.fracdiff fracdiff
```

```
library(caret)
```

```
## Loading required package: lattice
```

```
##
## Attaching package: 'caret'
```

```
## The following object is masked from 'package:purrr':
##
##     lift
```

```
library(arules)
```

```
## Loading required package: Matrix
```

```
##
## Attaching package: 'Matrix'
```

```
## The following objects are masked from 'package:tidyr':
##
##     expand, pack, unpack
```

```
##
## Attaching package: 'arules'
```

```
## The following object is masked from 'package:dplyr':
##
##     recode
```

```
## The following objects are masked from 'package:base':
##
##     abbreviate, write
```

```
library(arulesViz)
```

```
## Loading required package: grid
```

```
## Registered S3 method overwritten by 'seriation':
##   method         from
##   reorder.hclust gclus
```

```
library(randomForest)
```

```
## randomForest 4.6-14
```

```
## Type rfNews() to see new features/changes/bug fixes.
```

```
##
## Attaching package: 'randomForest'
```

```
## The following object is masked from 'package:ggplot2':
##
##     margin
```

```
## The following object is masked from 'package:dplyr':
##
##     combine
```

```
library(gridExtra)
```

```
##
## Attaching package: 'gridExtra'
```

```
## The following object is masked from 'package:randomForest':
##
##     combine
```

```
## The following object is masked from 'package:dplyr':
##
##     combine
```

```
library(tidymodels)
```

```
## ── Attaching packages ─────────────────────────────────────────────────────── tidymodels 0.0.3 ──
```

```
## ✔ broom     0.5.2     ✔ recipes   0.1.7
## ✔ dials     0.0.4     ✔ rsample   0.0.5
## ✔ infer     0.5.1     ✔ yardstick 0.0.5
## ✔ parsnip   0.0.5
```

```
## ── Conflicts ────────────────────────────────────────────────────────── tidymodels_conflicts() ──
## ✖ gridExtra::combine()   masks randomForest::combine(), dplyr::combine()
## ✖ scales::discard()      masks purrr::discard()
## ✖ recipes::discretize()  masks arules::discretize()
## ✖ dplyr::filter()        masks stats::filter()
## ✖ recipes::fixed()       masks stringr::fixed()
## ✖ dplyr::lag()           masks stats::lag()
## ✖ caret::lift()          masks purrr::lift()
## ✖ dials::margin()        masks randomForest::margin(), ggplot2::margin()
## ✖ yardstick::precision() masks caret::precision()
## ✖ yardstick::recall()    masks caret::recall()
## ✖ arules::recode()       masks dplyr::recode()
## ✖ yardstick::spec()      masks readr::spec()
## ✖ recipes::step()        masks stats::step()
```

```
library(curl)
```

```
##
## Attaching package: 'curl'
```

```
## The following object is masked from 'package:readr':
##
##     parse_date
```

#Reading file using read.csv

```
#knitr::opts_knit$set(root.dir = '/Users/juilee81/Desktop/jsalunkh14')
e_Data <- read.csv("/Users/juilee81/Desktop/DA/DA_HW_01/employee_attrition.csv",header=FALSE,stringsAsFactor
s=FALSE)
View(e_Data)
```

#Renaming Column names

```r
colnames(e_Data)[1] <- "Age"
colnames(e_Data)[2] <- "Attrition"
colnames(e_Data)[3] <- "BusinessTravel"
colnames(e_Data)[4] <- "DailyRate"
colnames(e_Data)[5] <- "Department"
colnames(e_Data)[6] <- "DistanceFromHome"
colnames(e_Data)[7] <- "Education"
colnames(e_Data)[8] <- "EducationField"
colnames(e_Data)[9] <- "EmployeeCount   "
colnames(e_Data)[10] <- "EmployeeNumber"
colnames(e_Data)[11] <- "EnvironmentSatisfaction"
colnames(e_Data)[12] <- "Gender"
colnames(e_Data)[13] <- "HourlyRate"
colnames(e_Data)[14] <- "JobInvolvement"
colnames(e_Data)[15] <- "JobLevel"
colnames(e_Data)[16] <- "JobRole"
colnames(e_Data)[17] <- "JobSatisfaction"
colnames(e_Data)[18] <- "MaritalStatus"
colnames(e_Data)[19] <- "MonthlyIncome"
colnames(e_Data)[20] <- "MonthlyRate"
colnames(e_Data)[21] <- "NumCompaniesWorked"
colnames(e_Data)[22] <- "Over18"
colnames(e_Data)[23] <- "OverTime"
colnames(e_Data)[24] <- "PercentSalaryHike"
colnames(e_Data)[25] <- "PerformanceRating"
colnames(e_Data)[26] <- "RelationshipSatisfaction"
colnames(e_Data)[27] <- "StandardHours"
colnames(e_Data)[28] <- "StockOptionLevel"
colnames(e_Data)[29] <- "TotalWorkingYears"
colnames(e_Data)[30] <- "TrainingTimesLastYear"
colnames(e_Data)[31] <- "WorkLifeBalance"
colnames(e_Data)[32] <- "YearsAtCompany"
colnames(e_Data)[33] <- "YearsInCurrentRole"
colnames(e_Data)[34] <- "YearsSinceLastPromotion"
colnames(e_Data)[35] <- "YearsWithCurrManager"
emp = e_Data[-c(1, 2),]
View(emp)
str(emp)
```

```
## 'data.frame':    1176 obs. of  35 variables:
##  $ Age                     : chr  "30" "52" "42" "55" ...
##  $ Attrition               : chr  "No" "No" "No" "No" ...
##  $ BusinessTravel          : chr  "Travel_Rarely" "Travel_Rarely" "Travel_Rarely" "Non-Travel" ...
##  $ DailyRate               : chr  "1358" "1325" "462" "177" ...
##  $ Department              : chr  "Sales" "Research & Development" "Sales" "Research & Development" ...
##  $ DistanceFromHome        : chr  "16" "11" "14" "8" ...
##  $ Education               : chr  "1" "4" "2" "1" ...
##  $ EducationField          : chr  "Life Sciences" "Life Sciences" "Medical" "Medical" ...
##  $ EmployeeCount          : chr  "1" "1" "1" "1" ...
##  $ EmployeeNumber          : chr  "1479" "813" "936" "1278" ...
##  $ EnvironmentSatisfaction : chr  "4" "4" "3" "4" ...
##  $ Gender                  : chr  "Male" "Female" "Female" "Male" ...
##  $ HourlyRate              : chr  "96" "82" "68" "37" ...
##  $ JobInvolvement          : chr  "3" "3" "2" "2" ...
##  $ JobLevel                : chr  "2" "2" "2" "4" ...
##  $ JobRole                 : chr  "Sales Executive" "Laboratory Technician" "Sales Executive" "Healthcar
e Representative" ...
##  $ JobSatisfaction         : chr  "3" "3" "3" "2" ...
##  $ MaritalStatus           : chr  "Married" "Married" "Single" "Divorced" ...
##  $ MonthlyIncome           : chr  "5301" "3149" "6244" "13577" ...
##  $ MonthlyRate             : chr  "2939" "21821" "7824" "25592" ...
##  $ NumCompaniesWorked      : chr  "8" "8" "7" "1" ...
##  $ Over18                  : chr  "Y" "Y" "Y" "Y" ...
##  $ OverTime                : chr  "No" "No" "No" "Yes" ...
##  $ PercentSalaryHike       : chr  "15" "20" "17" "15" ...
##  $ PerformanceRating       : chr  "3" "4" "3" "3" ...
##  $ RelationshipSatisfaction: chr  "3" "2" "1" "4" ...
##  $ StandardHours           : chr  "80" "80" "80" "80" ...
##  $ StockOptionLevel        : chr  "2" "1" "0" "1" ...
##  $ TotalWorkingYears       : chr  "4" "9" "10" "34" ...
##  $ TrainingTimesLastYear   : chr  "2" "3" "6" "3" ...
##  $ WorkLifeBalance         : chr  "2" "3" "3" "3" ...
##  $ YearsAtCompany          : chr  "2" "5" "5" "33" ...
##  $ YearsInCurrentRole      : chr  "1" "2" "4" "9" ...
##  $ YearsSinceLastPromotion : chr  "2" "1" "0" "15" ...
##  $ YearsWithCurrManager    : chr  "2" "4" "3" "0" ...
```

Out of the 35 variables we have 34 independent variables and one dependent/target variable which is Attrition.

#Converting blanks to NA values

```
emp[emp==""]<-NA
```

#Converting column type to numeric and factor.

