EMPLOYEES ATTRITION ANALYSIS

ST-O13: Data Mining

NAMES: 1} PRAKRITI PARSAILA (PRN-2145)

2}MONIKA (PRN-2134)

MENTOR: DR. AKANKSHA KASHIKAR, SPPU, PUNE

1. Introduction

Problem Statement:

A large company named ABC, employs, at any given point of time, around 4000 employees. However, every year, around 15% of its employees leave the company and the company needs to be replace them with the talent pool available in the job market. The management believes that this level of attrition (employees leaving, either on their own or because they got fired) is bad for the company, because of the following reasons :

1. The former employees projects get delayed, making it difficult to meet timelines, resulting in a loss of reputation among clients and partners.

2. More often than not, the new employees have to be trained for the job and/or given time to acclimatise themselves to the company.

Hence, the management has contracted you to understand what factors they should focus on, in order to curb attrition. In other words, they want to know what changes they should make to their workplace, in order to get most of their employees to stay.

The goal of the case study: The main objectives for this project is to uncover the factors that lead to employee attrition. Development of retention strategies is important to the growth of a company. Employee turnover has many consequences such as low morale, increased recruitment costs, and decreased productivity. The goal of this analysis is to discover why employees leave, and create a model that predicts employee attrition to help alleviate attrition.

Data Importing and understanding:

Originally, we are considering the data in the folder named Employee Records which contains csv files of about 4410 employees. The variables present are-

Age, Attrition, Business travel, Department, Distance from home, Education, Education Field, Employee Count, Employee Number, Environment Satisfaction, Gender, Job Involvement, Job Level, Job Role, Job Satisfaction, Marital Status, Monthly Income, Number Companies Worked, Over18, Percent Salary Hike, Performance Rating, Relationship Satisfaction, Standard Hours, Stock Option Level ,Total Working Years, Training Times Last Year, Work Life Balance, Years At Company, Years Since Last Promotion, Years With Current Manager.

Names of variables are self-explanatory.

Data Preprocessing & Treatment-

In this we combined **'general_data','employee_survey_data'**and **'manager_survey_data'**. For in and out time sheet we have computed the average working hours . This might gives us some additional insight. Code is given below:

```
## Merge all the data frames and the new variable which is computed from the
in_time and out_time data frame:

Data=inner_join(general_data,manager_survey_data, by = 'EmployeeID') %>%
  inner_join(.,employee_survey_data, by = 'EmployeeID') %>%
  inner_join(.,Avg_work,by="EmployeeID")
dim(Data)
## [1] 4410   30


# Droping values that have the same value for all observations:


same_values = nearZeroVar(Data, names = TRUE)
data = Data%>%
  dplyr::select(-c(c('EmployeeID', same_values)))
```

Here, Employee ID, Employee Count, Over 18 and Standard Hours variables have same values and doesn't play any role in it.

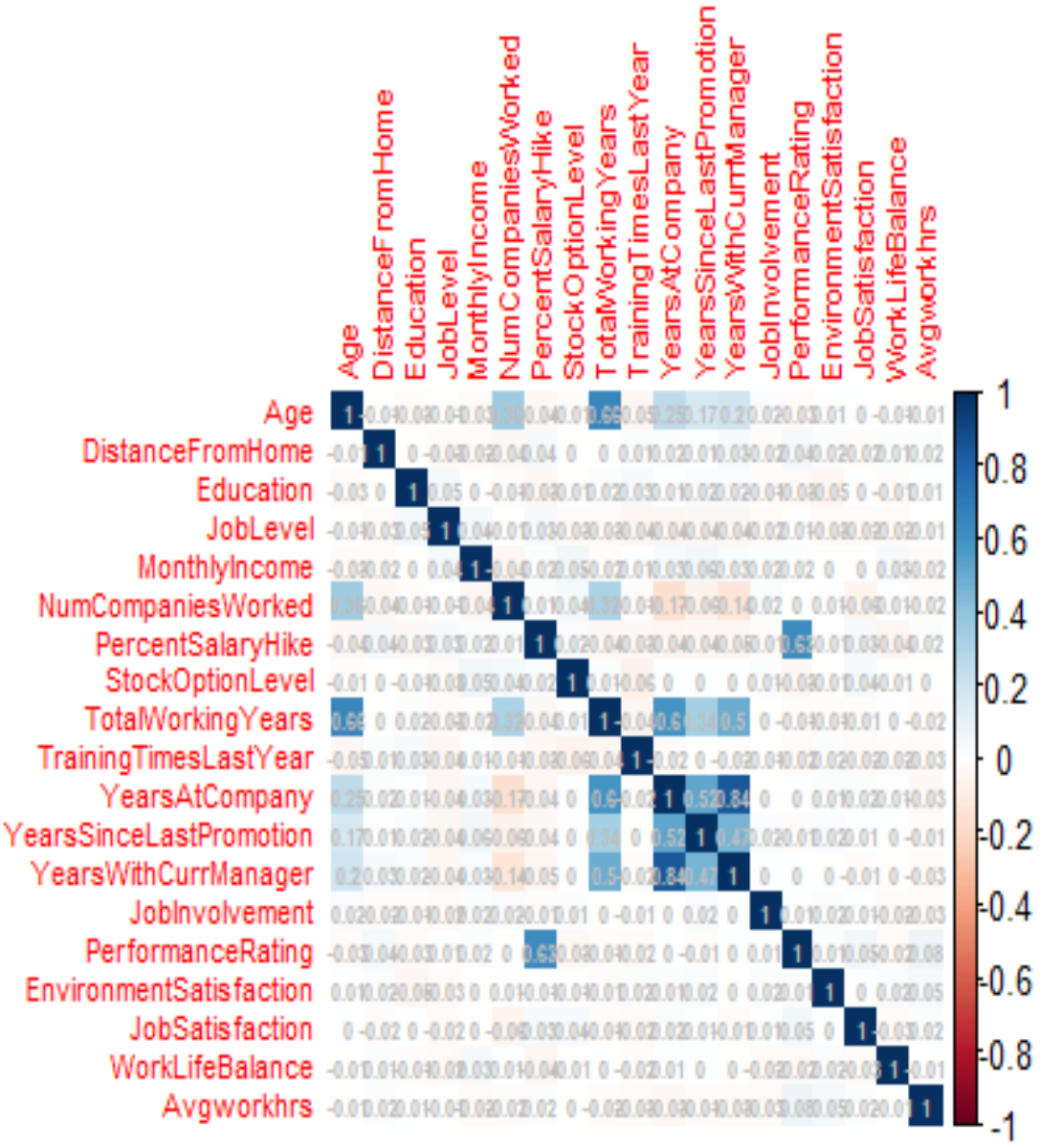
```
dim(data) ## Modified data for analysis
## [1] 4410   26
```

For missing values, we have removed 110 observations as the percentage for missing values is very low in comparison with 4410 observations.

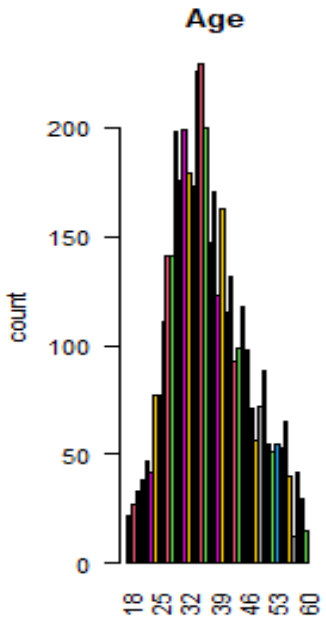
Data Exploration and Visualization:

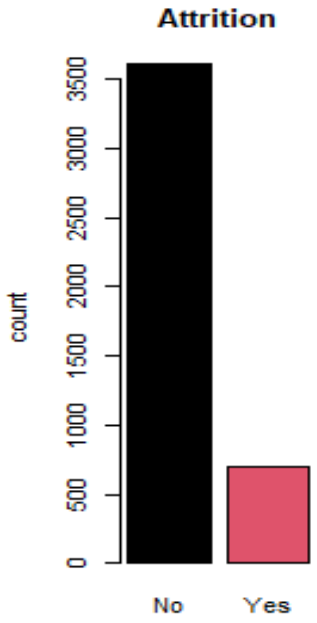
1) Correlations:

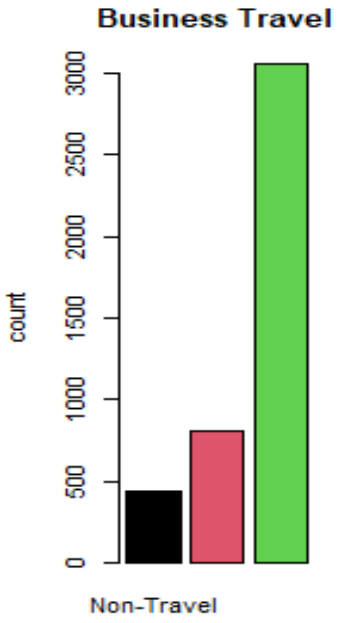
Years with current manger and years at company have a high correlation , total working year and age have a high correlation , percent salary hike and performance rating have a high correlation .

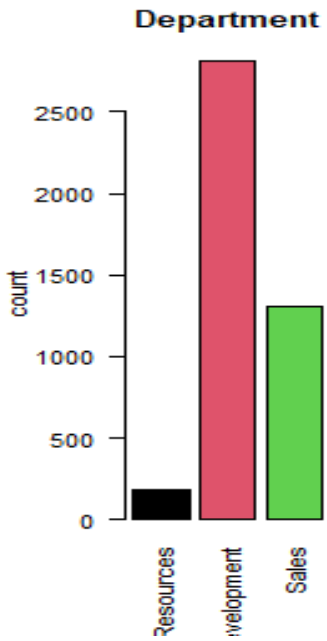


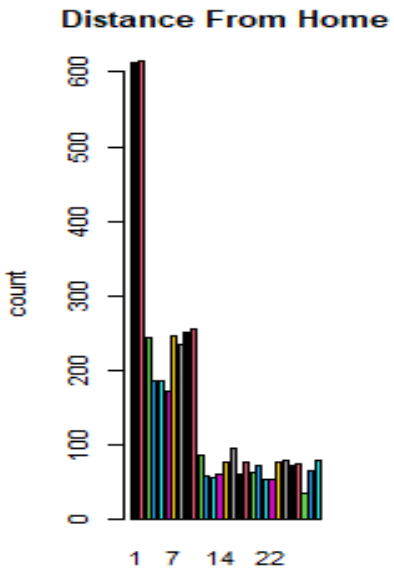
2) Barplot of categorical/count variables is plotted:

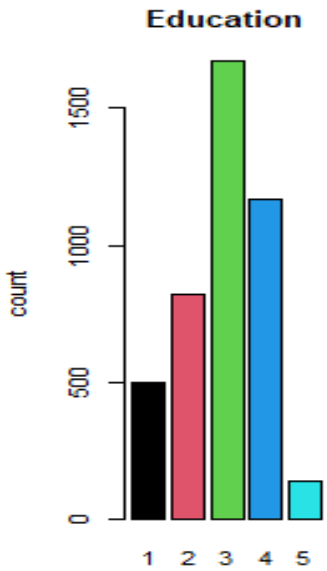


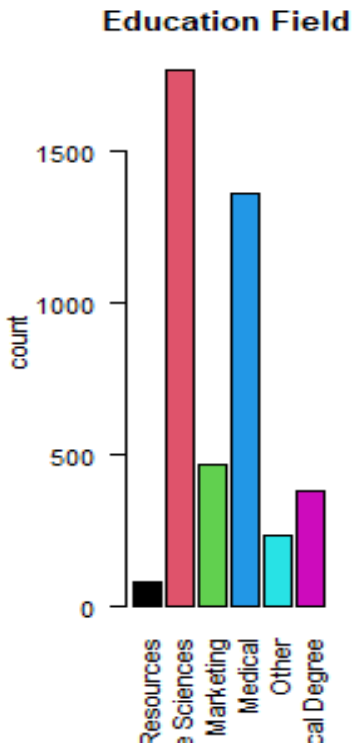


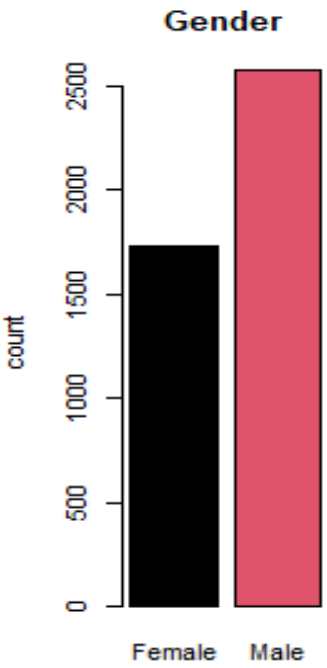


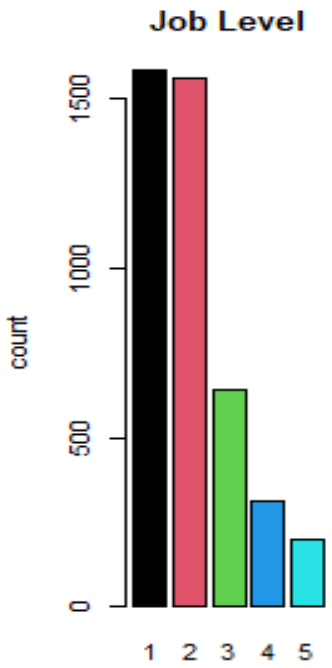


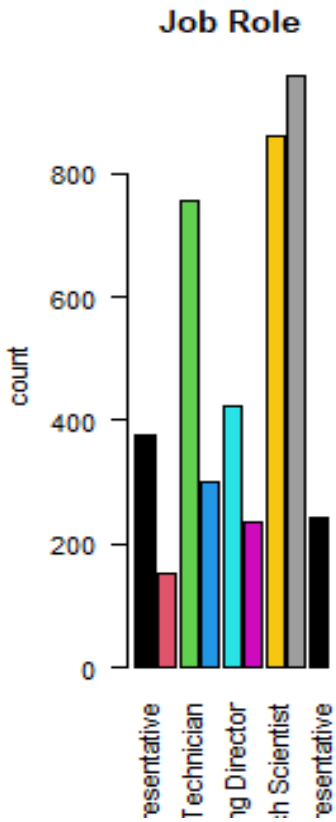


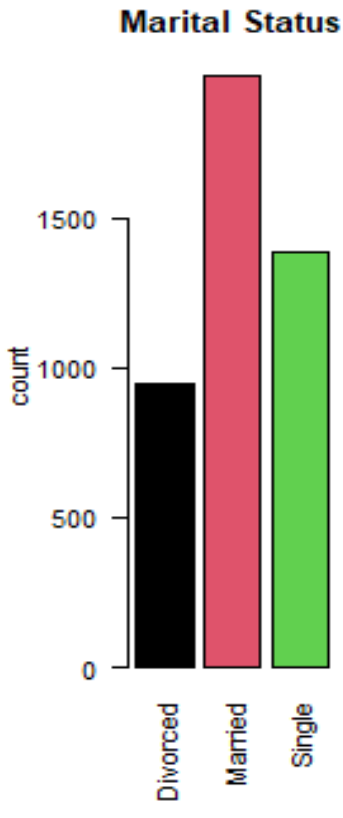


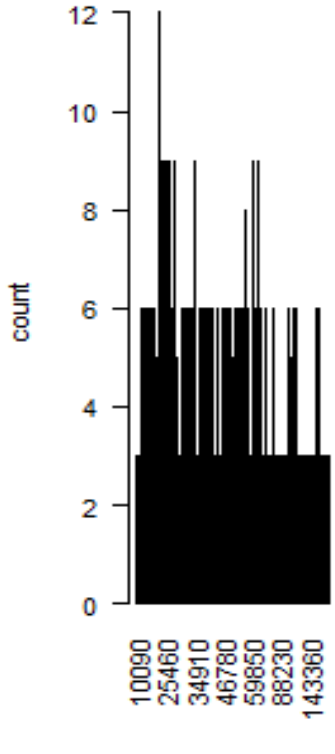

Education Field
count
0
500
1000
1500
Resources
Sciences
Marketing
Medical
Other
Degree

