EmployeeAttritionIBM

load packages

Hide

```
library(tidyverse)
```

```
 [30m--  [1mAttaching packages [22m --------------------------------------- t
idyverse 1.2.1 -- [39m

 [30m [32mv [30m  [34mggplot2 [30m 3.1.0      [32mv [30m  [34mpurrr   [30m 0.
2.5

 [32mv [30m  [34mtibble  [30m 1.4.2      [32mv [30m  [34mdplyr   [30m 0.7.8

 [32mv [30m  [34mtidyr   [30m 0.8.2      [32mv [30m  [34mstringr [30m 1.3.1

 [32mv [30m  [34mreadr   [30m 1.1.1      [32mv [30m  [34mforcats [30m 0.3.0 [
39m

 [30m--  [1mConflicts [22m ------------------------------------------ tidyver
se_conflicts() --

 [31mx [30m  [34mdplyr [30m:: [32mfilter() [30m masks  [34mstats [30m::filter
()

 [31mx [30m  [34mdplyr [30m:: [32mlag() [30m    masks  [34mstats [30m::lag()
[39m
```

Hide

```
library(caret)
```

```
Loading required package: lattice


Attaching package: <U+393C><U+3E31>caret<U+393C><U+3E32>


The following object is masked from <U+393C><U+3E31>package:purrr<U+393C><U+3
E32>:


    lift
```

Hide

```
library(rpart)
library(knitr) #Dynamic Report Generator including use of LateX, HTML
library(gridExtra)
```

```

Attaching package: <U+393C><U+3E31>gridExtra<U+393C><U+3E32>
```

```
The following object is masked from <U+393C><U+3E31>package:dplyr<U+393C><U+3
E32>:

    combine
```

Hide

```
library(corrplot)
```

```
corrplot 0.84 loaded
```

Hide

```
library(Boruta) #Feature selection
```

```
Loading required package: ranger
```

Hide

```
library(randomForest) #Random forest
```

```
randomForest 4.6-14

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


Attaching package: <U+393C><U+3E31>randomForest<U+393C><U+3E32>


The following object is masked from <U+393C><U+3E31>package:ranger<U+393C><U+
3E32>:

    importance


The following object is masked from <U+393C><U+3E31>package:gridExtra<U+393C>
<U+3E32>:

    combine


The following object is masked from <U+393C><U+3E31>package:dplyr<U+393C><U+3
E32>:

    combine


The following object is masked from <U+393C><U+3E31>package:ggplot2<U+393C><U
+3E32>:
```

```
    margin
```

Hide

```
library(ggRandomForests) #variable importance random forest
```

```
Loading required package: randomForestSRC


 randomForestSRC 2.7.0


 Type rfsrc.news() to see new features, changes, and bug fixes.





Attaching package: <U+393C><U+3E31>randomForestSRC<U+393C><U+3E32>


The following object is masked from <U+393C><U+3E31>package:purrr<U+393C><U+3
E32>:


    partial




Attaching package: <U+393C><U+3E31>ggRandomForests<U+393C><U+3E32>


The following object is masked from <U+393C><U+3E31>package:randomForestSRC<U
+393C><U+3E32>:


    partial.rfsrc
```

Hide

```
library(DMwR) #BINARY CLASSIFICATION
```

```
Loading required package: grid
```

Hide

```
library(pROC) #ROC PLOT
```

```
Type 'citation("pROC")' for a citation.


Attaching package: <U+393C><U+3E31>pROC<U+393C><U+3E32>
```

```

The following objects are masked from <U+393C><U+3E31>package:stats<U+393C><U
+3E32>:

    cov, smooth, var
```

Hide

```
library(shinydashboard)
```

```

Attaching package: <U+393C><U+3E31>shinydashboard<U+393C><U+3E32>

The following object is masked from <U+393C><U+3E31>package:graphics<U+393C><
U+3E32>:

    box
```

Hide

```
library(shiny)
library(readxl)
library(plotly)
```

```

Attaching package: <U+393C><U+3E31>plotly<U+393C><U+3E32>

The following object is masked from <U+393C><U+3E31>package:ggplot2<U+393C><U
+3E32>:

    last_plot

The following object is masked from <U+393C><U+3E31>package:stats<U+393C><U+3
E32>:

    filter

The following object is masked from <U+393C><U+3E31>package:graphics<U+393C><
U+3E32>:

    layout
```

Hide

```
library(ROCR)
```

```
Loading required package: gplots


Attaching package: <U+393C><U+3E31>gplots<U+393C><U+3E32>


The following object is masked from <U+393C><U+3E31>package:stats<U+393C><U+3
E32>:


    lowess
```

Hide

```
library(xgboost)
```

```


Attaching package: <U+393C><U+3E31>xgboost<U+393C><U+3E32>


The following object is masked from <U+393C><U+3E31>package:plotly<U+393C><U+
3E32>:


    slice


The following object is masked from <U+393C><U+3E31>package:dplyr<U+393C><U+3
E32>:


    slice
```

Import and read data

Hide

```
HR_Employee_Attrition_data <- read_excel("HR-Employee-Attrition-data.xlsx")
hr_data <- as.data.frame(HR_Employee_Attrition_data)
```

Summary of the data

Hide

```
head(hr_data)
```

	Employee Count <dbl>	Employee ID <dbl>	Department <chr>	Job Role <chr>
1	1	1	Sales	Sales Executive
2	1	2	Research & Development	Research Scientist
3	1	4	Research & Development	Laboratory Technician
4	1	5	Research & Development	Research Scientist
5	1	7	Research & Development	Laboratory Technician
6	1	8	Research & Development	Laboratory Technician

6 rows | 1-6 of 35 columns

Hide

```
summary(hr_data)
```

```
 Employee Count  Employee ID      Department          Job Role          Attrit
ion (Yes/No)
 Min.   :1      Min.   :   1.0   Length:1470        Length:1470        Length
:1470
 1st Qu.:1      1st Qu.: 491.2   Class :character   Class :character   Class
:character
 Median :1      Median :1020.5   Mode  :character   Mode  :character   Mode
:character
 Mean   :1      Mean   :1024.9
 3rd Qu.:1      3rd Qu.:1555.8
 Max.   :1      Max.   :2068.0
    Gender               Age           Over 18           Marital Status      Ed
ucation
 Length:1470        Min.   :18.00   Length:1470        Length:1470        Len
gth:1470
 Class :character   1st Qu.:30.00   Class :character   Class :character   Cla
ss :character
 Mode  :character   Median :36.00   Mode  :character   Mode  :character   Mod
e  :character
                    Mean   :36.92
                    3rd Qu.:43.00
                    Max.   :60.00
 Education Field    Business Travel    Distance From Home (kms) Job Involveme
nt
 Length:1470        Length:1470        Min.   : 1.000           Length:1470
```

```
 Class :character   Class :character   1st Qu.: 2.000           Class :charac
ter
 Mode  :character   Mode  :character   Median : 7.000           Mode  :charac
ter
                                       Mean   : 9.193
                                       3rd Qu.:14.000
                                       Max.   :29.000
   Job Level     Job Satisfaction   Hourly Rate (USD) Daily Rate (USD) Monthl
y Rate (USD)
 Min.   :1.000   Length:1470        Min.   : 30.00    Min.   : 102.0   Min.
: 2094
 1st Qu.:1.000   Class :character   1st Qu.: 48.00    1st Qu.: 465.0   1st Qu
.: 8047
 Median :2.000   Mode  :character   Median : 66.00    Median : 802.0   Median
:14236
 Mean   :2.064                      Mean   : 65.89    Mean   : 802.5   Mean
:14313
 3rd Qu.:3.000                      3rd Qu.: 83.75    3rd Qu.:1157.0   3rd Qu
.:20462
 Max.   :5.000                      Max.   :100.00    Max.   :1499.0   Max.
:26999
 Monthly Income (USD) Salary Hike (%) Stock Option Level Standard Hours  Over
Time
 Min.   : 1009        Min.   :11.00   Min.   :0.0000     Min.   :80     Lengt
h:1470
 1st Qu.: 2911        1st Qu.:12.00   1st Qu.:0.0000     1st Qu.:80     Class
:character
 Median : 4919        Median :14.00   Median :1.0000     Median :80     Mode
:character
 Mean   : 6503        Mean   :15.21   Mean   :0.7939     Mean   :80
 3rd Qu.: 8379        3rd Qu.:18.00   3rd Qu.:1.0000     3rd Qu.:80
 Max.   :19999        Max.   :25.00   Max.   :3.0000     Max.   :80
 No. of Companies Worked Total Working Years Years At Company Years In Curren
t Role
 Min.   :0.000           Min.   : 0.00       Min.   : 0.000   Min.   : 0.000
 1st Qu.:1.000           1st Qu.: 6.00       1st Qu.: 3.000   1st Qu.: 2.000
 Median :2.000           Median :10.00       Median : 5.000   Median : 3.000
 Mean   :2.693           Mean   :11.28       Mean   : 7.008   Mean   : 4.229
 3rd Qu.:4.000           3rd Qu.:15.00       3rd Qu.: 9.000   3rd Qu.: 7.000
 Max.   :9.000           Max.   :40.00       Max.   :40.000   Max.   :18.000
```

```
 Years Since Last Promotion Years With Curr Manager Environment Satisfaction
 Min.   : 0.000            Min.   : 0.000          Length:1470
 1st Qu.: 0.000            1st Qu.: 2.000          Class :character
 Median : 1.000            Median : 3.000          Mode  :character
 Mean   : 2.188            Mean   : 4.123
 3rd Qu.: 3.000            3rd Qu.: 7.000
 Max.   :15.000            Max.   :17.000
 Training Times Last Year Work Life Balance  Performance Rating Relationship 
Satisfaction
 Min.   :0.000            Length:1470        Length:1470        Length:1470
 1st Qu.:2.000            Class :character   Class :character   Class :charac
ter
 Median :3.000            Mode  :character   Mode  :character   Mode  :charac
ter
 Mean   :2.799
 3rd Qu.:3.000
 Max.   :6.000
```

Hide

```
str(hr_data)
```

```
'data.frame':   1470 obs. of  35 variables:
 $ Employee Count            : num  1 1 1 1 1 1 1 1 1 1 ...
 $ Employee ID               : num  1 2 4 5 7 8 10 11 12 13 ...
 $ Department                : chr  "Sales" "Research & Development" "Researc
h & Development" "Research & Development" ...
 $ Job Role                  : chr  "Sales Executive" "Research Scientist" "L
aboratory Technician" "Research Scientist" ...
 $ Attrition (Yes/No)        : chr  "Yes" "No" "Yes" "No" ...
 $ Gender                    : chr  "Female" "Male" "Male" "Female" ...
 $ Age                       : num  41 49 37 33 27 32 59 30 38 36 ...
 $ Over 18                   : chr  "Y" "Y" "Y" "Y" ...
 $ Marital Status            : chr  "Single" "Married" "Single" "Married" ...
 $ Education                 : chr  "College" "Below College" "College" "Mast
er" ...
 $ Education Field           : chr  "Life Sciences" "Life Sciences" "Other" "
Life Sciences" ...
 $ Business Travel           : chr  "Travel_Rarely" "Travel_Frequently" "Trav
el_Rarely" "Travel_Frequently" ...
```

```
 $ Distance From Home (kms)   : num  1 8 2 3 2 2 3 24 23 27 ...
 $ Job Involvement            : chr  "High" "Medium" "Medium" "High" ...
 $ Job Level                  : num  2 2 1 1 1 1 1 1 3 2 ...
 $ Job Satisfaction           : chr  "Very High" "Medium" "High" "High" ...
 $ Hourly Rate (USD)          : num  94 61 92 56 40 79 81 67 44 94 ...
 $ Daily Rate (USD)           : num  1102 279 1373 1392 591 ...
 $ Monthly Rate (USD)         : num  19479 24907 2396 23159 16632 ...
 $ Monthly Income (USD)       : num  5993 5130 2090 2909 3468 ...
 $ Salary Hike (%)            : num  11 23 15 11 12 13 20 22 21 13 ...
 $ Stock Option Level         : num  0 1 0 0 1 0 3 1 0 2 ...
 $ Standard Hours             : num  80 80 80 80 80 80 80 80 80 80 ...
 $ Over Time                  : chr  "Yes" "No" "Yes" "Yes" ...
 $ No. of Companies Worked    : num  8 1 6 1 9 0 4 1 0 6 ...
 $ Total Working Years        : num  8 10 7 8 6 8 12 1 10 17 ...
 $ Years At Company           : num  6 10 0 8 2 7 1 1 9 7 ...
 $ Years In Current Role      : num  4 7 0 7 2 7 0 0 7 7 ...
 $ Years Since Last Promotion: num  0 1 0 3 2 3 0 0 1 7 ...
 $ Years With Curr Manager    : num  5 7 0 0 2 6 0 0 8 7 ...
 $ Environment Satisfaction   : chr  "Medium" "High" "Very High" "Very High" .
..
 $ Training Times Last Year   : num  0 3 3 3 3 2 3 2 2 3 ...
 $ Work Life Balance          : chr  "Bad" "Better" "Better" "Better" ...
 $ Performance Rating         : chr  "Excellent" "Outstanding" "Excellent" "Ex
cellent" ...
 $ Relationship Satisfaction : chr  "Low" "Very High" "Medium" "High" ...
```

Hide

```
sum(is.na(hr_data)) # check numbers of missing values
```

```
[1] 0
```

Looking at the dataset, there are too many variables and we might not need all. For example. We will exclude “Over 18”, “Employee Count”, “Standard Hours”. Those variables are not informative and there is not variance in these variables

Hide

```
hr_data = hr_data[,!(names(hr_data) %in% c('Over 18','Employee Count','Standa
rd Hours','Employee ID'))]

str(hr_data)
```

```
'data.frame':	1470 obs. of  31 variables:
 $ Department                 : chr  "Sales" "Research & Development" "Researc
h & Development" "Research & Development" ...
 $ Job Role                   : chr  "Sales Executive" "Research Scientist" "L
aboratory Technician" "Research Scientist" ...
 $ Attrition (Yes/No)         : chr  "Yes" "No" "Yes" "No" ...
 $ Gender                     : chr  "Female" "Male" "Male" "Female" ...
 $ Age                        : num  41 49 37 33 27 32 59 30 38 36 ...
 $ Marital Status             : chr  "Single" "Married" "Single" "Married" ...
 $ Education                  : chr  "College" "Below College" "College" "Mast
er" ...
 $ Education Field            : chr  "Life Sciences" "Life Sciences" "Other" "
Life Sciences" ...
 $ Business Travel            : chr  "Travel_Rarely" "Travel_Frequently" "Trav
el_Rarely" "Travel_Frequently" ...
 $ Distance From Home (kms)   : num  1 8 2 3 2 2 3 24 23 27 ...
 $ Job Involvement            : chr  "High" "Medium" "Medium" "High" ...
 $ Job Level                  : num  2 2 1 1 1 1 1 1 3 2 ...
 $ Job Satisfaction           : chr  "Very High" "Medium" "High" "High" ...
 $ Hourly Rate (USD)          : num  94 61 92 56 40 79 81 67 44 94 ...
 $ Daily Rate (USD)           : num  1102 279 1373 1392 591 ...
 $ Monthly Rate (USD)         : num  19479 24907 2396 23159 16632 ...
 $ Monthly Income (USD)       : num  5993 5130 2090 2909 3468 ...
 $ Salary Hike (%)            : num  11 23 15 11 12 13 20 22 21 13 ...
 $ Stock Option Level         : num  0 1 0 0 1 0 3 1 0 2 ...
 $ Over Time                  : chr  "Yes" "No" "Yes" "Yes" ...
 $ No. of Companies Worked    : num  8 1 6 1 9 0 4 1 0 6 ...
 $ Total Working Years        : num  8 10 7 8 6 8 12 1 10 17 ...
 $ Years At Company           : num  6 10 0 8 2 7 1 1 9 7 ...
 $ Years In Current Role      : num  4 7 0 7 2 7 0 0 7 7 ...
 $ Years Since Last Promotion: num  0 1 0 3 2 3 0 0 1 7 ...
 $ Years With Curr Manager    : num  5 7 0 0 2 6 0 0 8 7 ...
 $ Environment Satisfaction   : chr  "Medium" "High" "Very High" "Very High" .
..
 $ Training Times Last Year   : num  0 3 3 3 3 2 3 2 2 3 ...
 $ Work Life Balance          : chr  "Bad" "Better" "Better" "Better" ...
```

```
 $ Performance Rating         : chr  "Excellent" "Outstanding" "Excellent" "Ex
cellent" ...
 $ Relationship Satisfaction : chr  "Low" "Very High" "Medium" "High" ...
```

Checking the attrition percentage

Hide

```
Attrition_ppl <- nrow(hr_data[hr_data$`Attrition (Yes/No)` == 'Yes',])
no_Attrition <- nrow(hr_data[hr_data$`Attrition (Yes/No)` == 'No',])
str(Attrition_ppl)
```

```
 int 237
```

Hide

```
hr_data$Attrition <- hr_data$`Attrition (Yes/No)`
(prop.table(table(hr_data$Attrition))*100)
```

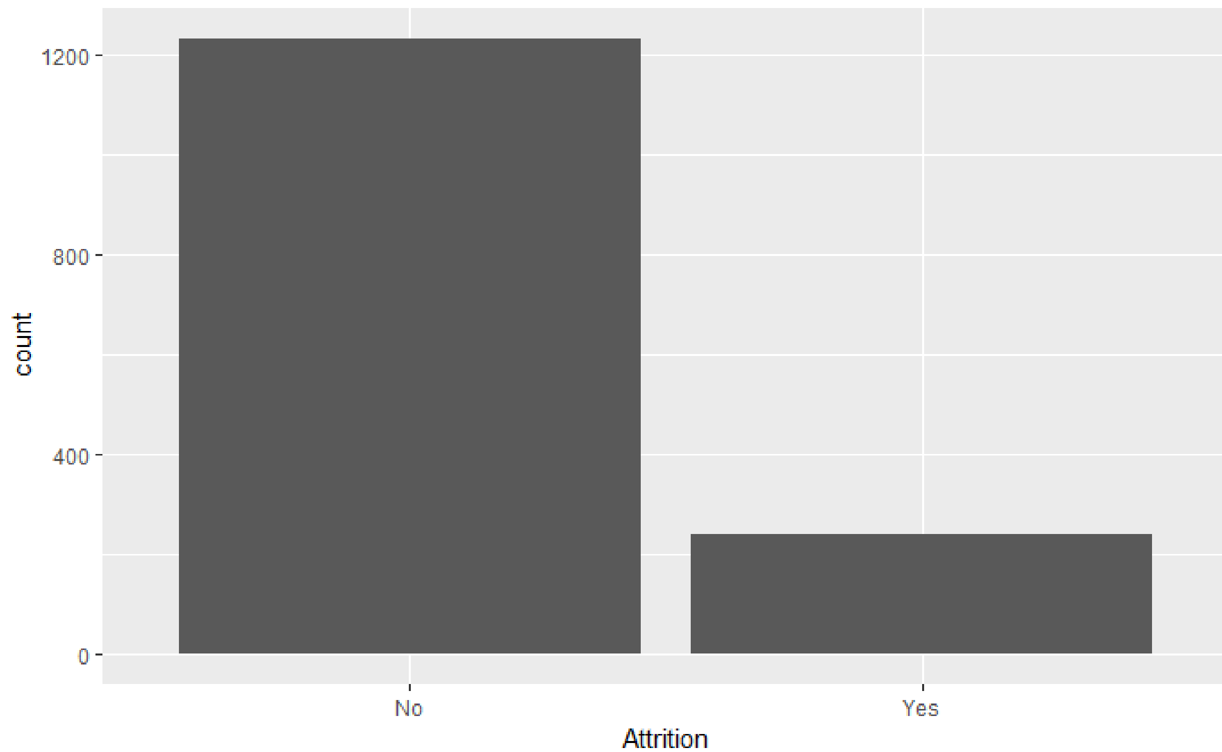
```

      No      Yes 
83.87755 16.12245 
```

Proceeding for Data Visualizing and Feature Extraction . Visualizing the different features will help to determine the features that might be important for our prediction.

