final_capstone_cyo

Muhammad Azeem

1/9/2020

```
library(knitr)
```

FINAL PROJECT - CYO [HarvardX Data Science Capstone (HarvardX: PH125.9x)]

INTRODUCTION

This document is comprised of the analysis report which was prepared as a part of HarvardX Data Science Capstone (HarvardX: PH125.9x)-CYO. The field chosen for this project is Human Resource Management. As the field of Artificial Intelligence (AI) is sneaking into many areas of human lives , e.g., engineering, product designing, education, space exploration, geo sciences, health sciences, etc., the case of Business studies is not very different. All functional areas of business management are quite vulnerable to such technological intrusions. Human resource management is one such field which will be adapted to AI. Many current challenges in manpower planning, job analysis and staffing, recruitment and selection, training and development, performance management, employee's relationship, good work environment, etc. One of the major issues that HR manger is facing is the problem of employees attrition. There are many terms used for this concept. For example; employees' turnover, termination, resignation and withdrawal, churn etc. The term attrition is covering all such concepts whether voluntary or involuntary. The study of the attrition problem can help mangers to understand who is vulnerable to such situations, and whether they can predict the potential cases for the leaving the organization. Such information will also guide to the workforce planning, staffing, and recruitment processes. In this study the attrition issue is examined. The data for the study was borrowed from the kaggle.com public data repository. The reference of the data page is given at the end of this document. The title of the data is "IBM-HR Analytics "IBM HR Analytics Employee Attrition & Performance". The goal in this study is to predict attrition of the employees. According to the information given on the page, the data set was created by the IBM data scientists"." Its main purpose is to demonstrate the Watson analytics tool for employee attrition". The data is considered quite appropriate for the current project as it has enough cases and multiple features to develop a classification model.

GOAL

The goal is to predict the employee’s attrition based on the known features.

METHODOLOGY

According to the goal of the study, the project is in the scope of the supervised learning model. The four ML algorithms are employed to predict the attrition from the given data. They are:

1: Decision Tree 2: Random Forest 3: Logistic Regression

Out of given 35 variables the 12 were removed to bring the data into a manageable rationale. The final data set has 1470 cases and 13 variables. The analysis included following sections:

1. Installing the required packages/libraries
2. Calling IBM-HR dataset
3. Preparing the data for analysis
4. Examine the structure of the final data set.
5. Replacing Yes and No in the label variable by 1 and 0 respectively.
6. Examine for the missing values
7. Univariate Analysis
8. Bivariate Analysis
9. Modeling
10. Conclusion
11. limitations
12. Future work
13. References/Bibliography

Installing the required packages/libraries

```
if(!require(tidyverse)) install.packages("tidyverse", repos = "http://cran.us.r-pro
ject.org")
```

```
## Loading required package: tidyverse
```

```
## -- Attaching packages ---------------------------------------------------
---------------------------- tidyverse 1.2.1 --
```

```
## v ggplot2 3.2.1     v purrr   0.3.2
## v tibble  2.1.3     v dplyr   0.8.3
## v tidyr   1.0.0     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0
```

```
## Warning: package 'ggplot2' was built under R version 3.6.2
```

```
## Warning: package 'tidyr' was built under R version 3.6.2
```

```
## -- Conflicts ------------------------------------------------------------
---------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
if(!require(caret)) install.packages("caret", repos = "http://cran.us.r-project.or
g")
```

```
## Loading required package: caret
```

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

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

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

```
if(!require(Amelia)) install.packages("Amelia", repos = "http://cran.us.r-project.o
rg")
```

```
## Loading required package: Amelia
```

```
## Warning: package 'Amelia' was built under R version 3.6.2
```

```
## Loading required package: Rcpp
```

```
## ##
## ## Amelia II: Multiple Imputation
## ## (Version 1.7.6, built: 2019-11-24)
## ## Copyright (C) 2005-2020 James Honaker, Gary King and Matthew Blackwell
## ## Refer to http://gking.harvard.edu/amelia/ for more information
## ##
```

```
if(!require(psych)) install.packages("psych", repos = "http://cran.us.r-project.org")
```

```
## Loading required package: psych
```

```
## Warning: package 'psych' was built under R version 3.6.2
```

```
##
## Attaching package: 'psych'
```

```
## The following objects are masked from 'package:ggplot2':
##
##     %+%, alpha
```

```
if(!require(corrplot)) install.packages("corrplot", repos = "http://cran.us.r-project.org")
```

```
## Loading required package: corrplot
```

```
## Warning: package 'corrplot' was built under R version 3.6.2
```

```
## corrplot 0.84 loaded
```

```
if(!require(GoodmanKruskal)) install.packages("GoodmanKruskal", repos = "http://cran.us.r-project.org")
```

```
## Loading required package: GoodmanKruskal
```

```
## Warning: package 'GoodmanKruskal' was built under R version 3.6.2
```

```
if(!require(rpart)) install.packages("rpart", repos = "http://cran.us.r-project.org")
```

```
## Loading required package: rpart
```

```
## Warning: package 'rpart' was built under R version 3.6.2
```

```
if(!require(rpart.plot)) install.packages("rpart.plot", repos = "http://cran.us.r-p
roject.org")
```

```
## Loading required package: rpart.plot
```

```
## Warning: package 'rpart.plot' was built under R version 3.6.2
```

```
if(!require(party)) install.packages("party", repos = "http://cran.us.r-project.or
g")
```

```
## Loading required package: party
```

```
## Warning: package 'party' was built under R version 3.6.2
```

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

```
## Loading required package: mvtnorm
```

```
## Loading required package: modeltools
```

```
## Loading required package: stats4
```

```
## Loading required package: strucchange
```

```
## Warning: package 'strucchange' was built under R version 3.6.2
```

```
## Loading required package: zoo
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
## Loading required package: sandwich
```

```
##
## Attaching package: 'strucchange'
```

```
## The following object is masked from 'package:stringr':
##
##     boundary
```

```
if(!require(randomForest)) install.packages("randomForest", repos = "http://cran.u
s.r-project.org")
```

```
## Loading required package: randomForest
```

```
## Warning: package 'randomForest' was built under R version 3.6.2
```

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

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

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

```
## The following object is masked from 'package:psych':
##
##     outlier
```

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

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

```
if(!require(ROCR)) install.packages("ROCR", repos = "http://cran.us.r-project.org")
```

```
## Loading required package: ROCR
```

```
## Warning: package 'ROCR' was built under R version 3.6.2
```

```
## Loading required package: gplots
```

```
## Warning: package 'gplots' was built under R version 3.6.2
```

```
##
## Attaching package: 'gplots'
```

```
## The following object is masked from 'package:stats':
##
##     lowess
```

```
if(!require(pscl)) install.packages("pscl", repos = "http://cran.us.r-project.org")
```

```
## Loading required package: pscl
```

```
## Warning: package 'pscl' was built under R version 3.6.2
```

```
## Classes and Methods for R developed in the
## Political Science Computational Laboratory
## Department of Political Science
## Stanford University
## Simon Jackman
## hurdle and zeroinfl functions by Achim Zeileis
```

```
if(!require(rcompanion)) install.packages("rcompanion", repos = "http://cran.us.r-p
roject.org")
```

```
## Loading required package: rcompanion
```

```
## Warning: package 'rcompanion' was built under R version 3.6.2
```

```
## Registered S3 method overwritten by 'DescTools':
##   method         from
##   reorder.factor gdata
```

```
##
## Attaching package: 'rcompanion'
```

```
## The following object is masked from 'package:psych':
##
##     phi
```

Calling IBM-HR dataset

```
df<- read.csv("WA_Fn-UseC_-HR-Employee-Attrition.csv")
```

Preperaing the data for analysis

Examining the dimensions and structure of the data set

```
dim(df)
```

```
## [1] 1470   35
```

> There are 1470 cases and 35 variables.

```
str(df)
```

```
## 'data.frame':    1470 obs. of  35 variables:
##  $ ï..EmployeeNumber       : int  1 2 4 5 7 8 10 11 12 13 ...
##  $ Age                     : int  41 49 37 33 27 32 59 30 38 36 ...
##  $ Over18                  : Factor w/ 1 level "Y": 1 1 1 1 1 1 1 1 1 1 ...
##  $ Gender                  : Factor w/ 2 levels "Female","Male": 1 2 2 1 2 2 1 2
2 2 ...
##  $ MaritalStatus           : Factor w/ 3 levels "Divorced","Married",..: 3 2 3 2
2 3 2 1 3 2 ...
##  $ Education               : int  2 1 2 4 1 2 3 1 3 3 ...
##  $ EducationField          : Factor w/ 6 levels "Human Resources",..: 2 2 5 2 4
2 4 2 2 4 ...
##  $ JobLevel                : int  2 2 1 1 1 1 1 1 3 2 ...
##  $ Department              : Factor w/ 3 levels "Human Resources",..: 3 2 2 2 2
2 2 2 2 2 ...
##  $ JobRole                 : Factor w/ 9 levels "Healthcare Representative",..:
8 7 3 7 3 3 3 3 5 1 ...
##  $ HourlyRate              : int  94 61 92 56 40 79 81 67 44 94 ...
##  $ DailyRate               : int  1102 279 1373 1392 591 1005 1324 1358 216 1299
...
##  $ MonthlyIncome           : int  5993 5130 2090 2909 3468 3068 2670 2693 9526 5
237 ...
##  $ MonthlyRate             : int  19479 24907 2396 23159 16632 11864 9964 13335
8787 16577 ...
##  $ NumCompaniesWorked      : int  8 1 6 1 9 0 4 1 0 6 ...
##  $ DistanceFromHome        : int  1 8 2 3 2 2 3 24 23 27 ...
##  $ EmployeeCount           : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ EnvironmentSatisfaction : int  2 3 4 4 1 4 3 4 4 3 ...
##  $ JobInvolvement          : int  3 2 2 3 3 3 4 3 2 3 ...
##  $ BusinessTravel          : Factor w/ 3 levels "Non-Travel","Travel_Frequentl
y",..: 3 2 3 2 3 2 3 3 2 3 ...
##  $ OverTime                : Factor w/ 2 levels "No","Yes": 2 1 2 2 1 1 2 1 1 1
...
##  $ 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 ...
##  $ JobSatisfaction         : int  4 2 3 3 2 4 1 3 3 3 ...
##  $ Attrition               : Factor w/ 2 levels "No","Yes": 2 1 2 1 1 1 1 1 1 1
...
```

Data frame of 1470 observations and 35 variables.

Removing the irrelavant and those with constant values.

```
df<- df[-c(1,3,7,8, 9, 10, 11,12,14,15, 16, 17,18, 19, 20, 21, 24, 25,27,28, 32,3
3)]
str(df)
```

```
## 'data.frame':    1470 obs. of  13 variables:
##  $ Age               : int  41 49 37 33 27 32 59 30 38 36 ...
##  $ Gender            : Factor w/ 2 levels "Female","Male": 1 2 2 1 2 2 1 2 2 2
...
##  $ MaritalStatus     : Factor w/ 3 levels "Divorced","Married",..: 3 2 3 2 2 3 2
1 3 2 ...
##  $ Education         : int  2 1 2 4 1 2 3 1 3 3 ...
##  $ MonthlyIncome     : int  5993 5130 2090 2909 3468 3068 2670 2693 9526 5237
...
##  $ PercentSalaryHike : int  11 23 15 11 12 13 20 22 21 13 ...
##  $ PerformanceRating : int  3 4 3 3 3 3 4 4 4 3 ...
##  $ StockOptionLevel  : int  0 1 0 0 1 0 3 1 0 2 ...
##  $ 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 ...
##  $ JobSatisfaction   : int  4 2 3 3 2 4 1 3 3 3 ...
##  $ Attrition         : Factor w/ 2 levels "No","Yes": 2 1 2 1 1 1 1 1 1 1 ...
```

list of columns names.

```
names(df)
```

```
##  [1] "Age"                "Gender"             "MaritalStatus"
##  [4] "Education"          "MonthlyIncome"      "PercentSalaryHike"
##  [7] "PerformanceRating"  "StockOptionLevel"   "WorkLifeBalance"
## [10] "YearsAtCompany"     "YearsInCurrentRole" "JobSatisfaction"
## [13] "Attrition"
```

Renaming the columns to the shorter form.

> Renamung “Age” as “age”, “Gender” as “gen”, “MaritalStatus” as “marital”, “Education” as “edu”, “MonthlyIncome” as “income”, “PercentSalaryHike” as “salHike” , “PerformanceRating” as “perRate”, “StockOptionLevel” as “stock”, “WorkLifeBalance” as “workLifeBal” , “YearsAtCompany” as “expComp” , “YearsInCurrentRole” as “expCurrRole” , “JobSatisfaction” as “jobSat” , and “Attrition” as “attr”.

```
colnames(df) <- c("age","gen","marital","edu", "income","salHike", "perRate", "stoc
k", "workLifeBal","expComp","expCurrRole","jobSat","attr")
```

Reading names of the variables in the data set

```
names(df)
```

```
##  [1] "age"         "gen"         "marital"     "edu"         "income"
##  [6] "salHike"     "perRate"     "stock"       "workLifeBal" "expComp"
## [11] "expCurrRole" "jobSat"      "attr"
```

Changing the order of columns

```
df <- df[,c(1,2,3,4,10,11,5,6,8,9,7,12,13)]
```

Examine the structure of the data.

```
str(df)
```

```
## 'data.frame':    1470 obs. of  13 variables:
##  $ age        : int  41 49 37 33 27 32 59 30 38 36 ...
##  $ gen        : Factor w/ 2 levels "Female","Male": 1 2 2 1 2 2 1 2 2 2 ...
##  $ marital    : Factor w/ 3 levels "Divorced","Married",..: 3 2 3 2 2 3 2 1 3 2
...
##  $ edu        : int  2 1 2 4 1 2 3 1 3 3 ...
##  $ expComp    : int  6 10 0 8 2 7 1 1 9 7 ...
##  $ expCurrRole: int  4 7 0 7 2 7 0 0 7 7 ...
##  $ income     : int  5993 5130 2090 2909 3468 3068 2670 2693 9526 5237 ...
##  $ salHike    : int  11 23 15 11 12 13 20 22 21 13 ...
##  $ stock      : int  0 1 0 0 1 0 3 1 0 2 ...
##  $ workLifeBal: int  1 3 3 3 3 2 2 3 3 2 ...
##  $ perRate    : int  3 4 3 3 3 3 4 4 4 3 ...
##  $ jobSat     : int  4 2 3 3 2 4 1 3 3 3 ...
##  $ attr       : Factor w/ 2 levels "No","Yes": 2 1 2 1 1 1 1 1 1 1 ...
```

> There are 13 variables of 1470 cases in df The above output shows that some variables are categorical but read as integers. For example, edu is categorial variable, but appearining as integer. Therfore, variables should be fixed for the proper data type.