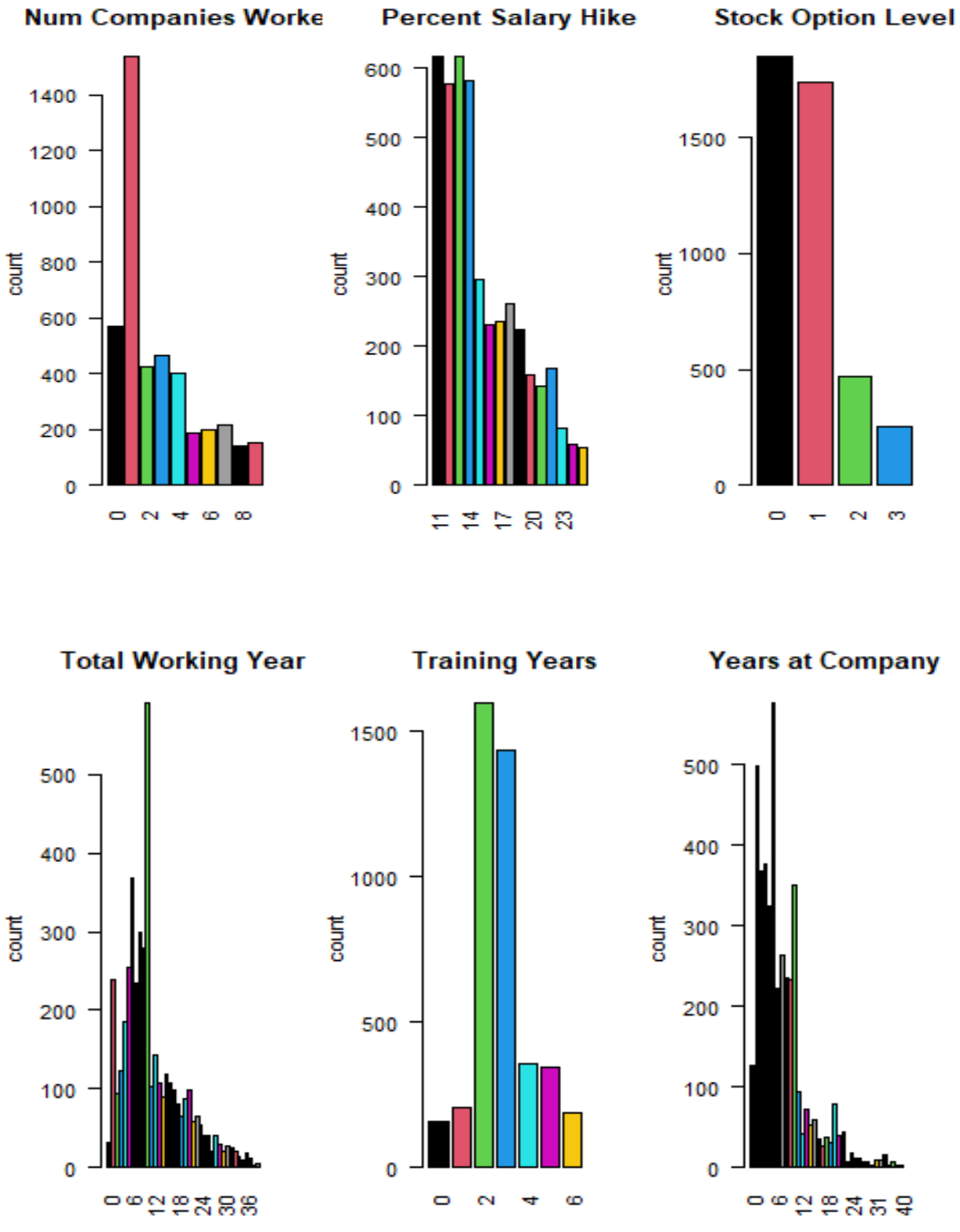

Gender
count
0
500
1000
1500
2000
2500
Female
Male

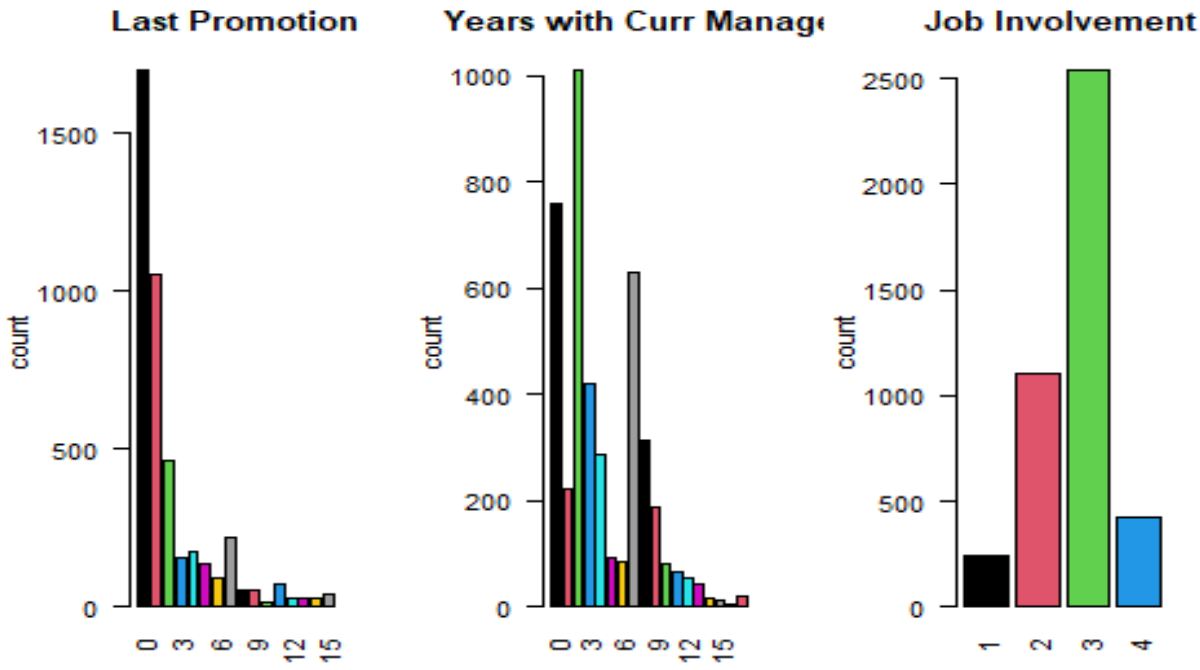

Job Level
count
0
500
1000
1500
1
2
3
4
5

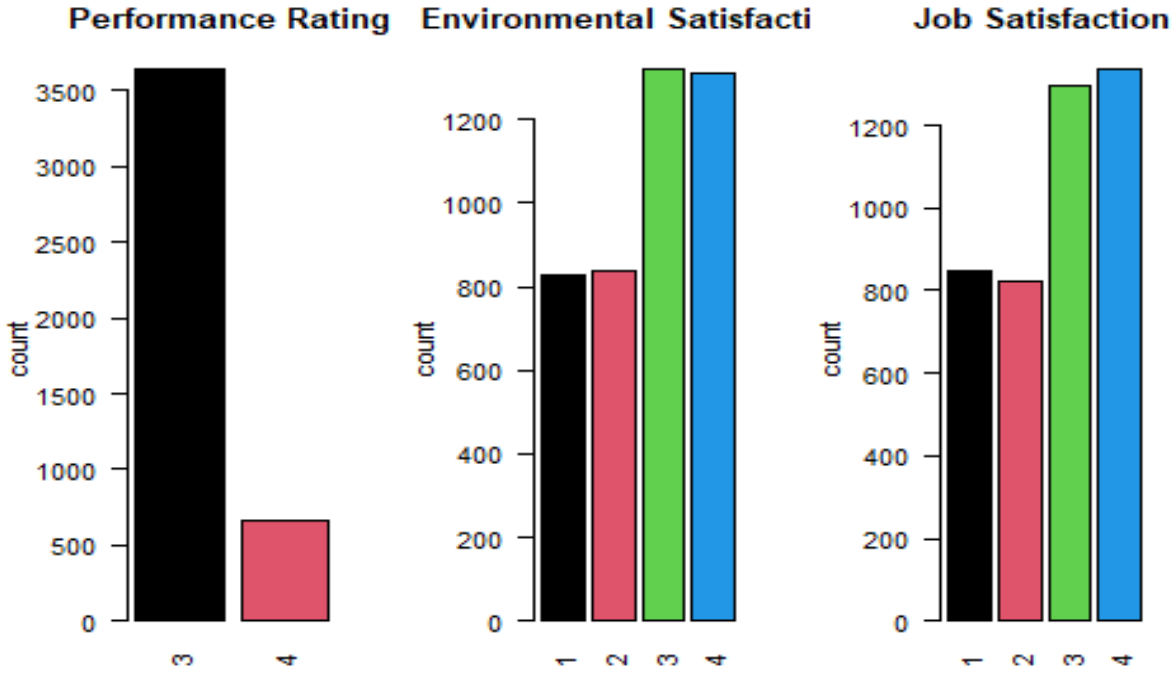

Job Role
count
0
200
400
600
800
Technician
Director
Scientist


Marital Status
count
0
500
1000
1500
Divorced
Married
Single

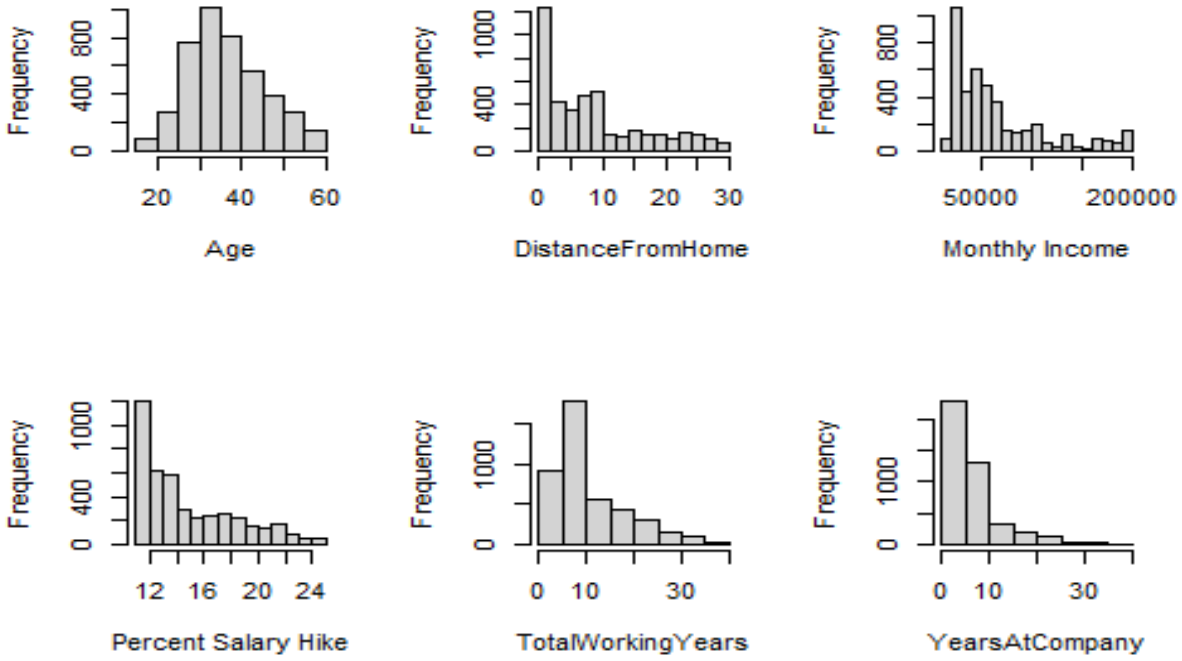

Monthly Income
count
0
2
4
6
8
10
12
10090
25460
34910
46780
59850
88230
143360


Num Companies Worke
count
0
200
400
600
800
1000
1200
1400
0
2
4
6
8
Percent Salary Hike
count
0
100
200
300
400
500
600
11
14
17
20
23
Stock Option Level
count
0
500
1000
1500
0
1
2
3
Total Working Year
count
0
100
200
300
400
500
0
6
12
18
24
30
36
Training Years
count
0
500
1000
1500
0
2
4
6
Years at Company
count
0
100
200
300
400
500
0
6
12
18
24
31
40


Last Promotion
count
0
500
1000
1500
0
3
6
9
12
15
Years with Curr Manage
count
0
200
400
600
800
1000
0
3
6
9
12
15
Job Involvement
count
0
500
1000
1500
2000
2500
1
2
3
4

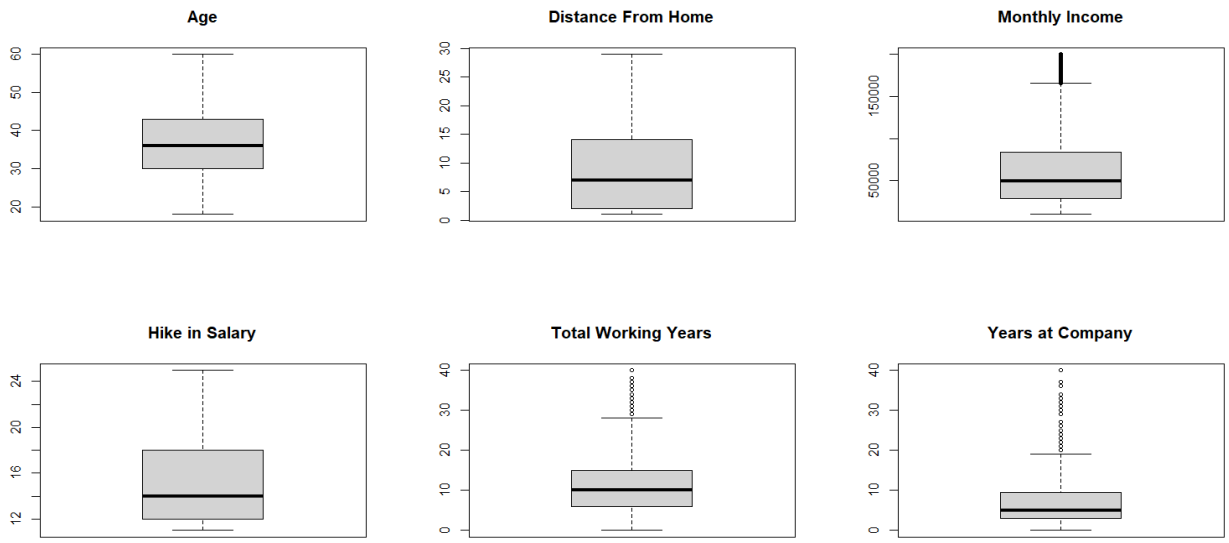

Performance Rating
count
0
500
1000
1500
2000
2500
3000
3500
3
4
Environmental Satisfacti
count
0
200
400
600
800
1000
1200
1
2
3
4
Job Satisfaction
count
0
200
400
600
800
1000
1200
1
2
3
4

3) Histogram of numeric variables is plotted:



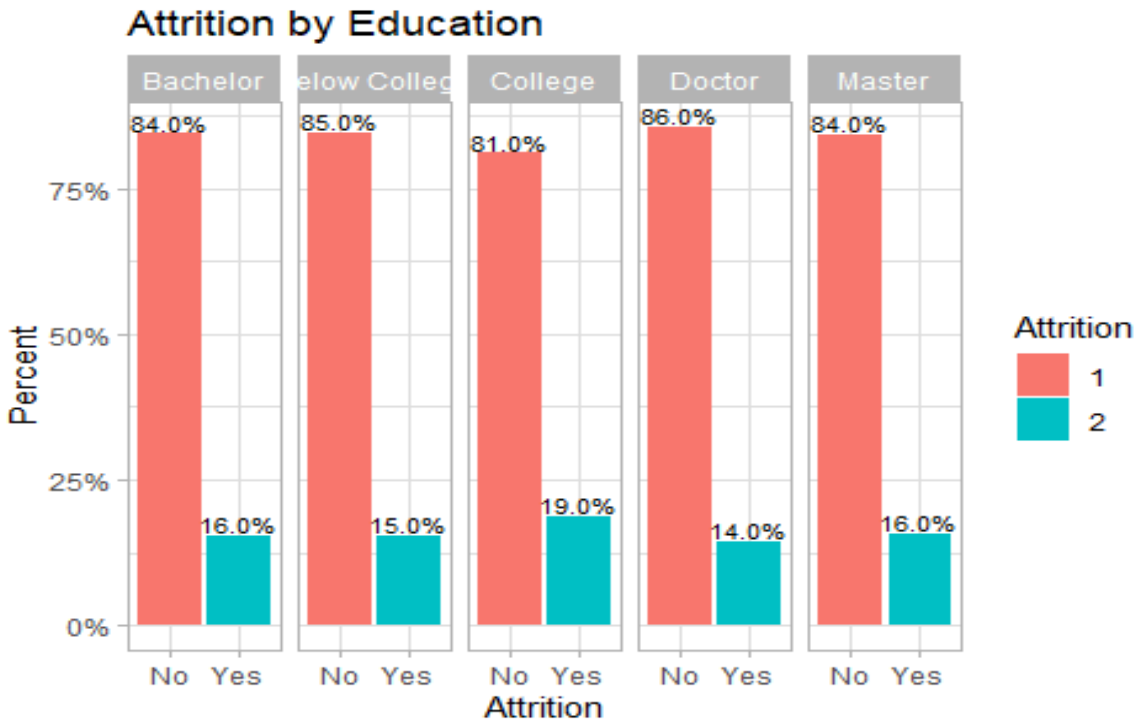
Except AGE, most of the variables are skewedly distributed. Age is almost normally distributed.

4) Boxplot (to find outliers in the data):



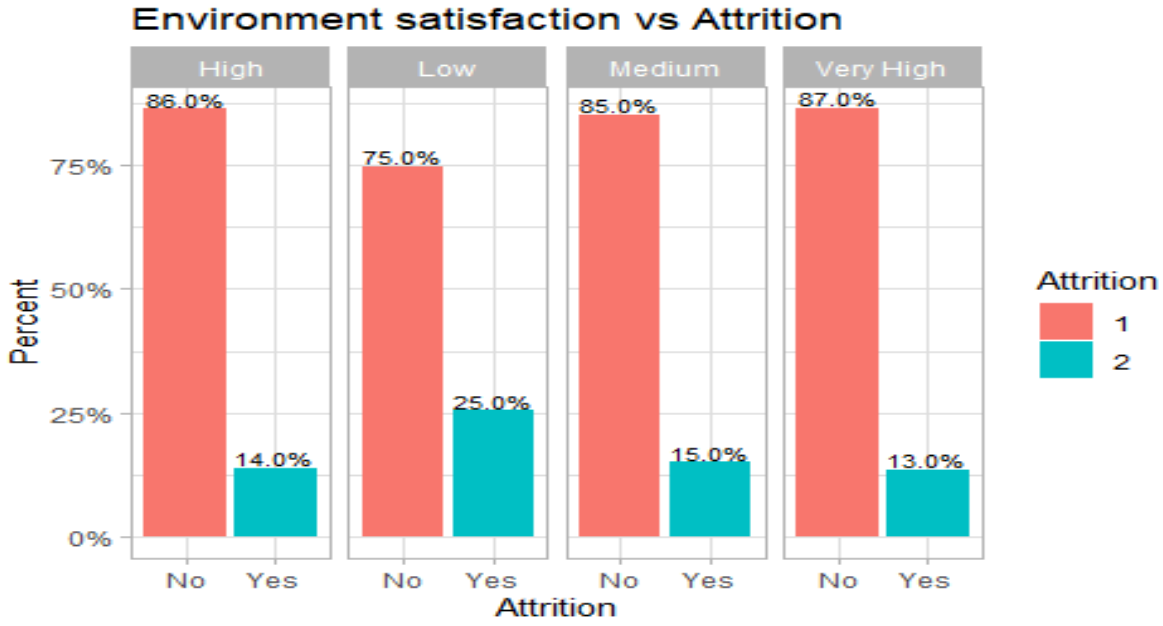
5) Education and Attrition:

This graph shows us that employees with college education the highest attrition rate.



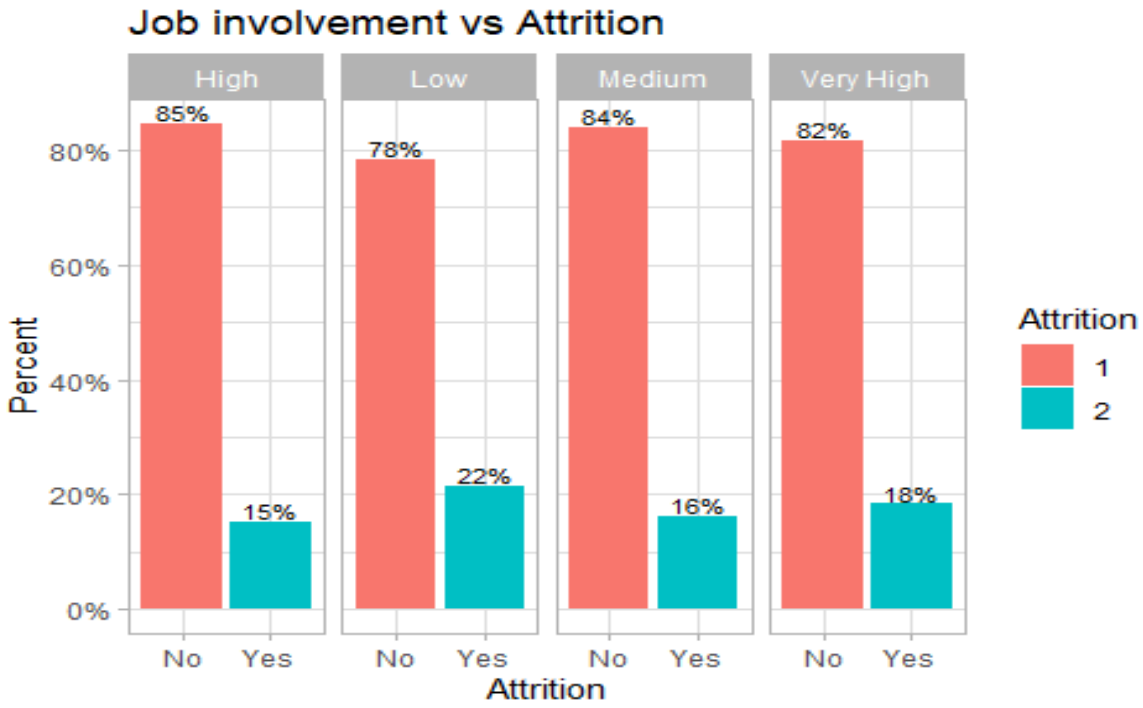
6) Environment satisfaction and Attrition:

Employees with low environment satisfaction leave more than employees with medium, high, and very high environment satisfaction.



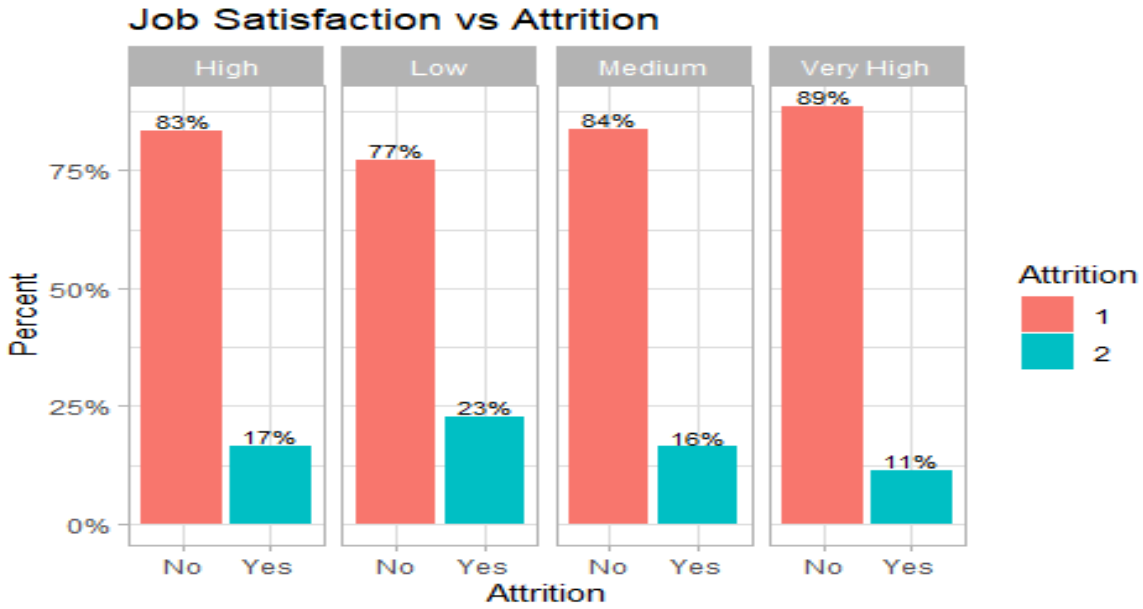
7) Job Involvment and Attrition:

This graph shows us that employees with low job involvement has the highest attrition rate.



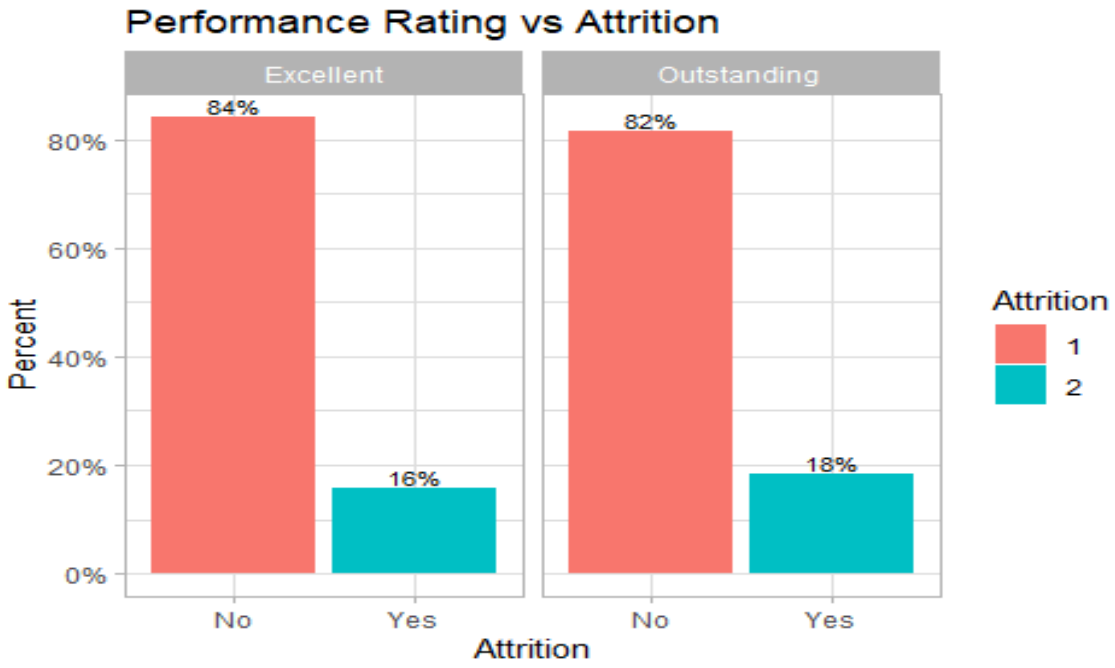
8) Job Satisfaction and Attrition:

Employees with low job satisfaction leave more than employees with medium, high, and very high job satisfaction.



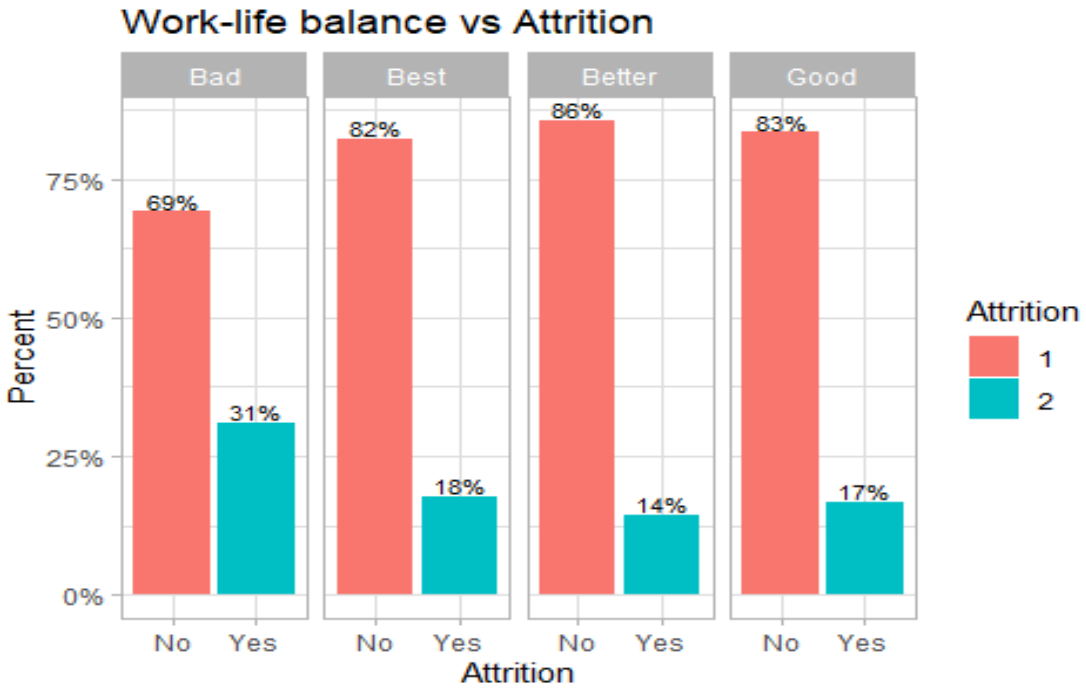
9) Performance Rating and Attrition:

Employees with outstanding performance leave more than employees with excellent performance.