Checking the attirition percentage of the IBM organisation

Hide

```
hr_data$Attrition <- hr_data$`Attrition (Yes/No)`
ggplot(hr_data, aes(Attrition)) + geom_bar()
```



In 1470 obervations of 31 variables, we see that about 84% of the population stayed at the organization and about 16% of the population left

Deparment and Attrition

Visualizing the Department of the employee ad the Relationship to attrition

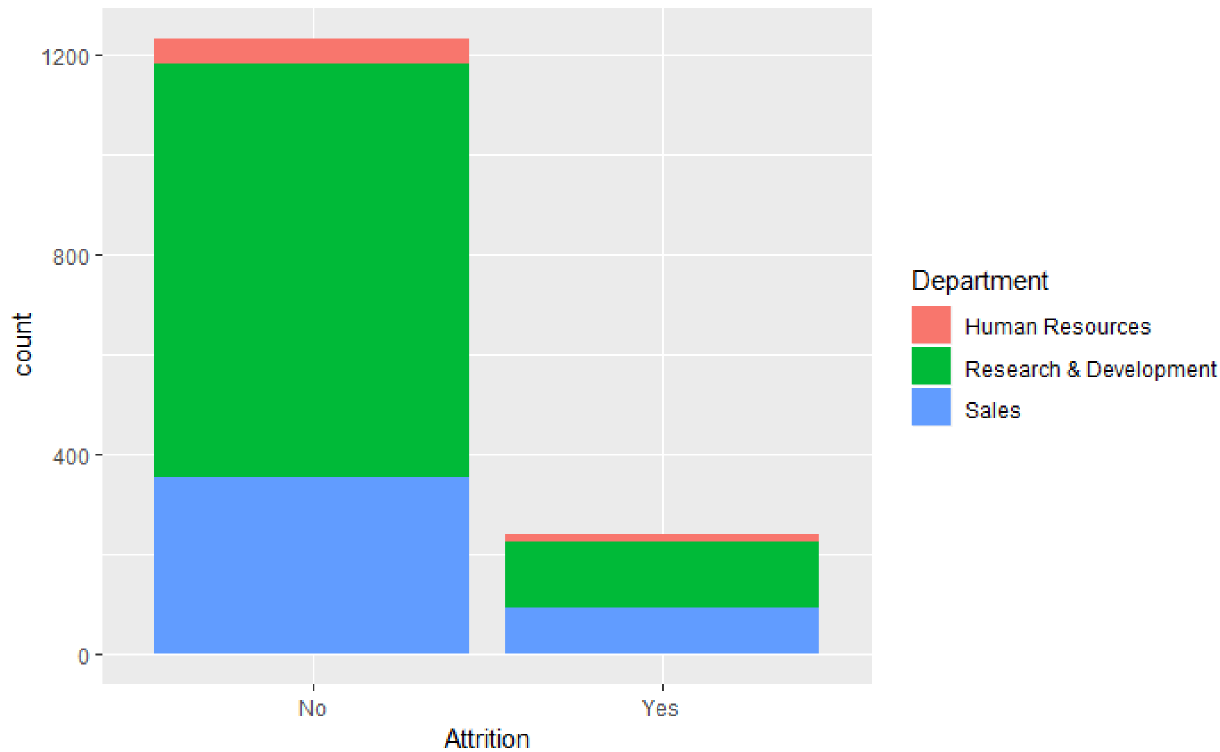
Hide

```
table(hr_data$Department)
```

```

       Human Resources Research & Development                  Sales 
                    63                    961                    446 
```

Hide

```
ggplot(hr_data, aes(Attrition, fill = Department)) + geom_bar()
```



Hide

```
# Most of the employees are from the Research and Development department
```

Hide

```
Dep_att <- hr_data %>%group_by(Department)%>%summarize(attrition_rate=mean(At
trition=="Yes"))%>% ggplot(aes(x=Department,y=attrition_rate,fill=Department)
) + geom_bar(stat='identity', alpha = 0.5)

Dep_att
```

Hide

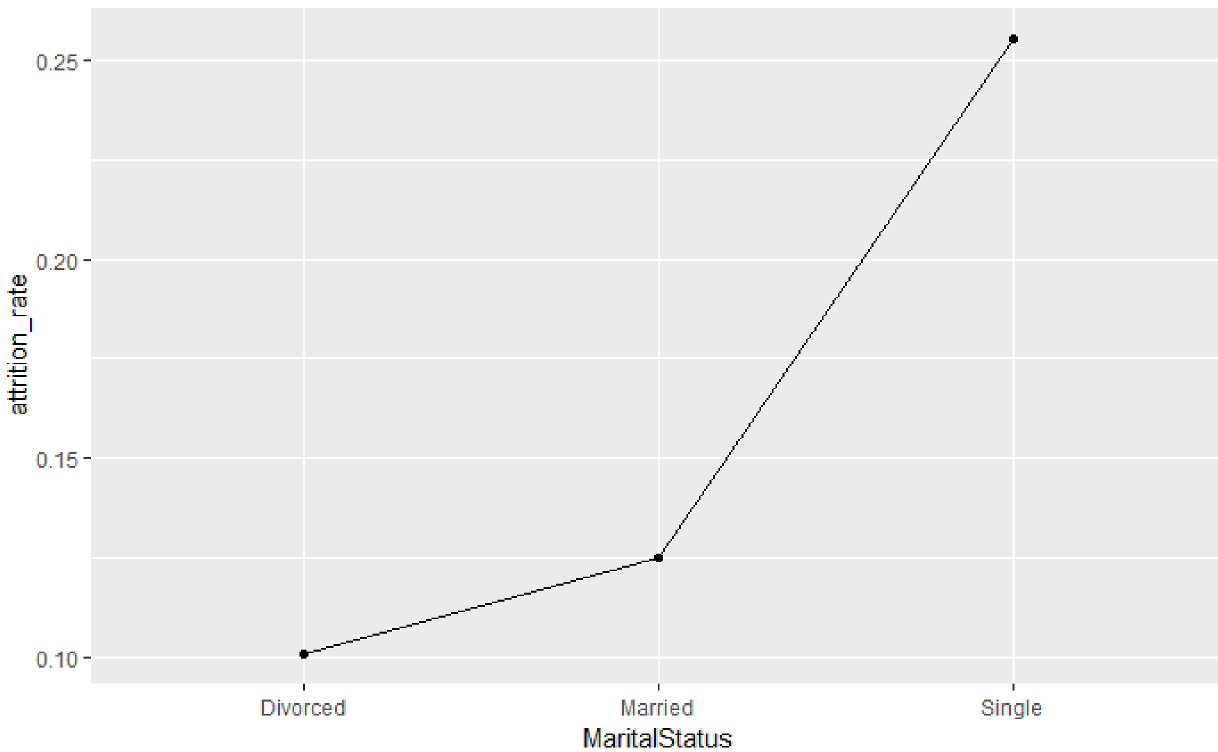
```
ggplot(hr_data, aes(Attrition, group=Department)) +
  geom_bar(aes(y = ..prop.., fill= factor(..x..))) +
  labs(y="Percentage", fill = "Attrition") +
  facet_grid(~Department)
```

It is evident that from the visualized bar chart that Sales has a higher attriton rate.

Marital status and Attrition

Hide

```
hr_data$MaritalStatus = hr_data$`Marital Status`

mar_status <-hr_data %>% group_by(MaritalStatus)%>%

  summarize(attrition_rate=mean(Attrition=="Yes"))%>%

  ggplot(aes(x=MaritalStatus,y=attrition_rate,group=2)) + geom_line(stat='ide
ntity') +geom_point()

mar_status
```

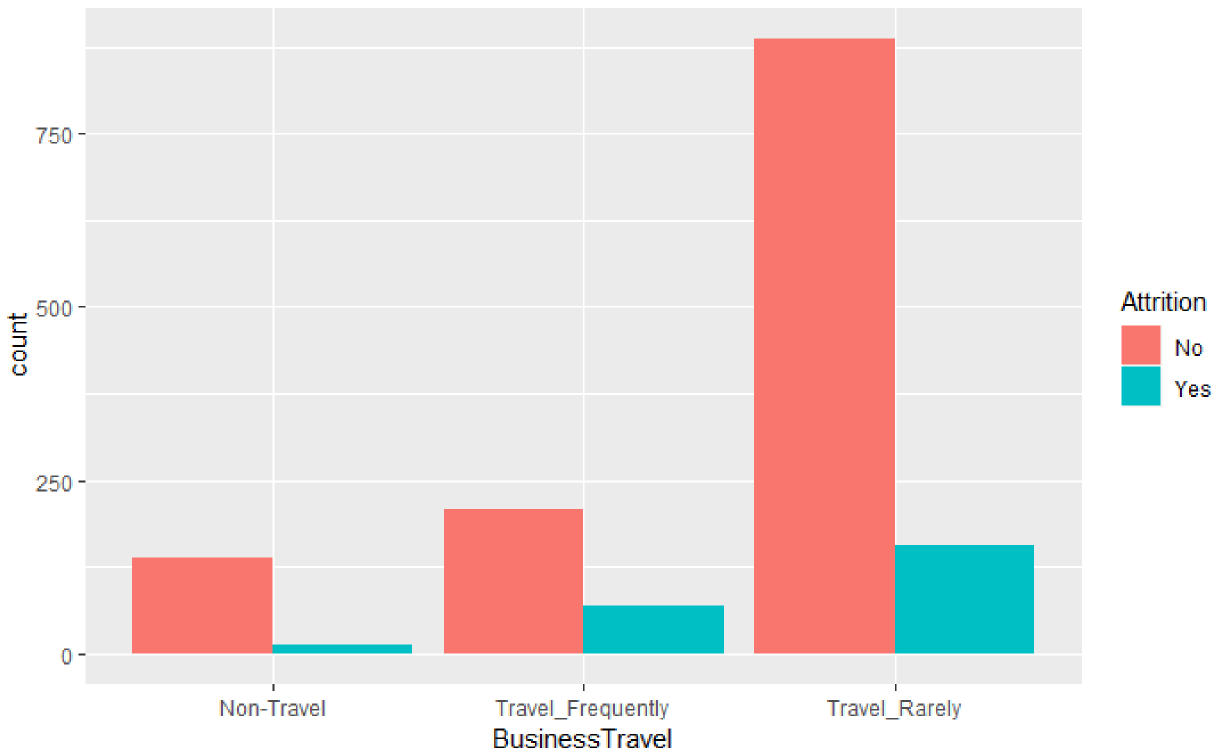


Attrition rate was far more for people who were single than married. Large people as compared to single persons might not necessarily leave the company. The marital status might be a weak predictor of attrition in this case.

Attrition and Business Travel

Hide

```
hr_data$BusinessTravel <- hr_data$`Business Travel`

ggplot(hr_data, aes(BusinessTravel, fill = Attrition)) + geom_bar(stat= "count", position =  position_dodge())
```

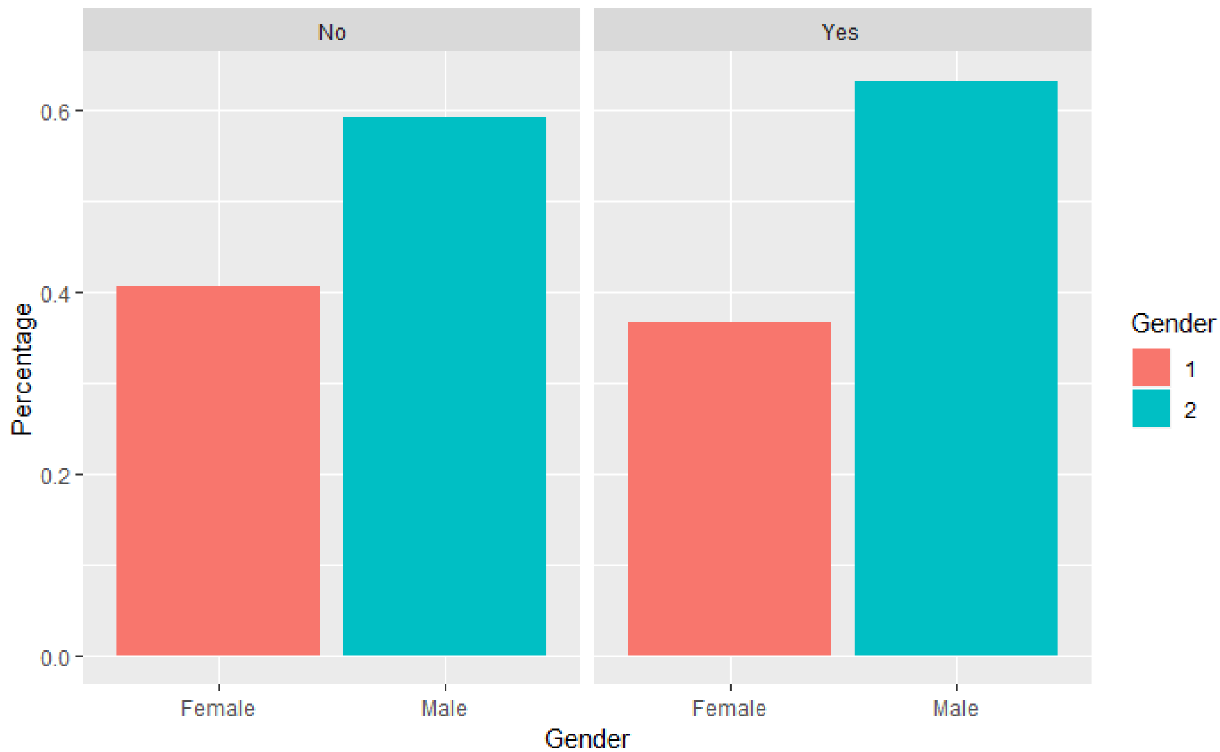


We observe that there are more people in the organization who travels rarely as compared to those who travel frequently. It also appears that those who travel rarely might have a likelihood of staying in the organization, however the Business Travel Variable does not appear to be a significant predictor of attrition rate.

Attrition and Gender

Hide

```
ggplot(hr_data, aes(Gender, group= Attrition)) +
  geom_bar(aes(y = ..prop.., fill= factor(..x..)), stat = "count") +
  labs(y="Percentage", fill = "Gender") +
  facet_grid(~Attrition)
```

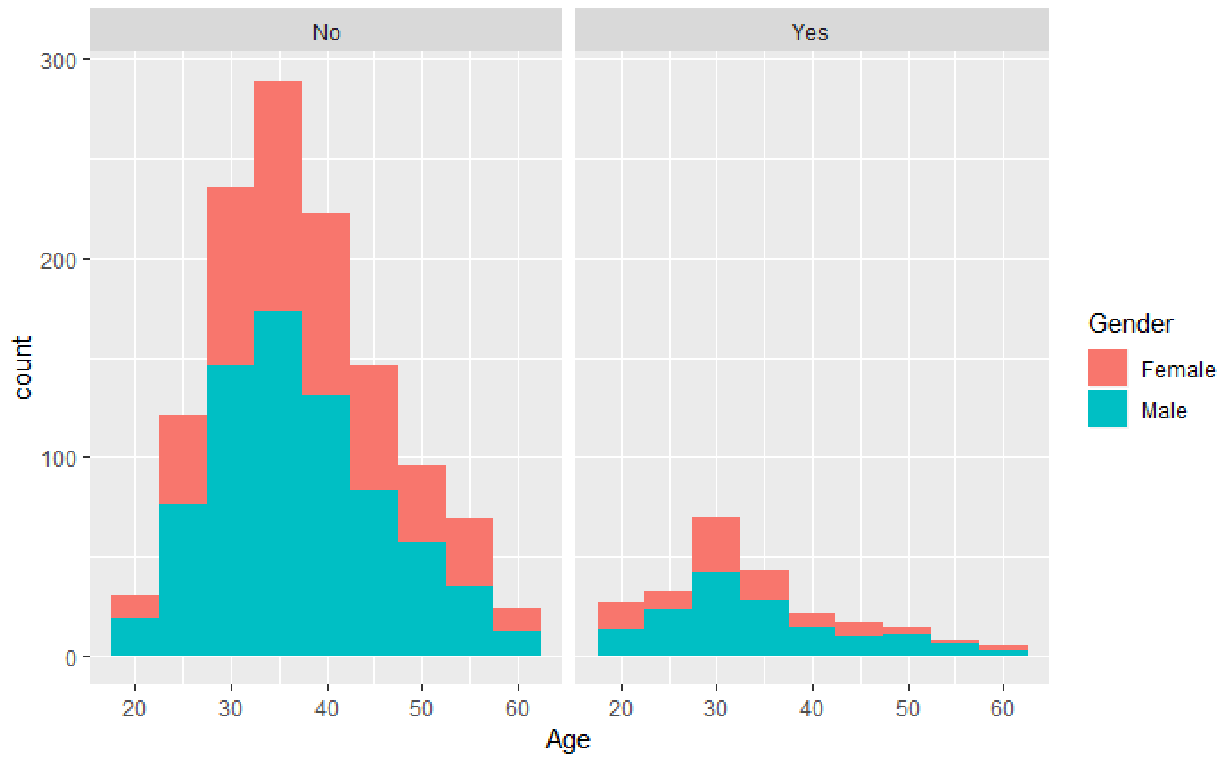


The data visualization shows that there are more males than females in this organization. Gender is not significant in respective to attrition

Attrition and Age

Hide

```
ggplot(hr_data, aes(Age, fill = Gender)) +
  geom_histogram(binwidth = 5) +
  facet_grid(~Attrition)
```



It is seen from the data visualization that the median age of the organization between 30-40 years. Also a the people who leave the organization are between 30-40 years old, likewise a significant number of people who doesnt leave the organization.

We will do a feature extraction of the age seperating the older people from the younger people.