Converting the integer variables “edu”,“stock”,“workLifeBal”,“perRate”, and “jobSat” into factor variable.

```
table(df$edu)
```

```
##
##   1   2   3   4   5
## 170 282 572 398  48
```

```
df$edu <-  as.factor(df$edu)
```

```
table(df$stock)
```

```
##
##   0   1   2   3
## 631 596 158  85
```

```
df$stock <-  as.factor(df$stock)
```

```
table(df$workLifeBal)
```

```
##
##   1   2   3   4
##  80 344 893 153
```

```
df$workLifeBal <-  as.factor(df$workLifeBal)
```

```
table(df$perRate)
```

```
##
##    3    4
## 1244  226
```

```
df$perRate <-  as.factor(df$perRate)
```

```
table(df$jobSat)
```

```
##
##   1   2   3   4
## 289 280 442 459
```

```
df$jobSat <-  as.factor(df$jobSat)
```

Replacing Yes and No in the label (attr) variable by 1 and 0 respectively.

```
class(df$attr)
```

```
## [1] "factor"
```

```
df$attr <-  as.character(df$attr)
df$attr <-ifelse(df$attr=="Yes","1","0")
table(df$attr)
```

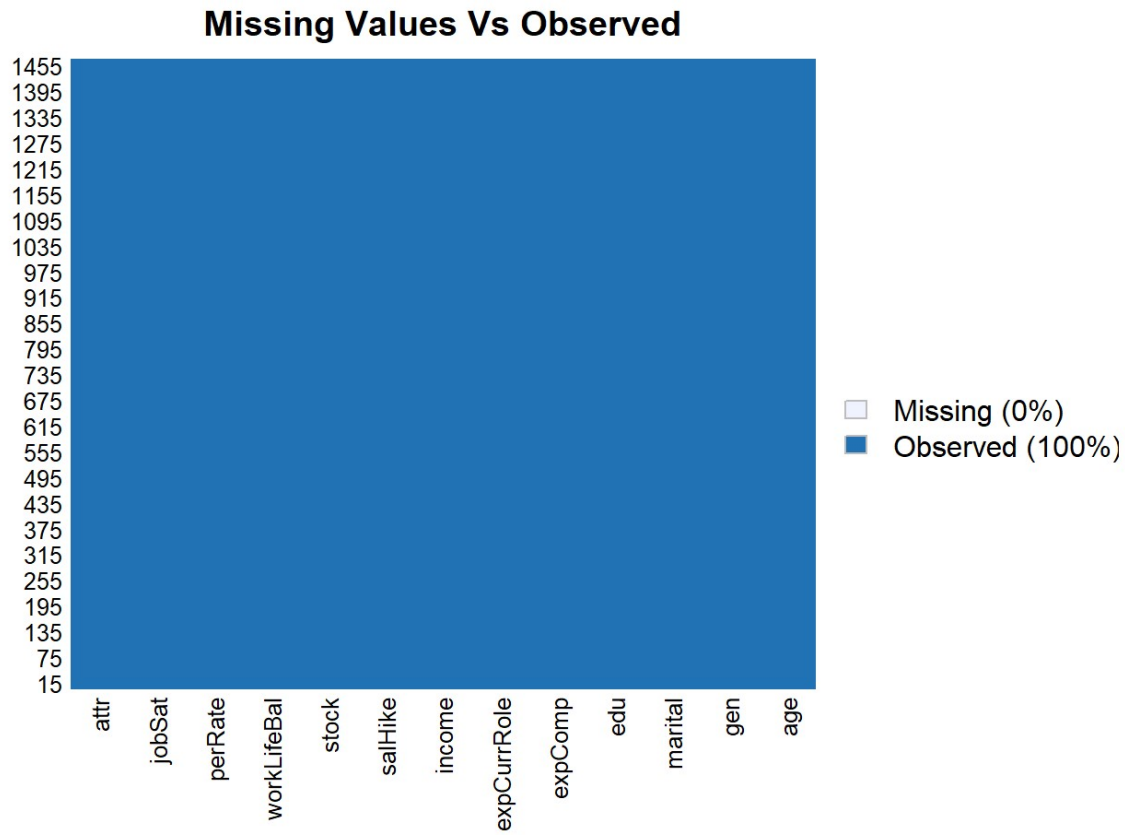
```
##
##    0    1
## 1233  237
```

```
df$attr <-  as.factor(df$attr)
str(df)
```

```
## 'data.frame':    1470 obs. of  13 variables:
##  $ age        : int  41 49 37 33 27 32 59 30 38 36 ...
##  $ gen        : Factor w/ 2 levels "Female","Male": 1 2 2 1 2 2 1 2 2 2 ...
##  $ marital    : Factor w/ 3 levels "Divorced","Married",..: 3 2 3 2 2 3 2 1 3 2
...
##  $ edu        : Factor w/ 5 levels "1","2","3","4",..: 2 1 2 4 1 2 3 1 3 3 ...
##  $ expComp    : int  6 10 0 8 2 7 1 1 9 7 ...
##  $ expCurrRole: int  4 7 0 7 2 7 0 0 7 7 ...
##  $ income     : int  5993 5130 2090 2909 3468 3068 2670 2693 9526 5237 ...
##  $ salHike    : int  11 23 15 11 12 13 20 22 21 13 ...
##  $ stock      : Factor w/ 4 levels "0","1","2","3": 1 2 1 1 2 1 4 2 1 3 ...
##  $ workLifeBal: Factor w/ 4 levels "1","2","3","4": 1 3 3 3 3 2 2 3 3 2 ...
##  $ perRate    : Factor w/ 2 levels "3","4": 1 2 1 1 1 1 2 2 2 1 ...
##  $ jobSat     : Factor w/ 4 levels "1","2","3","4": 4 2 3 3 2 4 1 3 3 3 ...
##  $ attr       : Factor w/ 2 levels "0","1": 2 1 2 1 1 1 1 1 1 1 ...
```

Examine for the missing values

```
missmap(df,main = "Missing Values Vs Observed")
```



The above diagram shows that there is no missing data set (no gray spot found).

Univariate Analysis

```
describe(df)
```

```
##              vars    n    mean      sd median trimmed     mad  min   max
## age             1 1470   36.92    9.14     36   36.47    8.90   18    60
## gen*            2 1470    1.60    0.49      2    1.62    0.00    1     2
## marital*        3 1470    2.10    0.73      2    2.12    1.48    1     3
## edu*            4 1470    2.91    1.02      3    2.98    1.48    1     5
## expComp         5 1470    7.01    6.13      5    5.99    4.45    0    40
## expCurrRole     6 1470    4.23    3.62      3    3.85    4.45    0    18
## income          7 1470 6502.93 4707.96   4919 5667.24 3260.24 1009 19999
## salHike         8 1470   15.21    3.66     14   14.80    2.97   11    25
## stock*          9 1470    1.79    0.85      2    1.67    1.48    1     4
## workLifeBal*   10 1470    2.76    0.71      3    2.77    0.00    1     4
## perRate*       11 1470    1.15    0.36      1    1.07    0.00    1     2
## jobSat*        12 1470    2.73    1.10      3    2.79    1.48    1     4
## attr*          13 1470    1.16    0.37      1    1.08    0.00    1     2
##              range  skew kurtosis     se
## age             42  0.41    -0.41   0.24
## gen*             1 -0.41    -1.83   0.01
## marital*         2 -0.15    -1.12   0.02
## edu*             4 -0.29    -0.56   0.03
## expComp         40  1.76     3.91   0.16
## expCurrRole     18  0.92     0.47   0.09
## income       18990  1.37     0.99 122.79
## salHike         14  0.82    -0.31   0.10
## stock*           3  0.97     0.35   0.02
## workLifeBal*     3 -0.55     0.41   0.02
## perRate*         1  1.92     1.68   0.01
## jobSat*          3 -0.33    -1.22   0.03
## attr*            1  1.84     1.39   0.01
```

> We can aso use summary() function, but it does not provide standard deviation value.

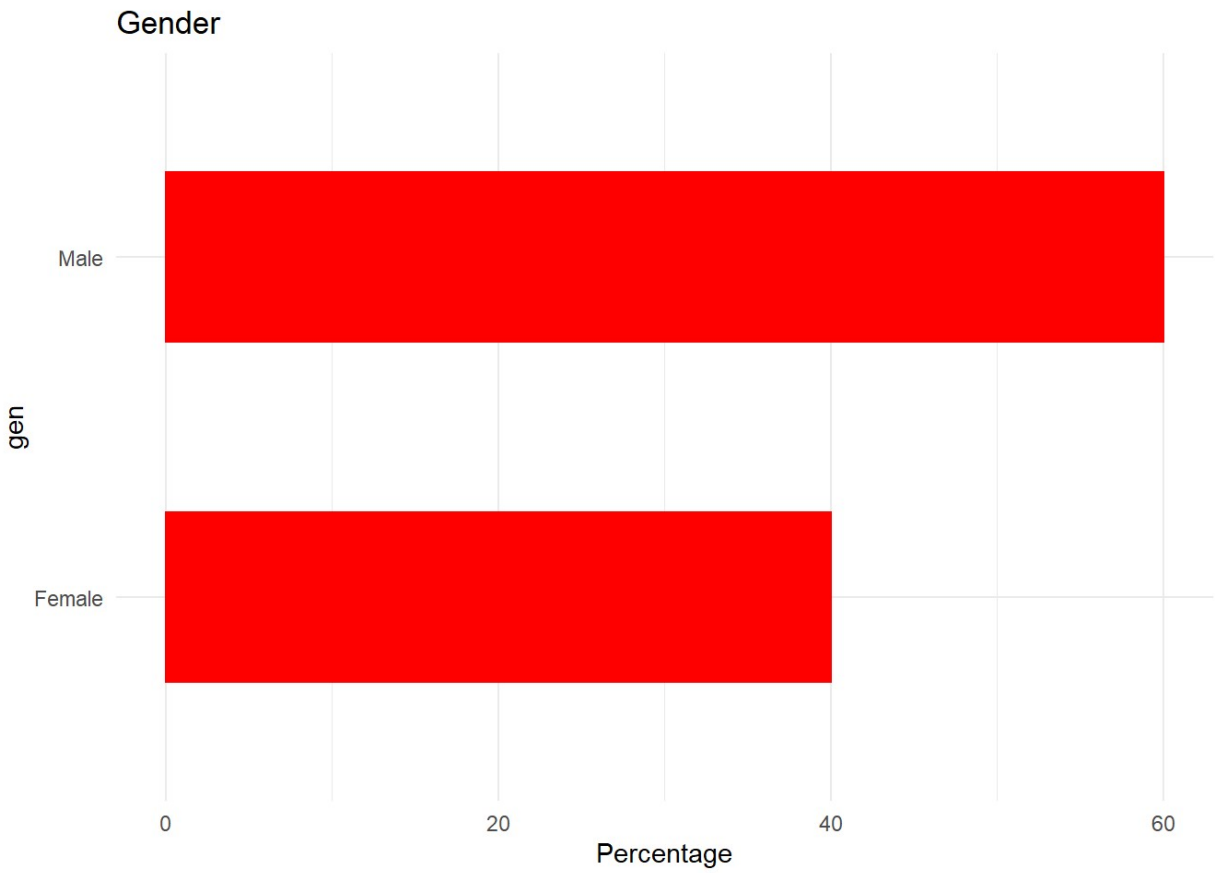
Visual display of the variables

Age

```
ggplot(df, aes(x=age)) + ggtitle("Age of Employee") + xlab("age") +
  geom_histogram(bins = 20, aes(y=..density..), color = "red", fill = "red") + geom
_density(alpha=.2, fill="#FF6666") +     ylab("Frequency")
```

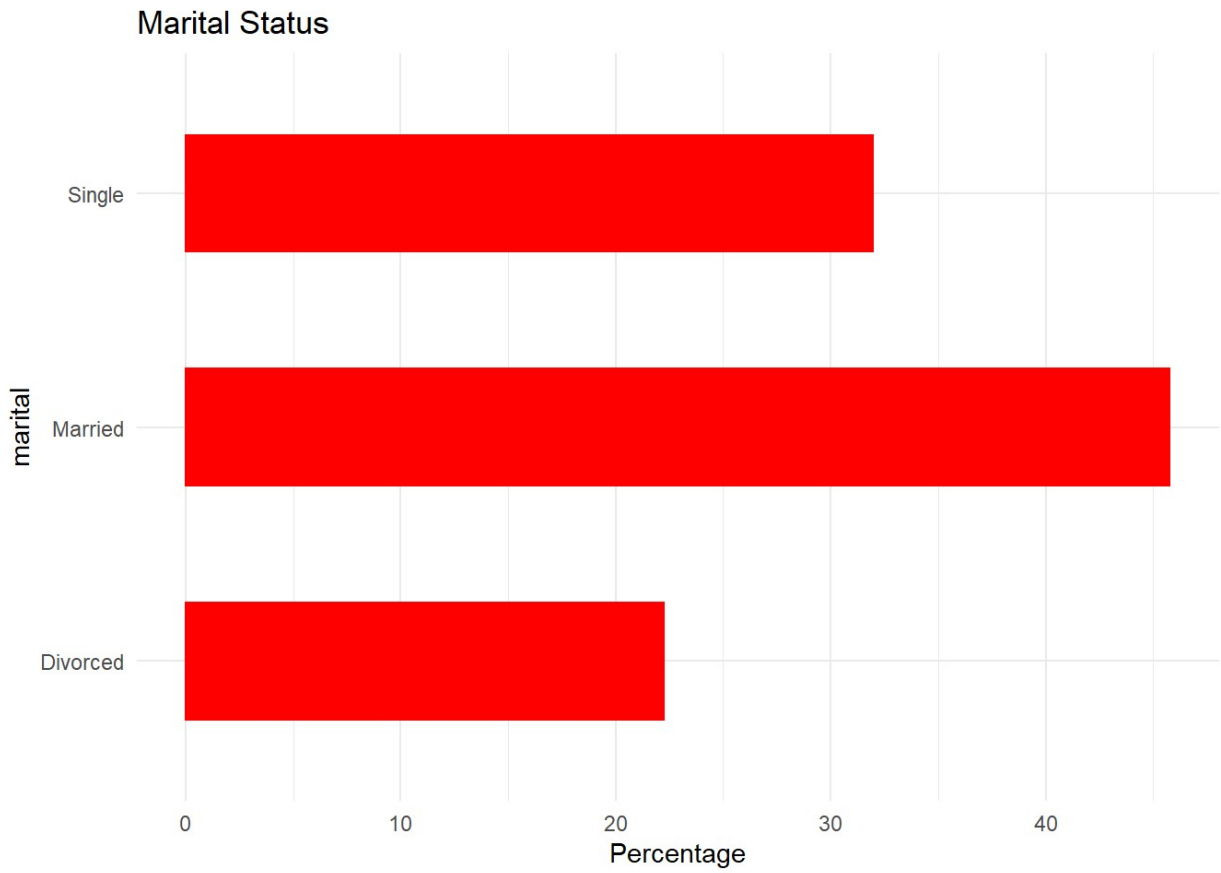
Gender

```
ggplot(df, aes(x=gen)) + ggtitle("Gender") + xlab("gen") +
  geom_bar(aes(y = 100*(..count..)/sum(..count..)), width = 0.5, color = "red", fil
l = "red") + ylab("Percentage") + coord_flip() + theme_minimal()
```