```
emp$Age<-as.numeric(emp$Age)
emp$DailyRate<-as.numeric(emp$DailyRate)
emp$DistanceFromHome<-as.numeric(emp$DistanceFromHome)
emp$EmployeeCount<-as.numeric(emp$EmployeeCount)
emp$EmployeeNumber<-as.numeric(emp$EmployeeNumber)
emp$EnvironmentSatisfaction<-as.numeric(emp$EnvironmentSatisfaction)
emp$HourlyRate<-as.numeric(emp$HourlyRate)
emp$JobInvolvement<-as.numeric(emp$JobInvolvement)
emp$JobLevel<-as.numeric(emp$JobLevel)
emp$JobSatisfaction<-as.numeric(emp$JobSatisfaction)
emp$MonthlyIncome<-as.numeric(emp$MonthlyIncome)
emp$MonthlyRate<-as.numeric(emp$MonthlyRate)
emp$NumCompaniesWorked<-as.numeric(emp$NumCompaniesWorked)
emp$PercentSalaryHike<-as.numeric(emp$PercentSalaryHike)
emp$PerformanceRating<-as.numeric(emp$PerformanceRating)
emp$RelationshipSatisfaction<-as.numeric(emp$RelationshipSatisfaction)
emp$StandardHours<-as.numeric(emp$StandardHours)
emp$StockOptionLevel<-as.numeric(emp$StockOptionLevel)
emp$TotalWorkingYears<-as.numeric(emp$TotalWorkingYears)
emp$TrainingTimesLastYear<-as.numeric(emp$TrainingTimesLastYear)
emp$WorkLifeBalance<-as.numeric(emp$WorkLifeBalance)
emp$YearsAtCompany<-as.numeric(emp$YearsAtCompany)
emp$YearsInCurrentRole<-as.numeric(emp$YearsInCurrentRole)
emp$YearsSinceLastPromotion<-as.numeric(emp$YearsSinceLastPromotion)
emp$YearsWithCurrManager<-as.numeric(emp$YearsWithCurrManager)
emp$Education<-as.numeric(emp$Education)

emp$Attrition<-as.factor(emp$Attrition)
emp$BusinessTravel<-as.factor(emp$BusinessTravel)
emp$Department<-as.factor(emp$Department)
emp$EducationField<-as.factor(emp$EducationField)
emp$Gender<-as.factor(emp$Gender)
emp$JobRole <-as.factor(emp$JobRole)
emp$MaritalStatus<-as.factor(emp$MaritalStatus)
emp$Over18<-as.factor(emp$Over18)
emp$OverTime<-as.factor(emp$OverTime)

emp = emp[,-c(9)]
str(emp)
```

```
## 'data.frame':    1176 obs. of  35 variables:
##  $ Age                     : num  30 52 42 55 35 51 42 23 38 27 ...
##  $ Attrition               : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 1 1 1 2 ...
##  $ BusinessTravel          : Factor w/ 3 levels "Non-Travel","Travel_Frequently",..: 3 3 3 1 3 3 3 3 3 3
...
##  $ DailyRate               : num  1358 1325 462 177 1029 ...
##  $ Department              : Factor w/ 3 levels "Human Resources",..: 3 2 3 2 2 3 2 2 2 3 ...
##  $ DistanceFromHome        : num  16 11 14 8 16 26 1 20 6 2 ...
##  $ Education               : num  1 4 2 1 3 4 2 1 2 1 ...
##  $ EducationField          : Factor w/ 6 levels "Human Resources",..: 2 2 4 4 2 3 2 2 5 3 ...
##  $ EmployeeNumber          : num  1479 813 936 1278 1529 ...
##  $ EnvironmentSatisfaction : num  4 4 3 4 4 1 4 1 4 3 ...
##  $ Gender                  : Factor w/ 2 levels "Female","Male": 2 1 1 2 1 1 1 2 1 2 ...
##  $ HourlyRate              : num  96 82 68 37 91 66 43 97 40 85 ...
##  $ JobInvolvement          : num  3 3 2 2 2 3 2 3 2 3 ...
##  $ JobLevel                : num  2 2 2 4 3 4 2 2 1 1 ...
##  $ JobRole                 : Factor w/ 9 levels "Healthcare Representative",..: 8 3 8 1 1 4 5 3 3 9 ...
##  $ JobSatisfaction         : num  3 3 3 2 2 3 4 3 3 1 ...
##  $ MaritalStatus           : Factor w/ 3 levels "Divorced","Married",..: 2 2 3 1 3 2 2 3 2 1 ...
##  $ MonthlyIncome           : num  5301 3149 6244 13577 8606 ...
##  $ MonthlyRate             : num  2939 21821 7824 25592 21195 ...
##  $ NumCompaniesWorked      : num  8 8 7 1 1 2 9 0 1 0 ...
##  $ Over18                  : Factor w/ 1 level "Y": 1 1 1 1 1 1 1 1 1 1 ...
##  $ OverTime                : Factor w/ 2 levels "No","Yes": 1 1 1 2 1 1 2 1 1 1 ...
##  $ PercentSalaryHike       : num  15 20 17 15 19 14 13 14 11 11 ...
##  $ PerformanceRating       : num  3 4 3 3 3 3 3 3 NA ...
##  $ RelationshipSatisfaction: num  3 2 1 4 4 3 4 2 2 2 ...
##  $ StandardHours           : num  80 80 80 80 80 80 80 80 80 80 ...
##  $ StockOptionLevel        : num  2 1 0 1 0 1 1 0 1 1 ...
##  $ TotalWorkingYears       : num  4 9 10 34 11 29 8 5 5 5 ...
##  $ TrainingTimesLastYear   : num  2 3 6 3 3 2 4 2 3 3 ...
##  $ WorkLifeBalance         : num  2 3 3 3 1 2 3 3 3 3 ...
##  $ YearsAtCompany          : num  2 5 5 33 11 20 4 4 5 4 ...
##  $ YearsInCurrentRole      : num  1 2 4 9 8 6 3 3 4 3 ...
##  $ YearsSinceLastPromotion : num  2 1 0 15 3 4 0 1 0 0 ...
##  $ YearsWithCurrManager    : num  2 4 3 0 3 17 2 2 4 2 ...
##  $ EmployeeCount           : num  1 1 1 1 1 1 1 1 1 1 ...
```

#**********Data Cleaning: Replacing NA values**********

#Checking no of NA values by columns

```
colSums(is.na(emp))
```

```
##                     Age               Attrition           BusinessTravel
##                       0                       0                        0
##               DailyRate              Department         DistanceFromHome
##                       0                       0                        2
##               Education           EducationField          EmployeeNumber
##                       0                       0                        0
##  EnvironmentSatisfaction                  Gender              HourlyRate
##                       0                       1                        0
##           JobInvolvement                JobLevel                  JobRole
##                       0                       1                        0
##          JobSatisfaction           MaritalStatus            MonthlyIncome
##                       0                       0                        0
##              MonthlyRate       NumCompaniesWorked                   Over18
##                       0                       0                        0
##                 OverTime        PercentSalaryHike        PerformanceRating
##                       1                       1                        1
## RelationshipSatisfaction           StandardHours         StockOptionLevel
##                       1                       0                        0
##        TotalWorkingYears    TrainingTimesLastYear          WorkLifeBalance
##                       2                       0                        0
##           YearsAtCompany       YearsInCurrentRole  YearsSinceLastPromotion
##                       0                       0                        1
##     YearsWithCurrManager           EmployeeCount
##                       0                       0
```

As shown, we have total 10 NA values #Substituing NA values using interpolation

```
emp$DistanceFromHome<-na_interpolation(emp$DistanceFromHome)
emp$JobLevel<-na_interpolation(emp$JobLevel)
emp$PercentSalaryHike<-na_interpolation(emp$PercentSalaryHike)
emp$PerformanceRating<-na_interpolation(emp$PerformanceRating)
emp$RelationshipSatisfaction<-na_interpolation(emp$RelationshipSatisfaction)
emp$TotalWorkingYears<-na_interpolation(emp$TotalWorkingYears)
emp$YearsSinceLastPromotion<-na_interpolation(emp$YearsSinceLastPromotion)
```

Using interpolation method we have replaced NA values with new interpolated values in the numeric columns.

#Replacing by Mode