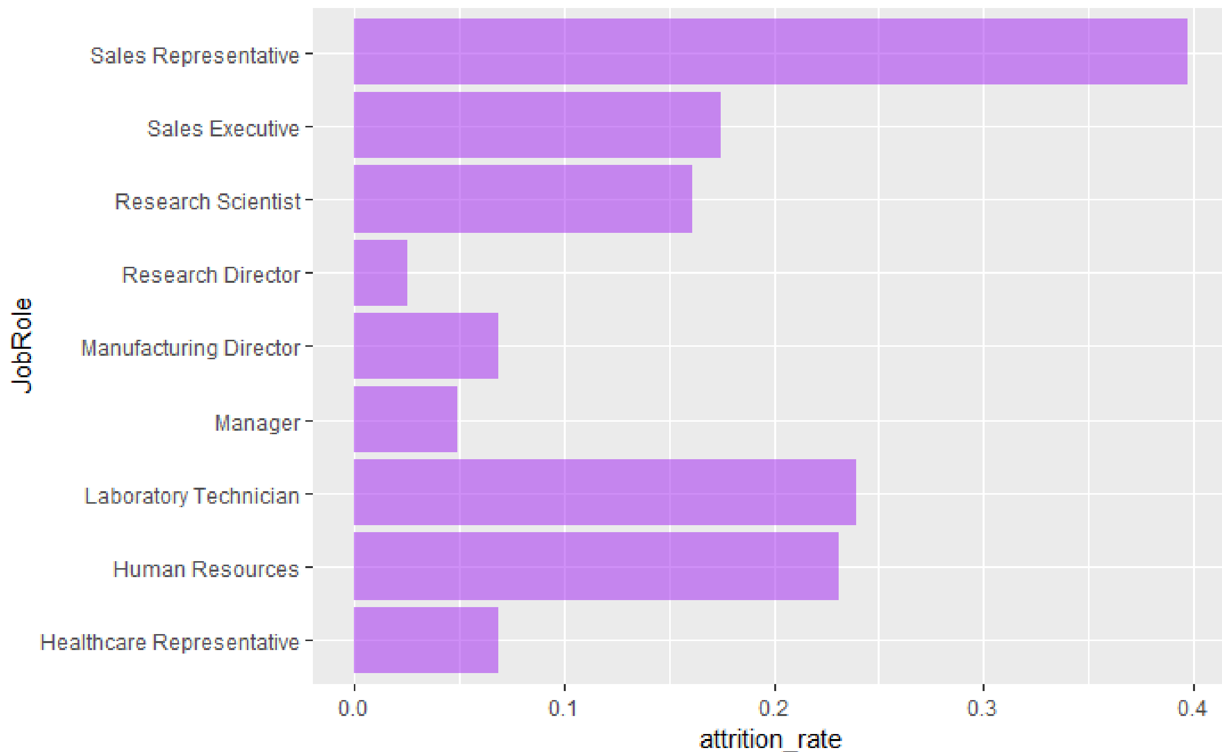
Job Role and Attrition

We know that work and stress levels might make an employee leave an organization, and that might depend on the job role. We want to visualize the job role and attrition to know the relationship between job roles and attrition.

Hide

```
hr_data$JobRole <- hr_data$`Job Role`

job_att <-hr_data %>%

  group_by(JobRole)%>%

  summarize(attrition_rate=mean(Attrition=="Yes"))%>%
```

```
  ggplot(aes(x=JobRole,y=attrition_rate)) + geom_bar(stat='identity',alpha=0.5,fill="purple") +
  coord_flip()
job_att
```



We see that the sales representatives have more attrition rate than any other department. The Stress level of the sales representative might make it a more likely factor of an employee leaving the organization. It also seems that the managers and leaders have a lower attrition rate.

We will extract features of Mangers and staff of the company

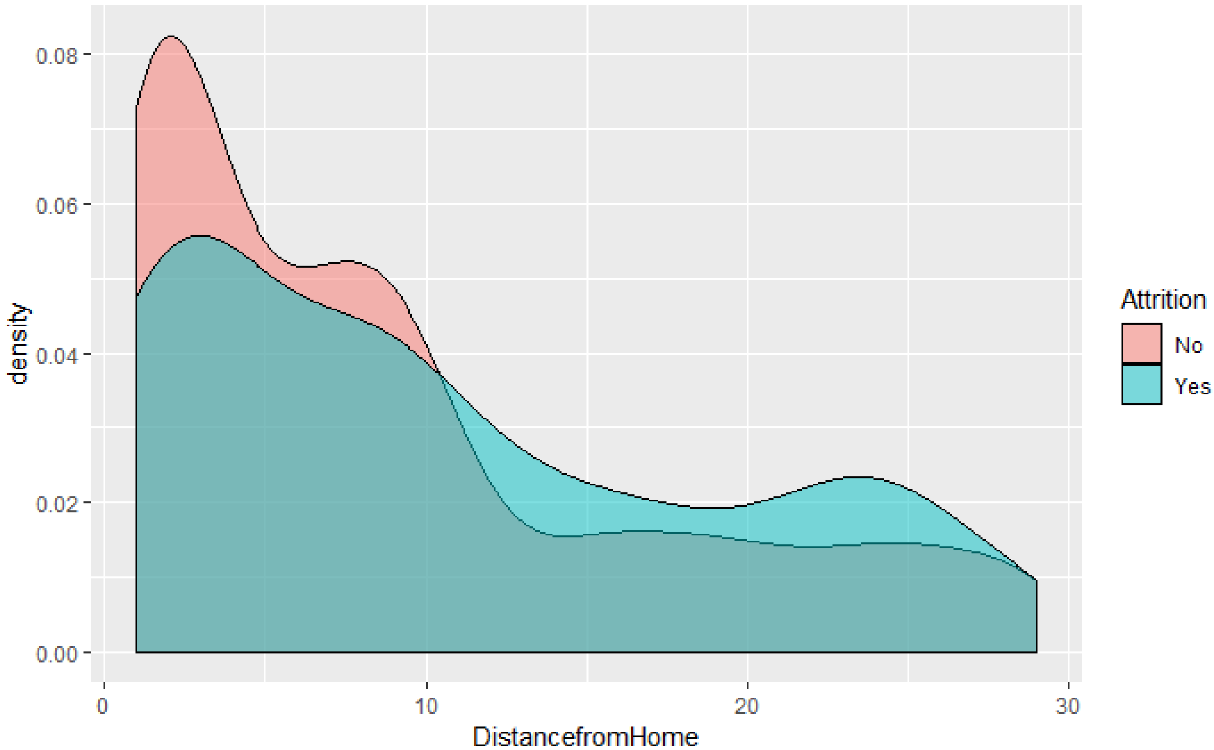
Attrition and Distance from Home

The likelihood that an employee will leave an organization might likely depend on the distance to the office. If the distance is too far, the employee might be looking to leave. We wil visuaize the relationship between distance to home and attrition

Hide

```
hr_data$DistancefromHome <- hr_data$`Distance From Home (kms)`
ggplot(hr_data,aes(DistancefromHome,fill=Attrition)) +
```

```
  geom_density(alpha=0.5)
```



Hide

```
NA
```

There doesnt a great deal in people staying far away from the office. There are a number of people staying closer to the office, the attrition rate are quite lower for those who stay within 10km away from the office. For those who stay farther away from the office, the attriton rate is quite higher.

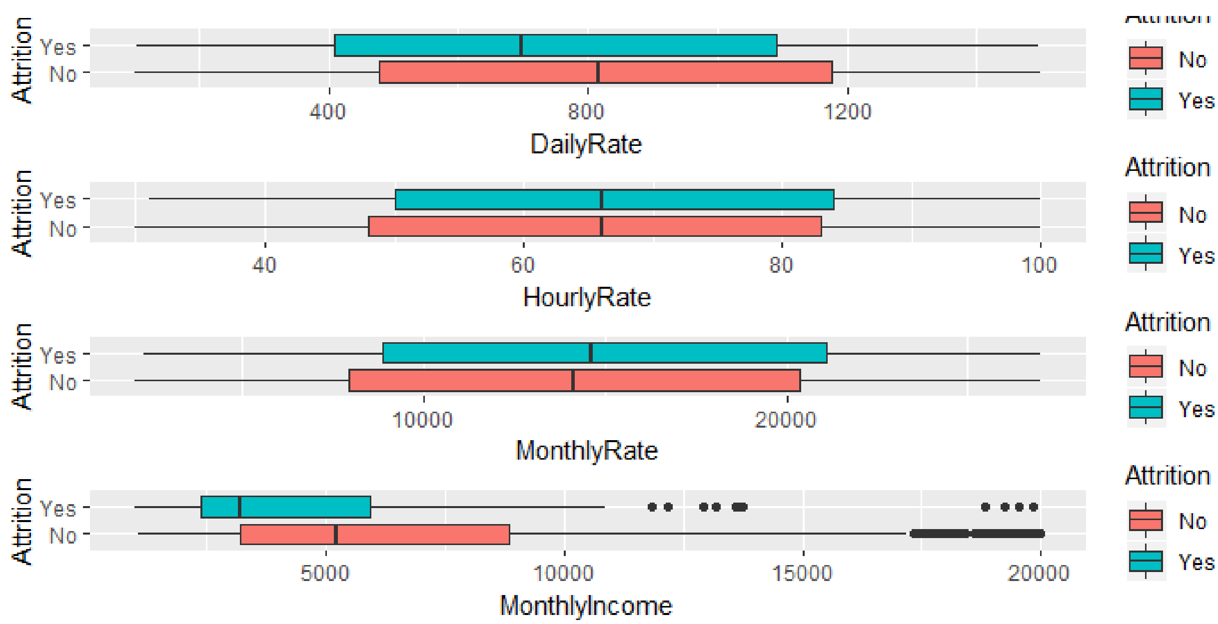
Attrition and Payrates

Visualzing the relationship between attrition and the different payrates using a boxplot.

Hide

```
hr_data$DailyRate <- hr_data$`Daily Rate (USD)`

hr_data$HourlyRate <- hr_data$`Hourly Rate (USD)`

hr_data$MonthlyRate <- hr_data$`Monthly Rate (USD)`

hr_data$MonthlyIncome <- hr_data$`Monthly Income (USD)`

  dr <- ggplot(hr_data,aes(Attrition,DailyRate, fill = Attrition)) + geom_box
plot() + coord_flip()

  hr <- ggplot(hr_data,aes(Attrition,HourlyRate, fill = Attrition)) + geom_bo
xplot() + coord_flip()
```

```
mr <- ggplot(hr_data,aes(Attrition,MonthlyRate, fill = Attrition)) + geom_b
oxplot() + coord_flip()

mi <- ggplot(hr_data,aes(Attrition,MonthlyIncome, fill = Attrition)) + geom
_boxplot() + coord_flip()

#feature extraction of rates


grid.arrange(dr,hr,mr,mi,nrow = 5)
```



Hide

```
NA
```

The pay rates doesnt give much information on the attriton rate. There is no much significant mean difference in the totalrate as well. #Other than the daily rate, attrition is present for those with lower rate and monthly income

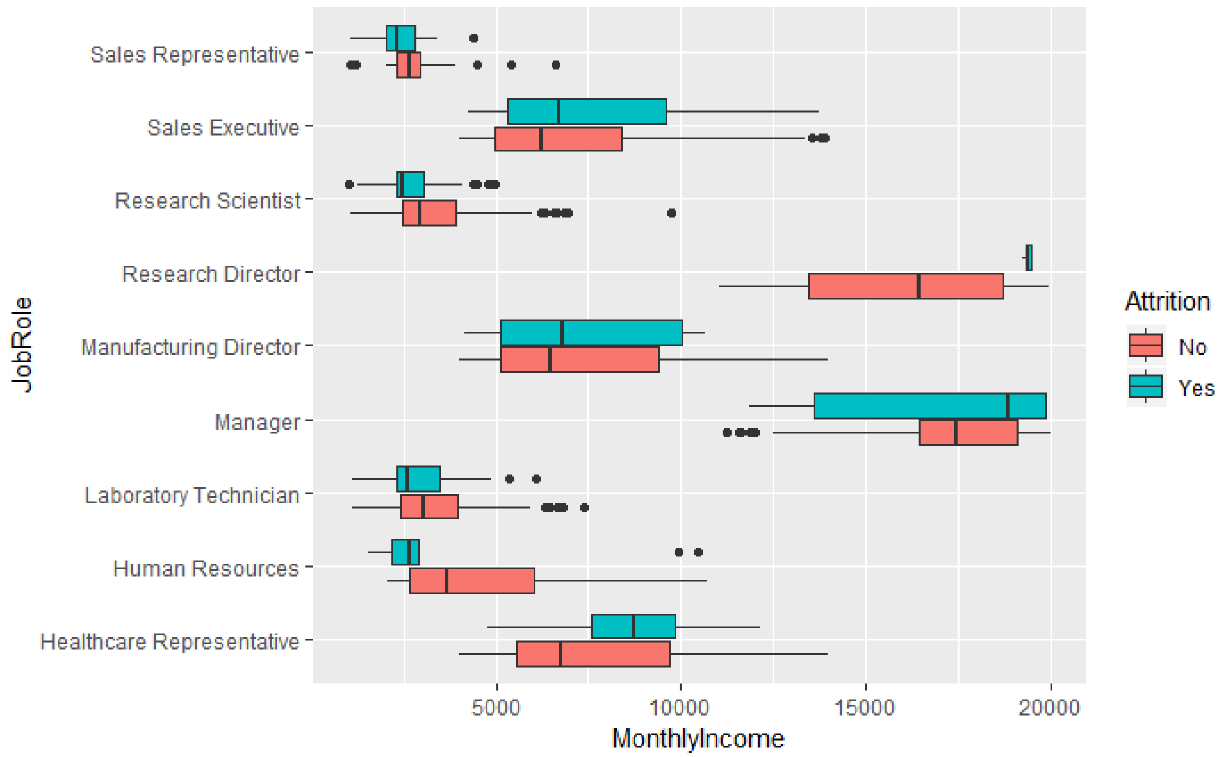
Monthly income and Job roles

Sales representatives employees tend to leave the organization most. We want to visualize the relationship between the monthly income and the job roles(Which Job is least paying?)

Hide

```
ggplot(hr_data, aes(JobRole,MonthlyIncome, fill= Attrition)) + geom_boxplot()
+
```

```
  coord_flip()
```



Hide

```
NA
```

we can see that Sales Representatives, Research Scientists and Labouratory Technicians are the lower job levels based on monthly income. The mean of those who leave is less than those who do not.

Correlation analysis

Visualizing the correlation between numerical variables, and checking for colinearity.

Hide

```
str(hr_data)
```

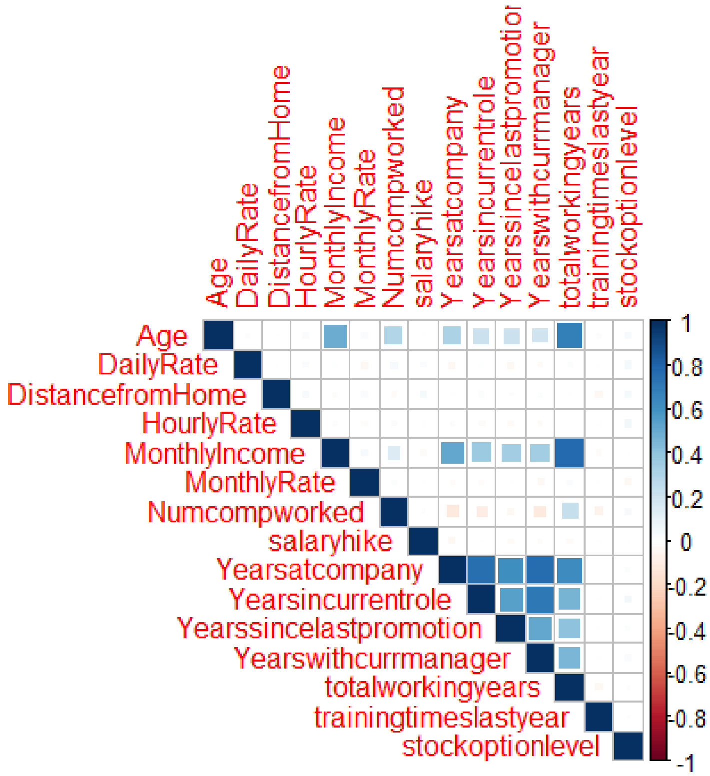
```
'data.frame':   1470 obs. of  40 variables:
 $ Department                : chr  "Sales" "Research & Development" "Researc
h & Development" "Research & Development" ...
 $ Job Role                  : chr  "Sales Executive" "Research Scientist" "L
aboratory Technician" "Research Scientist" ...
 $ Attrition (Yes/No)        : chr  "Yes" "No" "Yes" "No" ...
 $ Gender                    : chr  "Female" "Male" "Male" "Female" ...
 $ Age                       : num  41 49 37 33 27 32 59 30 38 36 ...
```

```
 $ Marital Status            : chr  "Single" "Married" "Single" "Married" ...
 $ Education                 : chr  "College" "Below College" "College" "Mast
er" ...
 $ Education Field           : chr  "Life Sciences" "Life Sciences" "Other" "
Life Sciences" ...
 $ Business Travel           : chr  "Travel_Rarely" "Travel_Frequently" "Trav
el_Rarely" "Travel_Frequently" ...
 $ Distance From Home (kms)  : num  1 8 2 3 2 2 3 24 23 27 ...
 $ Job Involvement           : chr  "High" "Medium" "Medium" "High" ...
 $ Job Level                 : num  2 2 1 1 1 1 1 1 3 2 ...
 $ Job Satisfaction          : chr  "Very High" "Medium" "High" "High" ...
 $ Hourly Rate (USD)         : num  94 61 92 56 40 79 81 67 44 94 ...
 $ Daily Rate (USD)          : num  1102 279 1373 1392 591 ...
 $ Monthly Rate (USD)        : num  19479 24907 2396 23159 16632 ...
 $ Monthly Income (USD)      : num  5993 5130 2090 2909 3468 ...
 $ Salary Hike (%)           : num  11 23 15 11 12 13 20 22 21 13 ...
 $ Stock Option Level        : num  0 1 0 0 1 0 3 1 0 2 ...
 $ Over Time                 : chr  "Yes" "No" "Yes" "Yes" ...
 $ No. of Companies Worked   : num  8 1 6 1 9 0 4 1 0 6 ...
 $ Total Working Years       : num  8 10 7 8 6 8 12 1 10 17 ...
 $ Years At Company          : num  6 10 0 8 2 7 1 1 9 7 ...
 $ Years In Current Role     : num  4 7 0 7 2 7 0 0 7 7 ...
 $ Years Since Last Promotion: num  0 1 0 3 2 3 0 0 1 7 ...
 $ Years With Curr Manager   : num  5 7 0 0 2 6 0 0 8 7 ...
 $ Environment Satisfaction  : chr  "Medium" "High" "Very High" "Very High" .
..
 $ Training Times Last Year  : num  0 3 3 3 3 2 3 2 2 3 ...
 $ Work Life Balance         : chr  "Bad" "Better" "Better" "Better" ...
 $ Performance Rating        : chr  "Excellent" "Outstanding" "Excellent" "Ex
cellent" ...
 $ Relationship Satisfaction : chr  "Low" "Very High" "Medium" "High" ...
 $ Attrition                 : chr  "Yes" "No" "Yes" "No" ...
 $ MaritalStatus             : chr  "Single" "Married" "Single" "Married" ...
 $ BusinessTravel            : chr  "Travel_Rarely" "Travel_Frequently" "Trav
el_Rarely" "Travel_Frequently" ...
 $ JobRole                   : chr  "Sales Executive" "Research Scientist" "L
aboratory Technician" "Research Scientist" ...
```

```
 $ DistancefromHome          : num  1 8 2 3 2 2 3 24 23 27 ...
 $ DailyRate                 : num  1102 279 1373 1392 591 ...
 $ HourlyRate                : num  94 61 92 56 40 79 81 67 44 94 ...
 $ MonthlyRate               : num  19479 24907 2396 23159 16632 ...
 $ MonthlyIncome             : num  5993 5130 2090 2909 3468 ...
```