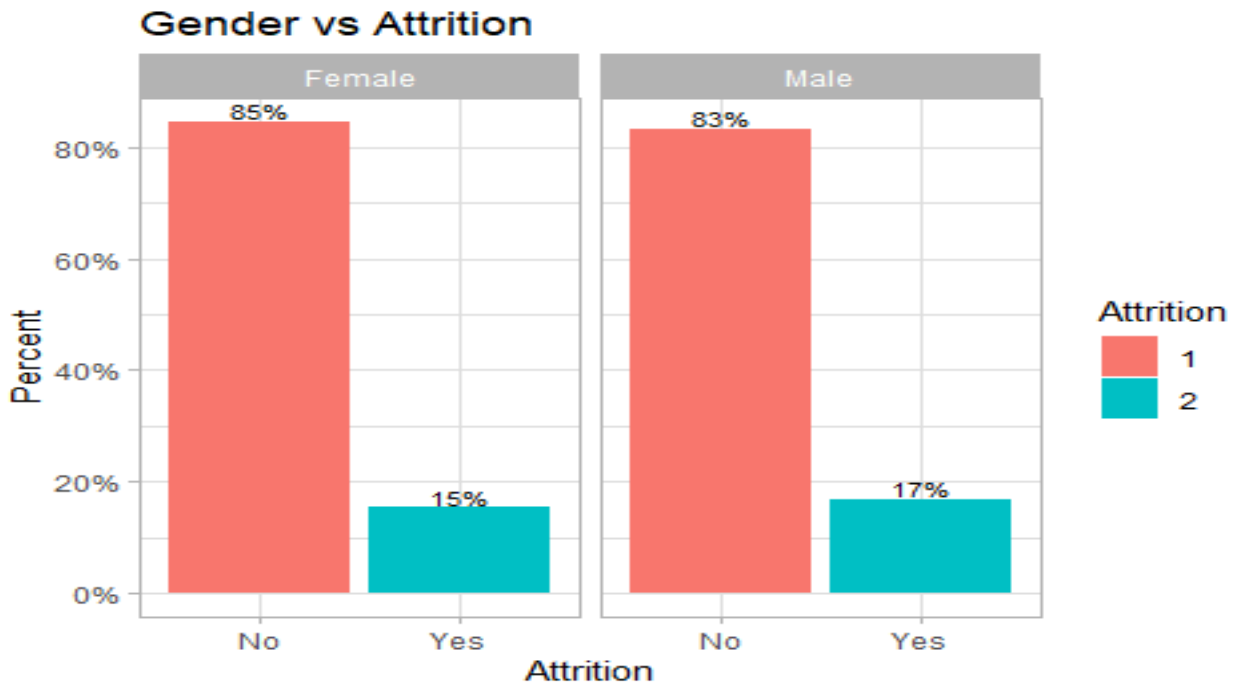
10) Work Life Balance and Attrition:

Employees with bad work life balance have the highest amount of turnover.



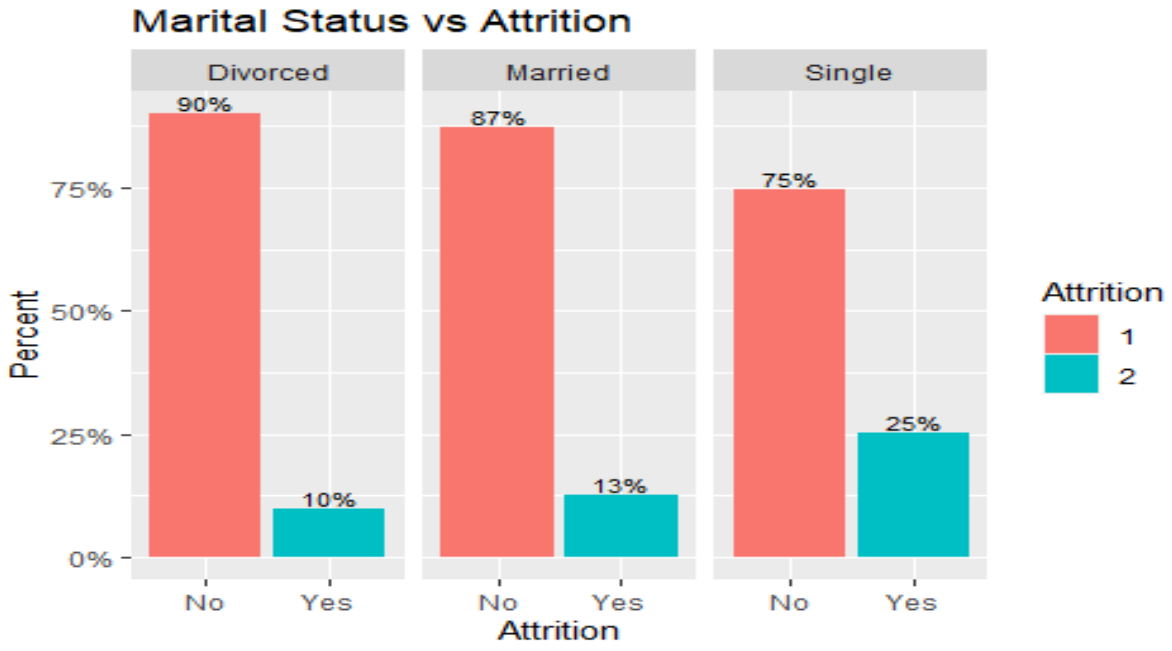
11) Gender and Attrition:

The attrition rates between men and women is very similar. Men have a slightly higher turnover rate.



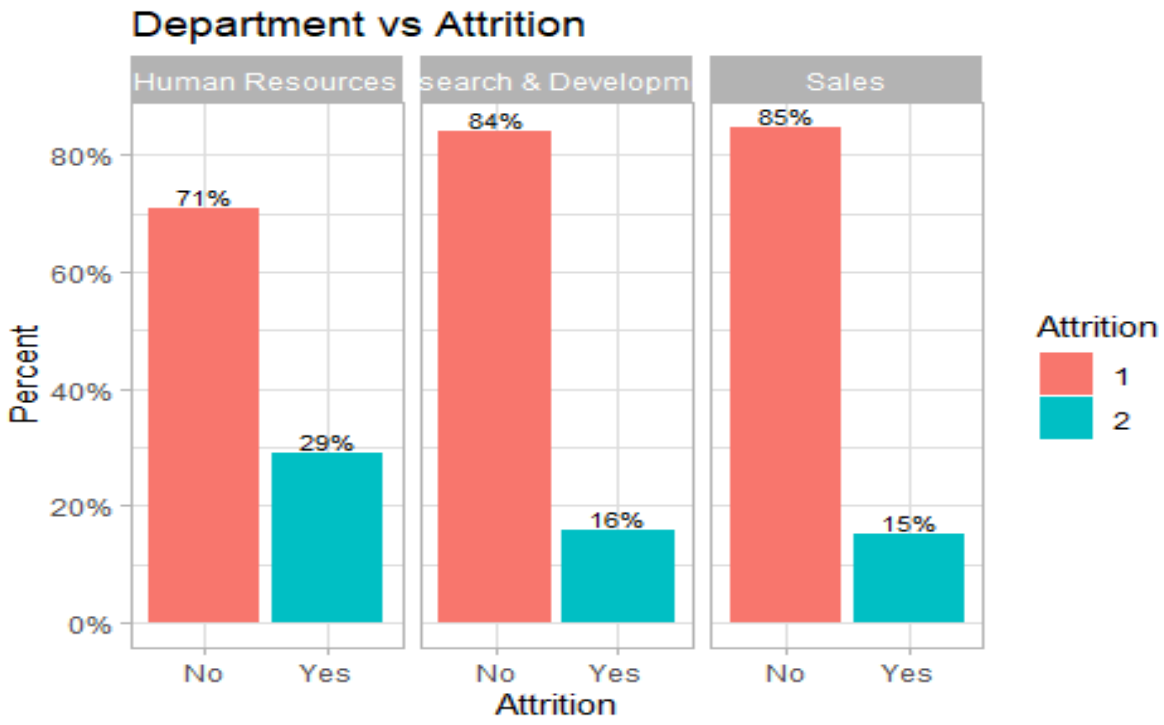
12) Martial Status and Attrition:

Employees that are single have the highest turnover.

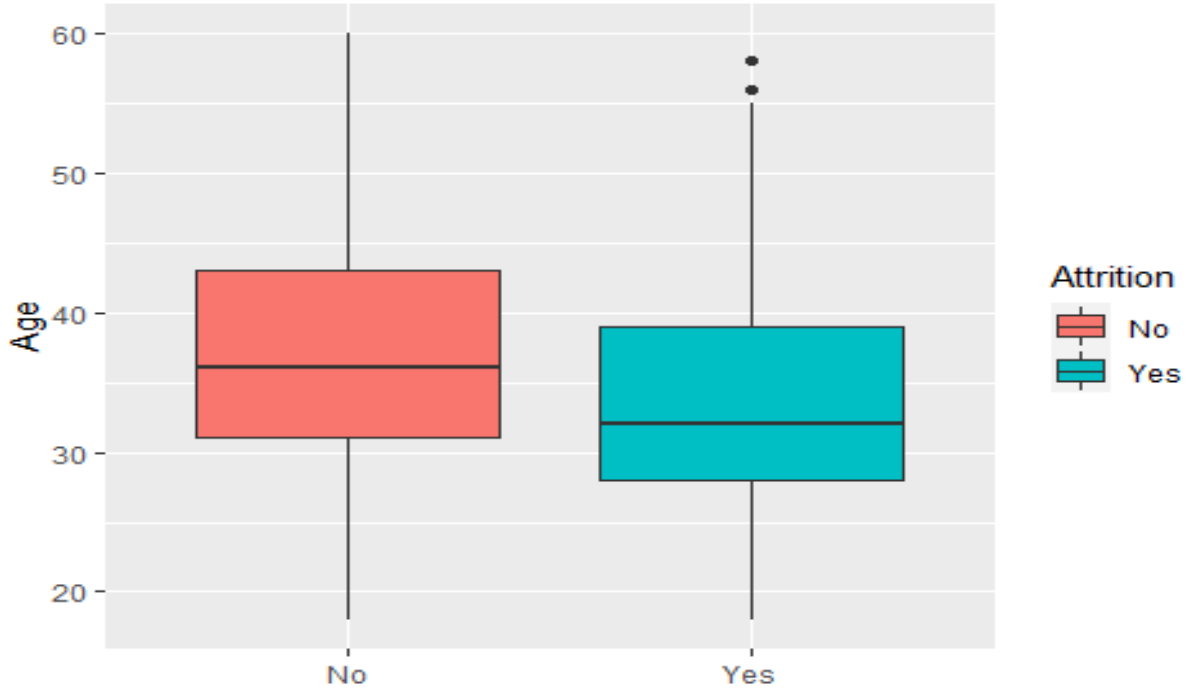


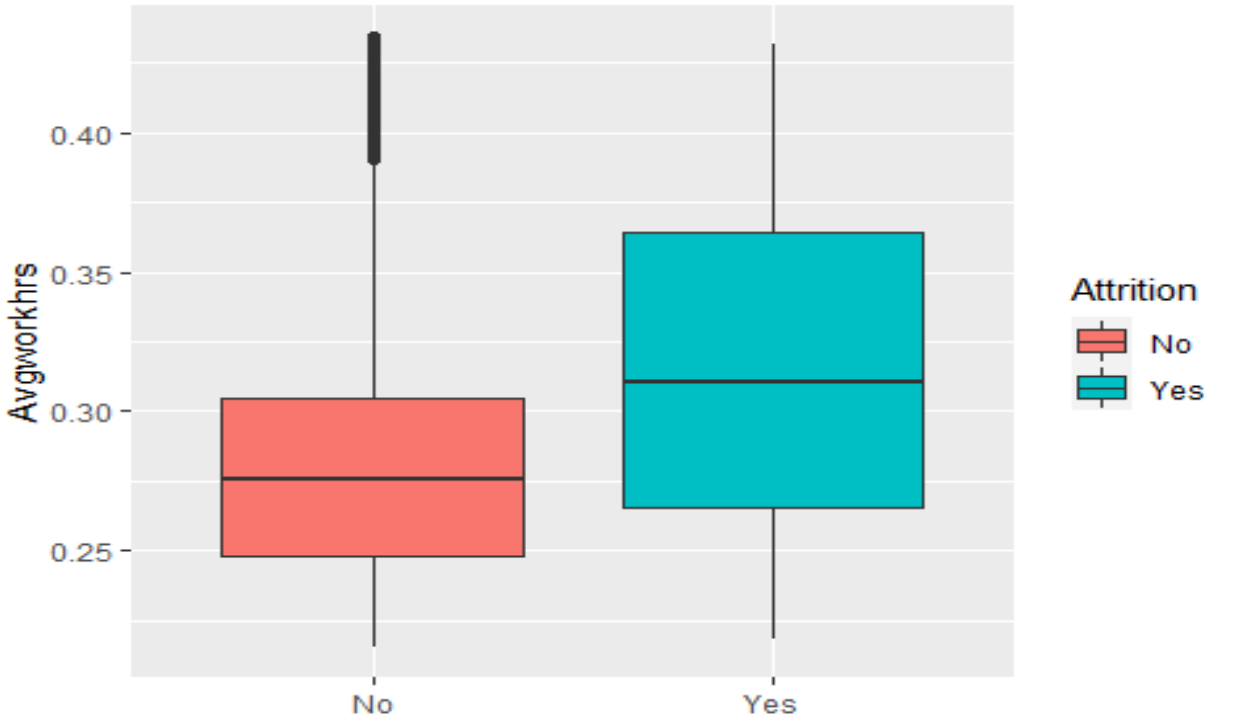
13) Department and Attrition:

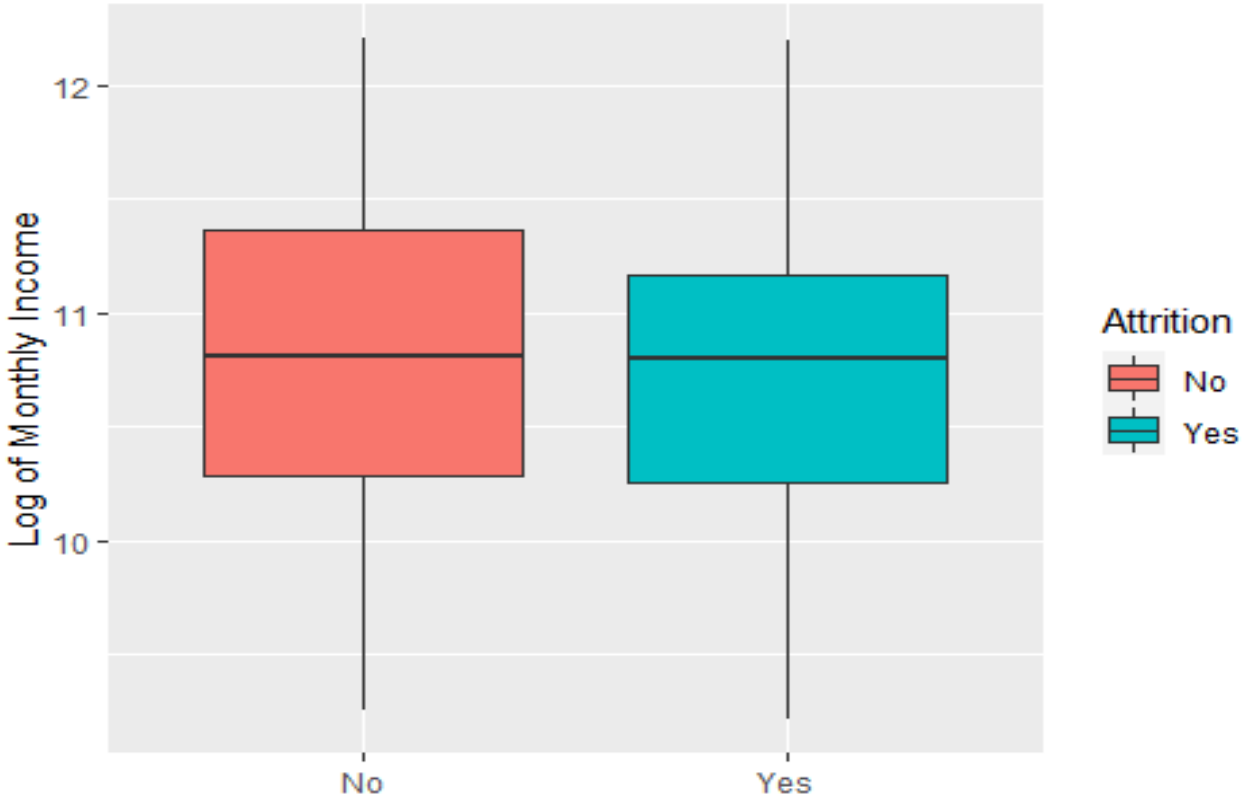
Employees from human resources department have the highest turnover.