```
#For gender replacing NA value with Mode
table(emp$Gender)
```

```
##
## Female   Male
##    482    693
```

```
emp$Gender[is.na(emp$Gender)] <- "Male"

#For OverTime replacing NA value with Mode
table(emp$OverTime)
```

```
##
##  No Yes
## 838 337
```

```
emp$OverTime[is.na(emp$OverTime)] <- "No"
```

Since we have more no of 'Males' and more no of 'No', replacing NA values by most occuring value.

Checking no of NA values by columns

```
colSums(is.na(emp))
```

```
##                       Age                 Attrition            BusinessTravel
##                         0                         0                         0
##                 DailyRate                Department          DistanceFromHome
##                         0                         0                         0
##                 Education             EducationField            EmployeeNumber
##                         0                         0                         0
##  EnvironmentSatisfaction                    Gender                HourlyRate
##                         0                         0                         0
##            JobInvolvement                  JobLevel                   JobRole
##                         0                         0                         0
##           JobSatisfaction             MaritalStatus             MonthlyIncome
##                         0                         0                         0
##               MonthlyRate         NumCompaniesWorked                   Over18
##                         0                         0                         0
##                  OverTime          PercentSalaryHike         PerformanceRating
##                         0                         0                         0
## RelationshipSatisfaction             StandardHours          StockOptionLevel
##                         0                         0                         0
##         TotalWorkingYears     TrainingTimesLastYear           WorkLifeBalance
##                         0                         0                         0
##            YearsAtCompany        YearsInCurrentRole  YearsSinceLastPromotion
##                         0                         0                         0
##     YearsWithCurrManager             EmployeeCount
##                         0                         0
```

As seen above, No NA values are left in the dataframe. Dataframe is ready for analysis.

#Exploratory Data Analysis

##Statistical summary of the data

```
summary(emp)
```

```
##       Age        Attrition           BusinessTravel   DailyRate
##  Min.   :18.00   No :991   Non-Travel       :110    Min.   : 102.0
##  1st Qu.:30.00   Yes:185   Travel_Frequently:227    1st Qu.: 461.8
##  Median :36.00             Travel_Rarely    :839    Median : 796.0
##  Mean   :36.96                                      Mean   : 800.4
##  3rd Qu.:43.00                                      3rd Qu.:1162.0
##  Max.   :60.00                                      Max.   :1499.0
##
##                  Department  DistanceFromHome    Education
##  Human Resources       : 54   Min.   :  1.000   Min.   :1.000
##  Research & Development:764   1st Qu.:  2.000   1st Qu.:2.000
##  Sales                 :358   Median :  7.000   Median :3.000
##                               Mean   :  9.507   Mean   :2.895
##                               3rd Qu.: 14.000   3rd Qu.:4.000
##                               Max.   :224.000   Max.   :5.000
##
##           EducationField EmployeeNumber   EnvironmentSatisfaction
##  Human Resources : 25    Min.   :   1.0   Min.   :1.000
##  Life Sciences   :477    1st Qu.: 499.8   1st Qu.:2.000
##  Marketing       :127    Median :1032.5   Median :3.000
##  Medical         :381    Mean   :1036.4   Mean   :2.705
##  Other           : 69    3rd Qu.:1574.5   3rd Qu.:4.000
##  Technical Degree: 97    Max.   :2068.0   Max.   :4.000
##
##     Gender      HourlyRate     JobInvolvement     JobLevel
##  Female:482   Min.   : 30.00   Min.   :1.000   Min.   :1.000
##  Male  :694   1st Qu.: 48.00   1st Qu.:2.000   1st Qu.:1.000
##               Median : 66.00   Median :3.000   Median :2.000
##               Mean   : 65.82   Mean   :2.741   Mean   :2.069
##               3rd Qu.: 83.00   3rd Qu.:3.000   3rd Qu.:3.000
##               Max.   :100.00   Max.   :4.000   Max.   :5.000
##
##                       JobRole    JobSatisfaction  MaritalStatus
##  Sales Executive          :263   Min.   :1.00    Divorced:266
##  Research Scientist       :220   1st Qu.:2.00    Married :545
##  Laboratory Technician    :209   Median :3.00    Single  :365
##  Manufacturing Director   :122   Mean   :2.71
##  Healthcare Representative:108   3rd Qu.:4.00
##  Manager                  : 79   Max.   :4.00
##  (Other)                  :175
##  MonthlyIncome    MonthlyRate    NumCompaniesWorked Over18   OverTime
##  Min.   : 1009   Min.   : 2094   Min.   :0.000      Y:1176   No :839
##  1st Qu.: 2954   1st Qu.: 8275   1st Qu.:1.000               Yes:337
##  Median : 4950   Median :14488   Median :2.000
##  Mean   : 6526   Mean   :14468   Mean   :2.709
##  3rd Qu.: 8354   3rd Qu.:20627   3rd Qu.:4.000
##  Max.   :19973   Max.   :26999   Max.   :9.000
##
##  PercentSalaryHike PerformanceRating RelationshipSatisfaction
##  Min.   :11.0      Min.   :3.000     Min.   :1.000
##  1st Qu.:12.0      1st Qu.:3.000     1st Qu.:2.000
##  Median :14.0      Median :3.000     Median :3.000
##  Mean   :15.3      Mean   :3.162     Mean   :2.719
##  3rd Qu.:18.0      3rd Qu.:3.000     3rd Qu.:4.000
##  Max.   :25.0      Max.   :4.000     Max.   :4.000
##
##  StandardHours StockOptionLevel TotalWorkingYears TrainingTimesLastYear
##  Min.   :80    Min.   :0.0000   Min.   :  0.00    Min.   :0.00
##  1st Qu.:80    1st Qu.:0.0000   1st Qu.:  6.00    1st Qu.:2.00
##  Median :80    Median :1.0000   Median : 10.00    Median :3.00
##  Mean   :80    Mean   :0.7959   Mean   : 11.41    Mean   :2.81
##  3rd Qu.:80    3rd Qu.:1.0000   3rd Qu.: 15.00    3rd Qu.:3.00
##  Max.   :80    Max.   :3.0000   Max.   :114.00    Max.   :6.00
##
##  WorkLifeBalance YearsAtCompany   YearsInCurrentRole
##  Min.   :1.000   Min.   : 0.000   Min.   : 0.000
##  1st Qu.:2.000   1st Qu.: 3.000   1st Qu.: 2.000
##  Median :3.000   Median : 5.000   Median : 3.000
##  Mean   :2.747   Mean   : 6.918   Mean   : 4.151
##  3rd Qu.:3.000   3rd Qu.: 9.000   3rd Qu.: 7.000
```

```
##  3rd Qu.:3.000   3rd Qu.: 9.000   3rd Qu.: 7.000
##  Max.   :4.000   Max.   :40.000   Max.   :18.000
##
##  YearsSinceLastPromotion YearsWithCurrManager EmployeeCount
##  Min.   : 0.000          Min.   :  0.000      Min.   :1
##  1st Qu.: 0.000          1st Qu.:  2.000      1st Qu.:1
##  Median : 1.000          Median :  3.000      Median :1
##  Mean   : 2.126          Mean   :  4.242      Mean   :1
##  3rd Qu.: 2.000          3rd Qu.:  7.000      3rd Qu.:1
##  Max.   :15.000          Max.   :219.000      Max.   :1
##
```

##Checking for columns having less variance. ##Columns those have less variance and hence can be removed from the dataset for further analysis.