Hide

```
hr_data$Numcompworked <- hr_data$`No. of Companies Worked`
hr_data$Yearsatcompany <- hr_data$`Years At Company`
hr_data$Yearsincurrentrole <- hr_data$`Years In Current Role`
hr_data$Yearswithcurrmanager <- hr_data$`Years With Curr Manager`
hr_data$Yearssincelastpromotion <- hr_data$`Years Since Last Promotion`
hr_data$totalworkingyears <- hr_data$`Total Working Years`
hr_data$trainingtimeslastyear <- hr_data$`Training Times Last Year`
hr_data$stockoptionlevel <- hr_data$`Stock Option Level`
hr_data$salaryhike <- hr_data$`Salary Hike (%)`
hr_data$joblevel <- hr_data$`Job Level`
data_corr = hr_data %>%
  dplyr::select(Age,DailyRate,DistancefromHome,HourlyRate,MonthlyIncome,MonthlyRate, Numcompworked,salaryhike,Yearsatcompany,Yearsincurrentrole,Yearssincelastpromotion,Yearswithcurrmanager,totalworkingyears,trainingtimeslastyear,stockoptionlevel)
corrplot(cor(data_corr), method = "square", type="upper")
```



From the correlation plot, we observe correlated features. We will exclude the variables that are correlated from the model. (Colinearity). The correlated variables are: Age and total working years Total working years and monthly income Years with current manager and years at company Years with current current manager and years in current role

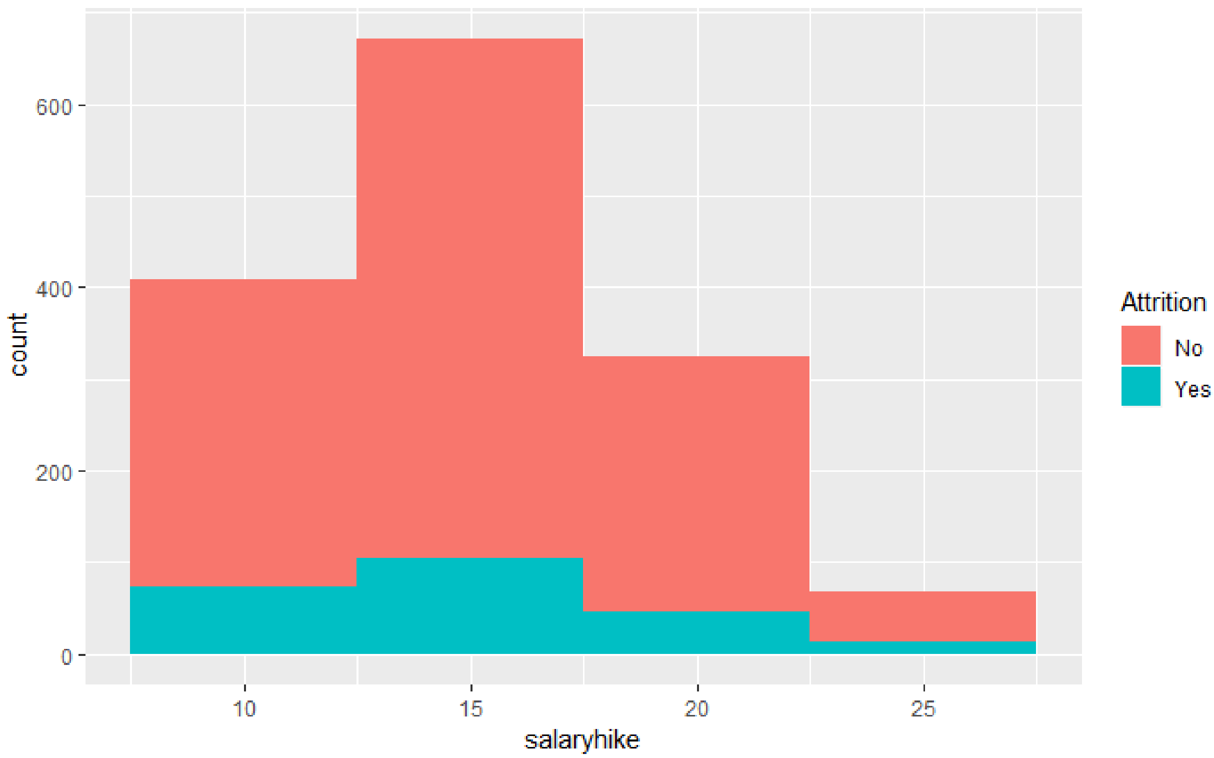
The variables we will exclude are Years with current manager and total working years

Attrition and Salary Hike

Visualizing salary hike and attrition

Hide

```
ggplot(hr_data,aes(salaryhike, fill = Attrition)) + geom_histogram(binwidth =
5)
```



Hide

```
#Salary Hike and Years at company

ggplot(hr_data,aes(Yearsatcompany,salaryhike,col=(Attrition),size = salaryhik
e)) +geom_point(alpha = 0.5)
```

Hide

```
#Salary Hike and Years at experience

#ggplot(hr_data,aes(totalworkingyears,salaryhike,col=(Attrition),colour = sal
aryhike))+ geom_point(alpha = 0.5)
```

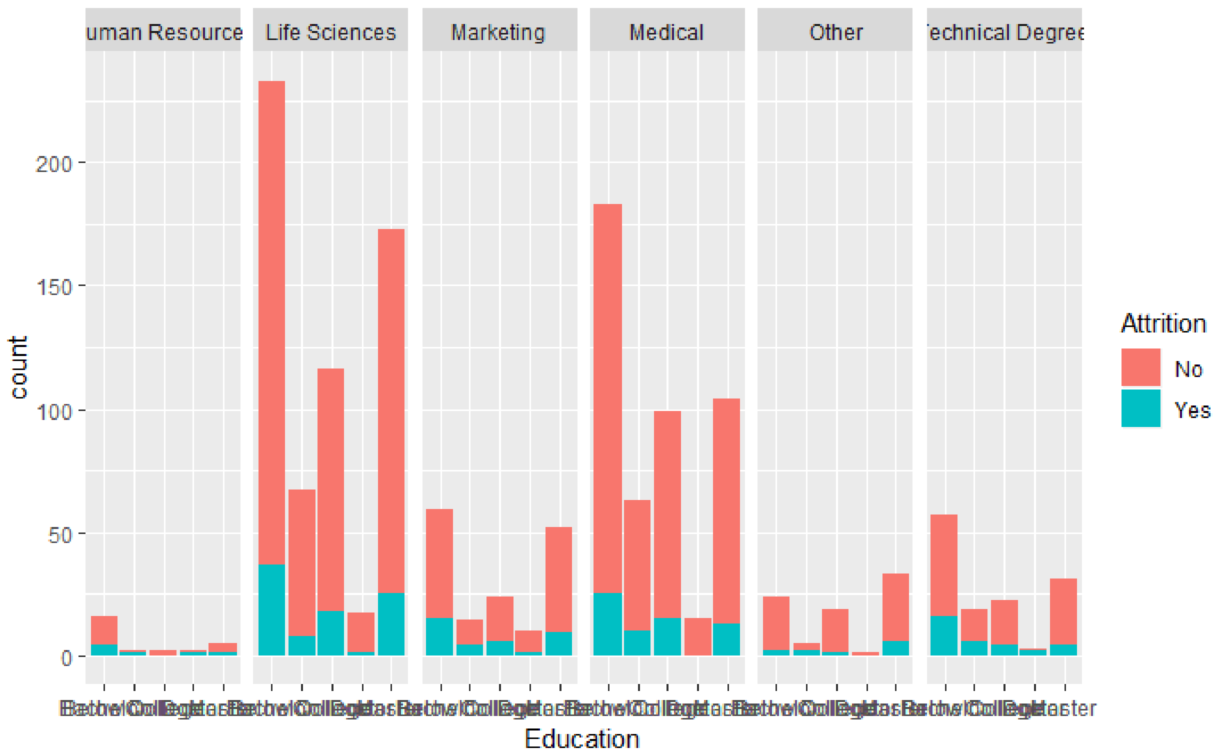
From the data visualization, we can see that there is no linear relationship between totalworkingyears

Attrition and Education

Visualizing attrition rate and education levels and fields

Hide

```
hr_data$educationfield <- hr_data$`Education Field`

ggplot(hr_data,aes(Education, fill = Attrition)) +geom_bar() + facet_grid(~ed
ucationfield)
```



EMployees mith a life sciences and medical education level seems more populated in the organization. There also seems more people with a bachelor degree in the organization. However educational background might not be related to attrition levels

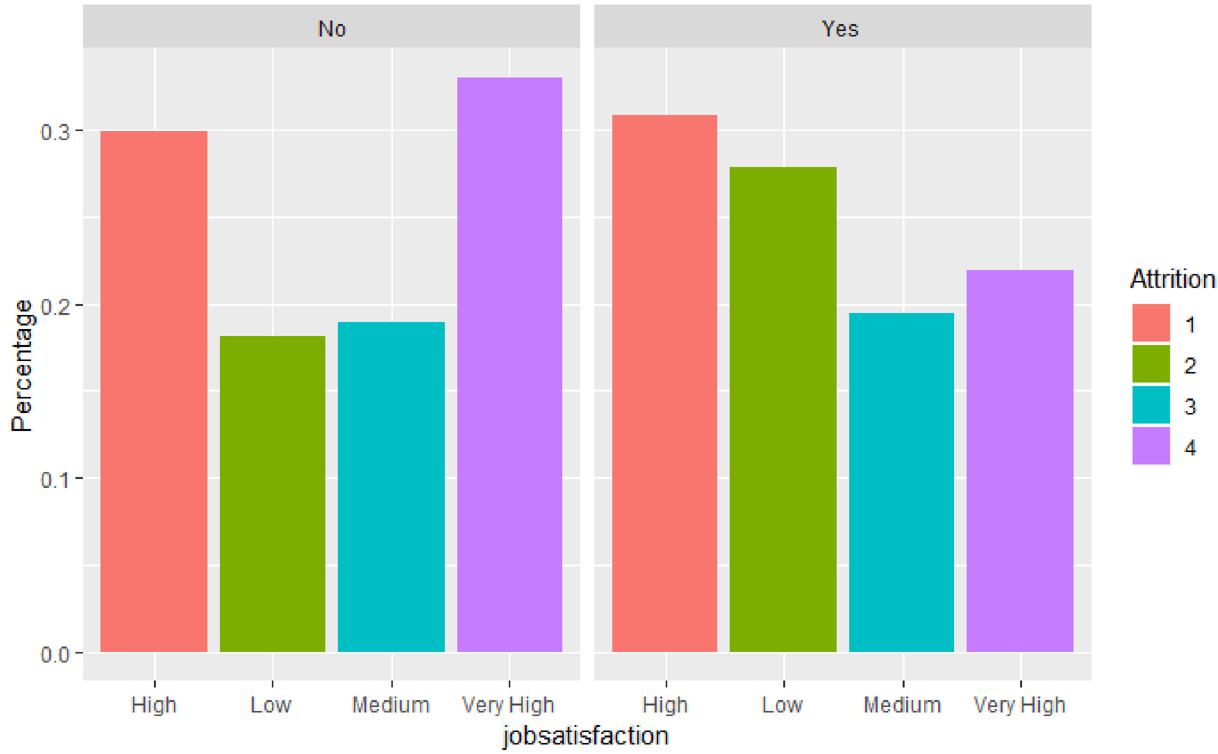
Attrition and categorical variables.

Attrition and JOb Satisfaction * Years with current manager

Hide

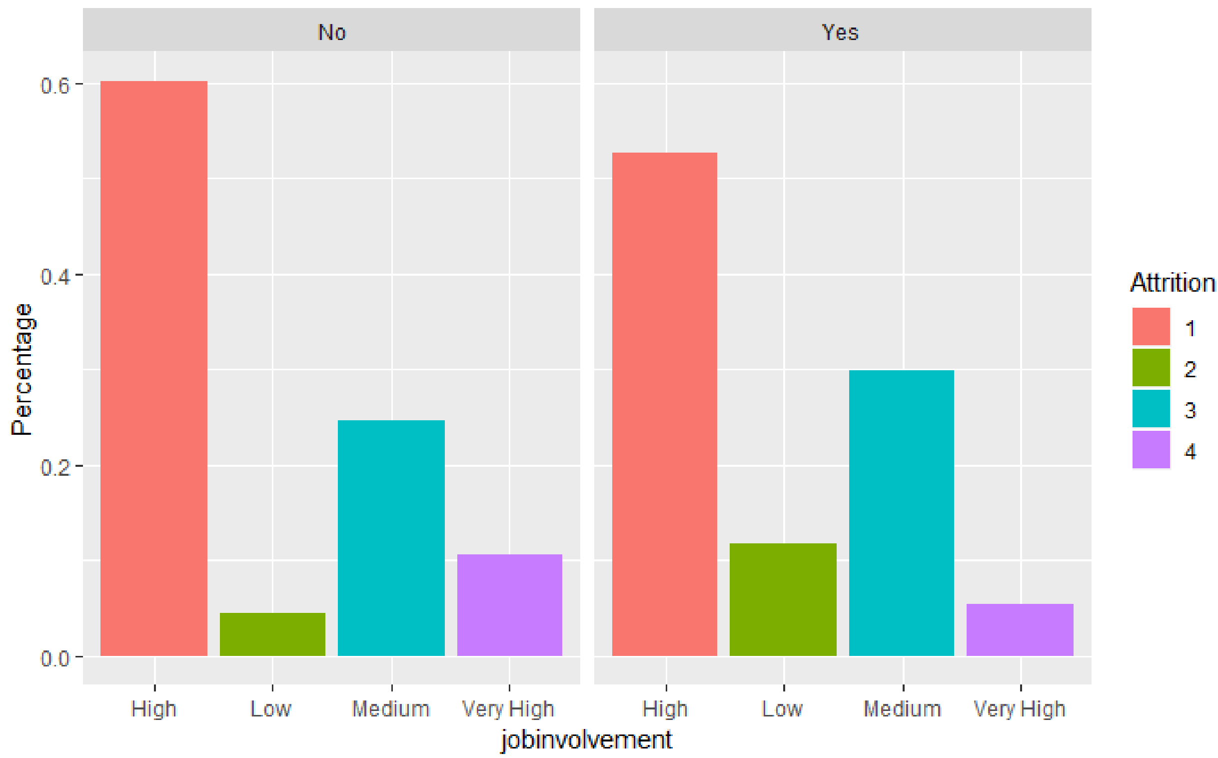
```
hr_data$jobsatisfaction <- hr_data$`Job Satisfaction`
hr_data$jobinvolvement <- hr_data$`Job Involvement`
hr_data$relationshipsatisfaction <- hr_data$`Relationship Satisfaction`
hr_data$worklifebalance <- hr_data$`Work Life Balance`
hr_data$environmentsatisfaction <- hr_data$`Environment Satisfaction`
hr_data$overtime <- hr_data$`Over Time`
hr_data$performancerating <- hr_data$`Performance Rating`
ggplot(hr_data,aes(x=jobsatisfaction,group=Attrition))+
  geom_bar(stat="count",aes(y=..prop..,fill=factor(..x..))) +
```

```
  labs(y="Percentage", fill = "Attrition") +
  facet_wrap(~Attrition)
```



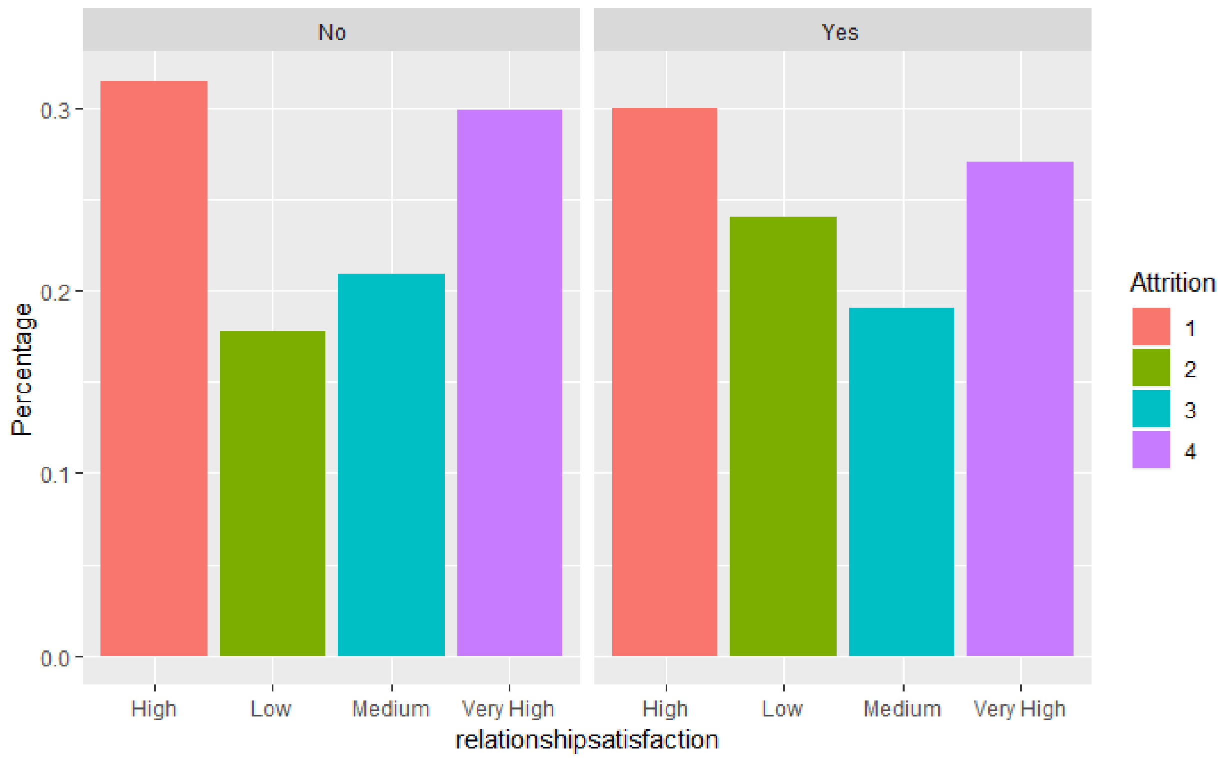
Hide

```
ggplot(hr_data,aes(x=jobinvolvement,group=Attrition))+
  geom_bar(stat="count",aes(y=..prop..,fill=factor(..x..))) +
  labs(y="Percentage", fill = "Attrition") +
  facet_wrap(~Attrition)
```