14) Boxplot of Attrition vs Age , Avg working hours and Monthly Income is plotted:

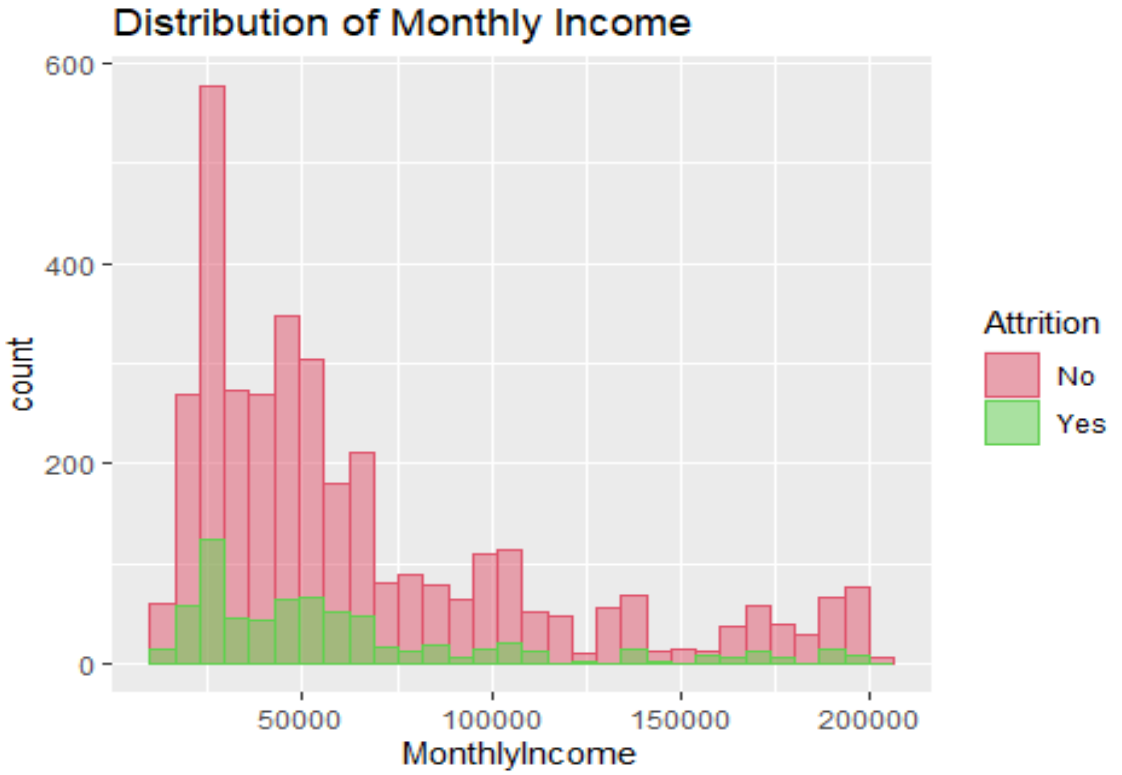



Avgworkhrs
0.40
0.35
0.30
0.25
No
Yes
Attrition
No
Yes


Log of Monthly Income
12
11
10
No
Yes
Attrition
No
Yes

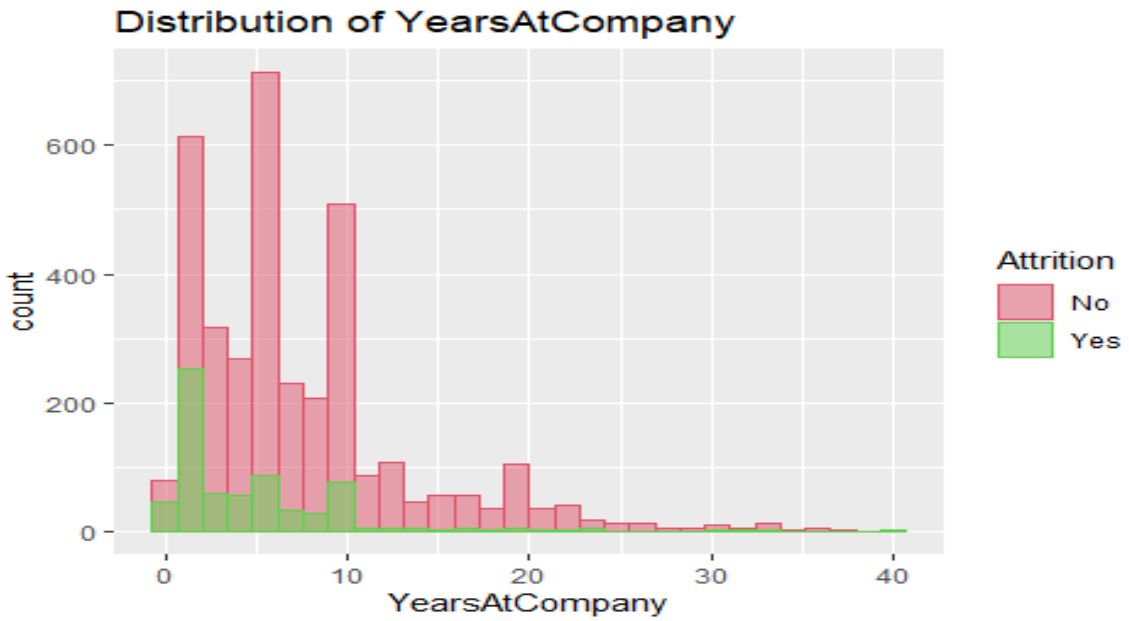
15) Monthly Income and Attrition:

Employees who left had a lower monthly income.



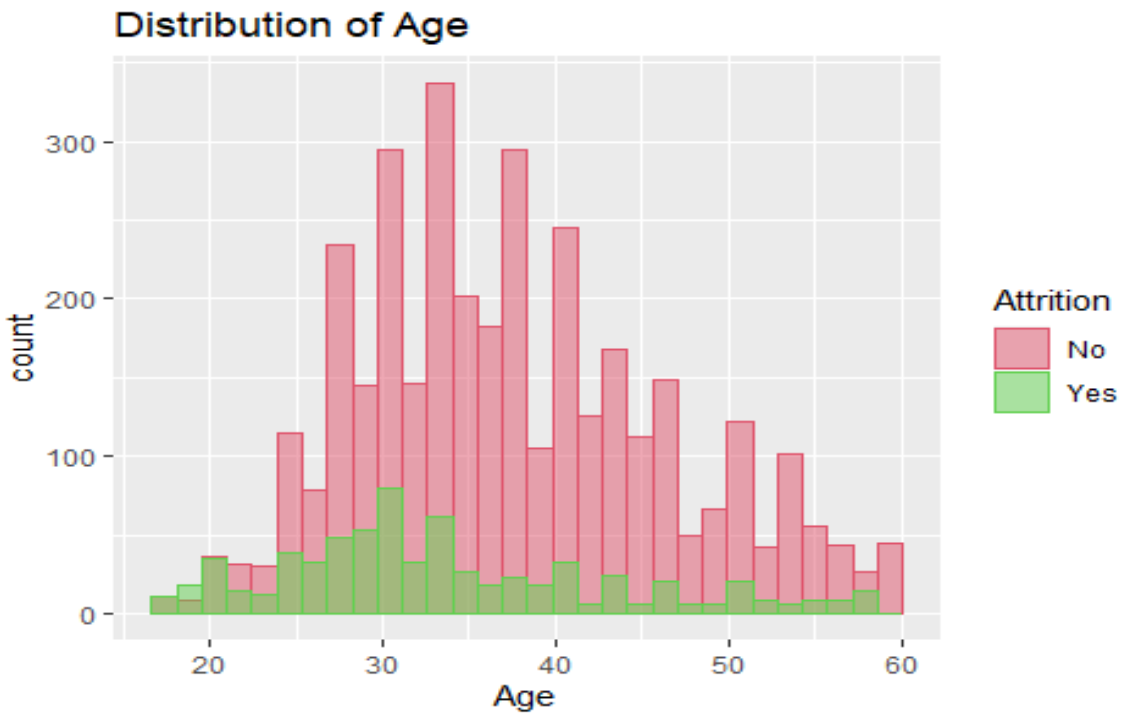
16) Years at company and Attrition:

Employees who left have spent less years working at the company.



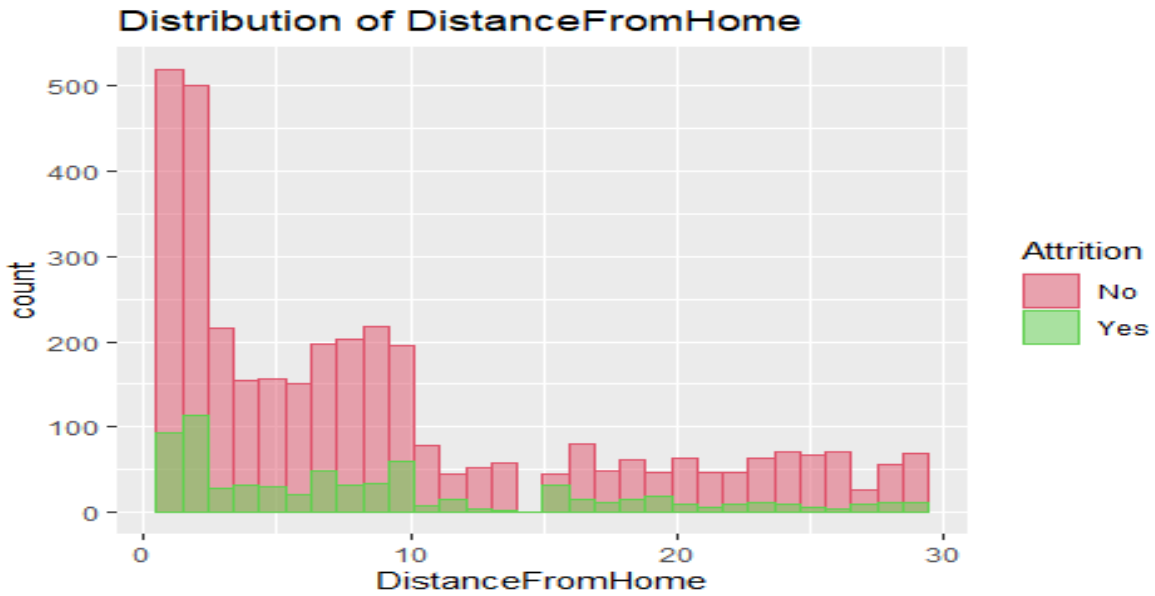
17) Age and Attrition:

Age distributions between genders are balanced.



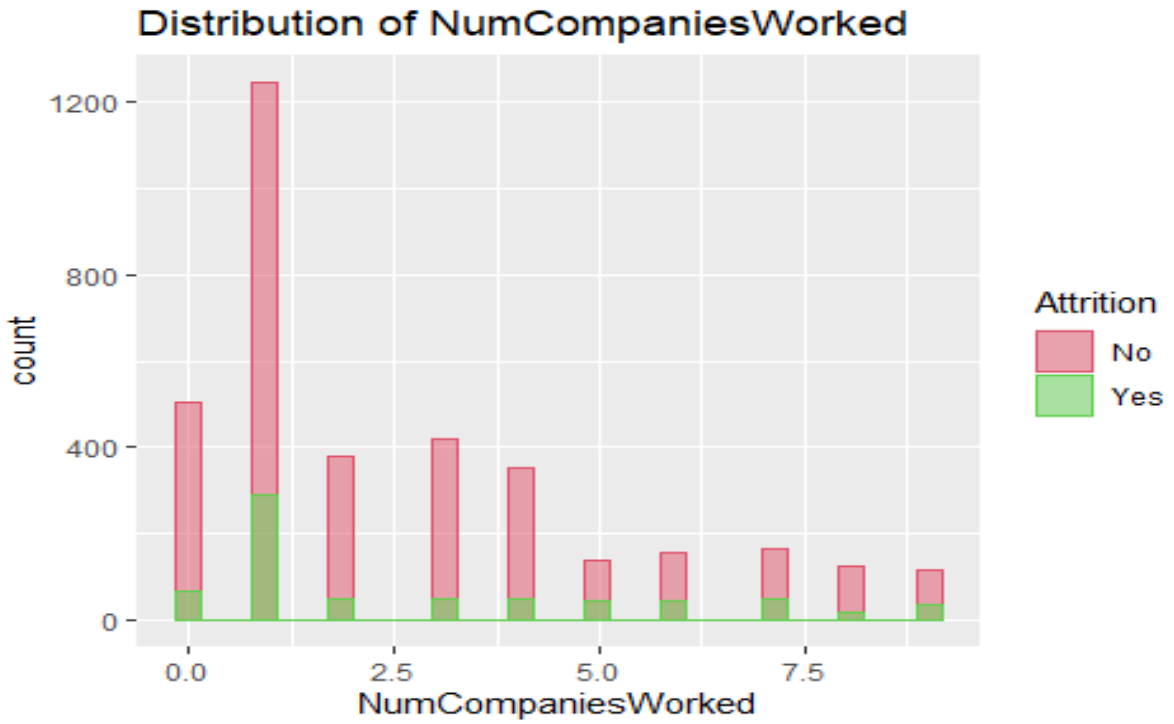
18) Distance from home and Attrition:

Employees with nearby home from office have left more in comparison with far home distance.



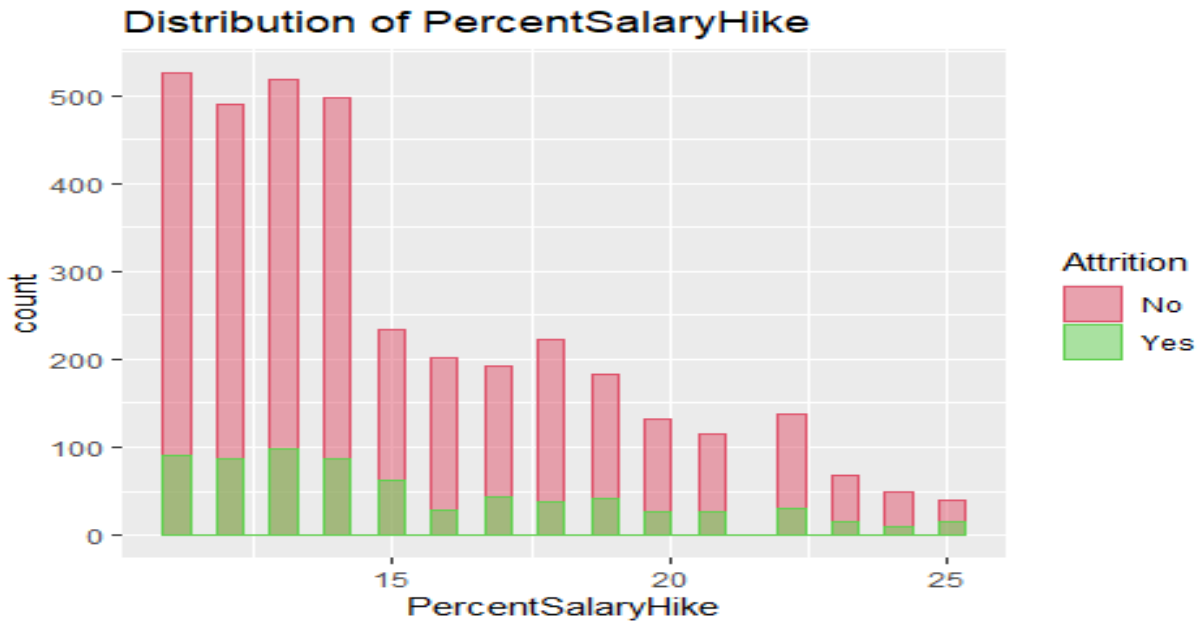
19) No. Of Companies worked and Attrition:

Employees with stability in company worked for have left more .