```
remove_cols <- nearZeroVar(emp, names = TRUE,
                           freqCut = 19, uniqueCut = 10)
remove_cols
```

```
## [1] "Over18"        "StandardHours" "EmployeeCount"
```

#Deleting above columns

```
emp2 = select(emp, -EmployeeCount, -Over18, -StandardHours)
View(emp2)
```

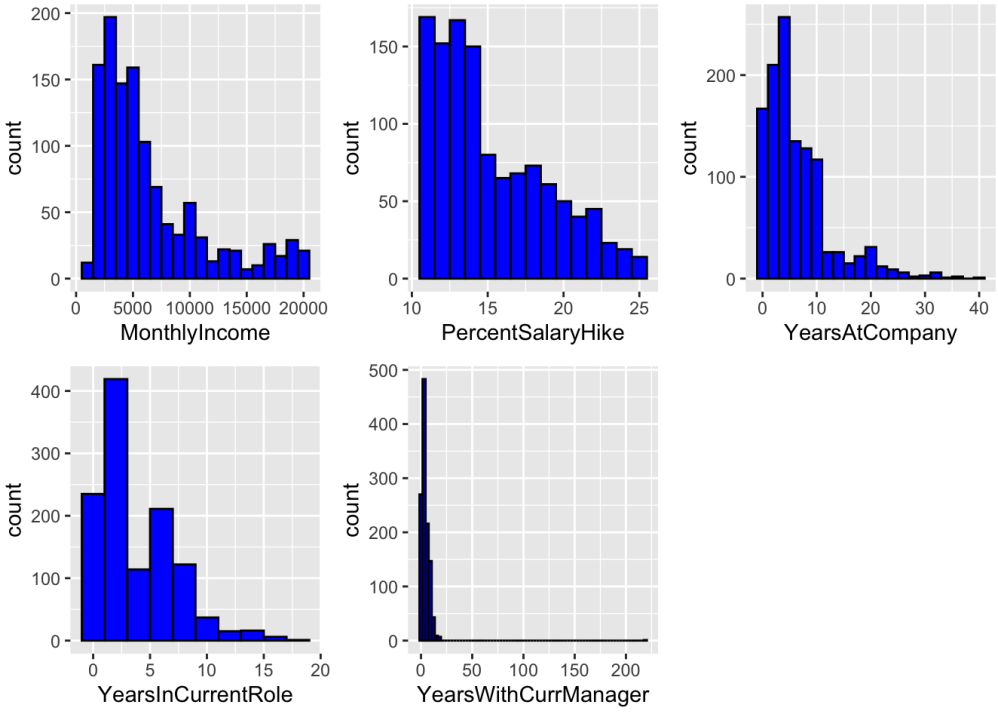
#Discretization of some attributes

```
emp2 <- emp2 %>%
mutate(Education = as.factor(if_else(Education == 1,"Below College",if_else(Education == 2,"College",if_else
(Education == 3, "Bachelor",if_else(Education == 4, "Master","Doctor")))))
         ,EnvironmentSatisfaction = as.factor(if_else(EnvironmentSatisfaction == 1,"Low",if_else(Environment
Satisfaction == 2, "Medium", if_else(EnvironmentSatisfaction == 3, "High","Very High"))))
         ,JobInvolvement = as.factor(if_else(JobInvolvement == 1,"Low",if_else(JobInvolvement == 2, "Medium"
,if_else(JobInvolvement == 3, "High", "Very High"))))
         ,JobSatisfaction = as.factor(if_else(JobSatisfaction == 1, "Low",if_else(JobSatisfaction == 2, "Med
ium",if_else(JobSatisfaction == 3, "High","Very High"))))
         ,PerformanceRating = as.factor(if_else(PerformanceRating == 1, "Low",if_else(PerformanceRating == 2
, "Good", if_else(PerformanceRating == 3, "Excellent", "Outstanding"))))
         ,RelationshipSatisfaction = as.factor(if_else(RelationshipSatisfaction == 1, "Low",if_else(Relation
shipSatisfaction == 2, "Medium", if_else(RelationshipSatisfaction == 3, "High", "Very High"))))
         ,WorkLifeBalance = as.factor(if_else(WorkLifeBalance == 1, "Bad",if_else(WorkLifeBalance == 2, "Goo
d", if_else(WorkLifeBalance == 3, "Better", "Best"))))
         ,JobLevel = as.factor(JobLevel),
       StockOptionLevel=as.factor(if_else(StockOptionLevel == 0,"None",if_else(StockOptionLevel == 1, "low",
if_else(StockOptionLevel == 2, "Medium",if_else(StockOptionLevel==4,"High","Very High")))))
  )
emp2$YearsSinceLastPromotion<-cut(emp2$YearsSinceLastPromotion, breaks = 5,
                                  labels = c("<5years", "5<Years<10", "10<Years<20","20<years<25",">25"), or
der = T)
```

#Data Visualisation as per various attributes

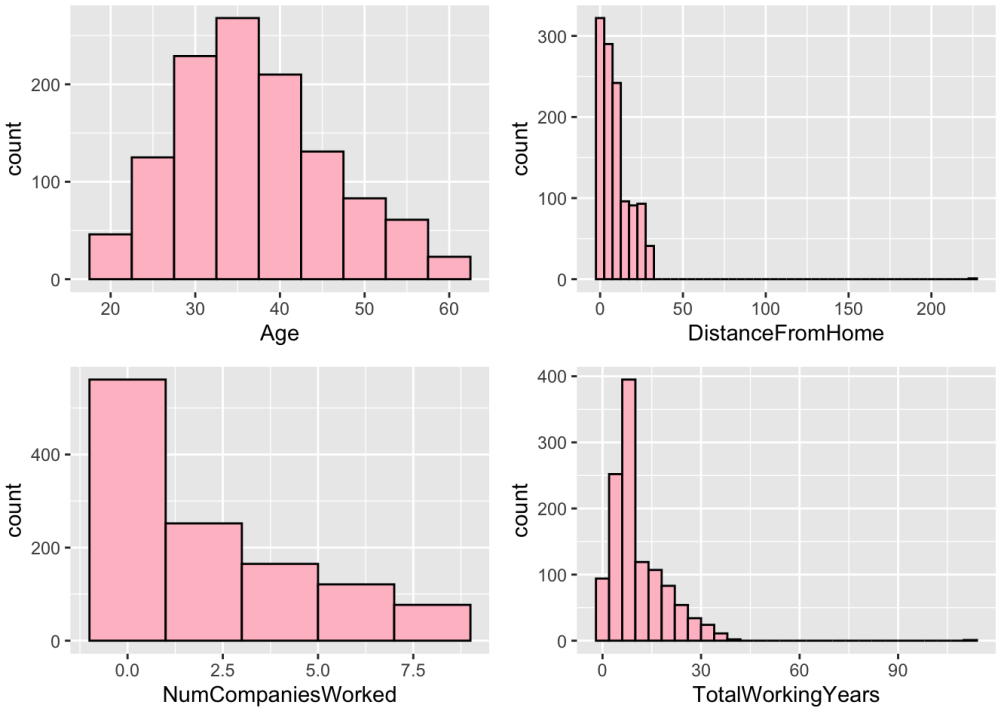
##Work Plots

```
p1 <- ggplot(emp2) + geom_histogram(aes(MonthlyIncome), binwidth = 1000, fill = "blue",col = "black")
p2 <- ggplot(emp2) + geom_histogram(aes(PercentSalaryHike), binwidth = 1, fill = "blue",col = "black")
p3 <- ggplot(emp2) + geom_histogram(aes(YearsAtCompany), binwidth = 2, fill = "blue",col = "black")
p4 <- ggplot(emp2) + geom_histogram(aes(YearsInCurrentRole), binwidth = 2, fill = "blue",col = "black")

p5 <- ggplot(emp2) + geom_histogram(aes(YearsWithCurrManager), binwidth = 3, fill = "blue",col = "black")