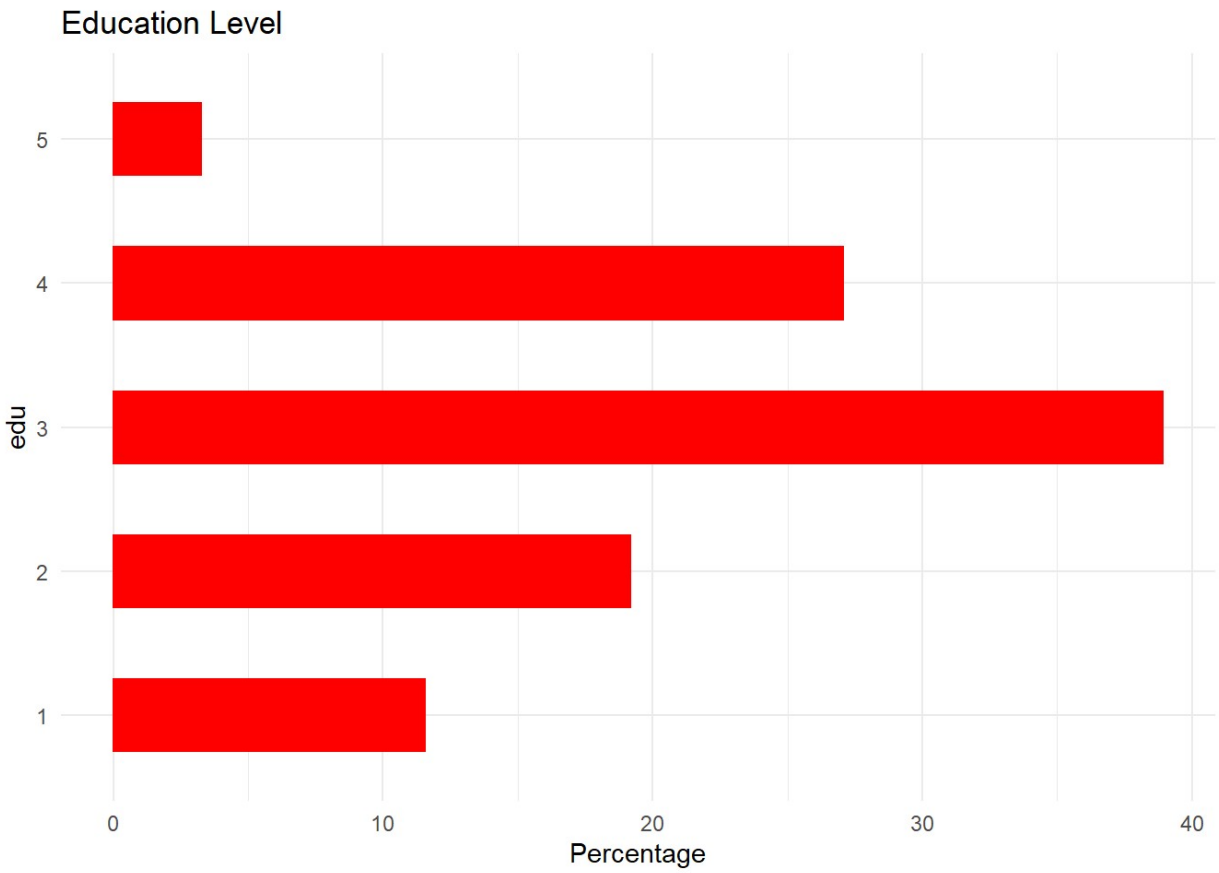
Marital Status

```
ggplot(df, aes(x=marital)) + ggtitle("Marital Status") + xlab("marital") +
  geom_bar(aes(y = 100*(..count..)/sum(..count..)), width = 0.5, color = "red", fil
l = "red") + ylab("Percentage") + coord_flip() + theme_minimal()
```



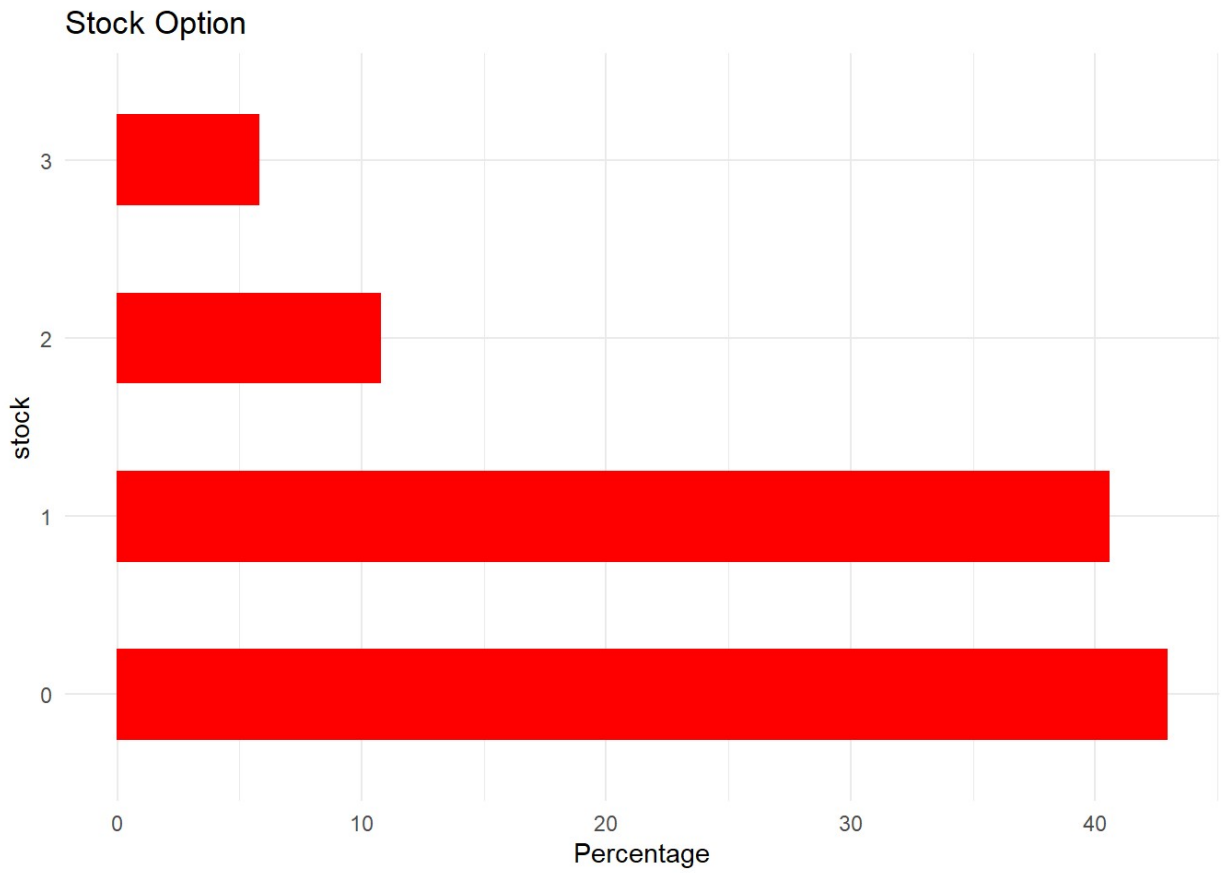
Education Level

```
ggplot(df, aes(x=edu)) + ggtitle("Education Level") + xlab("edu") +
  geom_bar(aes(y = 100*(..count..)/sum(..count..)), width = 0.5,color = "red", fill
= "red") + ylab("Percentage") + coord_flip() + theme_minimal()
```



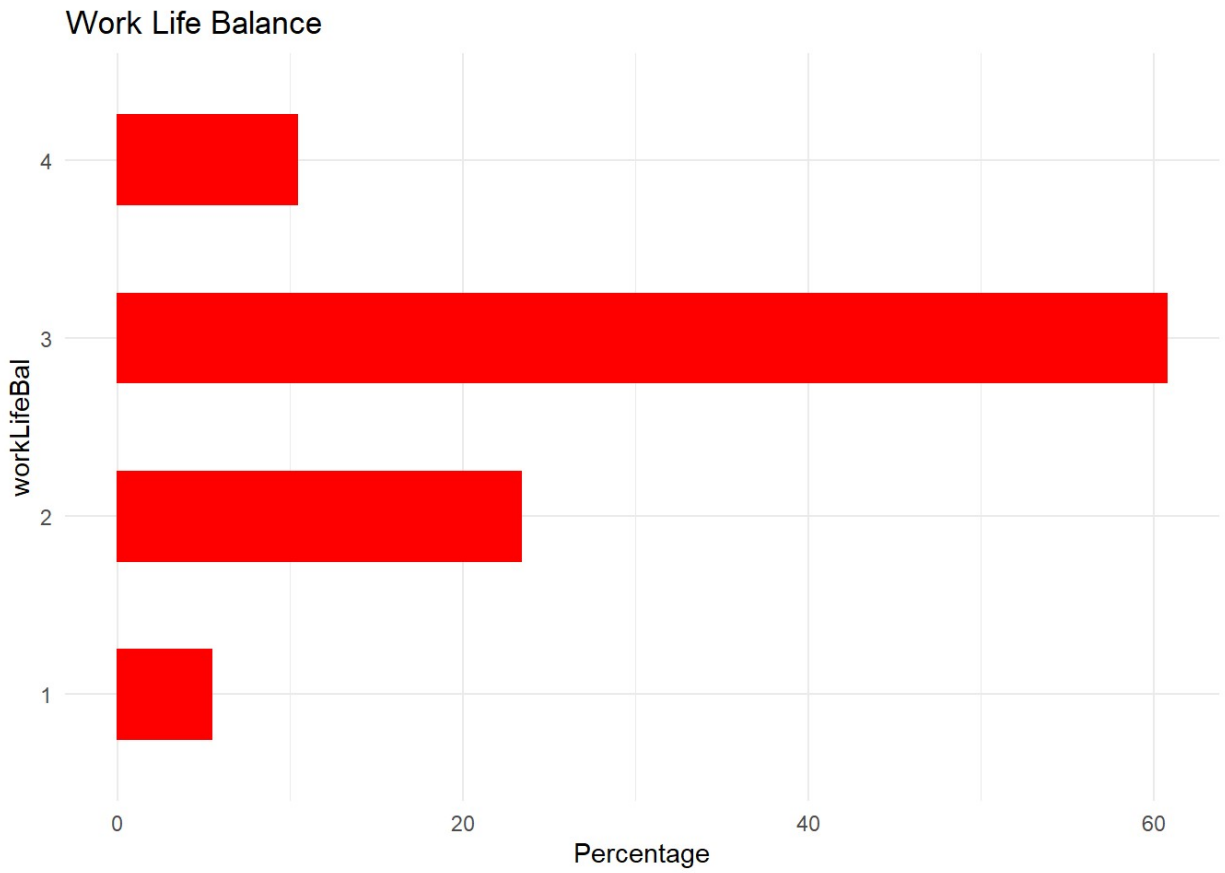
Stock Options

```
ggplot(df, aes(x=stock)) + ggtitle("Stock Option") + xlab("stock") +
  geom_bar(aes(y = 100*(..count..)/sum(..count..)), width = 0.5, color = "red", fil
l = "red") + ylab("Percentage") + coord_flip() + theme_minimal()
```



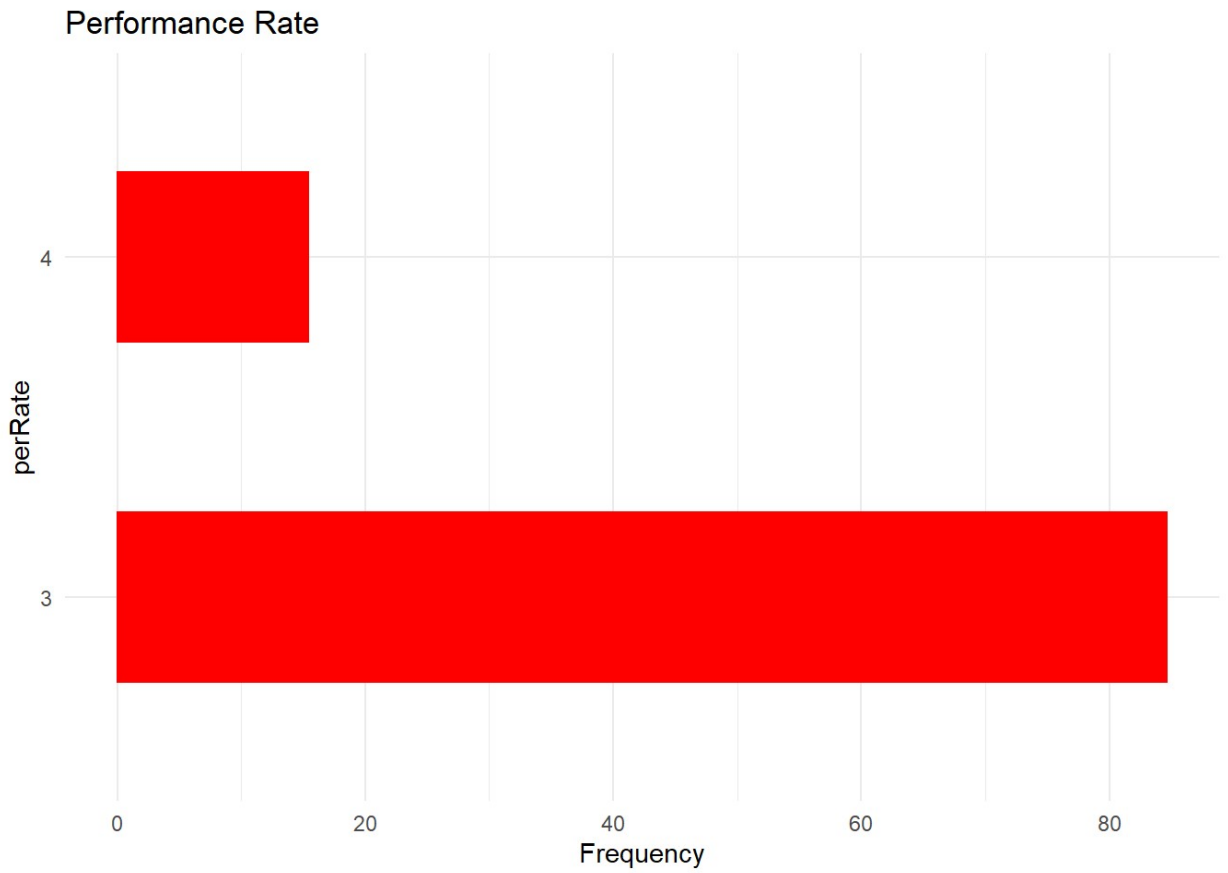
Work Life Balance

```
ggplot(df, aes(x=workLifeBal)) + ggtitle("Work Life Balance") + xlab("workLifeBal")
+
  geom_bar(aes(y = 100*(..count..)/sum(..count..)), width = 0.5, color = "red", fil
l = "red") + ylab("Percentage") + coord_flip() + theme_minimal()
```