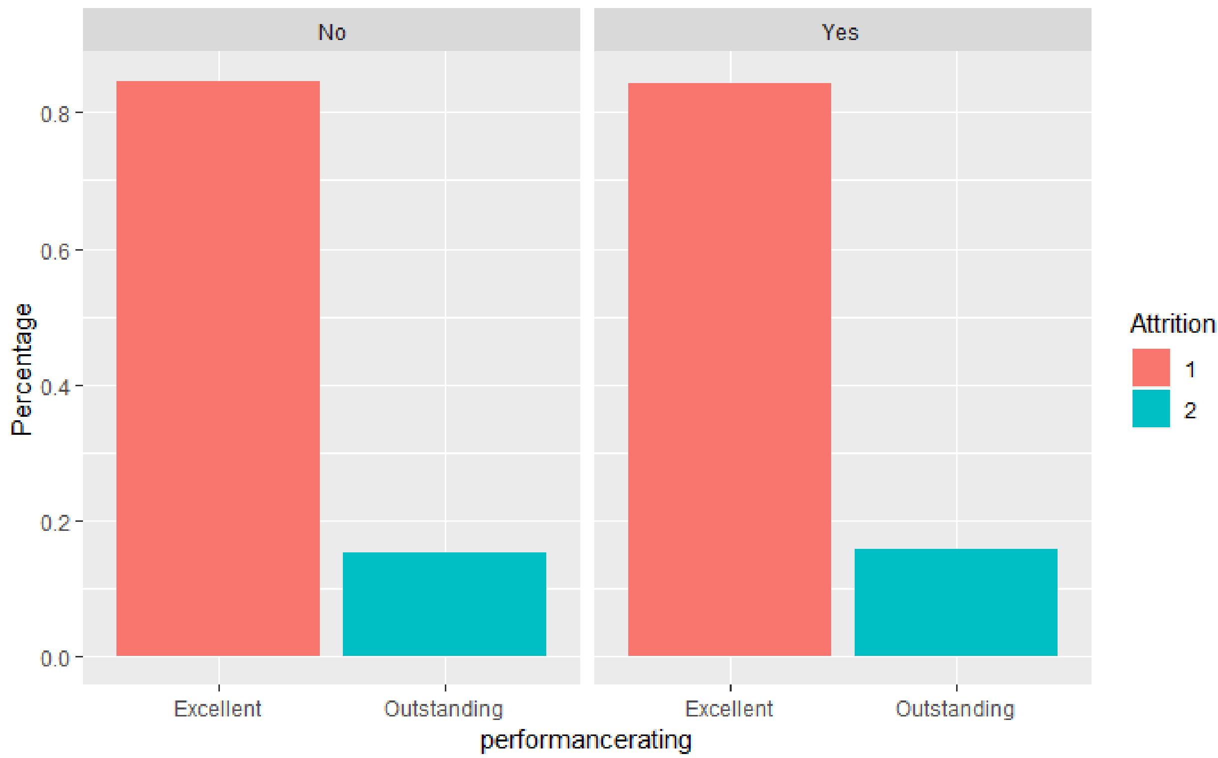
Hide

```
ggplot(hr_data,aes(x=relationshipsatisfaction,group=Attrition))+
  geom_bar(stat="count",aes(y=..prop..,fill=factor(..x..)),position = positio
n_dodge()) +
  labs(y="Percentage", fill = "Attrition") +
  facet_wrap(~Attrition)
```



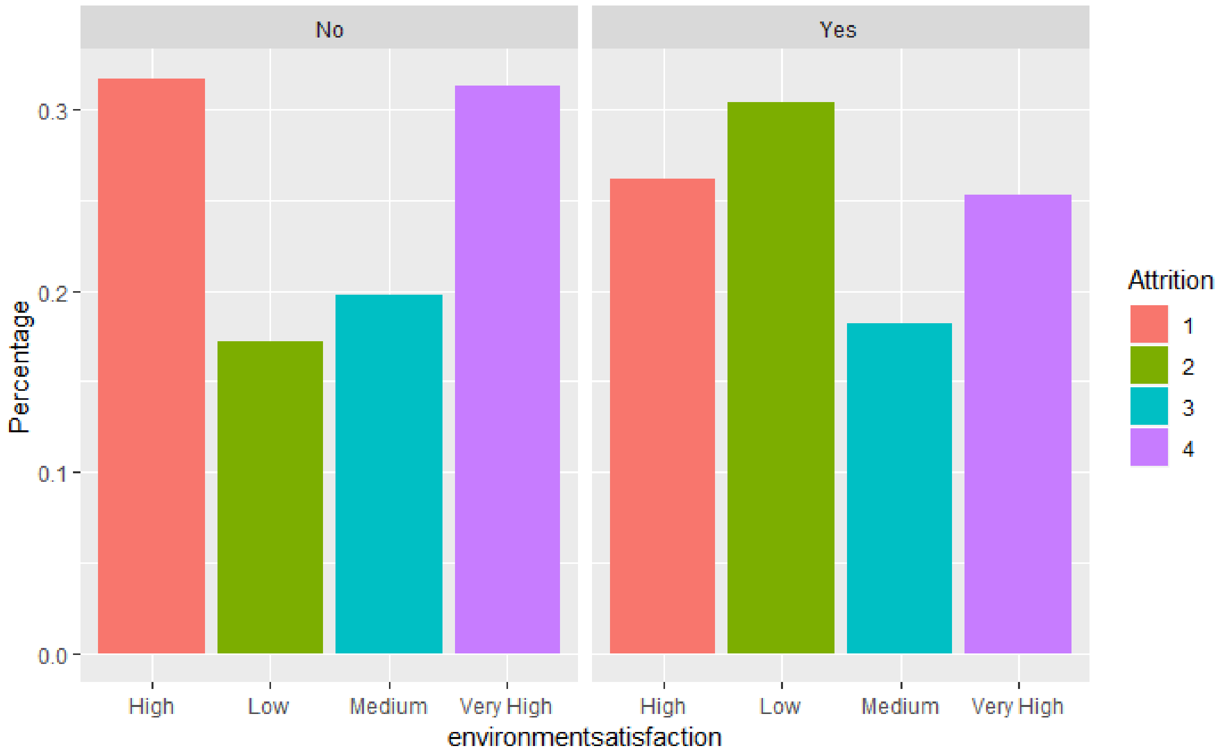
Hide

```
ggplot(hr_data,aes(x=performancerating,group=Attrition))+
  geom_bar(stat="count",aes(y=..prop..,fill=factor(..x..))) +
  labs(y="Percentage", fill = "Attrition") +
  facet_wrap(~Attrition)
```



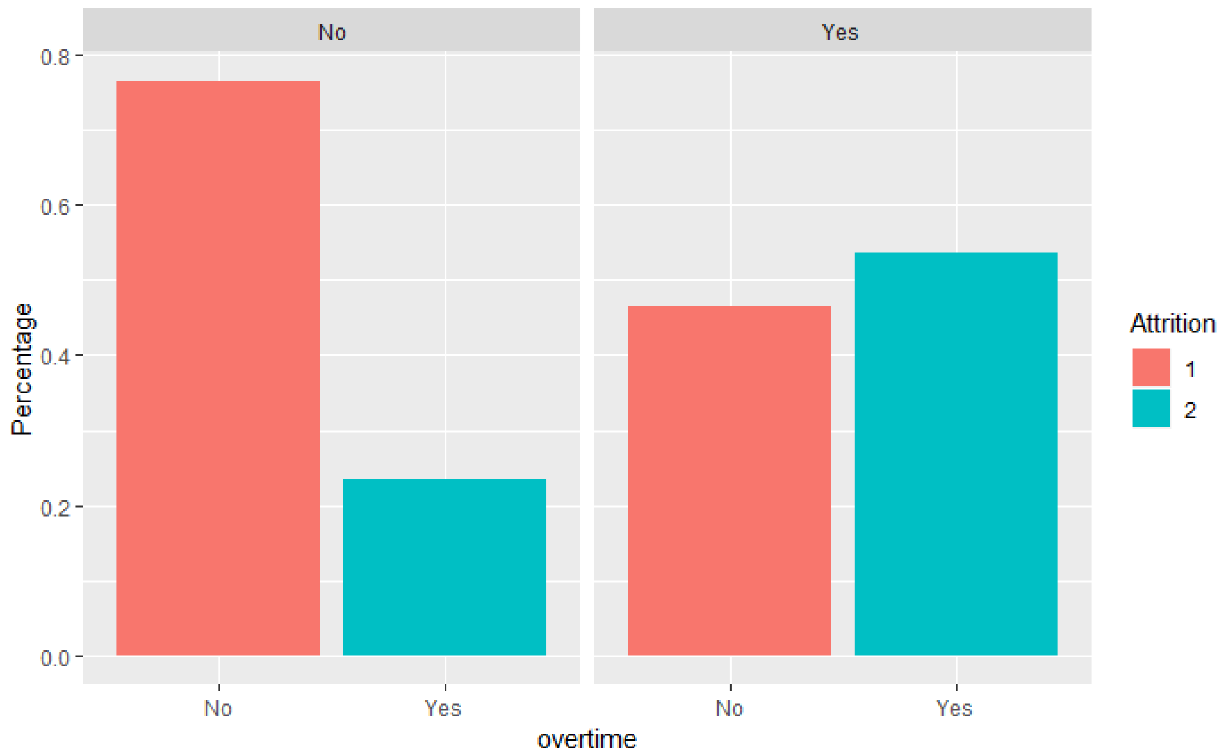
Hide

```
ggplot(hr_data,aes(x=environmentsatisfaction,group=Attrition))+
  geom_bar(stat="count",aes(y=..prop..,fill=factor(..x..))) +
  labs(y="Percentage", fill = "Attrition") +
  facet_wrap(~Attrition)
```



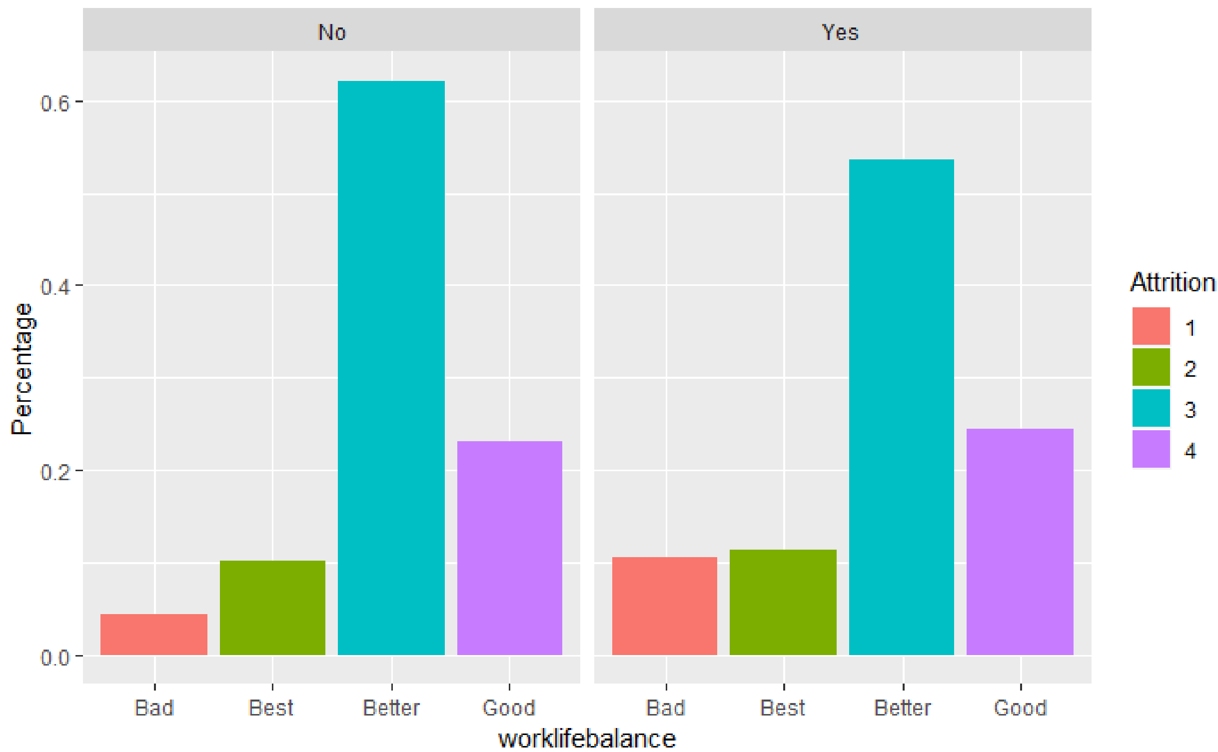
Hide

```
ggplot(hr_data,aes(x=overtime,group=Attrition))+
  geom_bar(stat="count",aes(y=..prop..,fill=factor(..x..))) +
  labs(y="Percentage", fill = "Attrition") +
  facet_wrap(~Attrition)
```



Hide

```
ggplot(hr_data,aes(x=worklifebalance,group=Attrition))+
  geom_bar(stat="count",aes(y=..prop..,fill=factor(..x..))) +
  labs(y="Percentage", fill = "Attrition") +
  facet_wrap(~Attrition)
```



We observe that people with low job satisfaction have higher attrition rate. Also it appears that people with high job satisfication tend to leave the company, however in people who do not leave. those who have very high job satisfaction then to stay.

Employees with higher job involvement tend to leave more, however people with high job invovlemnt have also reported no attrition rate

High relationship satisfaction have also reported staying as well as almost the same number of people have reported leaving

Feature extraction

Feature engineering from the variables. For age, we want to divide the age into 3 groups.

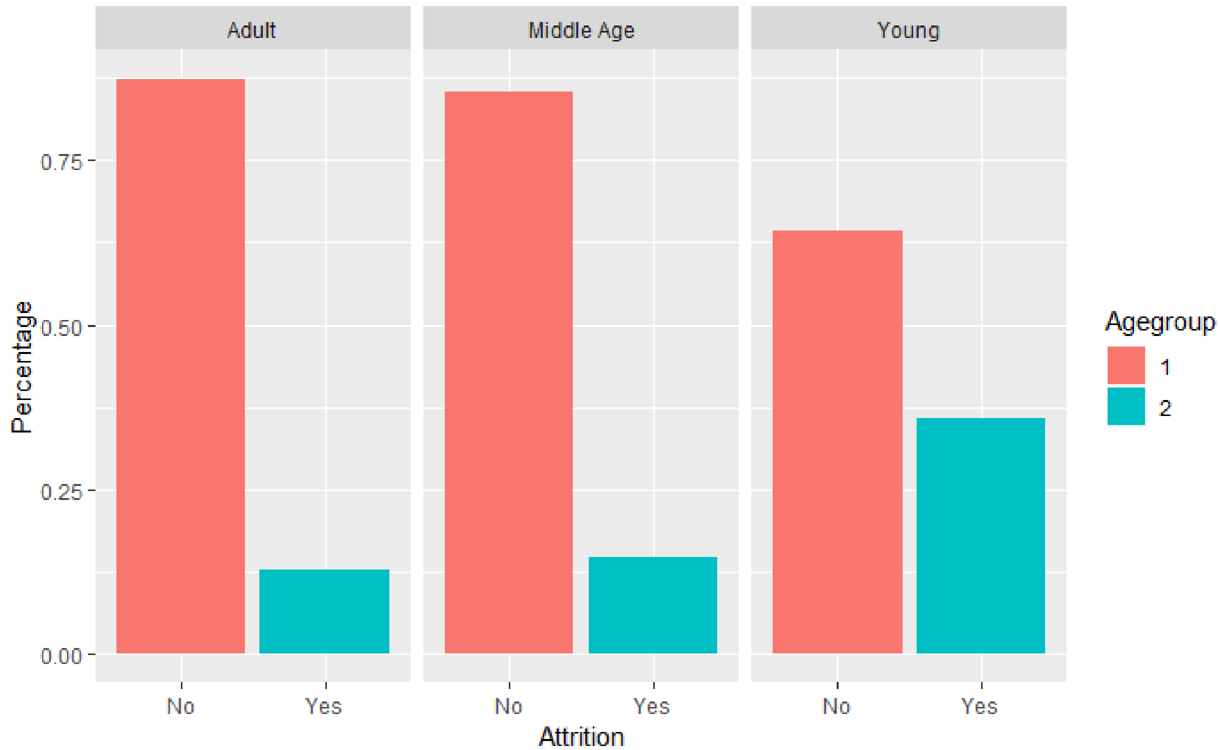
Hide

```
hr_data$AgeGroup <- as.factor(ifelse(hr_data$Age <= 25, "Young", ifelse(hr_da
ta$Age<=50, "Middle Age", "Adult")))

table(hr_data$AgeGroup,hr_data$Attrition)
```

```
              No  Yes
  Adult        125   18
  Middle Age 1029  175
  Young         79   44
```

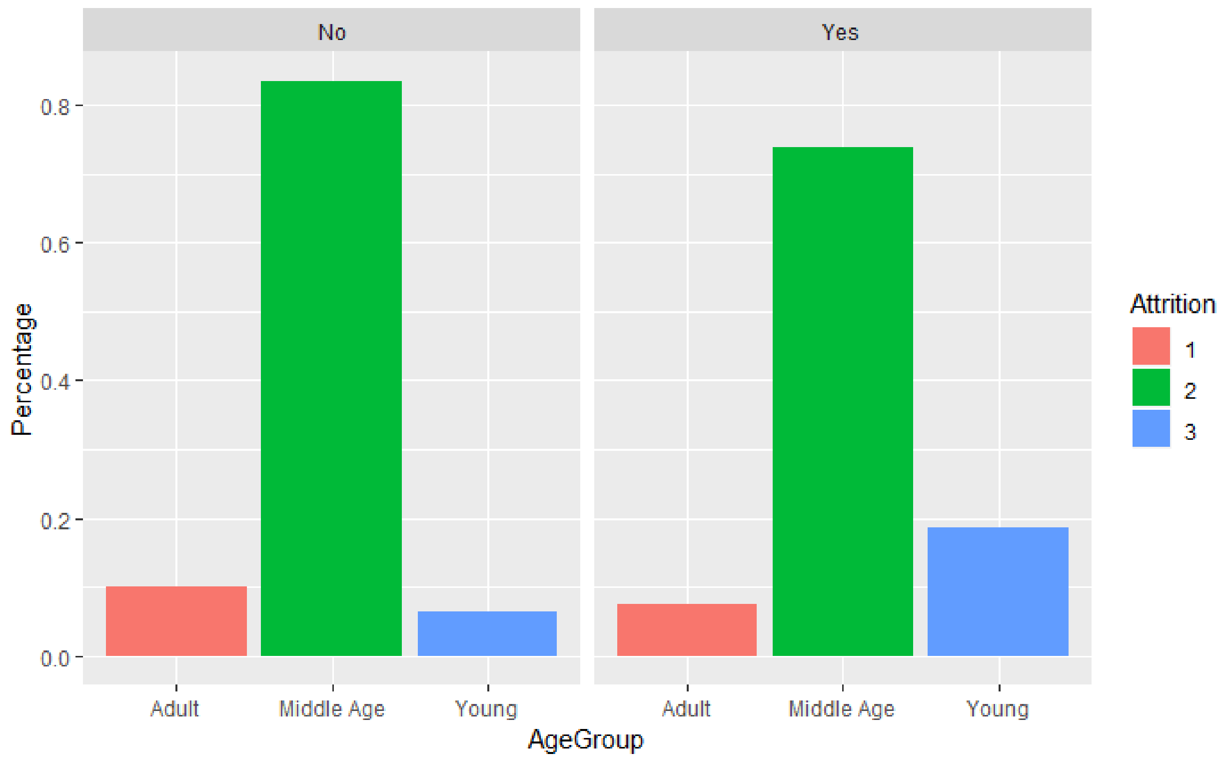
Hide

```
ggplot(hr_data,aes(x=Attrition,group=AgeGroup))+
  geom_bar(stat="count",aes(y=..prop..,fill=factor(..x..))) +
  labs(y="Percentage", fill = "Agegroup") +
  facet_wrap(~AgeGroup)
```