grid.arrange(p1, p2, p3, p4, p5, nrow = 2, ncol = 3)
```



##Personal facts Plots

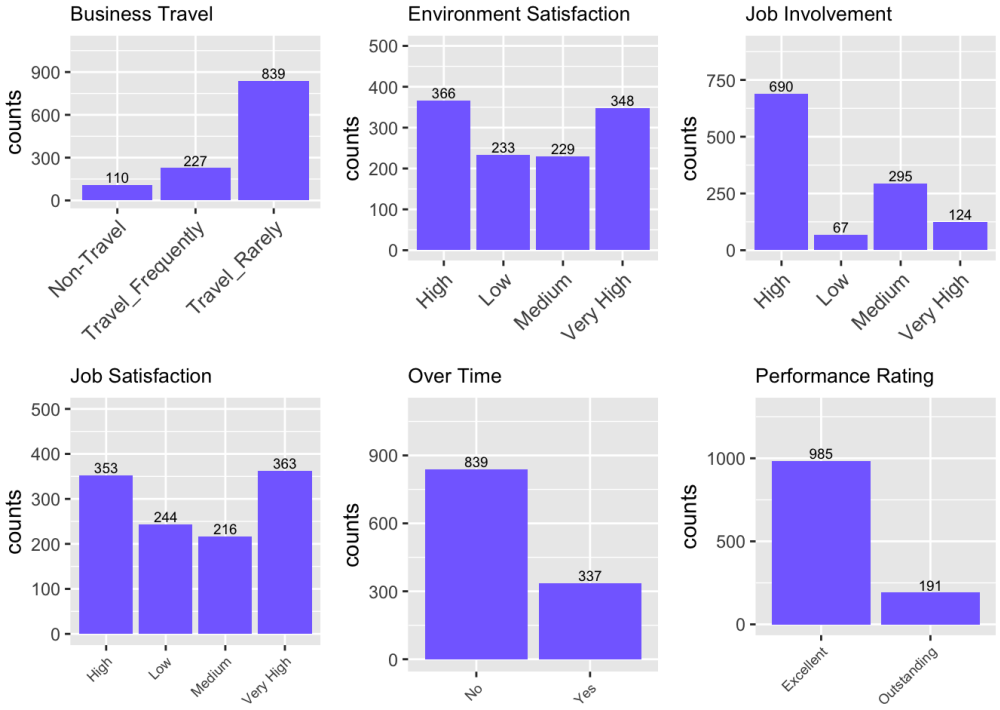
```
p1 <- ggplot(emp2) + geom_histogram(aes(Age), binwidth = 5, fill = "Pink",col = "black")
p2 <- ggplot(emp2) + geom_histogram(aes(DistanceFromHome), binwidth = 5, fill = "Pink",col = "black")
p3 <- ggplot(emp2) + geom_histogram(aes(NumCompaniesWorked), binwidth = 2, fill = "Pink",col = "black")
p4 <- ggplot(emp2) + geom_histogram(aes(TotalWorkingYears), binwidth = 4, fill = "Pink",col = "black")

grid.arrange(p1, p2, p3, p4, ncol = 2, nrow = 2)
```



Age can be seen normally distributed.

##Work Plots Category variables

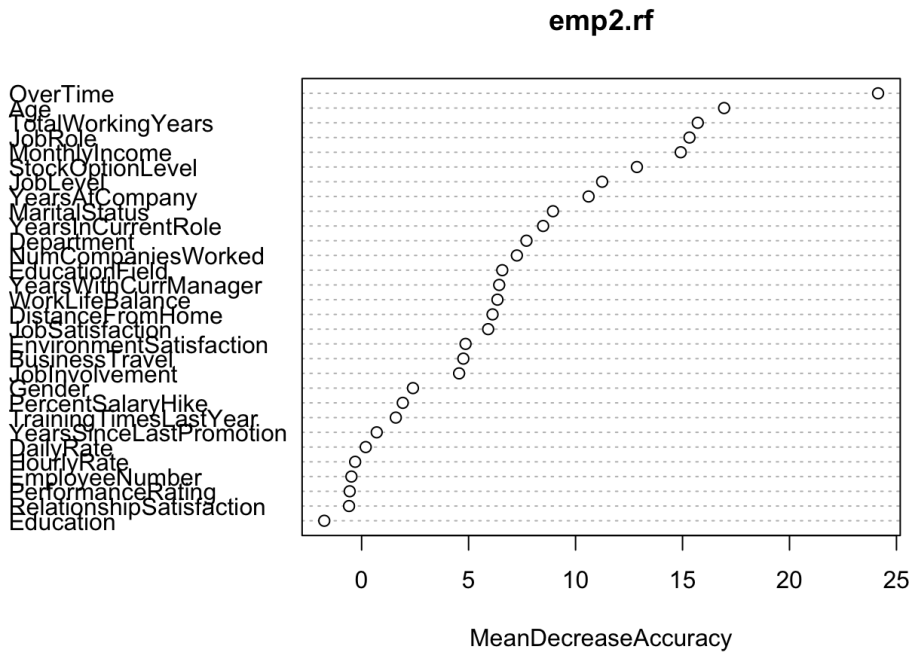
```
p1 <- emp2 %>%
  group_by(BusinessTravel) %>%
  summarise(counts = n()) %>%
  ggplot(aes(x = as.factor(BusinessTravel), y = counts)) + geom_bar(stat = 'identity', fill = "lightslateblu
e") + ggtitle("Business Travel") +geom_text(aes(label=counts), size = 2.5, position=position_dodge(width=0.2
), vjust=-0.25)+ theme(plot.title = element_text(size =10),axis.text.x = element_text(size =10,angle = 45, h
just = 1),axis.title.x=element_blank())+ scale_y_continuous(limits = c(0, 1100))

p2 <- emp2 %>%
  group_by(EnvironmentSatisfaction) %>%
  summarise(counts = n()) %>%
  ggplot(aes(x = as.factor(EnvironmentSatisfaction), y = counts)) + geom_bar(stat = 'identity', fill = "ligh
tslateblue") + ggtitle("Environment Satisfaction") + geom_text(aes(label=counts), size = 2.5, position=posit
ion_dodge(width=0.2), vjust=-0.25) + theme(plot.title = element_text(size =10),axis.text.x = element_text(si
ze =10,angle = 45, hjust = 1),axis.title.x=element_blank()) + scale_y_continuous(limits = c(0, 500))

p3 <- emp2 %>%
  group_by(JobInvolvement) %>%
  summarise(counts = n()) %>%
  ggplot(aes(x = as.factor(JobInvolvement), y = counts)) + geom_bar(stat = 'identity', fill = "lightslateblu
e") + ggtitle("Job Involvement") +geom_text(aes(label=counts), size = 2.5, position=position_dodge(width=0.2
), vjust=-0.25)+ theme(plot.title = element_text(size =10),axis.text.x = element_text(size =10,angle = 45, h
just = 1),axis.title.x=element_blank()) + scale_y_continuous(limits = c(0, 900))

p4 <- emp2 %>%
  group_by(JobSatisfaction) %>%
  summarise(counts = n()) %>%
  ggplot(aes(x = as.factor(JobSatisfaction), y = counts)) + geom_bar(stat = 'identity', fill = "lightsteblu
e") + ggtitle("Job Satisfaction") +geom_text(aes(label=counts), size = 2.5, position=position_dodge(width=0
.2), vjust=-0.25) + theme(plot.title = element_text(size =10),axis.text.x = element_text(size =7,angle = 45,
hjust = 1),axis.title.x=element_blank()) + scale_y_continuous(limits = c(0, 500))

p5 <- emp2 %>%
  group_by(OverTime) %>%
  summarise(counts = n()) %>%
  ggplot(aes(x = as.factor(OverTime), y = counts)) + geom_bar(stat = 'identity', fill = "lightslateblue") +
ggtitle("Over Time") +geom_text(aes(label=counts), size = 2.5, position=position_dodge(width=0.2), vjust=-0.
25)+ theme(plot.title = element_text(size =10),axis.text.x = element_text(size =7,angle = 45, hjust = 1),ax
is.title.x=element_blank()) + scale_y_continuous(limits = c(0, 1100))

p6 <- emp2 %>%
  group_by(PerformanceRating) %>%
  summarise(counts = n()) %>%
  ggplot(aes(x = as.factor(PerformanceRating), y = counts)) + geom_bar(stat = 'identity', fill = "lightslate
blue") + ggtitle("Performance Rating") +geom_text(aes(label=counts), size = 2.5, position=position_dodge(wid
th=0.2), vjust=-0.25)+ theme(plot.title = element_text(size =10),axis.text.x = element_text(size =7,angle =
45, hjust = 1),axis.title.x=element_blank()) + scale_y_continuous(limits = c(0, 1300))