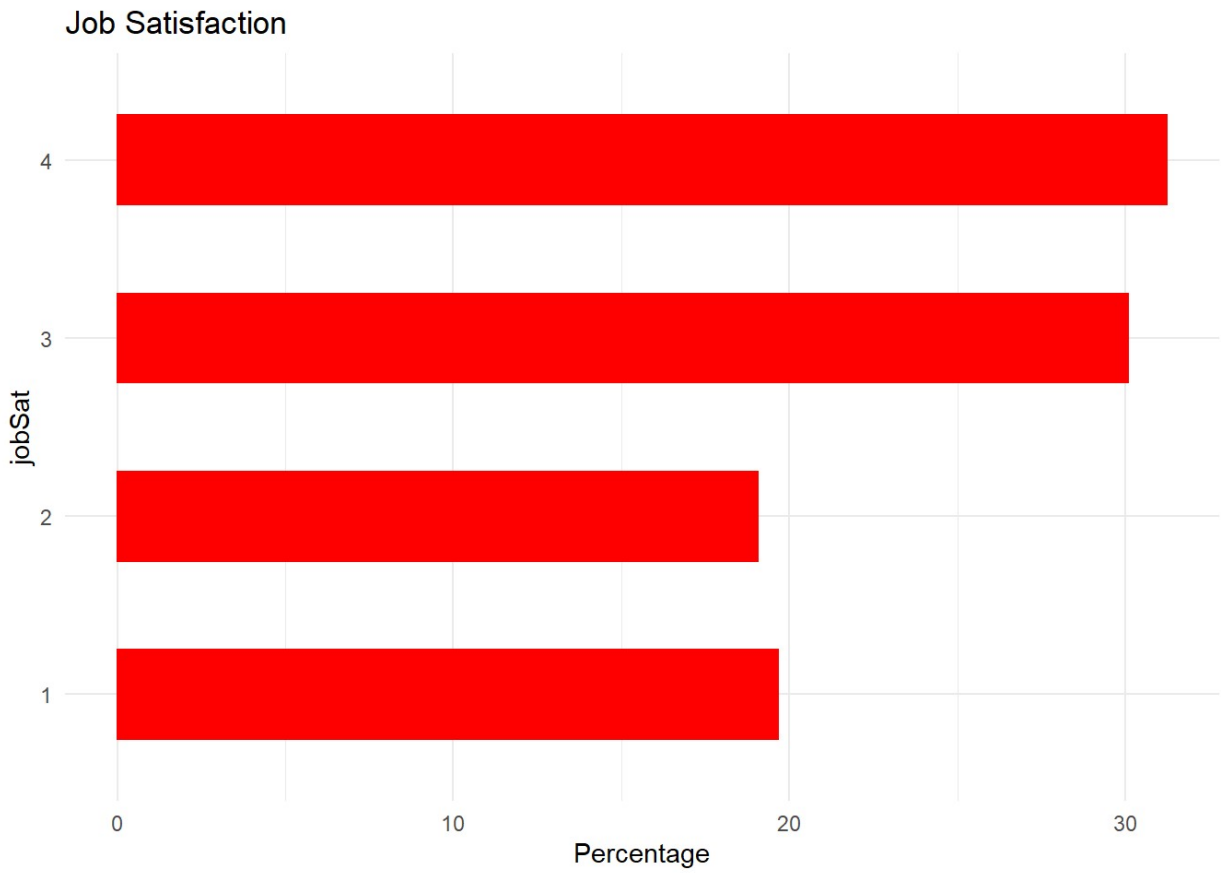
Performance Rate

```
ggplot(df, aes(x=perRate)) + ggtitle("Performance Rate") + xlab("perRate") +
  geom_bar(aes(y = 100*(..count..)/sum(..count..)), width = 0.5, color = "red", fil
l = "red") + ylab("Frequency") + coord_flip() + theme_minimal()
```



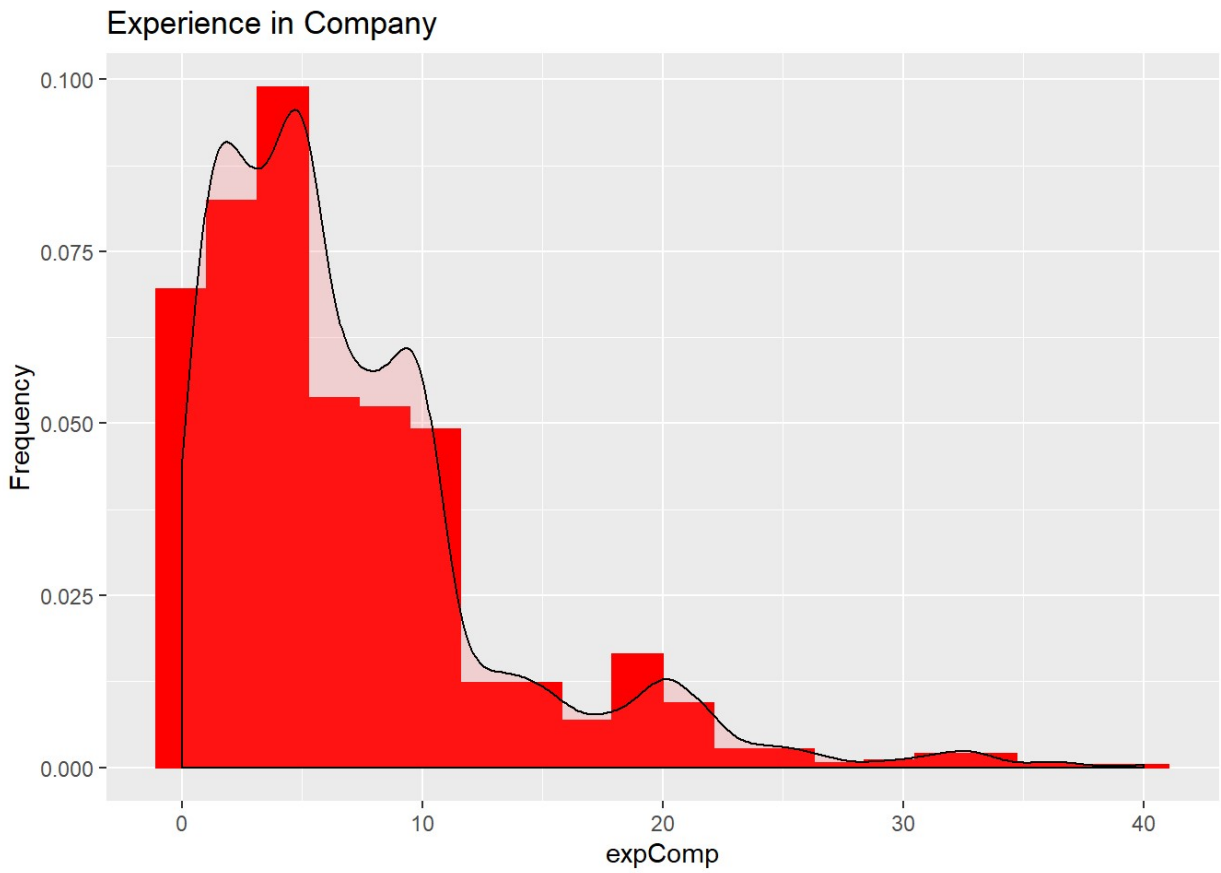
Job Satisfaction

```
ggplot(df, aes(x=jobSat)) + ggtitle("Job Satisfaction") + xlab("jobSat") +
  geom_bar(aes(y = 100*(..count..)/sum(..count..)), width = 0.5, color = "red", fil
l = "red") + ylab("Percentage") + coord_flip() + theme_minimal()
```



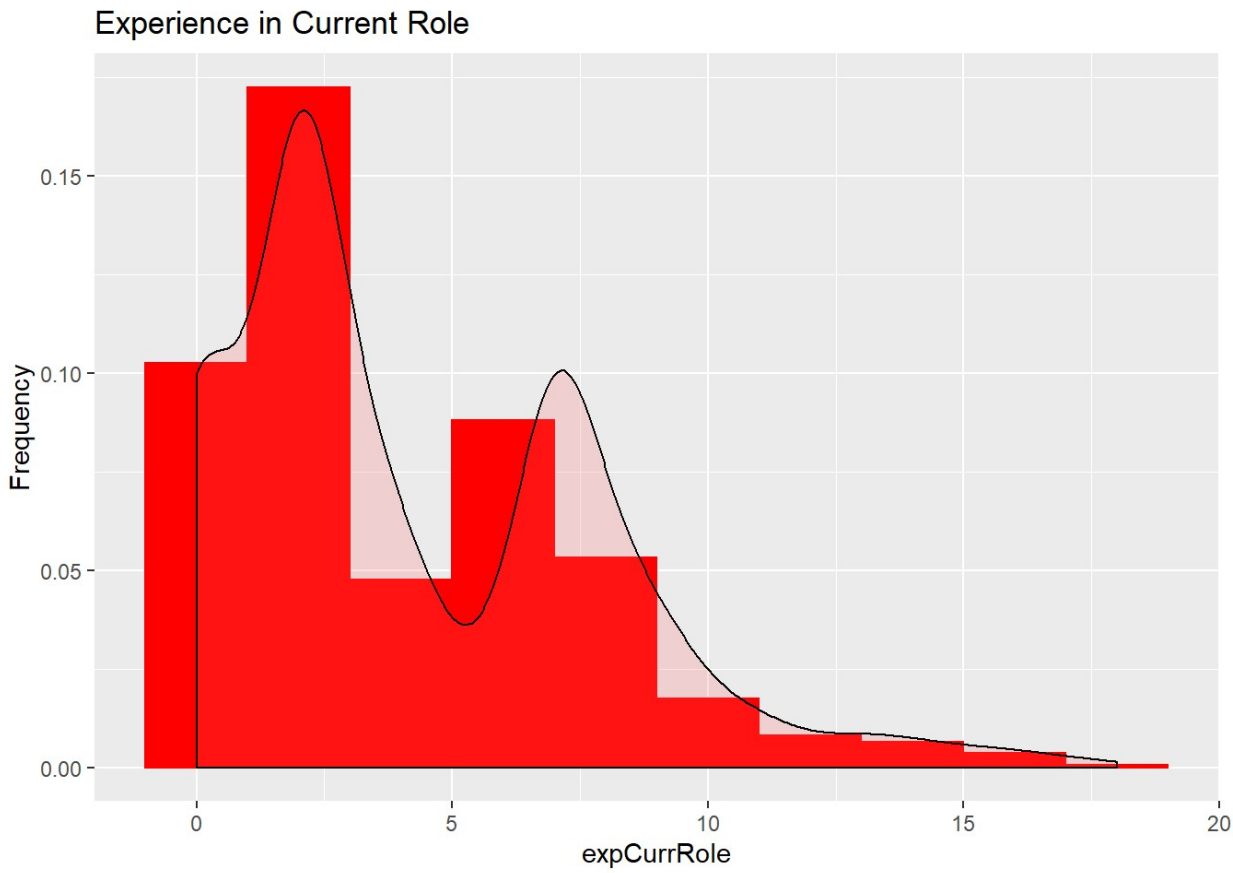
Experience in Company

```
ggplot(df, aes(x=expComp)) + ggtitle("Experience in Company") + xlab("expComp") +
  geom_histogram(bins = 20, aes(y=..density..), color = "red", fill = "red") + geom
_density(alpha=.2, fill="#FF6666") +     ylab("Frequency")
```



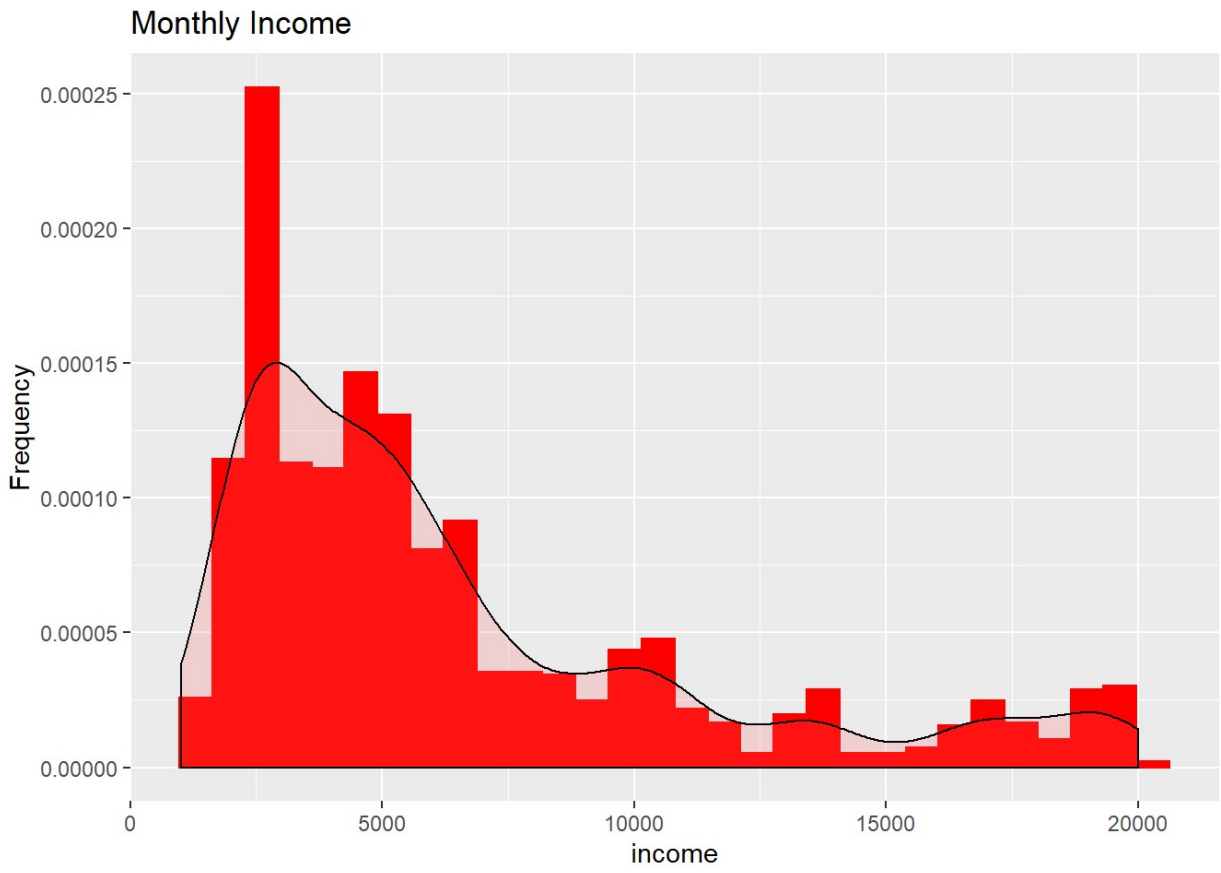
Experience in Current Role

```
ggplot(df, aes(x=expCurrRole)) + ggtitle("Experience in Current Role") + xlab("expC
urrRole") +
  geom_histogram(bins = 10, aes(y=..density..), color = "red", fill = "red") + geom
_density(alpha=.2, fill="#FF6666") +     ylab("Frequency")
```