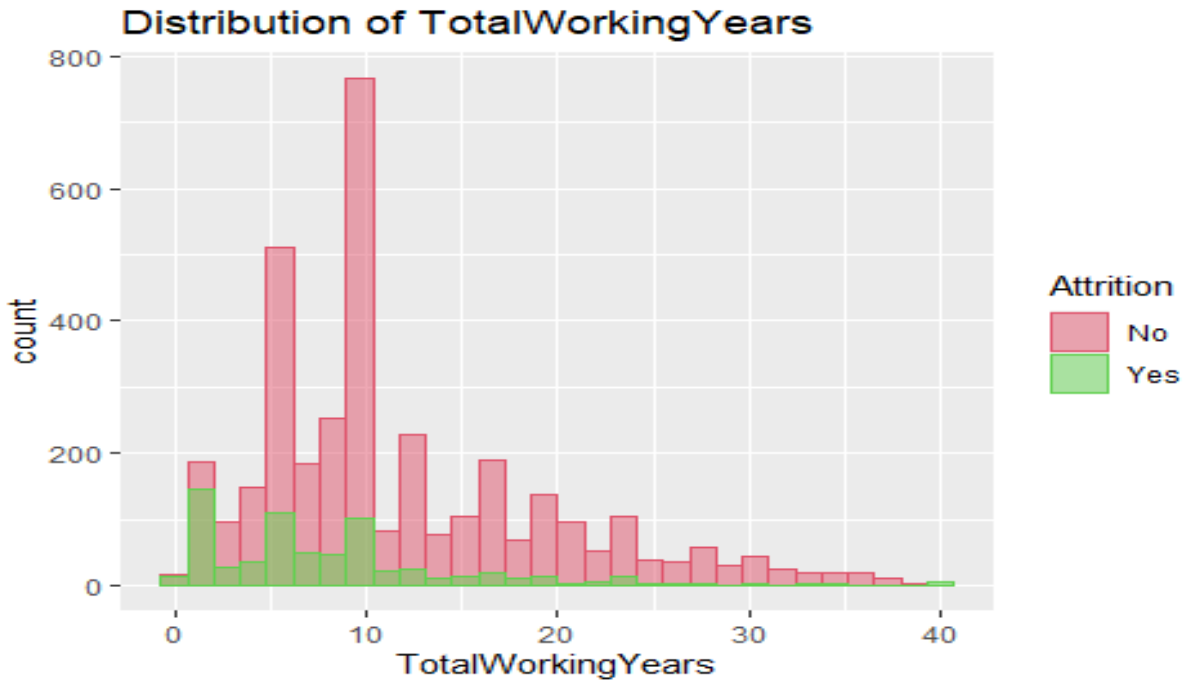
20) Salary hike and Attrition:

We observed that less hike in salary leads to employee turnover more.



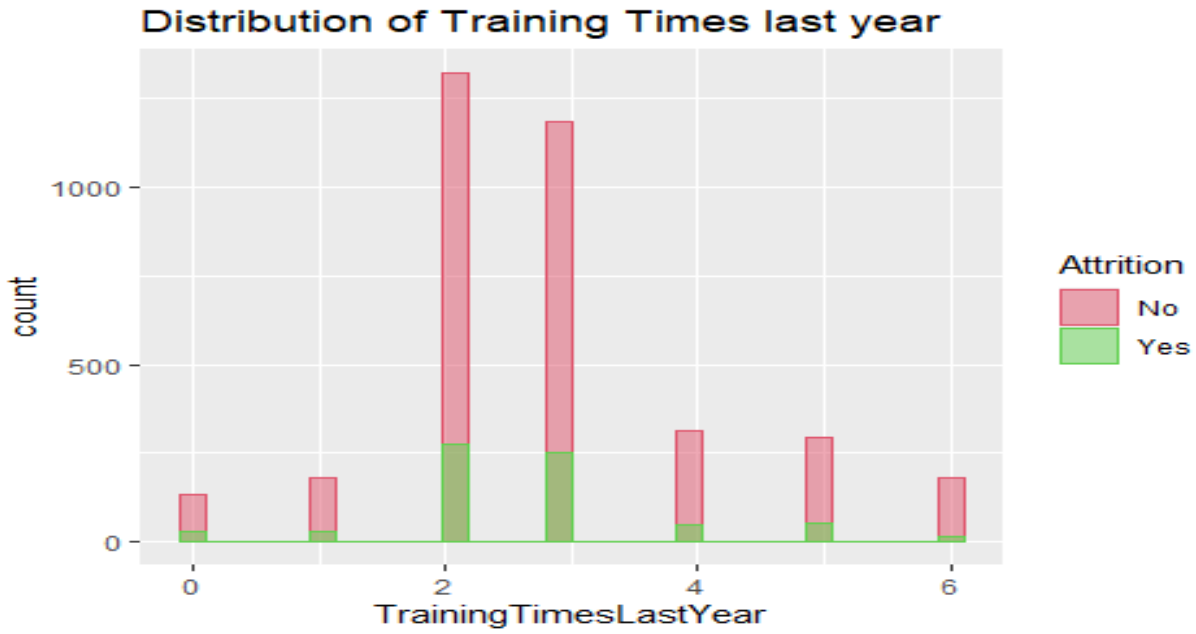
21) Total working years and Attrition:

Employees with less years of work experience have left more as compared to experienced employees.



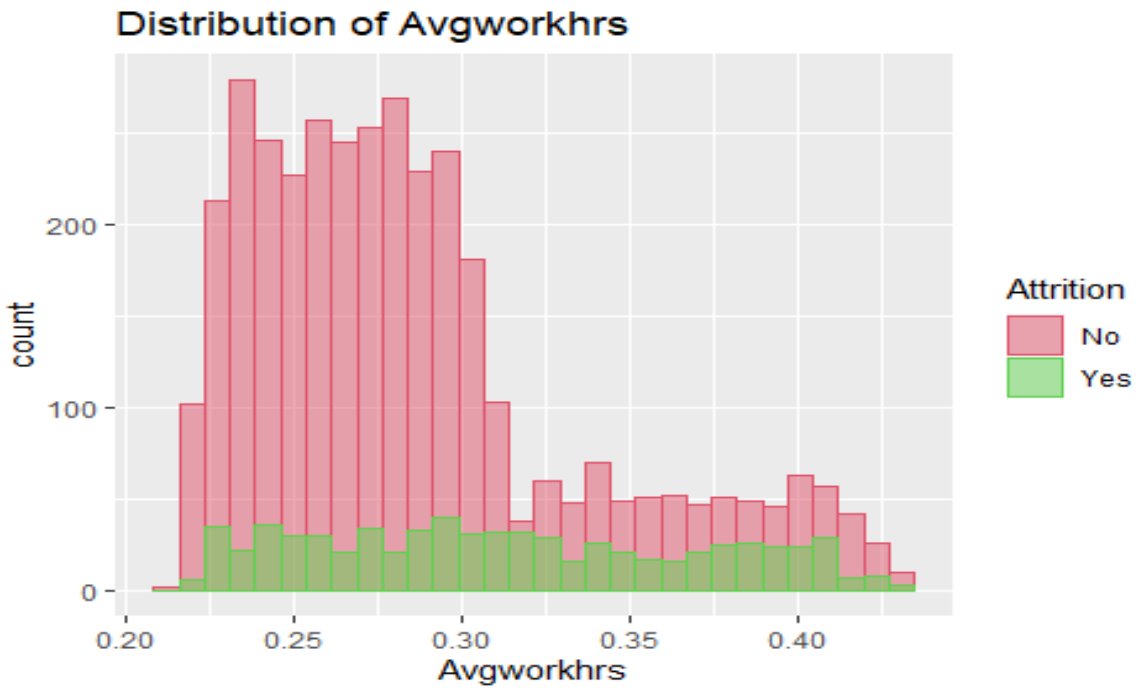
22) No. of trainings last year and Attrition:

Employees with 2-3 years of training last year have left more as others.



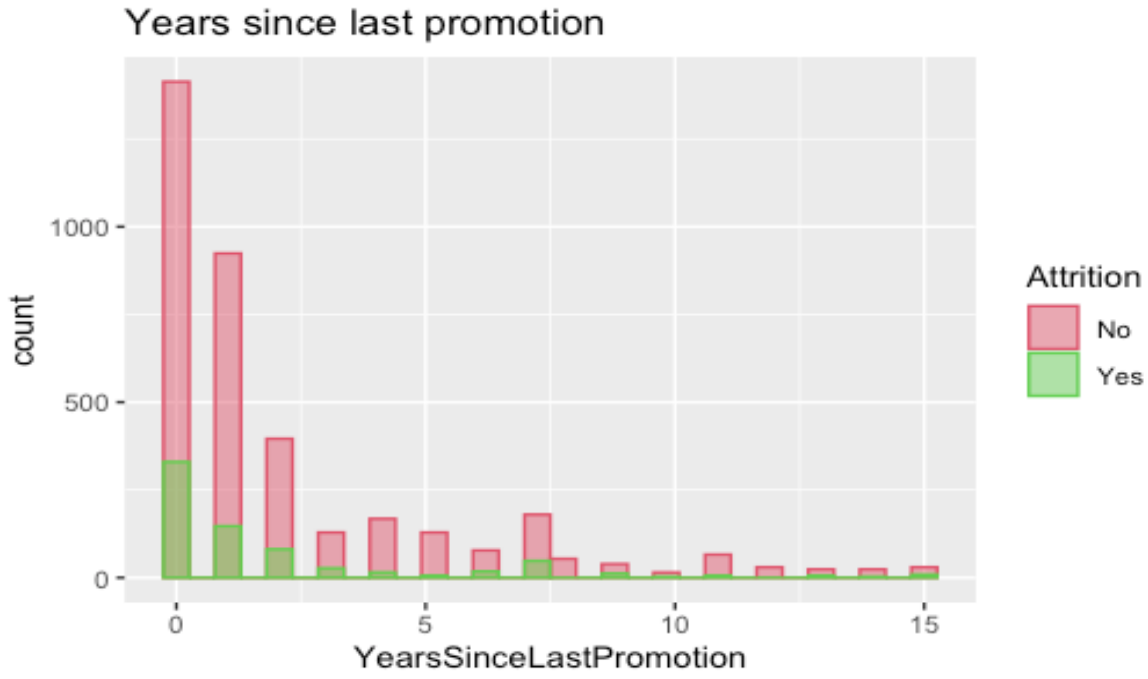
23) Average working hours {out time - in time} and Attrition:

There attrition rate is equally distributed through average working hours.



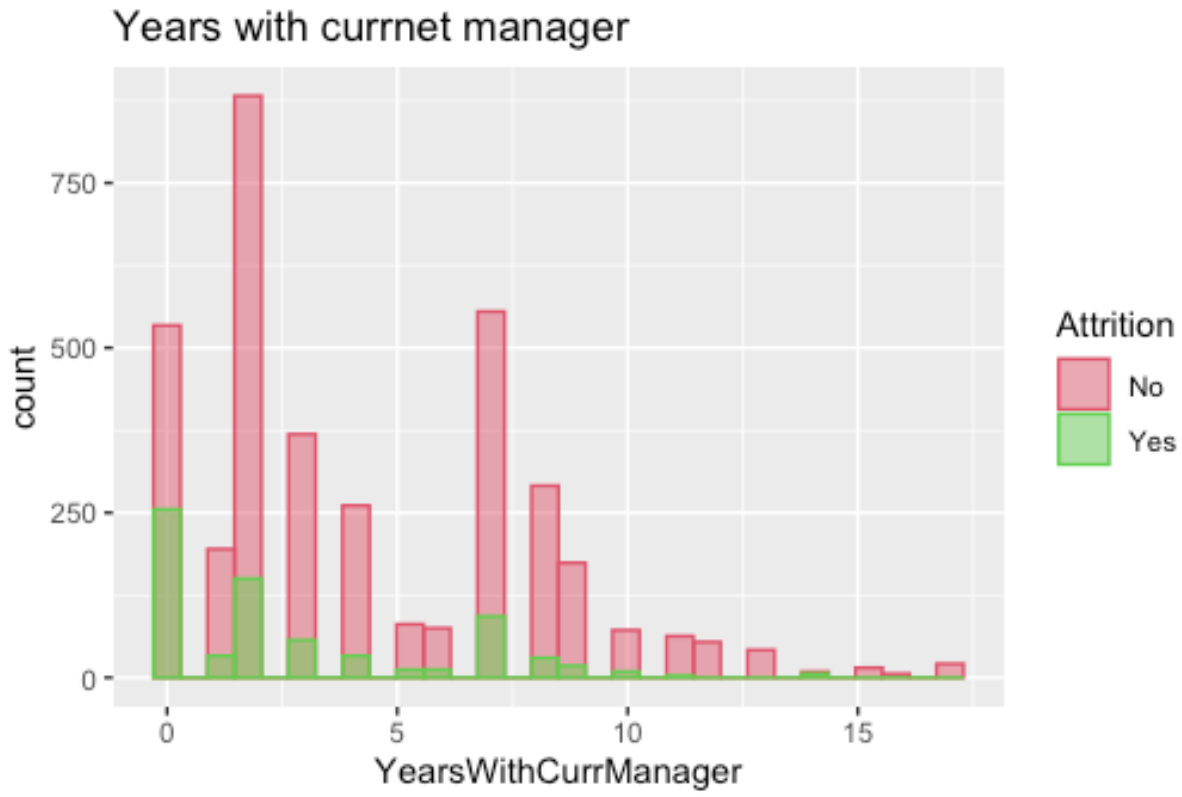
24) Years since last promotion and Attrition:

Attrition rate is high for employees that have got promotion in last 0-3 years.



24) Years since last promotion and Attrition:

Employees with 0-5 years with current manager have attrition rate.



Initial Conclusions:

1. Job quits rate is 15 percent.
2. Quits rate between genders is balanced.
3. Resign ratio is seriously high at human resources than others.
4. Research directors have a surprisingly high resigning ratio.
5. Singles are inclined to quit.
6. Employees with little experienced inclined to quit.
7. Frequently job changing employees inclined to quit.
8. Employees with low environment and job satisfaction are inclined to quit.

9. Employees with low work-life balance are inclined to quit.

10. We can see that attrition comes most from people with college education, perhaps they could be interns.

Chi-Square Test for Feature Selection:

To decide which categorical variables should be kept in the attrition model, Chi-Square will be used to test whether there is a relationship between the categorical variables and attrition. The null hypothesis for this test is the two variables are independent, and the alternative hypothesis is the variables are not independent. In order to reject the null hypothesis and keep variables in the model, the p-value of this test must have a p-value below 0.05.