grid.arrange(p1,p2,p3,p4,p5,p6,nrow = 2)
```



More no of employees are seen Traveling rarely for work.

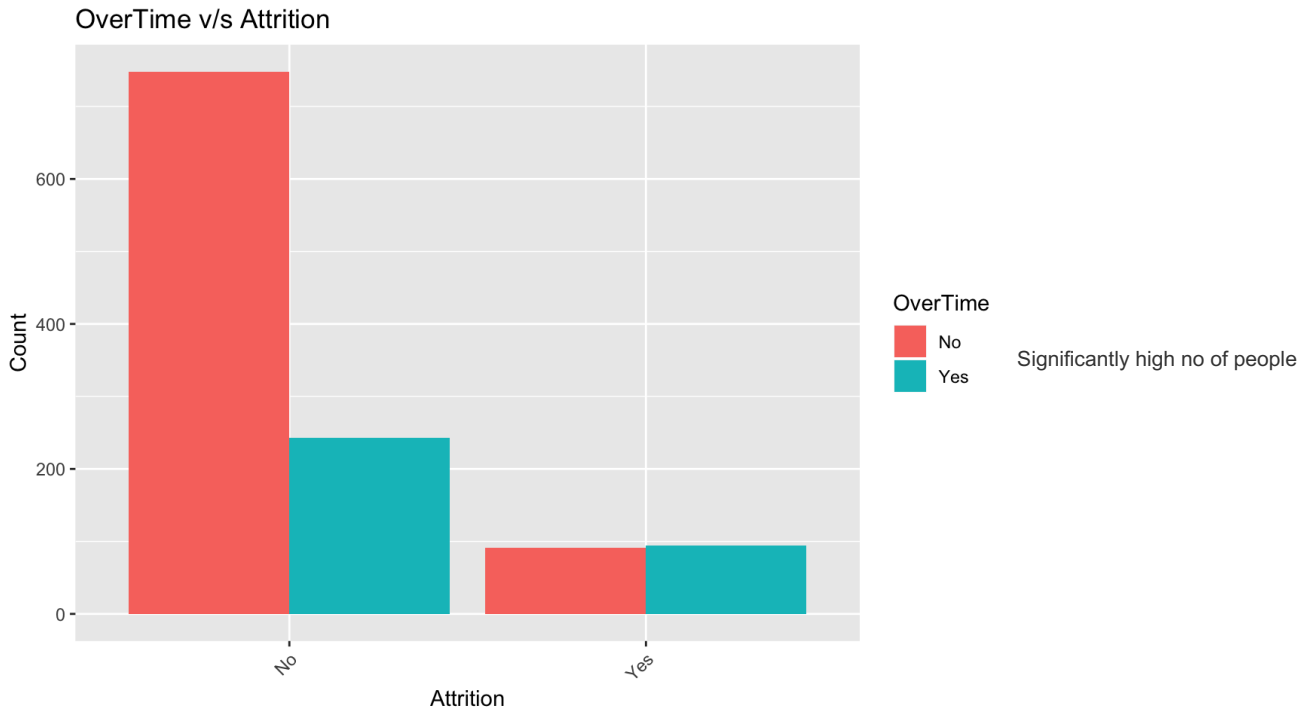
#Using variable importance graph to determine most impactful variables. ##variables with a large mean decrease in accuracy are more important for classification of the data.

```
set.seed(4543)
emp2.rf <- randomForest(emp2$Attrition ~ ., data=emp2, ntree=1000, keep.forest=FALSE,
                        importance=TRUE)
varImpPlot(emp2.rf,sort=TRUE,type=1,n.var=30)
```



#Graphs with respect to Attrition ##1. OverTime v/s Attrition

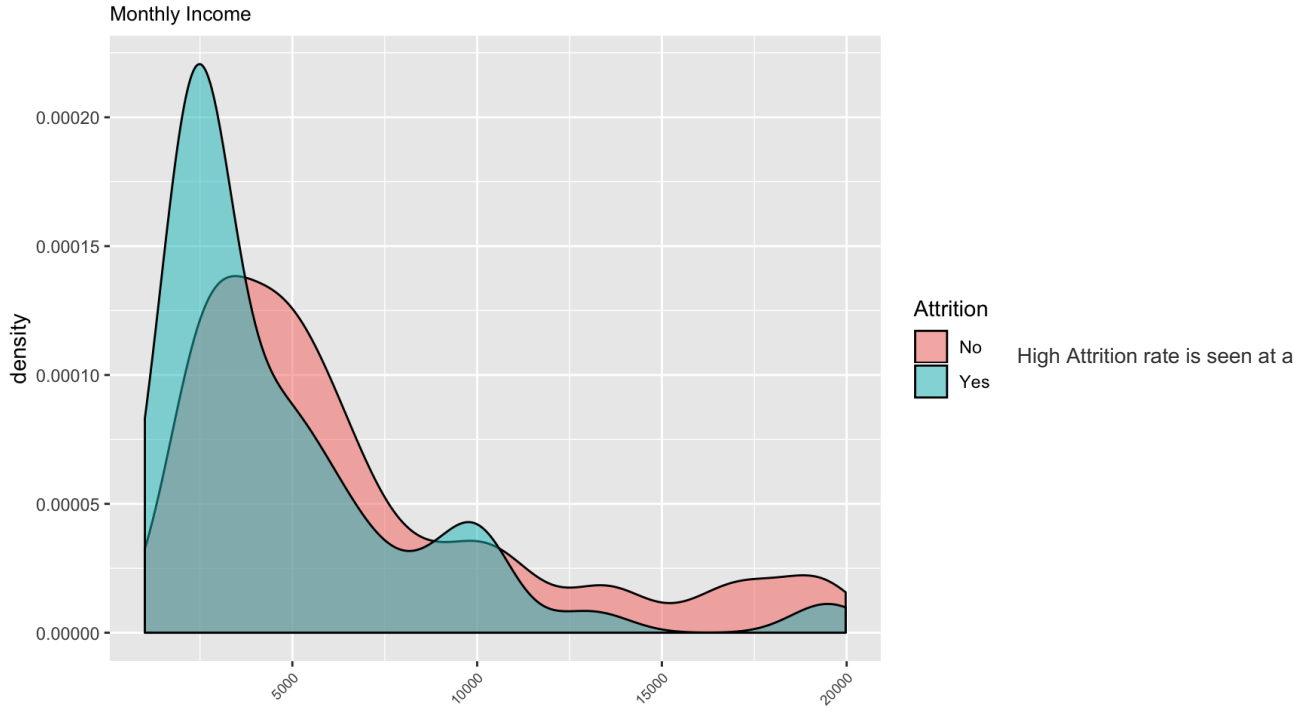
```
OverTime_Attrition <-  ggplot(emp2) +
                      aes(x = emp2$Attrition, fill = emp2$OverTime) +
                      geom_bar(position = "dodge") +
                      theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
                      ggtitle("OverTime v/s Attrition") +
                      labs(x = "Attrition", y = "Count", fill = "OverTime")
OverTime_Attrition
```



working Over Time result in attrition.

##2.MonthlyIncome v/s Attrition

```
Mi <- emp2 %>%
              ggplot(aes(x = MonthlyIncome, fill = Attrition)) +
              geom_density(alpha = 0.5) +
              ggtitle("Monthly Income") +
              theme(plot.title = element_text(size =10),axis.text.x = element_text(size =7,angle = 45, hjus
t = 1),axis.title.x=element_blank())
Mi
```



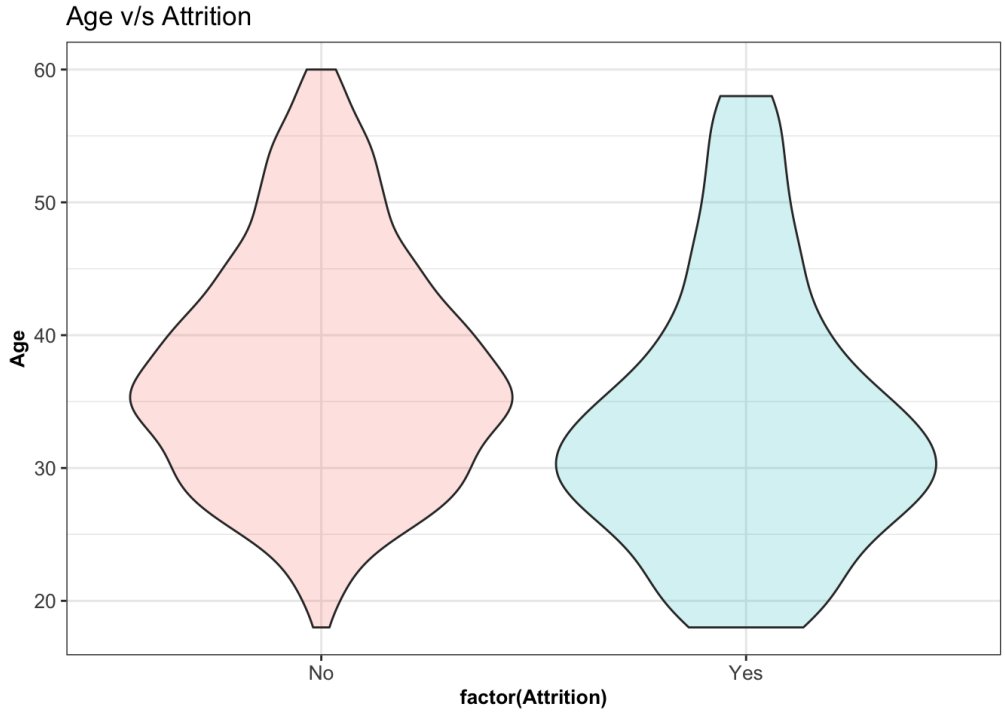
High Attrition rate is seen at a monthly income approx. 2500.

##3.Age v/s Attrition