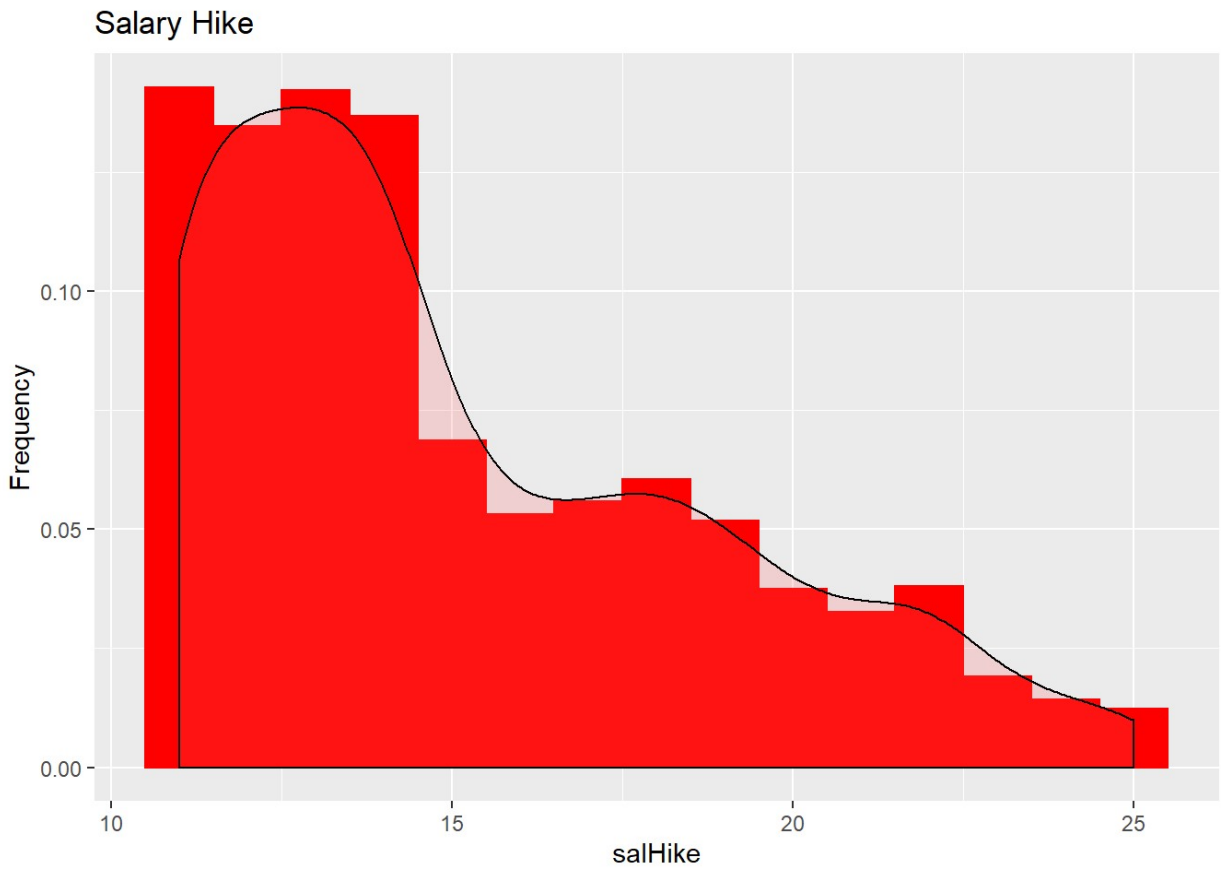
Monthly Income

```
ggplot(df, aes(x=income)) + ggtitle("Monthly Income") + xlab("income") +
  geom_histogram(bins = 30, aes(y=..density..), color = "red", fill = "red") + geom
_density(alpha=.2, fill="#FF6666") +     ylab("Frequency")
```



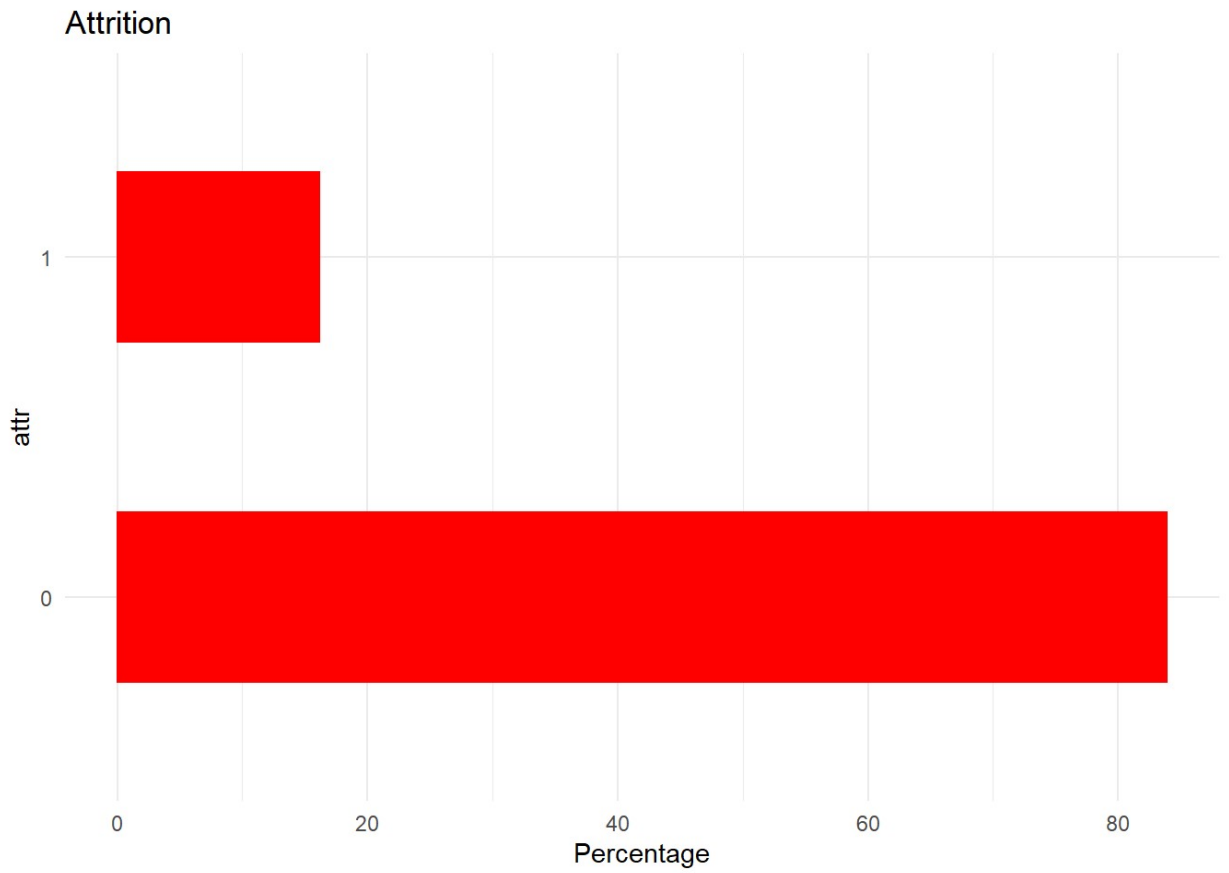
Salary Hike

```
ggplot(df, aes(x=salHike)) + ggtitle("Salary Hike") + xlab("salHike") +
  geom_histogram(bins = 15, aes(y=..density..), color = "red", fill = "red") + geom
_density(alpha=.2, fill="#FF6666") +     ylab("Frequency")
```



Attrition

```
ggplot(df, aes(x=attr)) + ggtitle("Attrition") + xlab("attr") +
  geom_bar(aes(y = 100*(..count..)/sum(..count..)), width = 0.5, color = "red", fil
l = "red") + ylab("Percentage") + coord_flip() + theme_minimal()
```



Examining the distribution of the categorical variables in the data for the factor levels representing employees with and without attrition (attr). This may lead to exclude variables that only have few (below 5) samples in any category.

```
table(df$attr)
```

```
##
##    0    1
## 1233  237
```

```
xtabs(~ attr + gen, data=df)
```

```
##     gen
## attr Female Male
##    0    501  732
##    1     87  150
```

```
xtabs(~ attr + marital, data=df)
```

```
##     marital
## attr Divorced Married Single
##    0      294     589    350
##    1       33      84    120
```

```
xtabs(~ attr + edu, data=df)
```

```
##     edu
## attr   1   2   3   4   5
##    0 139 238 473 340  43
##    1  31  44  99  58   5
```

```
xtabs(~ attr + stock, data=df)
```

```
##     stock
## attr   0   1   2   3
##    0 477 540 146  70
##    1 154  56  12  15
```

```
xtabs(~ attr + workLifeBal, data=df)
```

```
##     workLifeBal
## attr   1   2   3   4
##    0  55 286 766 126
##    1  25  58 127  27
```

```
xtabs(~ attr + perRate, data=df)
```

```
##     perRate
## attr    3    4
##    0 1044  189
##    1  200   37
```

```
xtabs(~ attr + jobSat, data=df)
```

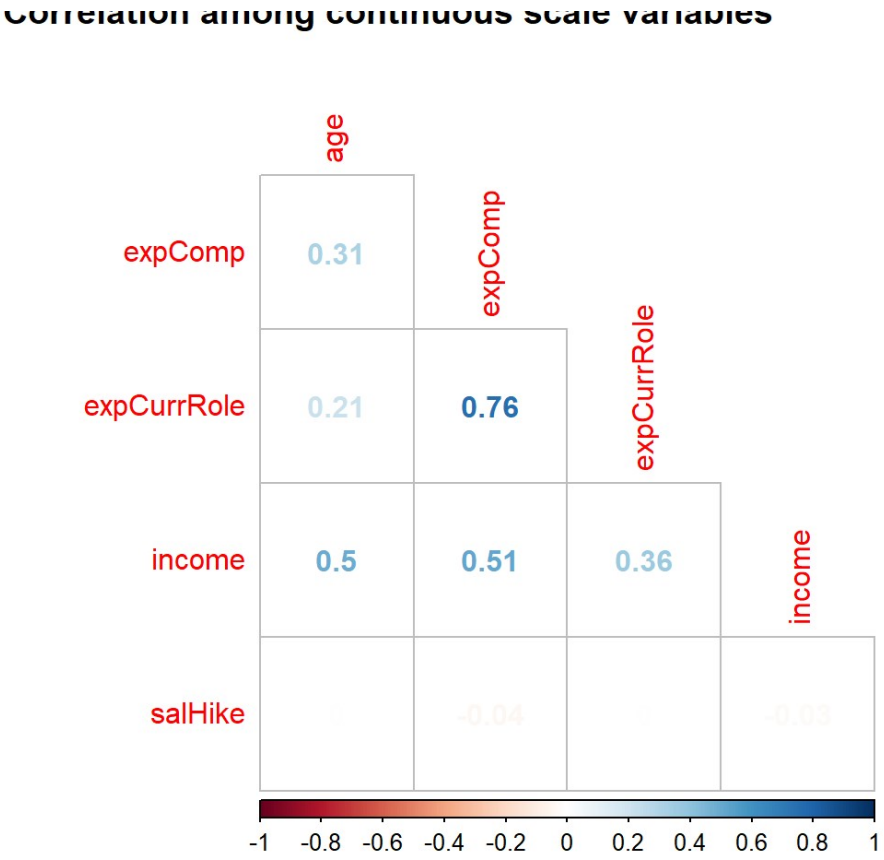
```
##     jobSat
## attr   1   2   3   4
##    0 223 234 369 407
##    1  66  46  73  52
```

The above output shows that all categorical variables have no category with frequency below 5.

Bivariate Analysiss

Examining the Correlation (continuous scale variables)

```
df_corr <- cor(df[c(1,5,6,7,8)])
corrplot(df_corr, method="number", diag=FALSE, type="lower",  title = "Correlation
among continuous scale variables")
```



There is evidence of moderate correlation between (age and income), and (age and expComp), (expComp and expCurrRole), (income and expComp), and (income and expCurrRole)

Examining the association (or independence) among categorical variables.

Q: Are Gender(gen) and Marital Status (marital) independent?

Chi-Square Test of Independence

Ho: They are independent H1: They are not independent

Applying chi-square test

```
table(df$gen, df$marital)
```

```
##
##          Divorced Married Single
##   Female      117     272    199
##   Male        210     401    271
```

```
chisq.test(df$gen, df$marital)
```

```
##
##  Pearson's Chi-squared test
##
## data:  df$gen and df$marital
## X-squared = 3.5478, df = 2, p-value = 0.1697
```

p-value is 0.169, which is above 0.05, thus Ho is accepted and that gen and marital are independent.

Q: Are Gender(gen) and Education Level (edu) independent?

Chi-Square Test of Independence

Ho: They are independent H1: They are not independent

Applying chi-square test

```
table(df$gen, df$edu)
```

```
##
##            1   2   3   4   5
##   Female  60 117 235 154  22
##   Male   110 165 337 244  26
```

```
chisq.test(df$gen, df$edu)
```

```
##
##  Pearson's Chi-squared test
##
## data:  df$gen and df$edu
## X-squared = 3.0729, df = 4, p-value = 0.5457
```

p-value is 0.54, which is above 0.05, thus Ho is accepted and that gen and edu are independent.

Q: Are Gender(gen) and Stock Option (stock) independent?

Chi-Square Test of Independence

Ho: They are independent H1: They are not independent

Applying chi-square test

```
table(df$gen, df$stock)
```

```
##
##            0   1   2   3
##   Female 255 241  58  34
##   Male   376 355 100  51
```

```
chisq.test(df$gen, df$stock)
```

```
##
##  Pearson's Chi-squared test
##
## data:  df$gen and df$stock
## X-squared = 0.80498, df = 3, p-value = 0.8483
```

p-value is 0.848, which is above 0.05, thus Ho is accepted and that gen and stock are independent.

Q: Are Gender(gen) and Job Role (jobRole) independent?

Chi-Square Test of Independence

Ho: They are independent H1: They are not independent

Applying chi-square test

```
table(df$gen, df$workLifeBal)
```

```
##
##            1   2   3   4
##   Female  30 136 365  57
##   Male    50 208 528  96
```

```
chisq.test(df$gen, df$workLifeBal)
```

```
##
##  Pearson's Chi-squared test
##
## data:  df$gen and df$workLifeBal
## X-squared = 1.0036, df = 3, p-value = 0.8004
```

p-value is 0.8004, which is above 0.05, thus Ho is accepted and that gen and workLifeBal are not independent.

Q: Are Gender(gen) and Overtime (overtime) independent?