Hide

```
ggplot(hr_data,aes(x=AgeGroup,group=Attrition))+
  geom_bar(stat="count",aes(y=..prop..,fill=factor(..x..))) +
  labs(y="Percentage", fill = "Attrition") +
  facet_wrap(~Attrition)
```

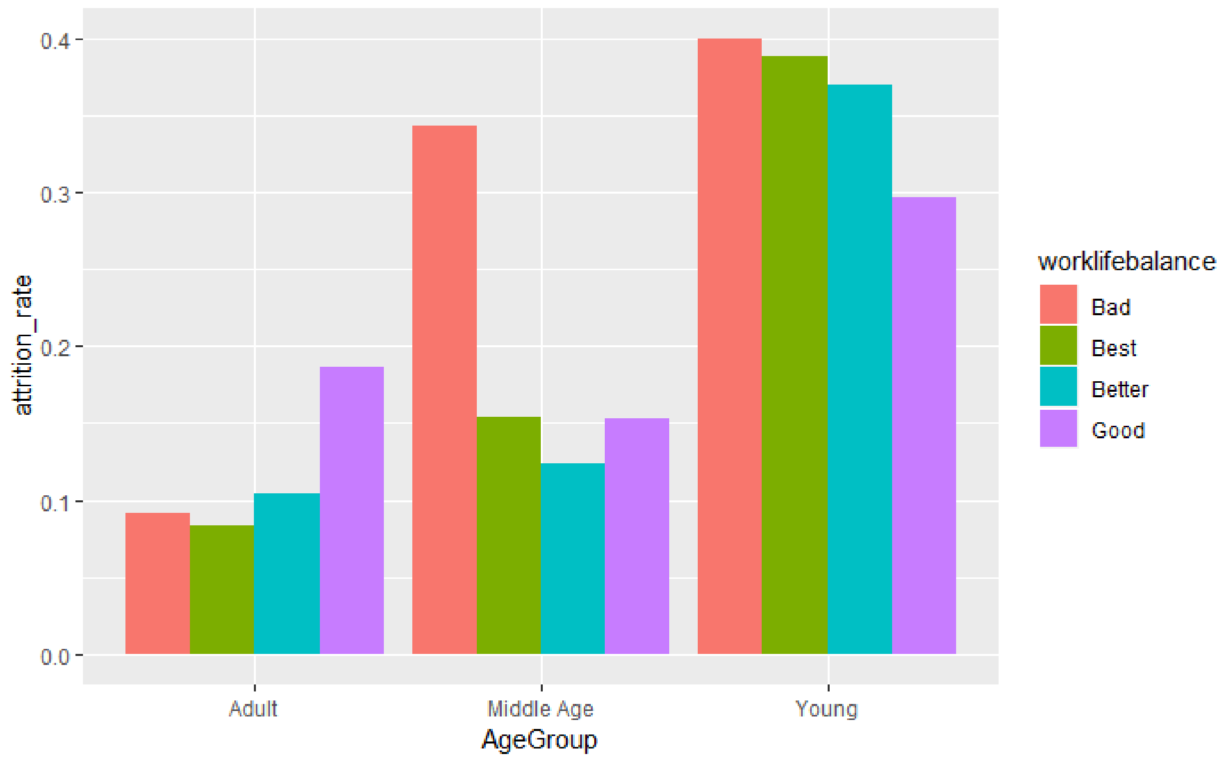


We can conclude that majority of the employees in the organization are middle aged. Also young people below 25 years tend to leave more.

Age group and work-life balance

Hide

```
age_work <- hr_data%>%group_by(AgeGroup,worklifebalance)%>%summarize(attritio
n_rate=mean(Attrition=="Yes")) %>%

ggplot(aes(x=AgeGroup,y=attrition_rate,fill=worklifebalance)) + geom_bar(stat
="identity",position = position_dodge())

age_work
```



Hide

```
#Change job level to factor

hr_data$joblevel <- as.factor(hr_data$joblevel)

#
```

Adult have the lowest attrition rate in percentage and it appears they have the the lowest reported bad work-life balance. Young people have the highest reported bad work-life environment. Probably that is why they leave more. ##Total Satisfaction

Hide

```
hr_data$environmentsatisfaction[hr_data$environmentsatisfaction=="Low"] <- 1

hr_data$environmentsatisfaction[hr_data$environmentsatisfaction=="Medium"] <-
2

hr_data$environmentsatisfaction[hr_data$environmentsatisfaction=="High"] <- 3

hr_data$environmentsatisfaction[hr_data$environmentsatisfaction=="Very High"]
<- 4

str(hr_data$environmentsatisfaction)
```

```
 chr [1:1470] "2" "3" "4" "4" "1" "4" "3" "4" "4" "3" "1" "4" "1" "2" "3" "2"
"1" "4" "1" "4" ...
```

Hide

```
hr_data$jobsatisfaction[hr_data$jobsatisfaction=="Low"] <- 1
```

```
hr_data$jobsatisfaction[hr_data$jobsatisfaction=="Medium"] <- 2
hr_data$jobsatisfaction[hr_data$jobsatisfaction=="High"] <- 3
hr_data$jobsatisfaction[hr_data$jobsatisfaction=="Very High"] <- 4
str(hr_data$jobsatisfaction)
```

```
 chr [1:1470] "4" "2" "3" "3" "2" "4" "1" "3" "3" "3" "2" "3" "3" "4" "3" "1"
"2" "4" "4" "4" ...
```

Hide

```
hr_data$relationshipsatisfaction[hr_data$relationshipsatisfaction=="Low"] <-
1
hr_data$relationshipsatisfaction[hr_data$relationshipsatisfaction=="Medium"]
<- 2
hr_data$relationshipsatisfaction[hr_data$relationshipsatisfaction=="High"] <-
3
hr_data$relationshipsatisfaction[hr_data$relationshipsatisfaction=="Very High
"] <- 4
str(hr_data$relationshipsatisfaction)
```

```
 chr [1:1470] "1" "4" "2" "3" "4" "3" "1" "2" "2" "2" "3" "4" "4" "3" "2" "3"
"4" "2" "3" "3" ...
```

Hide

```
hr_data$jobinvolvement[hr_data$jobinvolvement=="Low"] <- 1
hr_data$jobinvolvement[hr_data$jobinvolvement=="Medium"] <- 2
hr_data$jobinvolvement[hr_data$jobinvolvement=="High"] <- 3
hr_data$jobinvolvement[hr_data$jobinvolvement=="Very High"] <- 4
str(hr_data$jobinvolvement)
```

```
 chr [1:1470] "3" "2" "2" "3" "3" "3" "4" "3" "2" "3" "4" "2" "3" "3" "2" "4"
"4" "4" "2" "3" ...
```

Hide

```
hr_data$worklifebalance[hr_data$worklifebalance=="Bad"] <- 1
hr_data$worklifebalance[hr_data$worklifebalance=="Good"] <- 2
hr_data$worklifebalance[hr_data$worklifebalance=="Better"] <- 3
hr_data$worklifebalance[hr_data$worklifebalance=="Best"] <- 4
str(hr_data$worklifebalance)
```

```
 chr [1:1470] "1" "3" "3" "3" "3" "2" "2" "3" "3" "2" "3" "3" "2" "3" "3" "3"
"2" "2" "3" "3" ...
```

Hide

```
hr_data$OverallSatisfaction <- as.numeric(hr_data$environmentsatisfaction) + as.numeric(hr_data$jobsatisfaction) + as.numeric(hr_data$relationshipsatisfaction) + as.numeric(hr_data$jobinvolvement)

str(hr_data$OverallSatisfaction)
```

```
 num [1:1470] 10 11 11 13 10 14 9 12 11 11 ...
```

Hide

```
summary(hr_data$OverallSatisfaction)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   5.00   10.00   11.00   10.89   12.00   16.00 
```

Hide

```
hr_data$OverallSatisfactionlevel <- as.factor(ifelse (hr_data$OverallSatisfaction < ave(hr_data$OverallSatisfaction), "Low", "High"))

table(hr_data$OverallSatisfactionlevel,hr_data$Attrition)
```

```
      
        No Yes
  High 751  99
  Low  482 138
```

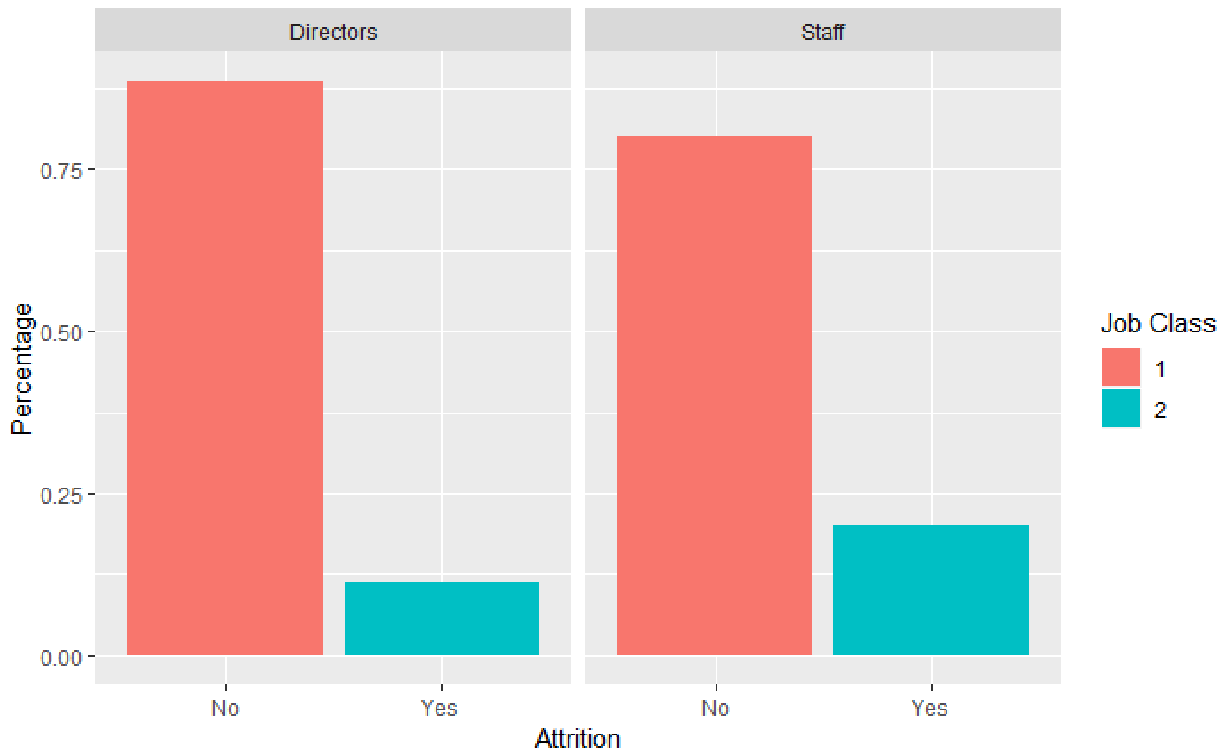
Hide

```
hr_data$jobclass <- hr_data$JobRole

directors <- c( 'Sales Executive', 'Manager','Research Director','Manufacturing Director')

staffs <- c('Research Scientist', 'Sales Representative', 'Laboratory Technician','Healthcare Representative','Human Resources')

hr_data$jobclass[hr_data$jobclass %in% directors]  <- 'Directors'

hr_data$jobclass[hr_data$jobclass %in% staffs]  <- 'Staff'

table(hr_data$jobclass)
```

```

Directors     Staff 
      653       817 
```

Hide

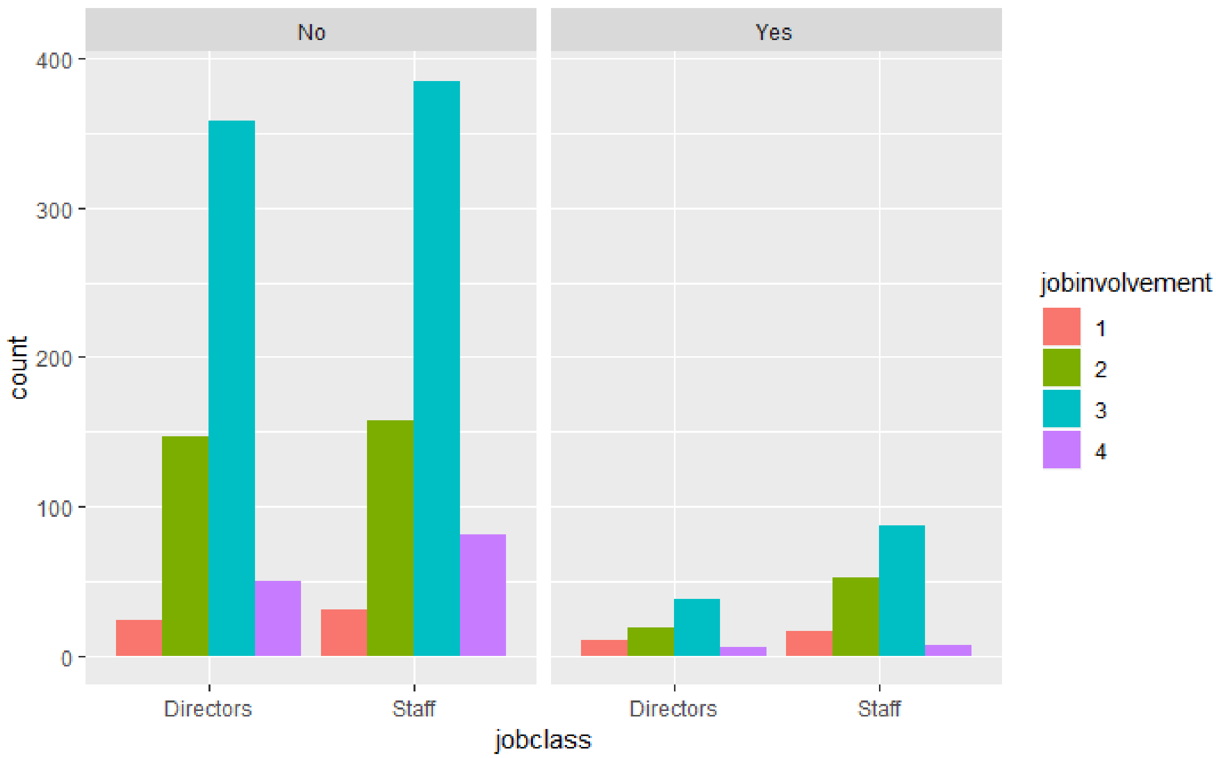
```
ggplot(hr_data,aes(x=Attrition,group=jobclass))+
  geom_bar(stat="count",aes(y=..prop..,fill=factor(..x..))) +
  labs(y="Percentage", fill = "Job Class") +
  facet_wrap(~jobclass)
```



As expected the staff of the company tend to leave more than the directors of the organization.

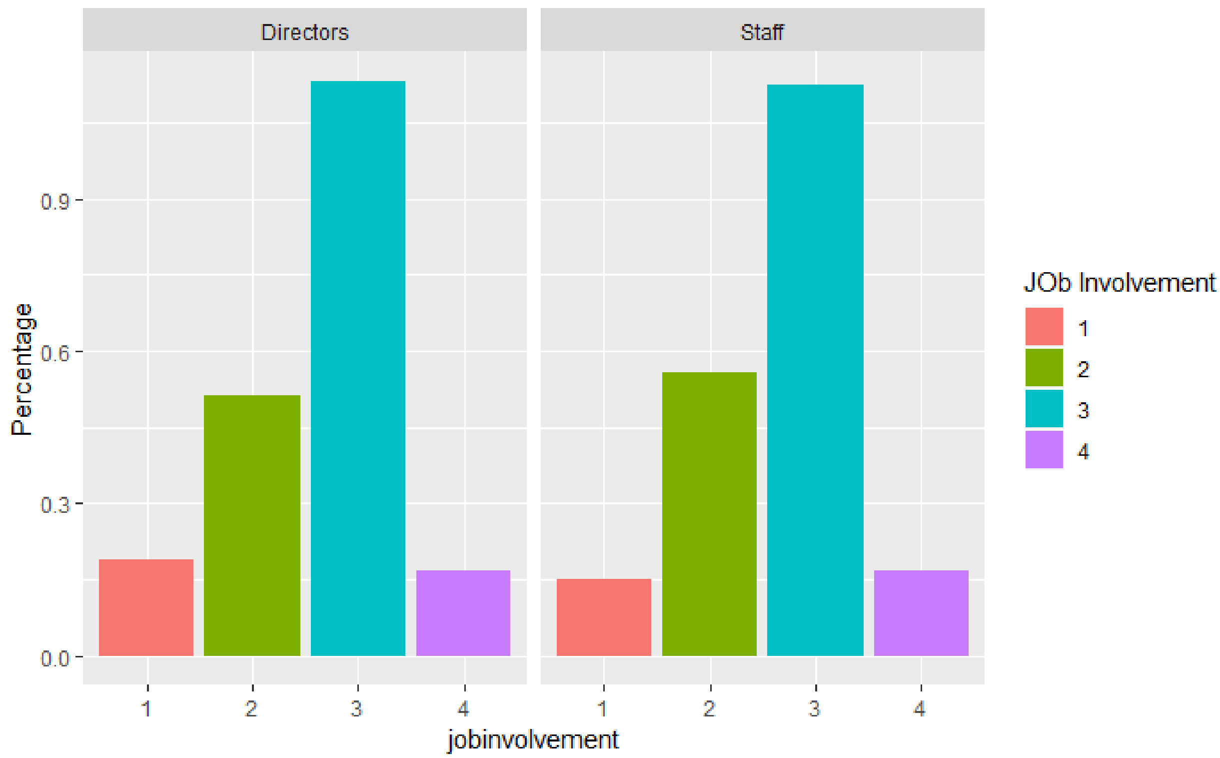
Hide

```
ggplot(hr_data, aes(jobclass, fill = jobinvolvement)) + geom_bar(stat= "count
", position =  position_dodge()) + facet_wrap(~Attrition)
```



Hide

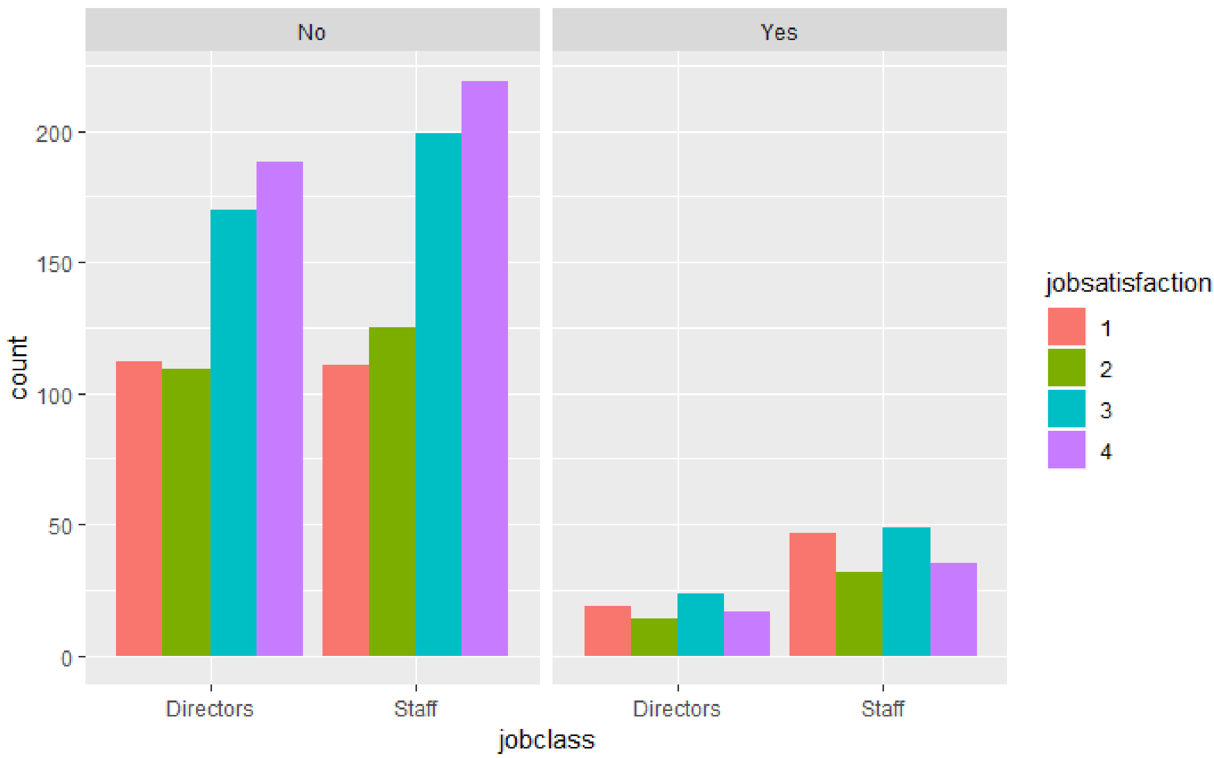
```
ggplot(hr_data,aes(x=jobinvolvement,group=Attrition))+
  geom_bar(stat="count",aes(y=..prop..,fill=factor(..x..))) +
  labs(y="Percentage", fill = "JOb Involvement") +
  facet_wrap(~jobclass)
```



The staffs have more job involvement than the directors. It appears that the with more job involvement, the employee is more likely to leave the organization.