```
Age<- ggplot(emp2, aes(factor(Attrition), Age))+ geom_violin(alpha = 0.2, aes(fill = factor(Attrition)))+
  theme_bw()+
  guides(fill=FALSE)+theme(axis.text=element_text(size=10),
                           axis.title=element_text(size=10,face="bold"),
                           legend.text=element_text(size=10),
                           legend.title=element_text(size=14),legend.position = "bottom")+
  ggtitle("Age v/s Attrition")
  xlab("Attrition")
```

```
## $x
## [1] "Attrition"
##
## attr(,"class")
## [1] "labels"
```

```
Age
```



Employees within 25-35 years have a higher attrition rate.

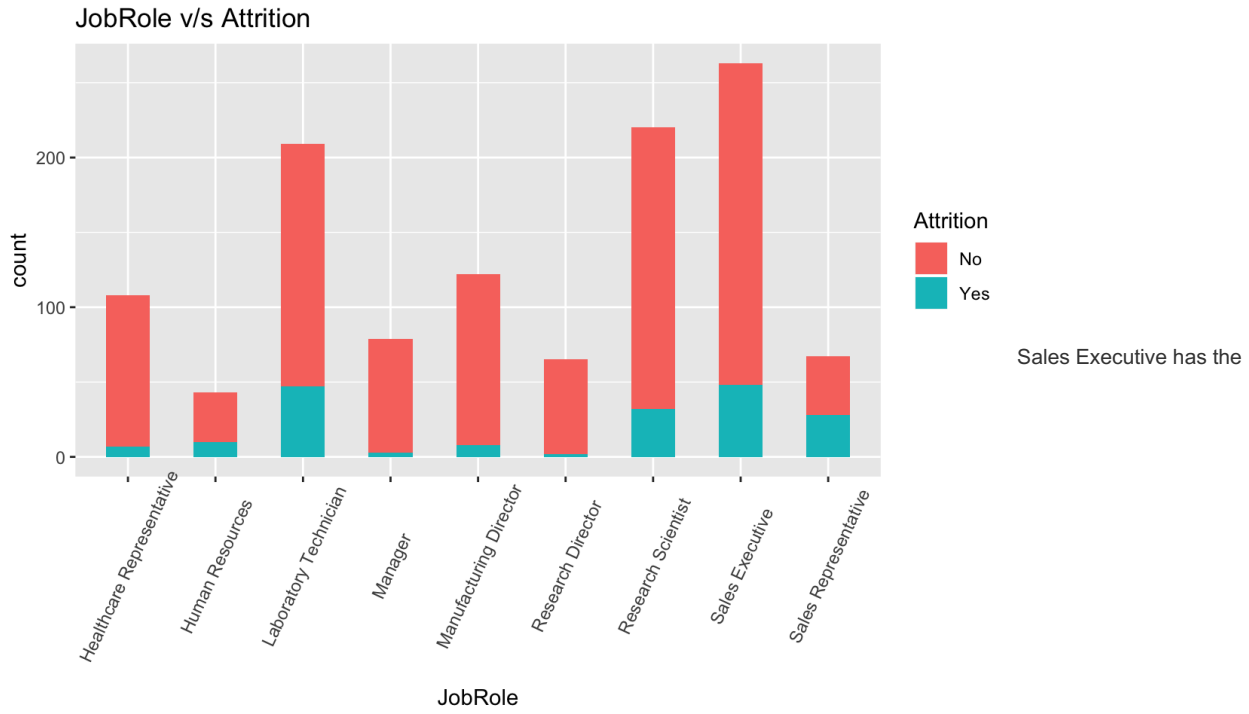
##3.YearsAtCompany v/s Attrition

```
Yac <- emp2 %>%
              ggplot(aes(x = YearsAtCompany, fill = Attrition)) +
              geom_density(alpha = 0.5) + ggtitle("Years At Company with respect to Attrition") +
              theme(plot.title = element_text(size =10),axis.text.x = element_text(size =7,angle = 45, hjus
t = 1),axis.title.x=element_blank())
Yac
```

Higher attrition rate is seen in when the employee is with the company for 0-2 years approx.

##5.JobRole v/s Attrition

```
JobRole <-  ggplot(emp2, aes(JobRole))
            JobRole + geom_bar(aes(fill=Attrition), width = 0.5) +
            theme(axis.text.x = element_text(angle=65, vjust=0.6)) +
            labs(title="JobRole v/s Attrition")
```



Sales Executive has the highest attrition rate.

#**********Association Rule Mining**********

```
Att_rules <- apriori(data=emp2)
```

```
## Warning: Column(s) 1, 4, 6, 9, 12, 18, 19, 20, 22, 26, 27, 29, 30, 32 not
## logical or factor. Applying default discretization (see '? discretizeDF').
```

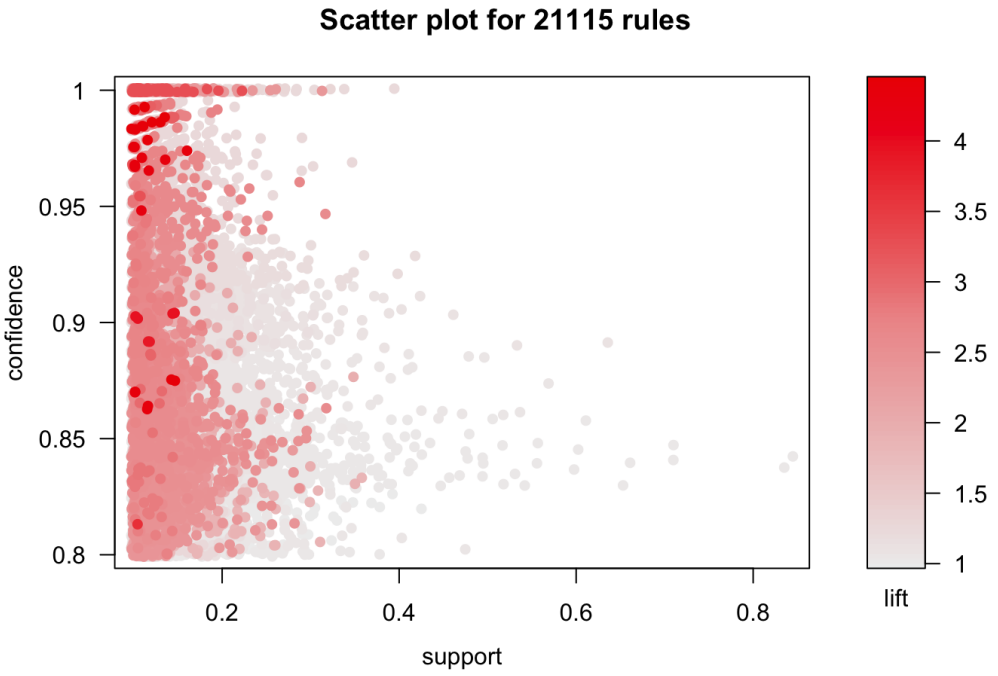
```
## Apriori
##
## Parameter specification:
##  confidence minval smax arem  aval originalSupport maxtime support minlen
##         0.8    0.1    1 none FALSE            TRUE       5     0.1      1
##  maxlen target   ext
##      10  rules FALSE
##
## Algorithmic control:
##  filter tree heap memopt load sort verbose
##     0.1 TRUE TRUE  FALSE TRUE    2    TRUE
##
## Absolute minimum support count: 117
##
## set item appearances ...[0 item(s)] done [0.00s].
## set transactions ...[113 item(s), 1176 transaction(s)] done [0.01s].
## sorting and recoding items ... [91 item(s)] done [0.00s].
## creating transaction tree ... done [0.00s].
## checking subsets of size 1 2 3 4 5 6 7 done [0.07s].
## writing ... [21115 rule(s)] done [0.00s].
## creating S4 object  ... done [0.01s].
```