COMMENT-The variables we will leave out of the model are education, gender, job level, performance rating and stock option level. These variables all have a p-value above .05 so they are independent from attrition.

```
a=chisq.test(final_data$BusinessTravel,final_data$Attrition)
b=chisq.test(final_data$Department,final_data$Attrition)
c=chisq.test(final_data$Education,final_data$Attrition)
d=chisq.test(final_data$EducationField,final_data$Attrition)
e=chisq.test(final_data$Gender,final_data$Attrition)
f=chisq.test(final_data$JobLevel,final_data$Attrition)
g=chisq.test(final_data$JobRole,final_data$Attrition)
h=chisq.test(final_data$MaritalStatus,final_data$Attrition)
i=chisq.test(final_data$WorkLifeBalance,final_data$Attrition)
j=chisq.test(final_data$JobSatisfaction,final_data$Attrition)
k=chisq.test(final_data$EnvironmentSatisfaction,final_data$Attrition)
l=chisq.test(final_data$PerformanceRating,final_data$Attrition)
m=chisq.test(final_data$JobInvolvement,final_data$Attrition)
n=chisq.test(final_data$StockOptionLevel,final_data$Attrition)

Stat_Sig = c(a$p.value<=0.05,b$p.value<=0.05,c$p.value<=0.05,
             d$p.value<=0.05,e$p.value<=0.05,f$p.value<=0.05,
             g$p.value<=0.05,h$p.value<=0.05,
             i$p.value<=0.05,j$p.value<=0.05,k$p.value<=0.05,
             l$p.value<=0.05,m$p.value<=0.05,n$p.value<=0.05)
Stat_Sig
##  [1]  TRUE  TRUE FALSE  TRUE FALSE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE FAL
SE
## [13]  TRUE FALSE
```

ANOVA for Feature Selection:

To decide if any numerical variables should be kept in the attrition model, we are using ANOVA to test if there is a significant difference between attrition and the numerical variables. The null hypothesis of ANOVA is there is no difference between means. The alternative hypothesis states there is a difference. If the p-value of the ANOVA is greater than .05, we can reject the null hypothesis and keep the variables in the model.

COMMENT-Distance from home is statistically insignificant, so it will be removed in the attrition model.

BORUTA for Feature Selection:

Boruta is a feature selection algorithm. Precisely, it works as a wrapper algorithm around Random Forest. Feature selection is a crucial step in predictive modelling . This technique achieves supreme importance when a data set comprised of several variables is given for model building.

Boruta can be your algorithm of choice to deal with such data sets. Particularly when one is interested in understanding the mechanisms related to the variable of interest, rather than just building a black box predictive model with good prediction accuracy.

How does it work?

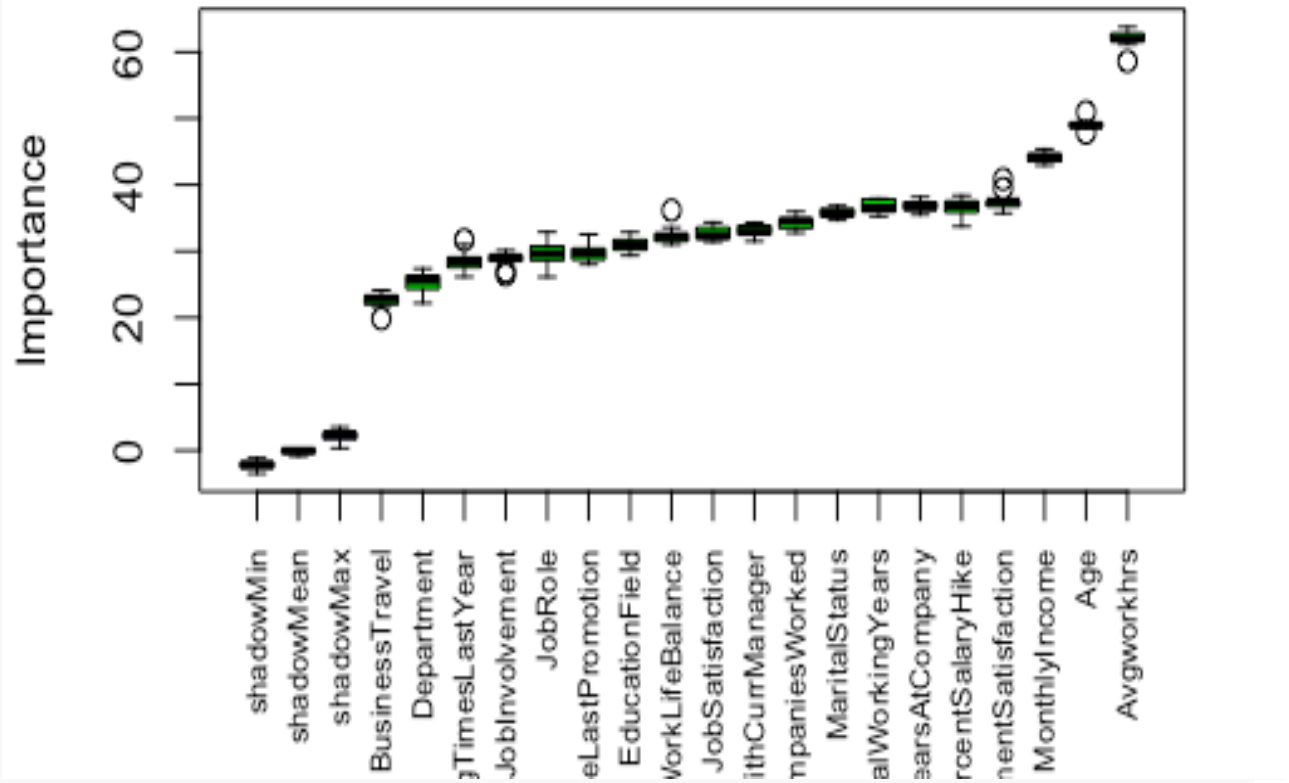
Below is the step wise working of boruta algorithm:

Firstly, it adds randomness to the given data set by creating shuffled copies of all features (which are called shadow features). Then, it trains a random forest classifier on the extended data set and applies a feature importance measure (the default is Mean Decrease Accuracy) to evaluate the importance of each feature where higher means more important.

At every iteration, it checks whether a real feature has a higher importance than the best of its shadow features (i.e. whether the feature has a higher Z score than the maximum Z score of its shadow features) and constantly removes features which are deemed highly unimportant.

Finally, the algorithm stops either when all features gets confirmed or rejected or it reaches a specified limit of random forest runs.

```
print(boruta.train)

##Boruta performed 11 iterations in 8.86976 secs.
 19 attributes confirmed important: Age, Avgworkhrs, BusinessTravel,
Department, EducationField and 14 more;
 No attributes deemed unimportant.
```




```
## Final Data after removing insignificant categorical variables ,Numerical
and count Variables:
Final_data=final_data[,-c(5,6,8,9,22,15)]
dim(Final_data)
## [1] 4300   20
```

Splitting Into Test and Train Data

Creating a test and a train dataset to evaluate and test the model.

Model building:

We want to understand the most important factors that lead to employee attrition. For this we use logistic regression to uncover which factors are the most relevant. We build three models: Logistic regression, decision trees and random forest and compare their results and Select the best model.

1) **FITTING LOGISTIC REGRESSION MODEL:**

By looking at the summary of logistic regression model we found out that out of 20 variables out of 26 are the most important .Accuracy with 50% threshold is 83.72% . We'll again use a better threshold based on ROC curve plotted below.

1}fitting logistic regression model-

```
model_1 = glm(Attrition ~ ., data = Train_data, family = "binomial")
summary(model_1)
##
## Call:
## glm(formula = Attrition ~ ., family = "binomial", data = Train_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7075  -0.5670  -0.3451  -0.1603   3.6901
##
## Coefficients:
##                                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)                     2.967e-04  7.056e-01   0.000 0.999664
## Age                            -3.197e-02  8.440e-03  -3.788 0.000152 **
*
## BusinessTravelTravel_Frequently  1.391e+00  2.420e-01   5.748 9.03e-09 **
*
## BusinessTravelTravel_Rarely      7.007e-01  2.249e-01   3.116 0.001835 **
## DepartmentResearch & Development -6.864e-01  3.095e-01  -2.218 0.026579 *
## DepartmentSales                 -6.013e-01  3.232e-01  -1.860 0.062844 .
## EducationFieldLife Sciences     -8.336e-01  4.320e-01  -1.930 0.053620 .
## EducationFieldMarketing         -1.065e+00  4.720e-01  -2.255 0.024118 *
## EducationFieldMedical           -8.441e-01  4.309e-01  -1.959 0.050111 .
## EducationFieldOther             -1.209e+00  4.859e-01  -2.489 0.012813 *
## EducationFieldTechnical Degree  -1.018e+00  4.590e-01  -2.219 0.026512 *
## JobRoleHuman Resources          -1.334e-01  3.566e-01  -0.374 0.708418
## JobRoleLaboratory Technician     8.214e-02  2.250e-01   0.365 0.715017
## JobRoleManager                  -3.882e-01  2.822e-01  -1.375 0.169022
## JobRoleManufacturing Director   -5.458e-01  2.617e-01  -2.086 0.036997 *
## JobRoleResearch Director         6.573e-01  2.733e-01   2.405 0.016162 *
## JobRoleResearch Scientist        9.754e-02  2.165e-01   0.451 0.652336
## JobRoleSales Executive           3.531e-01  2.121e-01   1.665 0.096003 .
## JobRoleSales Representative     -1.287e-01  2.887e-01  -0.446 0.655737
## MaritalStatusMarried             3.772e-01  1.620e-01   2.328 0.019931 *
## MaritalStatusSingle              1.247e+00  1.617e-01   7.709 1.27e-14 **
*
## MonthlyIncome                   -1.751e-07  1.142e-06  -0.153 0.878220
## NumCompaniesWorked               1.551e-01  2.291e-02   6.772 1.27e-11 **
*
## PercentSalaryHike                1.284e-02  1.456e-02   0.882 0.377795
## TotalWorkingYears               -8.135e-02  1.511e-02  -5.383 7.33e-08 **
*
## TrainingTimesLastYear           -1.740e-01  4.317e-02  -4.030 5.59e-05 **
*
## YearsAtCompany                   3.206e-02  2.205e-02   1.454 0.145968
## YearsSinceLastPromotion          1.789e-01  2.539e-02   7.046 1.85e-12 **
*
```

```
## YearsWithCurrManager              -1.770e-01  2.799e-02  -6.324 2.55e-10 **
*
## JobInvolvement                    -9.511e-02  7.392e-02  -1.287 0.198190
## EnvironmentSatisfaction           -3.732e-01  4.892e-02  -7.629 2.37e-14 **
*
## JobSatisfaction                   -3.319e-01  4.893e-02  -6.784 1.17e-11 **
*
## WorkLifeBalance                   -3.674e-01  7.388e-02  -4.973 6.60e-07 **
*
## Avgworkhrs                         1.085e+01  9.760e-01  11.112  < 2e-16 **
*
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2996.9  on 3439  degrees of freedom
## Residual deviance: 2335.3  on 3406  degrees of freedom
## AIC: 2403.3
##
## Number of Fisher Scoring iterations: 6
```

predicted probabilities of Attrition for test data

```
test_pred = predict(model_1, type = "response",
                    newdata = Test_data)
```

Let's use the probability cutoff of 50%.

```
test_pred_attrition =factor(ifelse(test_pred >= 0.50, "Yes", "No"))
test_actual_attrition = factor(ifelse(Test_data$Attrition==1,"Yes","No"))

test_conf = confusionMatrix(test_pred_attrition, test_actual_attrition, posit
ive = "Yes")
test_conf
```

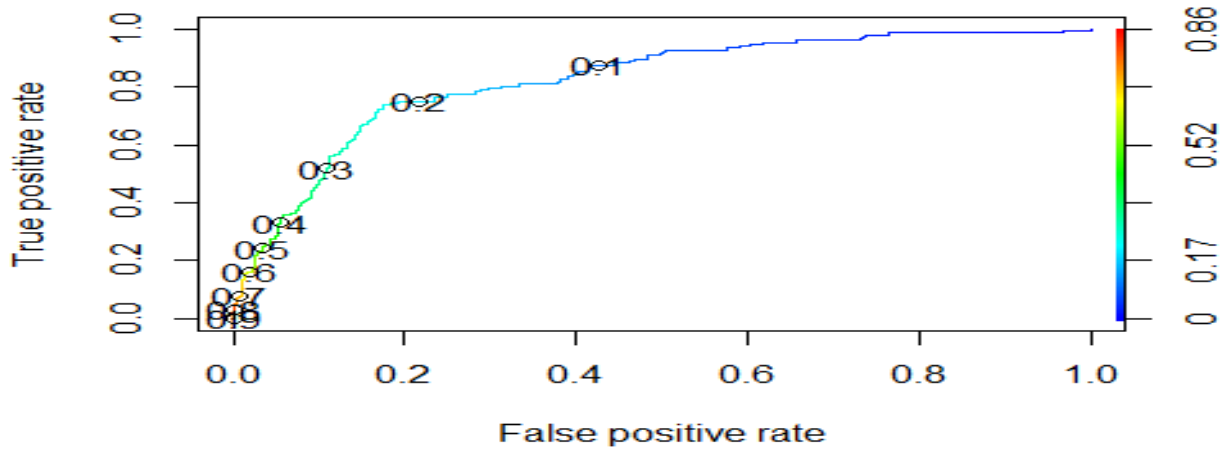
```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  No Yes
##        No  683 116
##        Yes  24  37
##
##                Accuracy : 0.8372
##                  95% CI : (0.8108, 0.8613)
##     No Information Rate : 0.8221
##     P-Value [Acc > NIR] : 0.132
##
```

```
##                   Kappa : 0.272
##
##  Mcnemar's Test P-Value : 1.461e-14
##
##             Sensitivity : 0.24183
##             Specificity : 0.96605
##          Pos Pred Value : 0.60656
##          Neg Pred Value : 0.85482
##              Prevalence : 0.17791
##          Detection Rate : 0.04302
##    Detection Prevalence : 0.07093
##       Balanced Accuracy : 0.60394
##
##        'Positive' Class : Yes
```

As we see, the results in our test dataset gives us accuracy of 84%.Futhermore , if we take a look at Sensitivity and Specificity we have values of 24% and 97% respectively. So if we want to predict people who might leave (Prediction = Yes and Reference = Yes) this is not a good model, since will only detect 24% of them .

Selecting a different threshold: Right now our threshold is 0.5, we could reduce it to increase the sensitivity value.

```
 Selecting a different threshold based on ROC curve

ROCR_prediction = prediction(test_pred, Test_data$Attrition)
ROCR_performance=performance(ROCR_prediction, 'tpr', 'fpr')

plot(ROCR_performance, colorize = TRUE, print.cutoffs.at = seq(0.1, by = 0.1)
)
```



Based on this we select the new threshold to be 0.2 and test our model again.