Hide

```
ggplot(hr_data, aes(jobclass, fill = jobsatisfaction)) + geom_bar(stat= "coun
t", position =  position_dodge()) + facet_wrap(~Attrition)
```



It appears that the staffs are more satisfied with their jobs than the directors and the job satisfaction is not necessarily the factor that the staffs are leaving the organization.

Income level

Hide

```
#Income level

hr_data$Incomelevel <- as.factor(ifelse (hr_data$MonthlyIncome < ave(hr_data$
MonthlyIncome), "Low", "High"))

table(hr_data$Incomelevel, hr_data$Attrition)
```

```
        No  Yes
  High 441   52
  Low  792  185
```

Years without employee change

Hide

```
hr_data$Yearswithoutchange <- hr_data$totalworkingyears - hr_data$Yearssincel
astpromotion
```

```
str(hr_data$Yearswithoutchange)
```

```
 num [1:1470] 8 9 7 5 4 5 12 1 9 10 ...
```

Hide

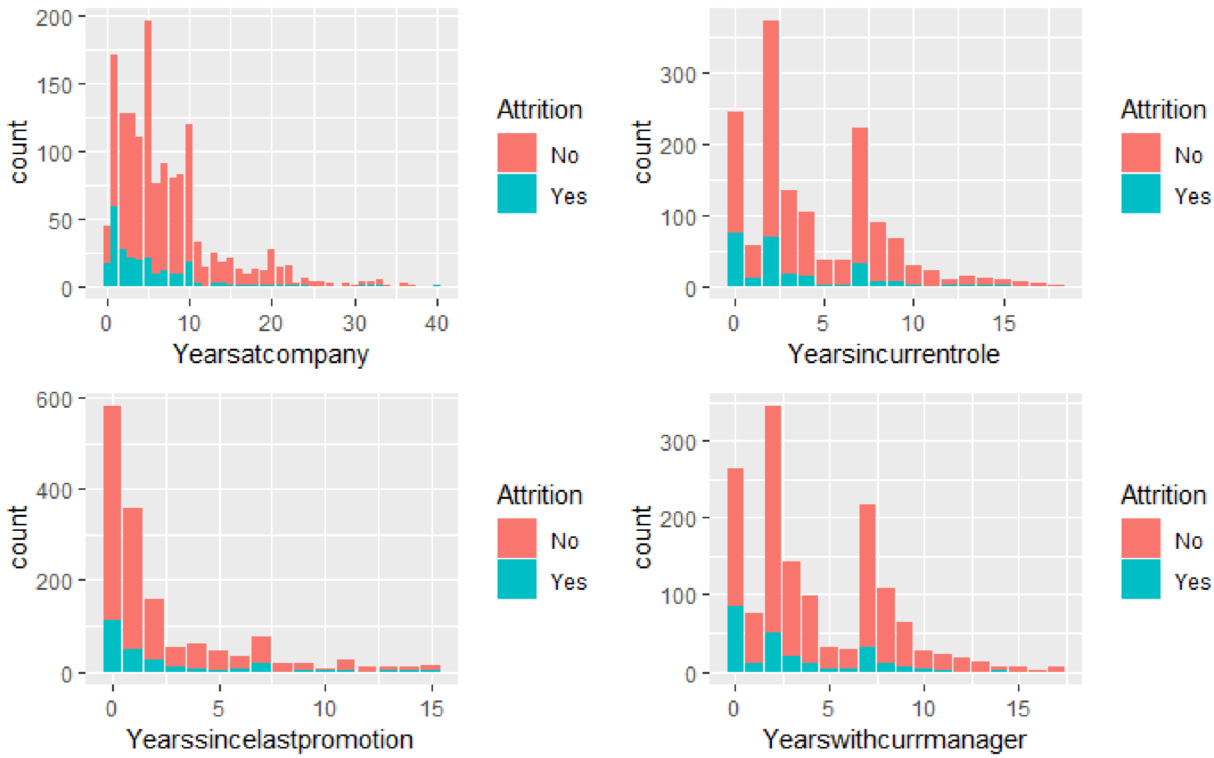
```
dist <- ggplot(hr_data,aes(DistancefromHome,fill=Attrition))+geom_bar()
edu <- ggplot(hr_data,aes(Education,fill=Attrition))+geom_bar()
edufield <- ggplot(hr_data,aes(educationfield,fill=Attrition))+geom_bar()
env <- ggplot(hr_data,aes(environmentsatisfaction,fill=Attrition))+geom_bar()
gen <- ggplot(hr_data,aes(Gender,fill=Attrition))+geom_bar()
grid.arrange(dist,edu,edufield,env,gen,ncol=2)
```

Hide

```
StockPlot <- ggplot(hr_data,aes(stockoptionlevel,fill = Attrition))+geom_bar(
)
workingYearsPlot <- ggplot(hr_data,aes(totalworkingyears,fill = Attrition))+g
eom_bar()
TrainTimesPlot <- ggplot(hr_data,aes(trainingtimeslastyear,fill = Attrition))
+geom_bar()
WLBPlot <- ggplot(hr_data,aes(worklifebalance,fill = Attrition))+geom_bar()
grid.arrange(StockPlot,workingYearsPlot,TrainTimesPlot,WLBPlot)
```

Hide

```
YearAtComPlot <- ggplot(hr_data,aes(Yearsatcompany,fill = Attrition))+geom_ba
r()

YearInCurrPlot <- ggplot(hr_data,aes(Yearsincurrentrole,fill = Attrition))+ge
om_bar()

YearsSinceProm <- ggplot(hr_data,aes(Yearssincelastpromotion,fill = Attrition
))+geom_bar()

YearsCurrManPlot <- ggplot(hr_data,aes(Yearswithcurrmanager,fill = Attrition)
)+geom_bar()

grid.arrange(YearAtComPlot,YearInCurrPlot,YearsSinceProm,YearsCurrManPlot,nco
l=2)
```



Data modeling

Divide the data into training and test dataset.

Data Preprocessing Convert characters to factors and remove

Hide

```
hr_dataclean <- hr_data

hr_dataclean = hr_dataclean[,!(names(hr_dataclean) %in% c('Marital Status','Job Role','Attrition (Yes/No)','Marital Status','Education Field','Business Travel','Job Involvement','Job Satisfaction','Job Level','Hourly Rate (USD)','Daily Rate (USD)','Monthly Rate (USD)','Monthly Income (USD)','Salary Hike (%)','Stock Option Level','Over Time','No. of Companies Worked','Total Working Years','Years At Company','Years In Current Role','Years Since Last Promotion','Years With Curr Manager','Environment Satisfaction','Training Times Last Year','Work Life Balance','Performance Rating','Relationship Satisfaction','Distance From Home (kms)'))]
```

Hide

```
hr_dataclean$Department <- as.factor(hr_dataclean$Department)

hr_dataclean$Education <- as.factor(hr_dataclean$Education)
```

```
hr_dataclean$Attrition <- as.factor(hr_dataclean$Attrition)

hr_dataclean$MaritalStatus <- as.factor(hr_dataclean$MaritalStatus)

hr_dataclean$BusinessTravel <- as.factor(hr_dataclean$BusinessTravel)

hr_dataclean$JobRole <- as.factor(hr_dataclean$JobRole)

hr_dataclean$educationfield <- as.factor(hr_dataclean$educationfield)

hr_dataclean$jobsatisfaction <- as.factor(hr_dataclean$jobsatisfaction)

hr_dataclean$jobinvolvement <- as.factor(hr_dataclean$jobinvolvement)

hr_dataclean$relationshipsatisfaction <- as.factor(hr_dataclean$relationships
atisfaction)

hr_dataclean$worklifebalance <- as.factor(hr_dataclean$worklifebalance)

hr_dataclean$environmentsatisfaction <- as.factor(hr_dataclean$environmentsat
isfaction)

hr_dataclean$overtime <- as.factor(hr_dataclean$overtime)

hr_dataclean$performancerating <- as.factor(hr_dataclean$performancerating)

hr_dataclean$jobclass <- as.factor(hr_dataclean$jobclass)

hr_dataclean$Gender <- as.factor(hr_dataclean$Gender)

hr_dataclean$stockoptionlevel <- as.factor(hr_dataclean$stockoptionlevel)

str(hr_dataclean)
```

```
'data.frame':   1470 obs. of  37 variables:

 $ Department              : Factor w/ 3 levels "Human Resources",..: 3 2 2 2
2 2 2 2 2 2 ...

 $ Gender                  : Factor w/ 2 levels "Female","Male": 1 2 2 1 2 2
1 2 2 2 ...

 $ Age                     : num  41 49 37 33 27 32 59 30 38 36 ...

 $ Education               : Factor w/ 5 levels "Bachelor","Below College",..
: 3 2 3 5 2 3 1 2 1 1 ...

 $ Attrition               : Factor w/ 2 levels "No","Yes": 2 1 2 1 1 1 1 1 1
1 ...

 $ MaritalStatus           : Factor w/ 3 levels "Divorced","Married",..: 3 2
3 2 2 3 2 1 3 2 ...

 $ BusinessTravel          : Factor w/ 3 levels "Non-Travel","Travel_Frequent
ly",..: 3 2 3 2 3 2 3 3 2 3 ...

 $ JobRole                 : Factor w/ 9 levels "Healthcare Representative",.
.: 8 7 3 7 3 3 3 3 5 1 ...

 $ DistancefromHome        : num  1 8 2 3 2 2 3 24 23 27 ...

 $ DailyRate               : num  1102 279 1373 1392 591 ...

 $ HourlyRate              : num  94 61 92 56 40 79 81 67 44 94 ...

 $ MonthlyRate             : num  19479 24907 2396 23159 16632 ...
```

```
 $ MonthlyIncome            : num  5993 5130 2090 2909 3468 ...
 $ Numcompworked            : num  8 1 6 1 9 0 4 1 0 6 ...
 $ Yearsatcompany           : num  6 10 0 8 2 7 1 1 9 7 ...
 $ Yearsincurrentrole       : num  4 7 0 7 2 7 0 0 7 7 ...
 $ Yearswithcurrmanager     : num  5 7 0 0 2 6 0 0 8 7 ...
 $ Yearssincelastpromotion  : num  0 1 0 3 2 3 0 0 1 7 ...
 $ totalworkingyears        : num  8 10 7 8 6 8 12 1 10 17 ...
 $ trainingtimeslastyear    : num  0 3 3 3 3 2 3 2 2 3 ...
 $ stockoptionlevel         : Factor w/ 4 levels "0","1","2","3": 1 2 1 1 2 1
4 2 1 3 ...
 $ salaryhike               : num  11 23 15 11 12 13 20 22 21 13 ...
 $ joblevel                 : Factor w/ 5 levels "1","2","3","4",..: 2 2 1 1 1
1 1 1 3 2 ...
 $ educationfield           : Factor w/ 6 levels "Human Resources",..: 2 2 5 2
4 2 4 2 2 4 ...
 $ jobsatisfaction          : Factor w/ 4 levels "1","2","3","4": 4 2 3 3 2 4
1 3 3 3 ...
 $ jobinvolvement           : Factor w/ 4 levels "1","2","3","4": 3 2 2 3 3 3
4 3 2 3 ...
 $ relationshipsatisfaction: Factor w/ 4 levels "1","2","3","4": 1 4 2 3 4 3
1 2 2 2 ...
 $ worklifebalance          : Factor w/ 4 levels "1","2","3","4": 1 3 3 3 3 2
2 3 3 2 ...
 $ environmentsatisfaction  : Factor w/ 4 levels "1","2","3","4": 2 3 4 4 1 4
3 4 4 3 ...
 $ overtime                 : Factor w/ 2 levels "No","Yes": 2 1 2 2 1 1 2 1 1
1 ...
 $ performancerating        : Factor w/ 2 levels "Excellent","Outstanding": 1
2 1 1 1 1 2 2 2 1 ...
 $ AgeGroup                 : Factor w/ 3 levels "Adult","Middle Age",..: 2 2
2 2 2 2 1 2 2 2 ...
 $ OverallSatisfaction      : num  10 11 11 13 10 14 9 12 11 11 ...
 $ OverallSatisfactionlevel: Factor w/ 2 levels "High","Low": 2 1 1 1 2 1 2 1
1 1 ...
 $ jobclass                 : Factor w/ 2 levels "Directors","Staff": 1 2 2 2
2 2 2 2 1 2 ...
 $ Incomelevel              : Factor w/ 2 levels "High","Low": 2 2 2 2 2 2 2 2
1 2 ...
 $ Yearswithoutchange       : num  8 9 7 5 4 5 12 1 9 10 ...
```

Partitioning the dataset

Hide

```
#Divide the data into training and test dataset.
set.seed(1337)
trainIndex <- createDataPartition(hr_dataclean$Attrition, p = 0.7, list = FAL
SE)
trainData <- hr_dataclean[trainIndex,]
testData  <- hr_dataclean[-trainIndex,]
```

uSING THE SMOTE METHOD TO balance classification

The data(Attrition is unbalanced)

Hide

```
prop.table(table(hr_data$Attrition))*100
```

```
      No      Yes
83.87755 16.12245
```

Hide

```
trainData <- as.data.frame(trainData)
smote_train <- SMOTE(Attrition ~ .,data=trainData)
smote_test <- SMOTE(Attrition ~ .,data=testData)
balanced_data =  prop.table(table(smote_train$Attrition))*100
cat("Balanced proportions is"); print(balanced_data, row.names=FALSE)
```

```
Balanced proportions is
      No      Yes
57.14286 42.85714
```

Hide

```
balanced_data1 =  prop.table(table(smote_train$Attrition))*100
cat("Balanced proportion of test is"); print(balanced_data1, row.names=FALSE)
```

```
Balanced proportion of test is
      No      Yes
```

```
57.14286 42.85714
```

The unbalanced data showed that 84% stayed as compared to 16% who left the organization, however doing a binary classification has balanced the data set and now we have 57% who did not leave as compared to 43% who left.

We will proceed to feature selection using the Boruta package. We can also use the lime package, but in this notebook, we will use the Boruta package.