```
inspect(head(sort(Att_rules, by='confidence'),5))
```

```
##     lhs                                  rhs                                    support confidence     li
ft count
## [1] {JobRole=Manufacturing Director} => {Department=Research & Development} 0.1037415          1 1.539267
122
## [2] {EducationField=Marketing}       => {Department=Sales}                   0.1079932          1 3.28491
6   127
## [3] {PerformanceRating=Outstanding}  => {PercentSalaryHike=[17,25]}          0.1624150          1 2.992366
191
## [4] {JobRole=Laboratory Technician}  => {Department=Research & Development} 0.1777211          1 1.539267
209
## [5] {JobRole=Research Scientist}     => {Department=Research & Development} 0.1870748          1 1.539267
220
```

```
plot(Att_rules)
```

```
## To reduce overplotting, jitter is added! Use jitter = 0 to prevent jitter.
```




```
Att_rules <- apriori(data=emp2, parameter=list (supp=0.3,conf =0.5, minlen= 4, maxtime=10, target = "rules")
)
```

```
## Warning: Column(s) 1, 4, 6, 9, 12, 18, 19, 20, 22, 26, 27, 29, 30, 32 not
## logical or factor. Applying default discretization (see '? discretizeDF').
```

```
## Apriori
##
## Parameter specification:
##  confidence minval smax arem  aval originalSupport maxtime support minlen
##         0.5    0.1    1 none FALSE            TRUE      10     0.3      4
##  maxlen target   ext
##      10  rules FALSE
##
## Algorithmic control:
##  filter tree heap memopt load sort verbose
##     0.1 TRUE TRUE  FALSE TRUE    2    TRUE
##
## Absolute minimum support count: 352
##
## set item appearances ...[0 item(s)] done [0.00s].
## set transactions ...[113 item(s), 1176 transaction(s)] done [0.00s].
## sorting and recoding items ... [60 item(s)] done [0.00s].
## creating transaction tree ... done [0.00s].
## checking subsets of size 1 2 3 4 5 done [0.00s].
## writing ... [105 rule(s)] done [0.00s].
## creating S4 object  ... done [0.00s].
```

```
inspect(head(sort(Att_rules, by='confidence'),5))
```

```
##     lhs                                    rhs                                 support confidence     l
ift count
## [1] {Attrition=No,

##      YearsInCurrentRole=[2,6),

##      YearsWithCurrManager=[2,6)}        => {YearsSinceLastPromotion=<5years} 0.3001701  0.9671233 1.22822
6   353
## [2] {BusinessTravel=Travel_Rarely,

##      Department=Research & Development,

##      OverTime=No}                       => {Attrition=No}                    0.3103741  0.9170854 1.0882
87   365
## [3] {Department=Research & Development,

##      OverTime=No,

##      PerformanceRating=Excellent}       => {Attrition=No}                    0.3477891  0.9149888 1.0857
99   409
## [4] {JobInvolvement=High,

##      OverTime=No,

##      PerformanceRating=Excellent}       => {Attrition=No}                    0.3163265  0.9117647 1.0819
73   372
## [5] {Department=Research & Development,

##      OverTime=No,

##      YearsSinceLastPromotion=<5years}   => {Attrition=No}                    0.3350340  0.9057471 1.0748
32   394
```

```
Association_rules <- apriori(data=emp2, parameter=list (supp=0.3,conf =0.5, minlen= 5, maxtime=10, target =
"rules"), appearance = list (rhs=c("Attrition=Yes")))
```

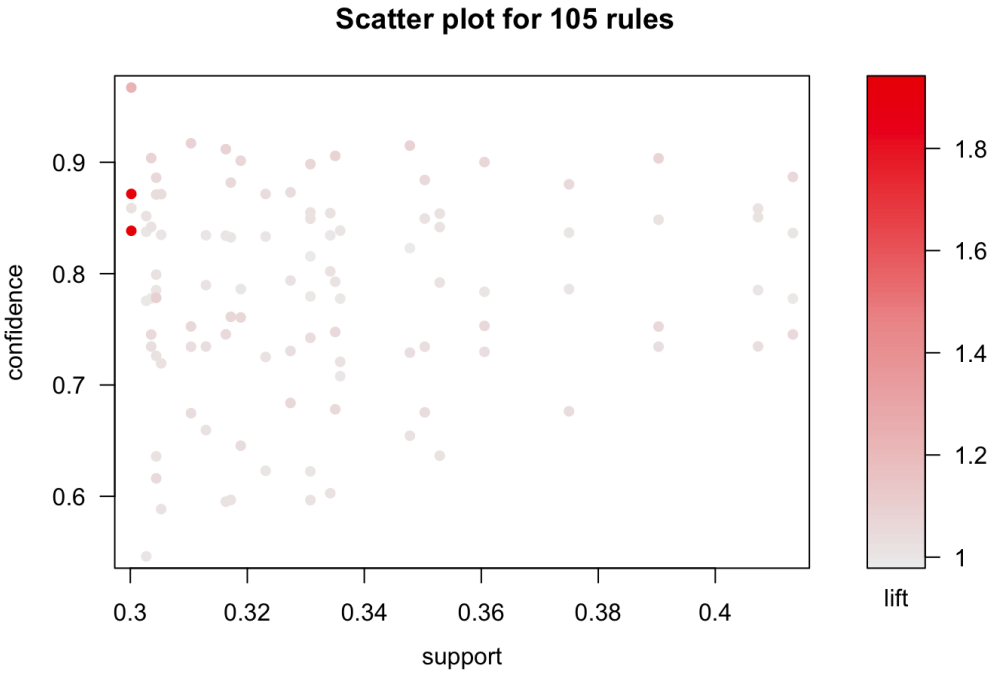
```
## Warning: Column(s) 1, 4, 6, 9, 12, 18, 19, 20, 22, 26, 27, 29, 30, 32 not
## logical or factor. Applying default discretization (see '? discretizeDF').
```

```
## Apriori
##
## Parameter specification:
##  confidence minval smax arem  aval originalSupport maxtime support minlen
##         0.5    0.1    1 none FALSE            TRUE      10     0.3      5
##  maxlen target   ext
##      10  rules FALSE
##
## Algorithmic control:
##  filter tree heap memopt load sort verbose
##     0.1 TRUE TRUE  FALSE TRUE    2    TRUE
##
## Absolute minimum support count: 352
##
## set item appearances ...[1 item(s)] done [0.00s].
## set transactions ...[113 item(s), 1176 transaction(s)] done [0.00s].
## sorting and recoding items ... [60 item(s)] done [0.00s].
## creating transaction tree ... done [0.00s].
## checking subsets of size 1 2 3 4 5 done [0.00s].
## writing ... [0 rule(s)] done [0.00s].
## creating S4 object  ... done [0.00s].
```

```
inspect(head(sort(Att_rules, by='confidence'),5))
```

```
##     lhs                                    rhs                                 support confidence     l
ift count
## [1] {Attrition=No,

##      YearsInCurrentRole=[2,6),

##      YearsWithCurrManager=[2,6)}        => {YearsSinceLastPromotion=<5years} 0.3001701  0.9671233 1.22822
6   353
## [2] {BusinessTravel=Travel_Rarely,

##      Department=Research & Development,

##      OverTime=No}                       => {Attrition=No}                    0.3103741  0.9170854 1.0882
87   365
## [3] {Department=Research & Development,

##      OverTime=No,

##      PerformanceRating=Excellent}       => {Attrition=No}                    0.3477891  0.9149888 1.0857
99   409
## [4] {JobInvolvement=High,

##      OverTime=No,

##      PerformanceRating=Excellent}       => {Attrition=No}                    0.3163265  0.9117647 1.0819
73   372
## [5] {Department=Research & Development,

##      OverTime=No,

##      YearsSinceLastPromotion=<5years}   => {Attrition=No}                    0.3350340  0.9057471 1.0748
32   394
```

```
plot(Att_rules)
```


Scatter plot for 105 rules
confidence
0.9
0.8
0.7
0.6
0.3
0.32
0.34
0.36
0.38
0.4
support
1.8
1.6
1.4
1.2
1
lift