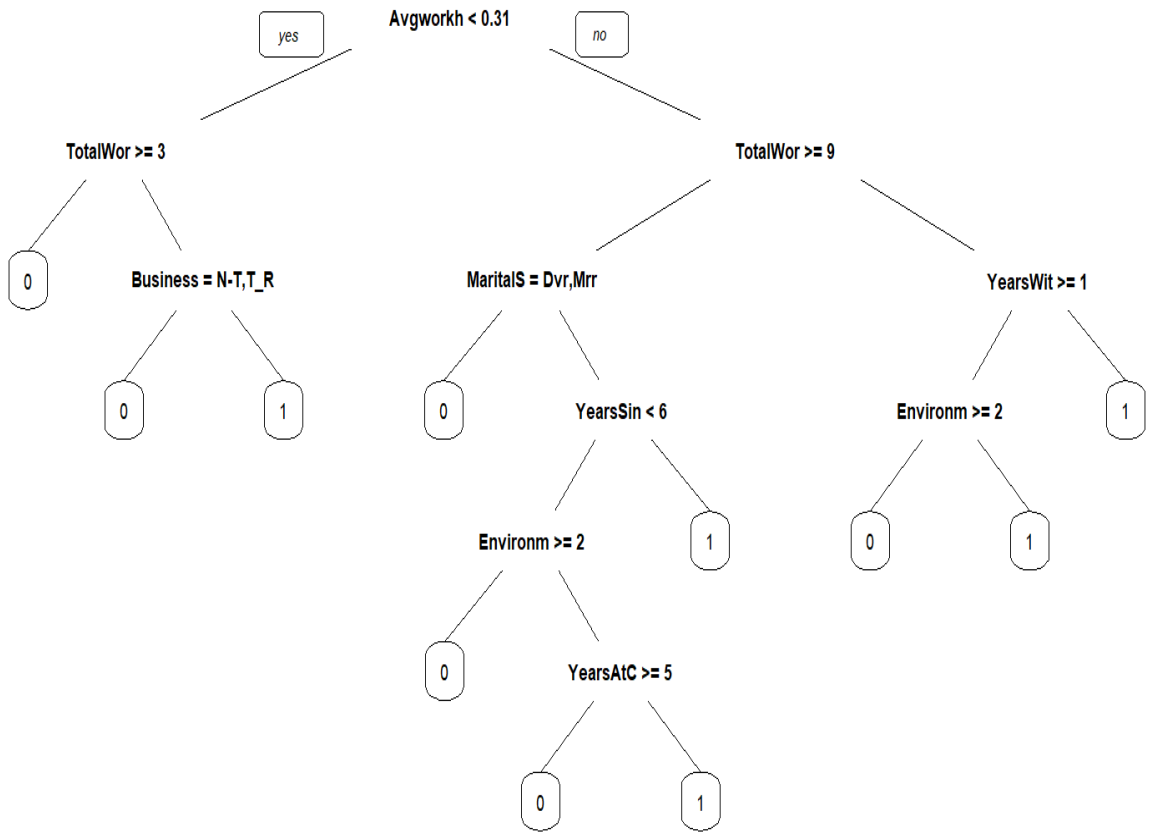
```
test_pred_attrition =factor(ifelse(test_pred >= 0.20, "Yes", "No"))
test_actual_attrition = factor(ifelse(Test_data$Attrition==1,"Yes","No"))

test_conf = confusionMatrix(test_pred_attrition, test_actual_attrition, posit
ive = "Yes")
test_conf
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  No Yes
##        No  554  38
##        Yes 153 115
##
##                Accuracy : 0.7779
##                  95% CI : (0.7486, 0.8053)
##     No Information Rate : 0.8221
##     P-Value [Acc > NIR] : 0.9996
##
##                   Kappa : 0.4135
##
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 0.7516
##             Specificity : 0.7836
##          Pos Pred Value : 0.4291
##          Neg Pred Value : 0.9358
##              Prevalence : 0.1779
##          Detection Rate : 0.1337
##    Detection Prevalence : 0.3116
##       Balanced Accuracy : 0.7676
##
##        'Positive' Class : Yes
##
```

Comment - From the 50% threshold criteria sensitivity was very low, hence ,we'll use 20% threshold based on the above ROC curve; this will provide us with 77.79% accuracy plus 75% sensitivity and 79% specificity. So now our model is predicting better the people who might leave the company.

2)**FITTING DECISION TREES:**

For the second model we create a decision tree to see if our results can improve.

```
2) Fitting Decision trees:

 Decision tree model

model.tree =rpart(formula = Attrition ~., data = Train_data, method = 'class'
)

Plotting decision tree
rpart.plot::prp(model.tree)
```



```
pred.tree=predict(model.tree,Test_data,type="class")

table(pred.tree,Test_data$Attrition)
##
## pred.tree   0   1
##         0 695 128
##         1  12  25
confusionMatrix(pred.tree,as.factor(Test_data$Attrition))

## Confusion Matrix and Statistics

##           Reference
## Prediction   0   1
##          0 695 128
##          1  12  25
##
##                Accuracy : 0.8372
##                  95% CI : (0.8108, 0.8613)
##     No Information Rate : 0.8221
##     P-Value [Acc > NIR] : 0.132
##
##                   Kappa : 0.2083
##
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 0.9830
##             Specificity : 0.1634
##          Pos Pred Value : 0.8445
##          Neg Pred Value : 0.6757
##              Prevalence : 0.8221
##          Detection Rate : 0.8081
##    Detection Prevalence : 0.9570
##       Balanced Accuracy : 0.5732
##
##        'Positive' Class : 0
```

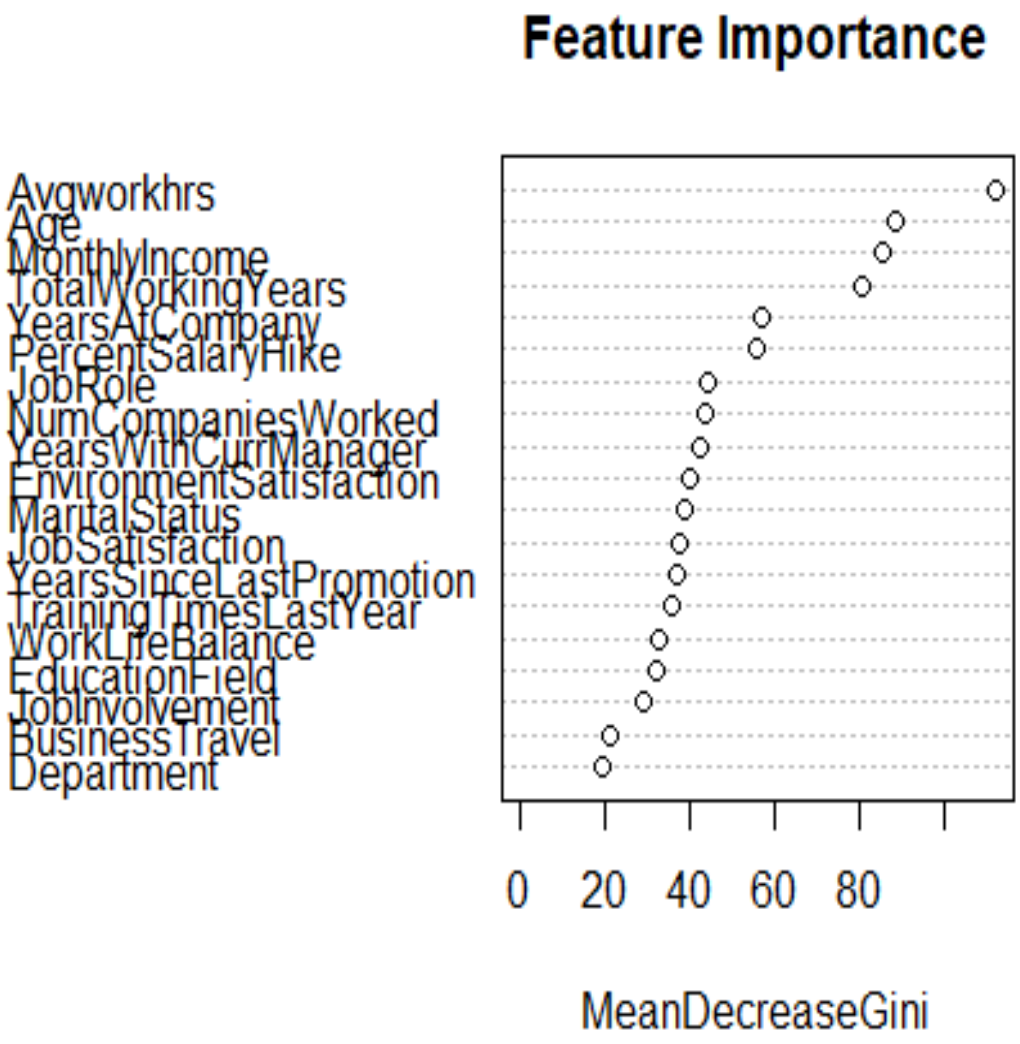
COMMENT- Accuracy given by decision trees is 83.72%

3) **FITTING RANDOM FOREST MODEL :**

```
3)Fitting Random Forest model:


library(randomForest)
```

```r
modelrf = randomForest(as.factor(Attrition) ~., data = Train_data)

# Printing the model

modelrf

## Call:
##  randomForest(formula = as.factor(Attrition) ~ ., data = Train_data)
##                Type of random forest: classification
##                      Number of trees: 500
## No. of variables tried at each split: 4
##
##         OOB estimate of  error rate: 1.16%
## Confusion matrix:
##      0   1  class.error
## 0 2884   1 0.0003466205
## 1   39 516 0.0702702703
predrf=predict(modelrf,Test_data)
confusionMatrix(predrf,as.factor(Test_data$Attrition))
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0 715   4
##          1   5 136
##
##                Accuracy : 0.9895
##                  95% CI : (0.9802, 0.9952)
##     No Information Rate : 0.8372
##     P-Value [Acc > NIR] : <2e-16
##
##                   Kappa : 0.9617
##
##  Mcnemar's Test P-Value : 1
##
##             Sensitivity : 0.9931
##             Specificity : 0.9714
##          Pos Pred Value : 0.9944
##          Neg Pred Value : 0.9645
##              Prevalence : 0.8372
##          Detection Rate : 0.8314
##    Detection Prevalence : 0.8360
##       Balanced Accuracy : 0.9822
##
##        'Positive' Class : 0
```

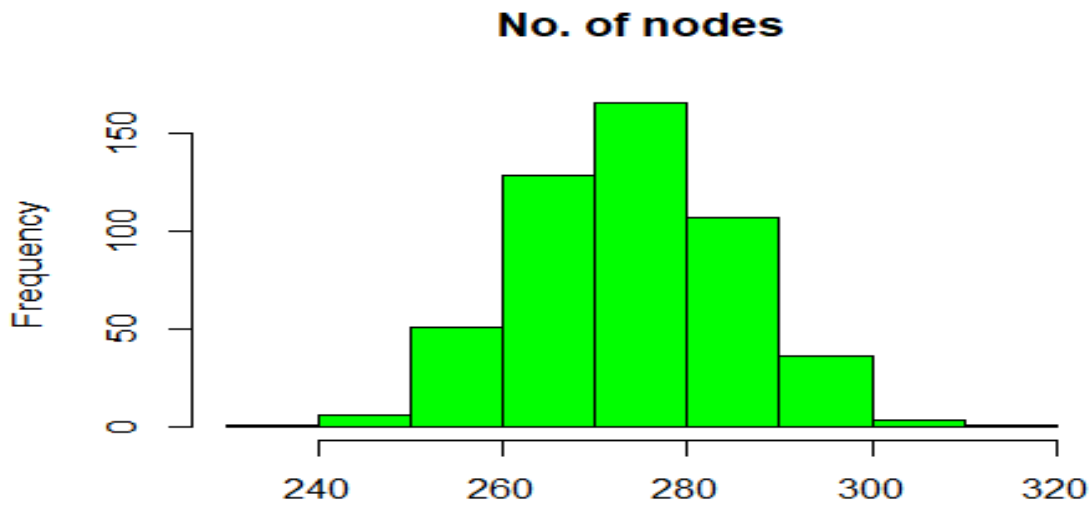
```r
# Plotting features based on importance

randomForest::varImpPlot(modelrf, main = 'Feature Importance')
```



COMMENT- From this plot, we observed that **top 5 features are-**

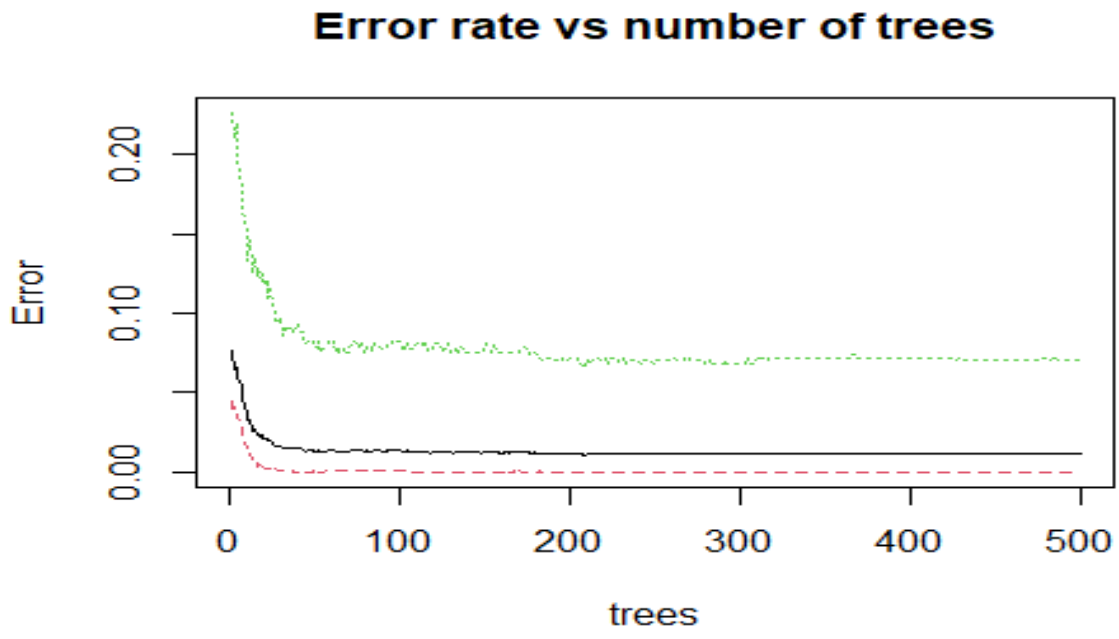
1. Average working hours
2. Age
3. Monthly Income
4. Total Working Years
5. Years at company

```
hist(treesize(modelrf), main = 'No. of nodes', xlab = '', col = 'green')
```



We also can take a look at the number of nodes by each tree.

```
# Plotting error rate
plot(modelrf, main = 'Error rate vs number of trees')#Based on this 100 trees
should be fine.
```



Based on this 100 trees should be fine.

COMMENT-

1.Random Forest is doing a pretty good job predicting attrition reaching a level of accuracy of 98%

2.The number of trees for this model is set by default to be 500.

3.We might want to decrease the number of trees based on the error rate. If we decrease the number of trees the time in computing the model should decrease as well.

CONCLUSIONS:

1) At the first part of the project, we have made a cleaning . We have put our target variable into the centre and made an exploration of data around of it. In that part, we have had some conclusions about features. In the Initial Conclusions section, we have expressed our conclusions clearly. After building models, we have seen that many of conclusions are covering our feature importance results which are obtained from random forest algorithm.

2) Based on the results of the random forest analysis we can conclude that the top 5 features are: **Average working hours, Age, Years at the company, Monthly Income and Total working Hours.** Those features could be important for the manager in order to reduce the level of attrition among employees.

3) For this particular dataset, we conclude that the best model to use in order to predict attrition is random forest. We also modify our random forest model by reducing the number of trees needed to construct the forest from 500 (default) to 100 based on the error rate for each tree without compromising our results.

4) The BORUTA PLOT clearly represents the factors which serve as the top reasons for attrition in a company: **Average working hours, Age, Monthly Income ,Employment satisfaction and Percent salary hike.**

5) Employees generally left when they are **underworked / overworked(Average working hours)**

6) Employee satisfaction plays a major role for employee turnover.
7) In the stacked bar charts, we saw employees who left were:

- In human resources department
- Less monthly income
- Worked over time
- Had low job satisfaction
- Had low environment satisfaction
- Had bad work life balance

8) Chi-square results revealed gender, education, stock option level and performance rating did not have a significant role in employee attrition. Similarly, ANOVA test revealed that Distance From Home also doesn't have a significant role in employee attrition.