Chi-Square Test of Independence

Ho: They are independent H1: They are not independent

Applying chi-square test

```
table(df$gen, df$perRate)
```

```
##
##            3   4
##   Female 494  94
##   Male   750 132
```

```
chisq.test(df$gen, df$perRate)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  df$gen and df$perRate
## X-squared = 0.20936, df = 1, p-value = 0.6473
```

p-value is 0.6473, which is above 0.05, thus Ho is accepted and that gen and perRate are independent.

Q: Are Gender(gen) and Business Travel (travel) independent?

Chi-Square Test of Independence

Ho: They are independent H1: They are not independent

Applying chi-square test

```
table(df$gen, df$jobSat)
```

```
##
##            1   2   3   4
##   Female 119 118 181 170
##   Male   170 162 261 289
```

```
chisq.test(df$gen, df$jobSat)
```

```
##
##  Pearson's Chi-squared test
##
## data:  df$gen and df$jobSat
## X-squared = 2.5477, df = 3, p-value = 0.4667
```

p-value is 0.4667, which is above 0.05, thus Ho is accepted and that gen and jobSat are independent.

Q: Are Marital(martial) and Education (edu) independent?

Chi-Square Test of Independence

Ho: They are independent H1: They are not independent

Applying chi-square test

```
table(df$marital, df$edu)
```

```
##
##               1   2   3   4   5
##   Divorced   32  72 125  90   8
##   Married    82 130 253 182  26
##   Single     56  80 194 126  14
```

```
chisq.test(df$marital, df$edu)
```

```
##
##  Pearson's Chi-squared test
##
## data:  df$marital and df$edu
## X-squared = 6.2687, df = 8, p-value = 0.6172
```

p-value is 0.6172, which is above 0.05, thus Ho is accepted and that marital and edu are independent.

Q: Are Marital(martial) and Stock Option (stock) independent?

Chi-Square Test of Independence

Ho: They are independent H1: They are not independent

Applying chi-square test

```
table(df$marital, df$stock)
```

```
##
##              0   1   2   3
##   Divorced   8 195  75  49
##   Married  153 401  83  36
##   Single   470   0   0   0
```

```
chisq.test(df$marital, df$stock)
```

```
##
##  Pearson's Chi-squared test
##
## data:  df$marital and df$stock
## X-squared = 998.1, df = 6, p-value < 2.2e-16
```

p-value is 2.2e-16, which is below 0.05, thus Ho is rejected and that marital and stock are not independent.

Q: Are Marital(martial) and Work Life Balance (workLifeBal) independent?

Chi-Square Test of Independence

Ho: They are independent H1: They are not independent

Applying chi-square test

```
table(df$marital, df$workLifeBal)
```

```
##
##              1   2   3   4
##   Divorced  13  88 194  32
##   Married   42 153 405  73
##   Single    25 103 294  48
```

```
chisq.test(df$marital, df$workLifeBal)
```

```
##
##  Pearson's Chi-squared test
##
## data:  df$marital and df$workLifeBal
## X-squared = 5.0476, df = 6, p-value = 0.5377
```

> p-value is 0.5377, which is above 0.05, thus Ho is accepted and that marital and workLifeBal are independent.

Q: Are Marital(martial) and Performance Rate (perRate) independent?

Chi-Square Test of Independence

> Chi-Square Test of Independence

> Ho: They are independent H1: They are not independent

> Applying chi-square test

```
table(df$marital, df$perRate)
```

```
##
##              3   4
##   Divorced 279  48
##   Married  567 106
##   Single   398  72
```

```
chisq.test(df$marital, df$perRate)
```

```
##
##  Pearson's Chi-squared test
##
## data:  df$marital and df$perRate
## X-squared = 0.1958, df = 2, p-value = 0.9067
```

p-value is 0.9067, which is above 0.05, thus Ho is accepted and that marital and perRate are independent.

Q: Are Marital(martial) and Job Satisfaction (jobSat) independent?

Chi-Square Test of Independence

Ho: They are independent H1: They are not independent

Applying chi-square test

```
table(df$marital, df$jobSat)
```

```
##
##              1   2   3   4
##   Divorced  70  61  94 102
##   Married  130 131 212 200
##   Single    89  88 136 157
```

```
chisq.test(df$marital, df$jobSat)
```

```
##
##  Pearson's Chi-squared test
##
## data:  df$marital and df$jobSat
## X-squared = 2.8416, df = 6, p-value = 0.8285
```

p-value is 0.8285, which is above 0.05, thus Ho is accepted and that marital and jobSat are independent.

Q: Are Education(edu) and Stock Option(stock)independent?

Chi-Square Test of Independence

Ho: They are independent H1: They are not independent

Applying chi-square test

```
table(df$edu, df$stock)
```

```
##
##       0   1   2   3
##   1  84  58  17  11
##   2 107 130  32  13
##   3 249 237  52  34
##   4 171 150  54  23
##   5  20  21   3   4
```

Since there are two categories having frequency below 5, therefore chi square approximation may not be appropriate

Q: Are Education(edu) and Work Life Balance(workLifeBal)independent?

Chi-Square Test of Independence

Ho: They are independent H1: They are not independent

Applying chi-square test

```
table(df$edu, df$workLifeBal)
```

```
##
##       1   2   3   4
##   1   6  43 104  17
##   2  16  65 170  31
##   3  34 142 341  55
##   4  22  86 243  47
##   5   2   8  35   3
```

Since one of the category has frequency below 5, therefore chi square approximation may not be appropriate

Q: Are Education(edu) and Performance Rate

(perRate)independent?

Chi-Square Test of Independence

Ho: They are independent H1: They are not independent

Applying chi-square test

```
table(df$edu, df$perRate)
```

```
##
##       3   4
##   1 140  30
##   2 234  48
##   3 490  82
##   4 341  57
##   5  39   9
```

```
chisq.test(df$edu, df$perRate)
```

```
##
##  Pearson's Chi-squared test
##
## data:  df$edu and df$perRate
## X-squared = 2.4965, df = 4, p-value = 0.6453
```

p-value is 0.645 , which is above 0.05, thus Ho is accepted and that edu and perRate are independent.

Q: Are Education(edu) and Job Satisfaction (jobSat) independent?

Chi-Square Test of Independence

Ho: They are independent H1: They are not independent

Applying chi-square test

```
table(df$edu, df$jobSat)
```

```
##
##       1   2   3   4
##   1  32  26  56  56
##   2  52  54  83  93
##   3 118 128 161 165
##   4  74  67 127 130
##   5  13   5  15  15
```

```
chisq.test(df$edu, df$jobSat)
```

```
##
##  Pearson's Chi-squared test
##
## data:  df$edu and df$jobSat
## X-squared = 13.03, df = 12, p-value = 0.3669
```

p-value is 0.3669 , which is above 0.05, thus Ho is accepted and that edu and jobSat are independent.

Q: Are Stock Option (stock) and Work Life Balance (workLifeBal) independent?

Chi-Square Test of Independence

Ho: They are independent H1: They are not independent

Applying chi-square test

```
table(df$stock, df$workLifeBal)
```

```
##
##       1   2   3   4
##   0  37 134 392  68
##   1  37 147 352  60
##   2   6  42  92  18
##   3   0  21  57   7
```

Since one of the category has frequency below 5, therefore we chi square approximation may not be appropriate

Q: Are Stock Option (stock) and Performance Rate (perRate) independent?

Chi-Square Test of Independence

Ho: They are independent H1: They are not independent

Applying chi-square test

```
table(df$stock, df$perRate)
```

```
##
##       3   4
##   0 540  91
##   1 495 101
##   2 136  22
##   3  73  12
```

```
chisq.test(df$stock, df$perRate)
```

```
##
##  Pearson's Chi-squared test
##
## data:  df$stock and df$perRate
## X-squared = 1.9309, df = 3, p-value = 0.5869
```

p-value is 0.586 , which is above 0.05, thus Ho is accepted and that edu and attr are independent.

Q: Are Stock Option (stock) and Job Satisfaction (jobSat) independent?

Chi-Square Test of Independence

Ho: They are independent H1: They are not independent

Applying chi-square test

```
table(df$stock, df$jobSat)
```

```
##
##       1   2   3   4
##   0 123 120 197 191
##   1 122 113 175 186
##   2  26  30  48  54
##   3  18  17  22  28
```

```
chisq.test(df$stock, df$jobSat)
```

```
##
##  Pearson's Chi-squared test
##
## data:  df$stock and df$jobSat
## X-squared = 2.7593, df = 9, p-value = 0.9731
```

p-value is 0.973 , which is above 0.05, thus Ho is accepted and that edu and attr are independent.

Q: Are Work Life Balance (workLifeBal) and Performance Rate (perRate) independent?

Chi-Square Test of Independence

Ho: They are independent H1: They are not independent

Applying chi-square test

```
table(df$workLifeBal, df$perRate)
```

```
##
##       3   4
##   1  66  14
##   2 296  48
##   3 752 141
##   4 130  23
```

```
chisq.test(df$workLifeBal, df$perRate)
```

```
##
##  Pearson's Chi-squared test
##
## data:  df$workLifeBal and df$perRate
## X-squared = 0.94363, df = 3, p-value = 0.8149
```

p-value is 0.814 , which is above 0.05, thus Ho is accepted and that workLifeBal and perRate are independent.

Q: Are Work Life Balance (workLifeBal) and Job Satisfaction (jobSat) independent?

Chi-Square Test of Independence

Ho: They are independent H1: They are not independent

Applying chi-square test

```
table(df$workLifeBal, df$jobSat)
```

```
##
##       1   2   3   4
##   1  17  18  21  24
##   2  61  58 101 124
##   3 182 175 273 263
##   4  29  29  47  48
```

```
chisq.test(df$workLifeBal, df$jobSat)
```

```
##
##  Pearson's Chi-squared test
##
## data:  df$workLifeBal and df$jobSat
## X-squared = 6.5765, df = 9, p-value = 0.6811
```

p-value is 0.681 , which is above 0.05, thus Ho is accepted and that workLifeBal and jobSat are independent.

Q: Are Performance Rate (perRate) and Job Satisfaction (jobSat) independent?

Chi-Square Test of Independence

Ho: They are independent H1: They are not independent

Applying chi-square test

```
table(df$perRate, df$jobSat)
```

```
##
##       1   2   3   4
##   3 241 237 386 380
##   4  48  43  56  79
```

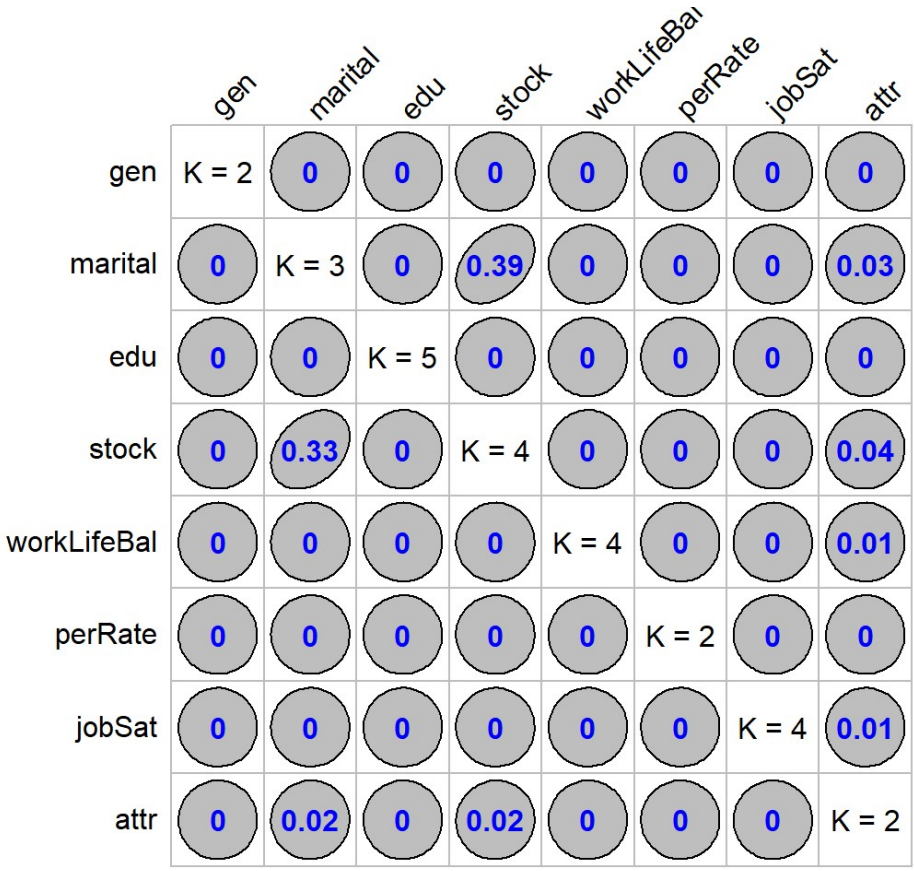
```
chisq.test(df$perRate, df$jobSat)
```

```
##
##  Pearson's Chi-squared test
##
## data:  df$perRate and df$jobSat
## X-squared = 4.0143, df = 3, p-value = 0.2599
```

p-value is 0.259 , which is above 0.05, thus Ho is accepted and that perRate and jobSat are independent.

The association can also be studied by the Googman Kruskal values.

```
fac_1<- c("gen","marital","edu","stock","workLifeBal","perRate","jobSat","attr")
df1<- subset(df, select = fac_1)
GKmatrix1<- GKtauDataframe(df1)
plot(GKmatrix1, corrColors = "blue")
```



In the 8×8 array plot, the diagonal entries are number of categories in the vaiable. The off-diagonal elements give the Goodman-Kruskal τ values represent the association measure τ(x,y) from the variable x (rows) to the variable y (columns). The close to zero means no association between two variables.

Examining the association (or independence) between Attrition (attr) and other categorical factors.

Q: Are Gender(gen) and Attrition (attr) independent?

Chi-Square Test of Independence

Ho: They are independent H1: They are not independent

Applying chi-square test

```
chisq.test(df$gen, df$attr)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  df$gen and df$attr
## X-squared = 1.117, df = 1, p-value = 0.2906
```

p-value is 0.2906, which is above 0.05, thus Ho is accepted and that gen and attr are independent.

Q: Which gender has more tendency towards Attiration?

```
table(df$gen, df$attr)
```

```
##
##            0   1
##   Female 501  87
##   Male   732 150
```

```
prop.table(table(df$gen, df$attr))
```

```
##
##                    0          1
##   Female 0.34081633 0.05918367
##   Male   0.49795918 0.10204082
```

The output table shows that the proportion of the male employees is larger than the female in favour of the attrition.

Marital Statuss (marit) and Attrition (attr)

Q: Are Marital Status (martial) and Attrition (attr) independent?

> Chi-Square Test of Independence

> Ho: They are independent H1: They are not independent

> Applying chi-square test

```
chisq.test(df$marital, df$attr)
```

```
##
##  Pearson's Chi-squared test
##
## data:  df$marital and df$attr
## X-squared = 46.164, df = 2, p-value = 9.456e-11
```

> p-value is 9.456e-11, which is below 0.05, thus Ho is rejected and that marital and attr are not independent.

Q: Whether marital status affects more the decision towards Attiration?

```
table(df$marital, df$attr)
```

```
##
##              0   1
##   Divorced 294  33
##   Married  589  84
##   Single   350 120
```

```
prop.table(table(df$marital, df$attr))
```

```
##
##                     0          1
##   Divorced 0.20000000 0.02244898
##   Married  0.40068027 0.05714286
##   Single   0.23809524 0.08163265
```

The output shows that in non attrition ase, the married employes have larger proportion. In case of attrition, the sigles have more trend towards the attrition.

Q: Whether gender influences the decision on the bases of the marital status towards Attiration?

```
table(df$gen, df$marital, df$attr)
```

```
## , ,  = 0
##
##
##          Divorced Married Single
##   Female      108     241    152
##   Male        186     348    198
##
## , ,  = 1
##
##
##          Divorced Married Single
##   Female        9      31     47
##   Male         24      53     73
```

```
prop.table(table(df$gen, df$marital, df$attr))
```

```
## , ,  = 0
##
##
##             Divorced     Married      Single
##   Female 0.073469388 0.163945578 0.103401361
##   Male   0.126530612 0.236734694 0.134693878
##
## , ,  = 1
##
##
##             Divorced     Married      Single
##   Female 0.006122449 0.021088435 0.031972789
##   Male   0.016326531 0.036054422 0.049659864
```

There are three levels of marital status level: 1 ‘Divorced’ 2 ‘Married’ 3 ‘Single’. The output shows that the male employees have greater tendency towards Attiration and non-attrition in all ctegories of marital status.

Education(edu) and Attrition (attr)

Q: Are Education(edu) and Attrition (attr) independent?

> Chi-Square Test of Independence

> Ho: They are independent H1: They are not independent

> Applying chi-square test

```
chisq.test(df$edu, df$attr)
```

```
##
##  Pearson's Chi-squared test
##
## data:  df$edu and df$attr
## X-squared = 3.074, df = 4, p-value = 0.5455
```

> p-value is 0.545 , which is above 0.05, thus Ho is accepted and that edu and attr are independent.

Q: Which qualification level has greater tendency towards attrition?

```
table(df$edu, df$attr)
```

```
##
##       0   1
##   1 139  31
##   2 238  44
##   3 473  99
##   4 340  58
##   5  43   5
```

```
prop.table(table(df$edu, df$attr))
```

```
##
##               0           1
##   1 0.094557823 0.021088435
##   2 0.161904762 0.029931973
##   3 0.321768707 0.067346939
##   4 0.231292517 0.039455782
##   5 0.029251701 0.003401361
```

There are five levels of education level: 1 ‘Below College’ 2 ‘College’ 3 ‘Bachelor’ 4 ‘Master’ 5 ‘Doctor’ # Bachelor degree holders are having greater tendency towards both the Attiration and non-attrition, followed by the master degree holders.
##

Q: Are Stock Option (stock) and Attrition (attr) independent?

Chi-Square Test of Independence

Ho: They are independent H1: They are not independent

Applying chi-square test

```
chisq.test(df$stock, df$attr)
```

```
##
##  Pearson's Chi-squared test
##
## data:  df$stock and df$attr
## X-squared = 60.598, df = 3, p-value = 4.379e-13
```

p-value is 4.379e-13 , which is below 0.05, thus Ho is not accepted and that stock and attr are not independent.

Q: Does Attiration depend on the stock options to the employees?

```
table(df$stock, df$attr)
```

```
##
##       0   1
##   0 477 154
##   1 540  56
##   2 146  12
##   3  70  15
```

```
prop.table(table(df$stock, df$attr))
```

```
##
##               0           1
##   0 0.324489796 0.104761905
##   1 0.367346939 0.038095238
##   2 0.099319728 0.008163265
##   3 0.047619048 0.010204082
```

There are four levels of stock options: 0 ‘No stock Option’ 1 ‘Few Stock Options’ 2 ‘normal Stock Options’ 4 ‘Good stock Options’. With low or no stock options, the attrition rate is high.

Q: Are Work Life Balance (workLifeBal) and Attrition (attr) independent?

Chi-Square Test of Independence

Ho: They are independent H1: They are not independent

Applying chi-square test

```
chisq.test(df$workLifeBal, df$attr)
```

```
##
##  Pearson's Chi-squared test
##
## data:  df$workLifeBal and df$attr
## X-squared = 16.325, df = 3, p-value = 0.0009726
```

p-value is 0.000 , which is below 0.05, thus Ho is not accepted and that workLifeBal and attr are not independent.

Q: Does Attiration has any relevance to the work life balance situation of the employees?

```
table(df$workLifeBal, df$attr)
```

```
##
##       0   1
##   1  55  25
##   2 286  58
##   3 766 127
##   4 126  27
```

```
prop.table(table(df$workLifeBal, df$attr))
```

```
##
##              0          1
##   1 0.03741497 0.01700680
##   2 0.19455782 0.03945578
##   3 0.52108844 0.08639456
##   4 0.08571429 0.01836735
```

There are four levels of WorkLifeBalance: 1 ‘Bad’ 2 ‘Good’ 3 ‘Better’ 4 ‘Best’. With high score in workLifeBal , the case is strong for no attrition, and with low workLifeBal, the attrition is evident

Q: Are Performance Rate (perRate) and Attrition (attr) independent?

Chi-Square Test of Independence

Ho: They are independent H1: They are not independent

Applying chi-square test

```
chisq.test(df$perRate, df$attr)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  df$perRate and df$attr
## X-squared = 0.00015475, df = 1, p-value = 0.9901
```

p-value is 0.99 , which is above 0.05, thus Ho is accepted and that perRate and attrition are independent.

Q: Does performance rate influence the Attiration?

```
table(df$perRate, df$attr)
```

```
##
##        0    1
##   3 1044  200
##   4  189   37
```

```
prop.table(table(df$perRate, df$attr))
```

```
##
##              0          1
##   3 0.71020408 0.13605442
##   4 0.12857143 0.02517007
```

There are four levels of Performance Rating: 1 ‘Low’ 2 ‘Good’ 3 ‘Excellent’ 4 ’Outstanding’With high performance rating the case is weak for attrition, and with weak performance rating the attrition is quite evident.

Q: Are Job Satisfaction (jobSat) and Attrition (attr) independent?

Chi-Square Test of Independence

Ho: They are independent H1: They are not independent

Applying chi-square test

```
chisq.test(df$jobSat, df$attr)
```

```
##
##  Pearson's Chi-squared test
##
## data:  df$jobSat and df$attr
## X-squared = 17.505, df = 3, p-value = 0.0005563
```

> p-value is 0.000 , which is below 0.05, thus Ho is not accepted and that attr and jobSat are not independent.

Q: Does job satisfaction influence the Attiration?

```
table(df$jobSat, df$attr)
```

```
##
##       0   1
##   1 223  66
##   2 234  46
##   3 369  73
##   4 407  52
```

```
prop.table(table(df$jobSat, df$attr))
```

```
##
##              0          1
##   1 0.15170068 0.04489796
##   2 0.15918367 0.03129252
##   3 0.25102041 0.04965986
##   4 0.27687075 0.03537415
```

> There are four levels of JobSatisfaction: 1 ‘Low’ 2 ‘Medium’ 3 ‘High’ 4 ‘Very High’# With high job satisfaction the case is weak for attrition, and with poor job satisfaction the attrition is quite evident.

Modeling (Classification)

> First We split the data into train and test

```
set.seed(1)
temp <- sample(2, nrow(df), replace=T, prob = c(0.8,0.2))
train <- df[temp==1,]
test <- df[temp==2,]
str(train)
```

```
## 'data.frame':    1182 obs. of  13 variables:
##  $ age        : int  41 49 37 27 30 38 36 35 29 31 ...
##  $ gen        : Factor w/ 2 levels "Female","Male": 1 2 2 2 2 2 2 2 1 2 ...
##  $ marital    : Factor w/ 3 levels "Divorced","Married",..: 3 2 3 2 1 3 2 2 3 1
...
##  $ edu        : Factor w/ 5 levels "1","2","3","4",..: 2 1 2 1 1 3 3 3 2 1 ...
##  $ expComp    : int  6 10 0 2 1 9 7 5 9 5 ...
##  $ expCurrRole: int  4 7 0 2 0 7 7 4 5 2 ...
##  $ income     : int  5993 5130 2090 3468 2693 9526 5237 2426 4193 2911 ...
##  $ salHike    : int  11 23 15 12 22 21 13 13 12 17 ...
##  $ stock      : Factor w/ 4 levels "0","1","2","3": 1 2 1 2 2 1 3 2 1 2 ...
##  $ workLifeBal: Factor w/ 4 levels "1","2","3","4": 1 3 3 3 3 3 2 3 3 2 ...
##  $ perRate    : Factor w/ 2 levels "3","4": 1 2 1 1 2 2 1 1 1 1 ...
##  $ jobSat     : Factor w/ 4 levels "1","2","3","4": 4 2 3 2 3 3 3 2 3 3 ...
##  $ attr       : Factor w/ 2 levels "0","1": 2 1 2 1 1 1 1 1 1 1 ...
```

```
str(test)
```

```
## 'data.frame':    288 obs. of  13 variables:
##  $ age        : int  33 32 59 22 24 44 24 35 28 32 ...
##  $ gen        : Factor w/ 2 levels "Female","Male": 1 2 1 2 1 1 2 2 2 2 ...
##  $ marital    : Factor w/ 3 levels "Divorced","Married",..: 2 3 2 1 1 2 2 1 3 2
...
##  $ edu        : Factor w/ 5 levels "1","2","3","4",..: 4 2 3 2 2 4 3 2 4 3 ...
##  $ expComp    : int  8 7 1 1 4 22 2 1 2 9 ...
##  $ expCurrRole: int  7 7 0 0 2 6 0 0 2 8 ...
##  $ income     : int  2909 3068 2670 2935 4011 10248 2293 1951 3441 6162 ...
##  $ salHike    : int  11 13 20 13 18 14 16 12 13 22 ...
##  $ stock      : Factor w/ 4 levels "0","1","2","3": 1 1 4 3 2 2 2 2 1 2 ...
##  $ workLifeBal: Factor w/ 4 levels "1","2","3","4": 3 2 2 2 2 3 2 3 2 3 ...
##  $ perRate    : Factor w/ 2 levels "3","4": 1 1 2 1 1 1 1 1 1 2 ...
##  $ jobSat     : Factor w/ 4 levels "1","2","3","4": 3 4 1 4 3 4 4 4 3 4 ...
##  $ attr       : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 2 1 2 1 ...
```

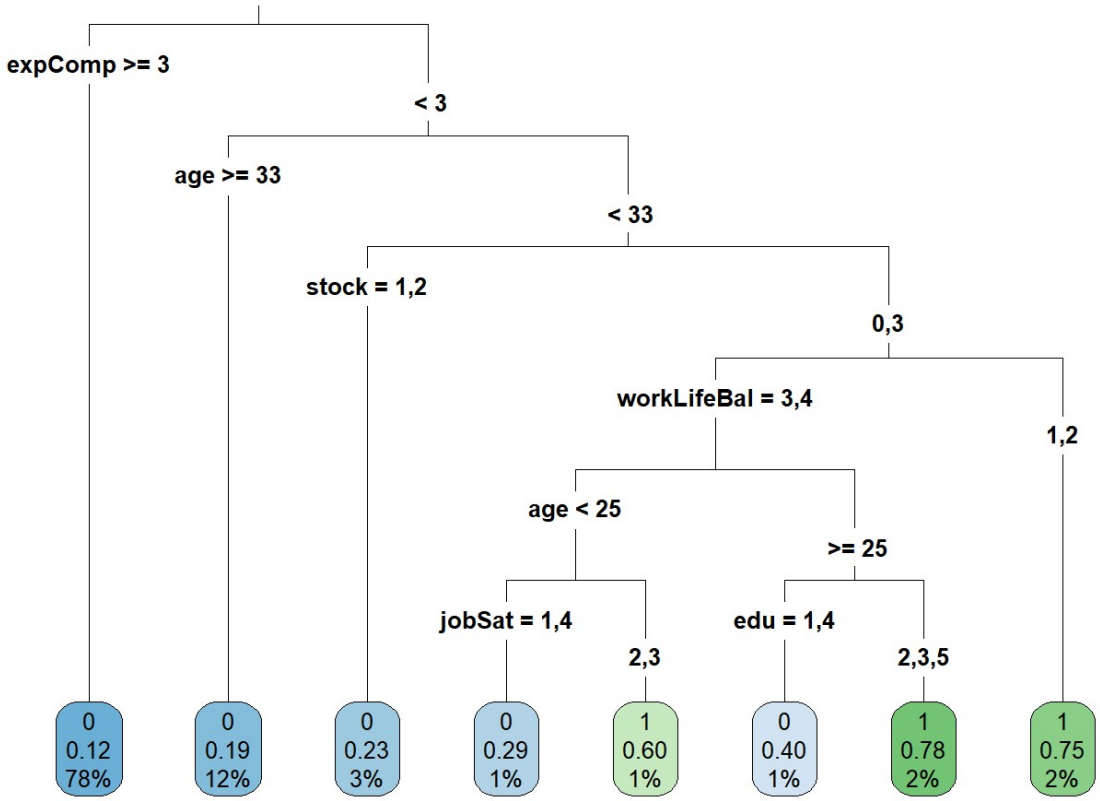
Decision Tree

```
m1 <- rpart(attr ~ .,
            data=train,
            method="class",
            parms=list(split="information"),
            control=rpart.control(usesurrogate=0,
                                  maxsurrogate=0))

m1
```

```
## n= 1182
##
## node), split, n, loss, yval, (yprob)
##       * denotes terminal node
##
##   1) root 1182 189 0 (0.8401015 0.1598985)
##     2) expComp>=2.5 922 107 0 (0.8839479 0.1160521) *
##     3) expComp< 2.5 260  82 0 (0.6846154 0.3153846)
##       6) age>=32.5 144  27 0 (0.8125000 0.1875000) *
##       7) age< 32.5 116  55 0 (0.5258621 0.4741379)
##        14) stock=1,2 39   9 0 (0.7692308 0.2307692) *
##        15) stock=0,3 77  31 1 (0.4025974 0.5974026)
##          30) workLifeBal=3,4 57  26 1 (0.4561404 0.5438596)
##            60) age< 24.5 29  13 0 (0.5517241 0.4482759)
##             120) jobSat=1,4 14   4 0 (0.7142857 0.2857143) *
##             121) jobSat=2,3 15   6 1 (0.4000000 0.6000000) *
##            61) age>=24.5 28  10 1 (0.3571429 0.6428571)
##             122) edu=1,4 10   4 0 (0.6000000 0.4000000) *
##             123) edu=2,3,5 18   4 1 (0.2222222 0.7777778) *
##          31) workLifeBal=1,2 20   5 1 (0.2500000 0.7500000) *
```

```
rpart.plot(m1, type = 3)
```



Model performance

```
pred_m1 <- predict(m1, test, type = 'class')
```

Obtaining the Confusion matrix

```
table(pred_m1,test$attr)
```

```
##
## pred_m1   0   1
##       0 236  41
##       1   4   7
```

Examine Accuracy

```
 mean(pred_m1 == test$attr)
```

```
## [1] 0.84375
```

Random Forest

```
m2 <- randomForest(attr ~ ., data = train, importance = TRUE)
m2
```

```
##
## Call:
##  randomForest(formula = attr ~ ., data = train, importance = TRUE)
##                Type of random forest: classification
##                      Number of trees: 500
## No. of variables tried at each split: 3
##
##         OOB estimate of  error rate: 15.06%
## Confusion matrix:
##     0  1 class.error
## 0 980 13  0.01309164
## 1 165 24  0.87301587
```

> The output of the above code produces the confusion matrix and error rate. At teh default 500 trees, the out of bag error 15.14% which shows that model (m2_1) has about 85% predicitve accuracy.

Model performance

```
pred_m2 <- predict(m2, test, type = "class")
```

Obtaining the Confusion matrix

```
table(pred_m2,test$attr)
```

```
##
## pred_m2   0   1
##       0 236  37
##       1   4  11
```

Examining Accuracy

```
mean(pred_m2 == test$attr)
```

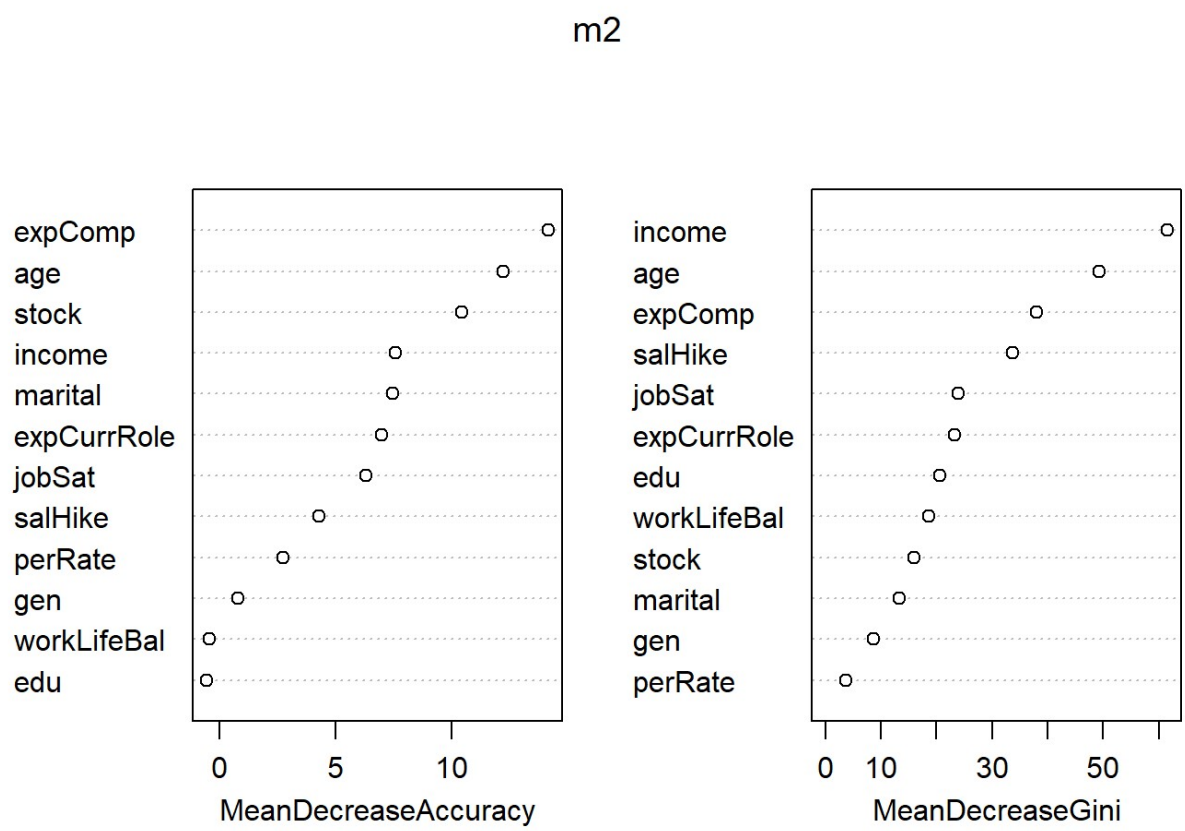
```
## [1] 0.8576389
```

Obtaining the key variables

```
importance(m2)
```

```
##                      0         1 MeanDecreaseAccuracy MeanDecreaseGini
## age          8.3531411 10.414677            12.2347156        49.335605
## gen         -0.3475908  2.202358             0.7678736         8.644153
## marital      5.6893335  4.736106             7.4624087        13.230093
## edu          0.7395710 -2.932664            -0.5717438        20.564210
## expComp      9.9620275  9.057927            14.1995746        37.920990
## expCurrRole  3.7099508  6.998061             6.9947750        23.222268
## income       4.5372072  6.765247             7.5758933        61.603574
## salHike      4.0105641  1.455511             4.2679993        33.715681
## stock        6.9389024  8.372038            10.4603301        15.902802
## workLifeBal  0.5476238 -2.145242            -0.4314838        18.542904
## perRate      1.1494169  3.201585             2.7383415         3.707558
## jobSat       4.0184449  6.241725             6.3242517        23.931732
```

```
varImpPlot(m2)
```



Logistic Regressopn

```
m4 <- glm(attr ~ ., family = binomial(link="logit"), data = train)
summary(m4)
```

```
##
## Call:
## glm(formula = attr ~ ., family = binomial(link = "logit"), data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.4103  -0.6273  -0.4513  -0.2839   3.1474
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)     1.135e+00  8.041e-01   1.411 0.158224
## age            -2.408e-02  1.128e-02  -2.135 0.032796 *
## genMale         2.337e-02  1.725e-01   0.136 0.892200
## maritalMarried  1.080e-01  2.644e-01   0.408 0.682934
## maritalSingle   3.374e-01  3.638e-01   0.927 0.353706
## edu2            6.895e-02  3.123e-01   0.221 0.825252
## edu3            1.509e-01  2.749e-01   0.549 0.583044
## edu4            1.937e-01  2.982e-01   0.650 0.515960
## edu5           -1.169e-01  6.052e-01  -0.193 0.846839
## expComp         9.974e-03  2.807e-02   0.355 0.722378
## expCurrRole    -1.071e-01  4.240e-02  -2.527 0.011509 *
## income         -6.184e-05  2.769e-05  -2.233 0.025533 *
## salHike        -5.893e-03  3.667e-02  -0.161 0.872315
## stock1         -8.091e-01  2.914e-01  -2.776 0.005495 **
## stock2         -8.980e-01  4.097e-01  -2.192 0.028368 *
## stock3         -8.754e-02  4.198e-01  -0.209 0.834794
## workLifeBal2   -5.929e-01  3.508e-01  -1.690 0.091002 .
## workLifeBal3   -7.215e-01  3.209e-01  -2.249 0.024530 *
## workLifeBal4   -7.501e-01  3.988e-01  -1.881 0.060005 .
## perRate4        2.077e-01  3.624e-01   0.573 0.566475
## jobSat2        -5.653e-01  2.561e-01  -2.208 0.027253 *
## jobSat3        -4.689e-01  2.250e-01  -2.084 0.037176 *
## jobSat4        -9.092e-01  2.374e-01  -3.830 0.000128 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1038.98  on 1181  degrees of freedom
## Residual deviance:  928.14  on 1159  degrees of freedom
## AIC: 974.14
##
## Number of Fisher Scoring iterations: 5
```

Model performance

goodness of fit (pseudo R-Square)

```
nagelkerke(m4)
```

```
## $Models
##
## Model: "glm, attr ~ ., binomial(link = \"logit\"), train"
## Null:  "glm, attr ~ 1, binomial(link = \"logit\"), train"
##
## $Pseudo.R.squared.for.model.vs.null
##                              Pseudo.R.squared
## McFadden                            0.106683
## Cox and Snell (ML)                  0.089512
## Nagelkerke (Cragg and Uhler)        0.153063
##
## $Likelihood.ratio.test
##  Df.diff LogLik.diff  Chisq    p.value
##      -22     -55.421 110.84 7.8063e-14
##
## $Number.of.observations
##
## Model: 1182
## Null:  1182
##
## $Messages
## [1] "Note: For models fit with REML, these statistics are based on refitting wit
h ML"
##
## $Warnings
## [1] "None"
```

The output of above function is McFadden = 0.1066, Cox and Snell (ML) = 0.0895, and Nege lkerke (Vragg and uhler) = 0.153. The liklihood test ratio (chi square) is 110.84 at p-value 0.000

We an also directly get above mentioed p-value by using following function

```
p_value <- with(m4,pchisq(null.deviance - deviance, df.null - df.residual, lower.ta
il = F))
p_value
```

```
## [1] 7.806279e-14
```

The result is 0.000 (less than 5% thus goodness of fit is significant)

We can also use following code to get above goodness of

fit statistics

```
pR2(m4)
```

```
##           llh       llhNull            G2      McFadden          r2ML
## -464.06997966 -519.49076475  110.84157019    0.10668291    0.08951203
##          r2CU
##    0.15306348
```

Examine the Accuracy

```
fitted.results <- predict(m4,newdata=test,type='response')
fitted.results <- ifelse(fitted.results > 0.5,1,0)

misClasificError <- mean(fitted.results != test$attr)
print(paste('Accuracy',1-misClasificError))
```

```
## [1] "Accuracy 0.847222222222222"
```

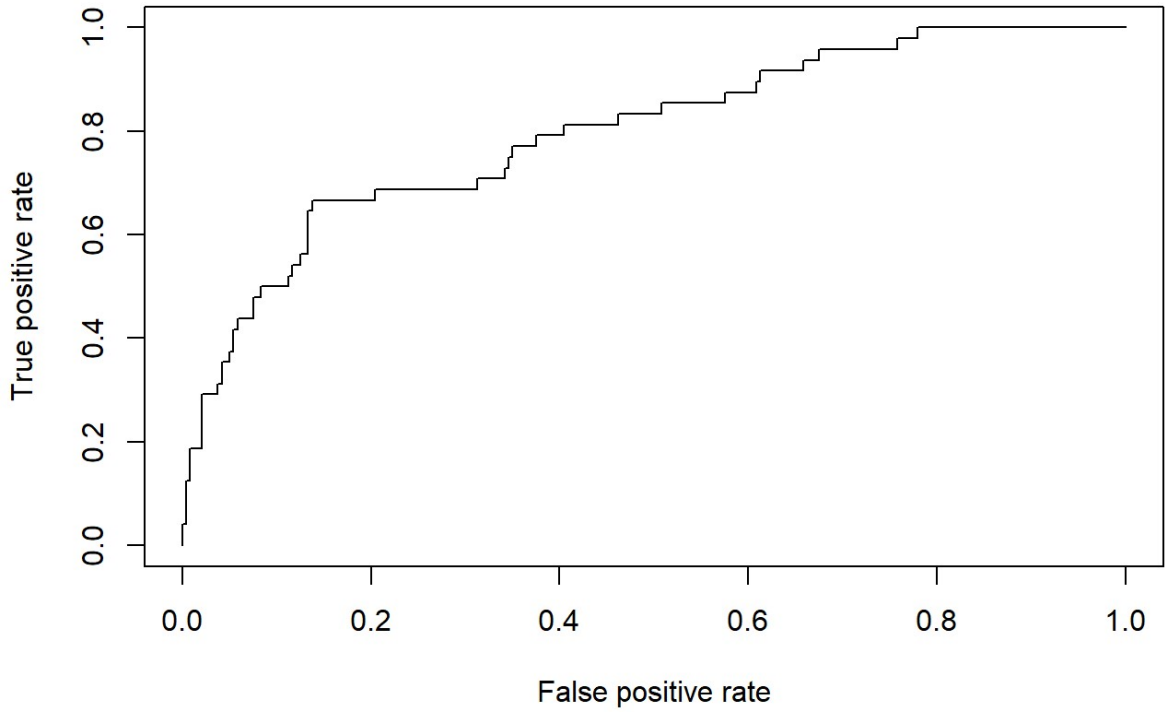
Accuracy is 0.85

Plotting the ROC curve

ROC plot and the AUC (area under the curve) are merformance measures for the binary classifier.

The ROC is a curve generated by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings while the AUC is the area under the ROC curve. As a rule of thumb, a model with good predictive ability should have an AUC closer to 1 (1 is ideal) than to 0.5.

```
p <- predict(m4, newdata=test, type="response")
pr <- prediction(p, test$attr)
prf <- performance(pr, measure = "tpr", x.measure = "fpr")
plot(prf)
```




```
auc <- performance(pr, measure = "auc")
auc <- auc@y.values[[1]]
auc
```

```
## [1] 0.7976562
```

The AUC is 0.79 which is reasonable.

Conclusion

The three classification ML-algorithms employed in this project were Decision Tree, Random Forest, and the Logistic Regression. The accuracy was almost similar. The predictability can be improved by including more features into the model and removing those which are not contributing to the variance in the outcome variable. The study has proved that the application of ML-algorithm can be an effective way to improve the HR operations. Employees attrition is a major concern for many organizations. Due to the rapid growth in technology and globalization the labor mobility has increased tremendously. In that case, the ML technique can be an important skill for the HR managers.

limitations

The HR data in the large size organizations is quite a lot and machine memory, and internet speed to download or link are the challenges. Many companies are also reluctant to share real HR data as it poses many challenges to them.

Future work

The project finding has shown that the ML-algorithms are an effective tool to improve the HR operations especially with reference to the Attrition issues. By including more variables, the further variance can be explained in the outcome variables and higher level of accuracy can be obtained. For example the variables like , education field, job level, department, job role, monthly, hourly and daily rates, number of companies an employee has worked previously, total Work experience, distance from home, business travel frequency, overtime options, training times in the previous years, promotion frequency, time spent with the current manager, satisfaction rates like environment satisfaction, job involvement, relationship satisfaction etc. Moreover, the predictive models can also be studied on the continuous scale outcome variable.

References/Bibliograpgy

Following resources were consulted

http://www.sthda.com/english/wiki/visualize-correlation-matrix-using-correlogram (http://www.sthda.com/english/wiki/visualize-correlation-matrix-using-correlogram) https://cran.r-project.org/web/packages/corrplot/vignettes/corrplot-intro.html (https://cran.r-project.org/web/packages/corrplot/vignettes/corrplot-intro.html) https://www.kaggle.com/rohitkumar06/let-s-reduce-attrition-logistic-reg-acc-88 (https://www.kaggle.com/rohitkumar06/let-s-reduce-attrition-logistic-reg-acc-88) https://www.kaggle.com/pavansubhasht/ibm-hr-analytics-attrition-dataset (https://www.kaggle.com/pavansubhasht/ibm-hr-analytics-attrition-dataset) Material provided in the HarvardX: PH125 courses especially in the Machine Learning course (HarvardX: PH125.8x).