Feature selection using Boruta

Hide

```
boruta_output=Boruta(Attrition~.,data=smote_train,doTrace=2)
```

```
 1. run of importance source...
 2. run of importance source...
 3. run of importance source...
 4. run of importance source...
 5. run of importance source...
 6. run of importance source...
 7. run of importance source...
 8. run of importance source...
 9. run of importance source...
 10. run of importance source...
 11. run of importance source...
 12. run of importance source...
After 12 iterations, +16 secs:
 confirmed 31 attributes: Age, AgeGroup, DailyRate, Department, DistancefromH
ome and 26 more;
 still have 5 attributes left.

 13. run of importance source...
 14. run of importance source...
 15. run of importance source...
 16. run of importance source...
After 16 iterations, +21 secs:
 confirmed 2 attributes: Education, educationfield;
 still have 3 attributes left.
```

```
 17. run of importance source...
 18. run of importance source...
 19. run of importance source...
 20. run of importance source...
 21. run of importance source...
 22. run of importance source...
 23. run of importance source...
 24. run of importance source...
 25. run of importance source...
 26. run of importance source...
 27. run of importance source...
 28. run of importance source...
 29. run of importance source...
After 29 iterations, +41 secs:
 confirmed 1 attribute: BusinessTravel;
 still have 2 attributes left.

 30. run of importance source...
 31. run of importance source...
 32. run of importance source...
 33. run of importance source...
 34. run of importance source...
 35. run of importance source...
 36. run of importance source...
 37. run of importance source...
 38. run of importance source...
 39. run of importance source...
 40. run of importance source...
 41. run of importance source...
 42. run of importance source...
 43. run of importance source...
 44. run of importance source...
 45. run of importance source...
 46. run of importance source...
```

```
 47. run of importance source...
 48. run of importance source...
 49. run of importance source...
 50. run of importance source...
 51. run of importance source...
 52. run of importance source...
 53. run of importance source...
 54. run of importance source...
 55. run of importance source...
 56. run of importance source...
 57. run of importance source...
 58. run of importance source...
 59. run of importance source...
 60. run of importance source...
 61. run of importance source...
 62. run of importance source...
After 62 iterations, +1.7 mins:
 confirmed 1 attribute: Gender;
 still have 1 attribute left.

 63. run of importance source...
 64. run of importance source...
 65. run of importance source...
 66. run of importance source...
 67. run of importance source...
 68. run of importance source...
 69. run of importance source...
 70. run of importance source...
 71. run of importance source...
 72. run of importance source...
 73. run of importance source...
 74. run of importance source...
 75. run of importance source...
 76. run of importance source...
```

```
77. run of importance source...
78. run of importance source...
79. run of importance source...
80. run of importance source...
81. run of importance source...
82. run of importance source...
83. run of importance source...
84. run of importance source...
85. run of importance source...
86. run of importance source...
87. run of importance source...
88. run of importance source...
89. run of importance source...
90. run of importance source...
91. run of importance source...
92. run of importance source...
93. run of importance source...
94. run of importance source...
95. run of importance source...
96. run of importance source...
97. run of importance source...
98. run of importance source...
99. run of importance source...
```

Hide

```
print(boruta_output)
```

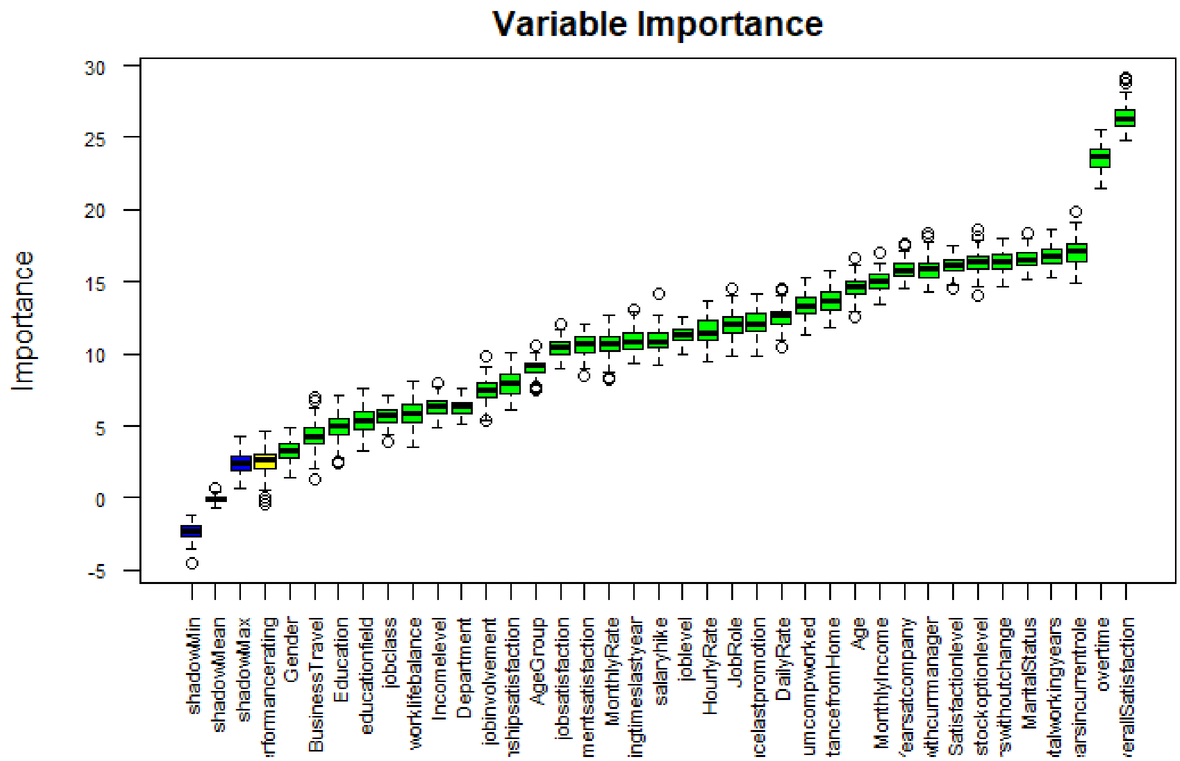
```
Boruta performed 99 iterations in 2.788494 mins.
 35 attributes confirmed important: Age, AgeGroup, BusinessTravel, DailyRate,
Department and 30 more;
 No attributes deemed unimportant.
 1 tentative attributes left: performancerating;
```

Tentative feature

Print out the new important features and display the boruta plot

Hide

```
#boruta_train <- TentativeRoughFix(boruta_output)

#cat("New important features", getSelectedAttributes(boruta_train), sep = "\n
")

plot(boruta_output, cex.axis=.7, las=2, xlab=" ", main="Variable Importance")
```



Display the boruta output statistics

Hide

```
boruta_stat <- attStats(boruta_output)
print(boruta_stat)
```

	meanImp <dbl>	medianImp <dbl>	minImp <dbl>	maxImp <dbl>	ne
Department	6.324061	6.345016	5.1756275	7.641933	1.(
Gender	3.178563	3.209117	1.4105569	4.834887	0.7
Age	14.611071	14.650566	12.5414386	16.587729	1.(
Education	4.866906	4.966679	2.4080745	7.077447	0.9
MaritalStatus	16.541453	16.453607	15.1490458	18.428802	1.(
BusinessTravel	4.292713	4.250748	1.3182773	7.029798	0.9

	meanImp	medianImp	minImp	maxImp	ne
	<dbl>	<dbl>	<dbl>	<dbl>	
JobRole	11.981308	12.051001	9.8576937	14.480192	1.0
DistancefromHome	13.646278	13.646047	11.8418852	15.806555	1.0
DailyRate	12.536528	12.632666	10.4782325	14.543952	1.0
HourlyRate	11.543864	11.406925	9.4358068	13.720417	1.0

Next
1234
Previous
1-10 of 36 rows

Removing unwanted features

Hide

```
#smote_trainrem = smote_train[,!(names(smote_train) %in% c("performancerating
"))]
```

Hide

```
set.seed(1337)

library(randomForest)

rf_varimportance <- randomForest(Attrition ~ Department + Age + Education + M
aritalStatus + BusinessTravel + JobRole + DistancefromHome + DailyRate + Hour
lyRate + MonthlyRate + MonthlyIncome +  Numcompworked + Yearsatcompany +  Yea
rssincelastpromotion +  trainingtimeslastyear + stockoptionlevel + salaryhike
+ educationfield + jobsatisfaction + jobinvolvement + relationshipsatisfactio
n + worklifebalance + environmentsatisfaction + overtime + AgeGroup + Overall
Satisfaction + OverallSatisfactionlevel + jobclass + Incomelevel + Yearswitho
utchange + joblevel + Gender , smote_trainrem, importance=TRUE,ntree=500)
```

Model fitting

basic Parameter tuning-Cross Validation

Hide

```
set.seed(1337)

trainControl <- trainControl(method = "cv", repeats = 10)
```

```
`repeats` has no meaning for this resampling method.
```

Hide

```
#Using the full dataset while ignoring the feature selection

##Logistic Regression
```

```r
fit_glm <- train(Attrition~. ,method="rf", data = smote_train, trControl = tr
ainControl)

confusionMatrix(fit_glm)
```

```
Cross-Validated (10 fold) Confusion Matrix


(entries are percentual average cell counts across resamples)


          Reference
Prediction   No  Yes
       No  54.0  7.9
       Yes  3.2 34.9


 Accuracy (average) : 0.889
```

Removing unimportant features

Hide

```r
#Logistic regression

fit_glm1 <- train(Attrition~. ,method="rf", data = smote_trainrem, trControl
= trainControl)

confusionMatrix(fit_glm1)
```

```
Cross-Validated (10 fold) Confusion Matrix


(entries are percentual average cell counts across resamples)


          Reference
Prediction   No  Yes
       No  53.7  7.4
       Yes  3.4 35.5


 Accuracy (average) : 0.8916
```

removing the correlated variables

Random Forest

Hide

```
# Removing total working years, years with current manager, years in current role

set.seed(1337)

model_rf <- train(Attrition ~ Department + Age + Education + MaritalStatus + BusinessTravel + JobRole + DistancefromHome + DailyRate + HourlyRate + MonthlyRate + MonthlyIncome + Numcompworked + Yearsatcompany + Yearssincelastpromotion +  trainingtimeslastyear + stockoptionlevel + salaryhike + educationfield + jobsatisfaction + jobinvolvement + relationshipsatisfaction + worklifebalance + environmentsatisfaction + overtime + AgeGroup + OverallSatisfaction + OverallSatisfactionlevel + jobclass + Incomelevel + Yearswithoutchange + joblevel + Gender, method = "rf", data = smote_trainrem, trControl = trainControl)

 confusionMatrix(model_rf)
```

```
Cross-Validated (10 fold) Confusion Matrix


(entries are percentual average cell counts across resamples)


          Reference
Prediction   No  Yes
       No  54.0  7.2
       Yes  3.2 35.6


 Accuracy (average) : 0.8959
```

Support vector machine

Hide

```
set.seed(1337)

model_svm <- train(Attrition ~ Department + Age + Education + MaritalStatus + BusinessTravel + JobRole + DistancefromHome + DailyRate + HourlyRate + MonthlyRate + MonthlyIncome + Numcompworked + Yearsatcompany + Yearssincelastpromotion +  trainingtimeslastyear + stockoptionlevel + salaryhike + educationfield + jobsatisfaction + jobinvolvement + relationshipsatisfaction + worklifebalance + environmentsatisfaction + overtime + AgeGroup + OverallSatisfaction + OverallSatisfactionlevel + jobclass + Incomelevel + Yearswithoutchange + joblevel + Gender, method = "svmRadial", data = smote_trainrem, trControl = trainControl)

 confusionMatrix(model_svm)
```

```
Cross-Validated (10 fold) Confusion Matrix


(entries are percentual average cell counts across resamples)
```

```
          Reference
Prediction   No  Yes
       No  51.0  8.9
       Yes  6.1 34.0

 Accuracy (average) : 0.8503
```

Logistic regression

Hide

```
set.seed(1337)

model_glm <- train(Attrition ~ Department + Age + Education + MaritalStatus +
BusinessTravel + JobRole + DistancefromHome + DailyRate + HourlyRate + Monthl
yRate + MonthlyIncome +  Numcompworked + Yearsatcompany  + Yearssincelastprom
otion +  trainingtimeslastyear + stockoptionlevel + salaryhike + educationfie
ld + jobsatisfaction + jobinvolvement + relationshipsatisfaction + worklifeba
lance + environmentsatisfaction + overtime + AgeGroup + OverallSatisfaction +
OverallSatisfactionlevel + jobclass + Incomelevel + joblevel + Yearswithoutch
ange + Gender , method = "glm", data = smote_trainrem, trControl = trainContr
ol, family = binomial(logit))

 confusionMatrix(model_glm)
```

```
Cross-Validated (10 fold) Confusion Matrix

(entries are percentual average cell counts across resamples)

          Reference
Prediction   No  Yes
       No  48.5 10.4
       Yes  8.6 32.4

 Accuracy (average) : 0.8098
```

Extreme Gradient Boost

Hide

```
library(xgboost)

set.seed(1337)

model_xgb <- train(Attrition ~ Department + Age + Education + MaritalStatus +
BusinessTravel + JobRole + DistancefromHome + DailyRate + HourlyRate + Monthl
```

```
yRate + MonthlyIncome +  Numcompworked + Yearsatcompany +  Yearssincelastprom
otion +  trainingtimeslastyear + stockoptionlevel + salaryhike + educationfie
ld + jobsatisfaction + jobinvolvement + relationshipsatisfaction + worklifeba
lance + environmentsatisfaction + overtime + AgeGroup + OverallSatisfaction +
OverallSatisfactionlevel + jobclass + Incomelevel + Yearswithoutchange + jobl
evel + Gender, method = "xgbTree", data = smote_trainrem, trControl = trainCo
ntrol)

 confusionMatrix(model_xgb)
```

```
Cross-Validated (10 fold) Confusion Matrix

(entries are percentual average cell counts across resamples)

          Reference
Prediction   No  Yes
       No  53.9  5.9
       Yes  3.3 36.9

 Accuracy (average) : 0.9079
```

Tuned Extreme gradient boost

Hide

```
set.seed(1337)

fitControl <- trainControl(method ="cv", number = 10)

xgbGrid <- expand.grid(nrounds = 50, max_depth = 12, eta = .03, gamma = 0.01,
colsample_bytree = .7, min_child_weight =1, subsample= 0.9)

model_xgb1 <- train(Attrition ~ Department + Age + Education + MaritalStatus
+ BusinessTravel + JobRole + DistancefromHome + DailyRate + HourlyRate + Mont
hlyRate + MonthlyIncome +  Numcompworked + Yearsatcompany +  Yearssincelastpr
omotion +  trainingtimeslastyear + stockoptionlevel + salaryhike + educationf
ield + jobsatisfaction + jobinvolvement + relationshipsatisfaction + worklife
balance + environmentsatisfaction + overtime + AgeGroup + OverallSatisfaction
+ OverallSatisfactionlevel + jobclass + Incomelevel + Yearswithoutchange + jo
blevel + Gender, method = "xgbTree", data = smote_trainrem, trControl = fitCo
ntrol, tuneGrid = xgbGrid)

 confusionMatrix(model_xgb1)
```

```
Cross-Validated (10 fold) Confusion Matrix

(entries are percentual average cell counts across resamples)
```

```
          Reference
Prediction   No  Yes
       No  54.4  8.0
       Yes  2.8 34.9

 Accuracy (average) : 0.8924
```

Hide

```
Predictions_xgb1 <- predict(model_xgb1, smote_test)

confusionMatrix(Predictions_xgb1, smote_test$Attrition)
```

```
Confusion Matrix and Statistics

          Reference
Prediction  No Yes
       No  260  64
       Yes  24 149

               Accuracy : 0.8229
                 95% CI : (0.7865, 0.8555)
    No Information Rate : 0.5714
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.6298
 Mcnemar's Test P-Value : 3.219e-05

            Sensitivity : 0.9155
            Specificity : 0.6995
         Pos Pred Value : 0.8025
         Neg Pred Value : 0.8613
             Prevalence : 0.5714
         Detection Rate : 0.5231
   Detection Prevalence : 0.6519
      Balanced Accuracy : 0.8075
```

```
       'Positive' Class : No
```

Hide

```
varImp(model_xgb)
```

```
xgbTree variable importance

  only 20 most important variables shown (out of 65)
```

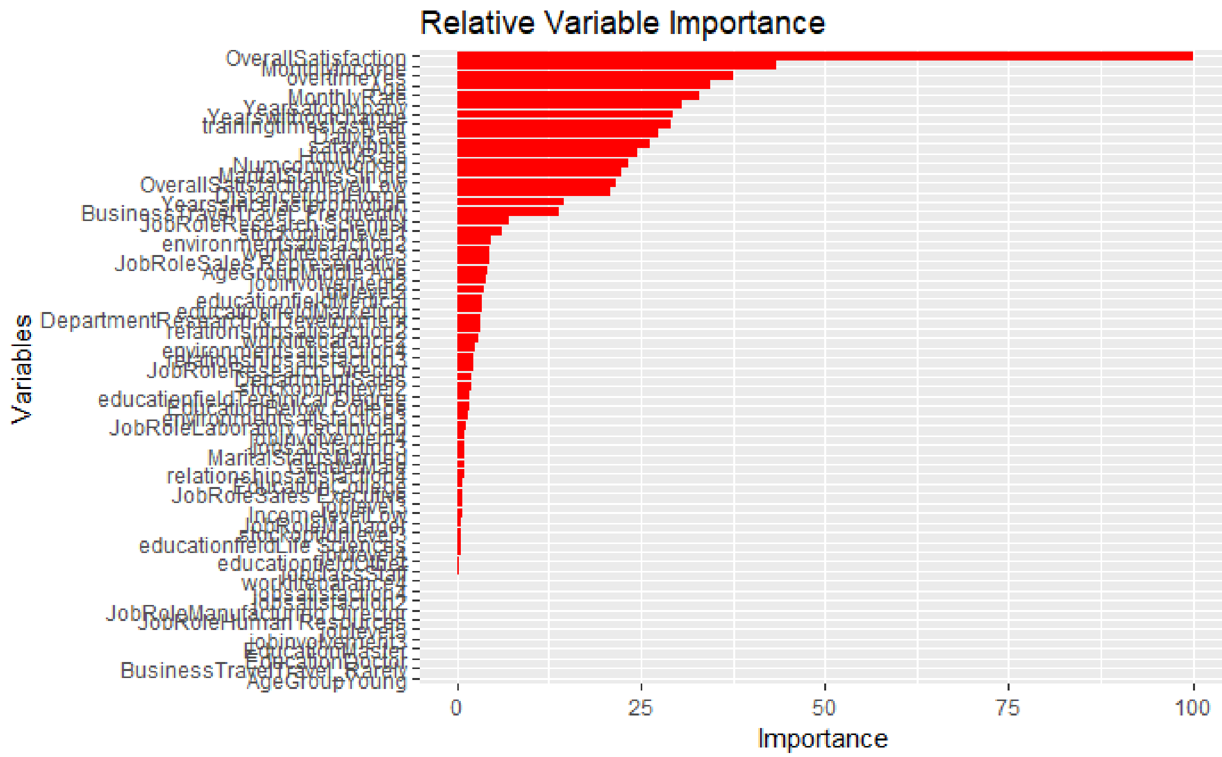
OverallSatisfaction
MonthlyIncome
overtimeYes
Age
MonthlyRate
Yearsatcompany
Yearswithoutchange
trainingtimeslastyear
DailyRate
salaryhike

Next
12
Previous
1-10 of 20 rows

Hide

```
importance <- varImp(model_xgb)
varImportance <- data.frame(Variables = row.names(importance[[1]]),
                            Importance = round(importance[[1]]$Overall,2))
rankImportance <- varImportance %>%
          mutate(Rank = paste0('Number',dense_rank(desc(Importance))))
          ggplot(rankImportance, aes(x = reorder(Variables, Importance), y =
Importance)) +
          geom_bar(stat='identity',fill = "red") +
        labs(x = 'Variables', title = 'Relative Variable Importance') +
```

```
        coord_flip()
```



Linear discriminant analysis

Hide

```
set.seed(1337)

model_lda <- train(Attrition ~ Department + Age + Education + MaritalStatus + BusinessTravel + JobRole + DistancefromHome + DailyRate + HourlyRate + MonthlyRate + MonthlyIncome + Numcompworked + Yearsatcompany + Yearssincelastpromotion +  trainingtimeslastyear + stockoptionlevel + salaryhike + educationfield + jobsatisfaction + jobinvolvement + relationshipsatisfaction + worklifebalance + environmentsatisfaction + overtime + AgeGroup + OverallSatisfaction + OverallSatisfactionlevel + jobclass + Incomelevel + Yearswithoutchange + joblevel + Gender, method = "lda", data = smote_trainrem, trControl = trainControl)

 confusionMatrix(model_lda)
```

```
Cross-Validated (10 fold) Confusion Matrix


(entries are percentual average cell counts across resamples)


          Reference

Prediction   No  Yes
```

```
       No   47.7 11.0
       Yes   9.5 31.8

 Accuracy (average) : 0.7952
```

Predictions

Hide

```
set.seed(1337)
Predictions_rf <- predict(model_rf, smote_test)
confusionMatrix(Predictions_rf, smote_test$Attrition)
```

```
Confusion Matrix and Statistics

          Reference
Prediction  No Yes
       No  259  70
       Yes  25 143

               Accuracy : 0.8089
                 95% CI : (0.7715, 0.8425)
    No Information Rate : 0.5714
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.5992
 Mcnemar's Test P-Value : 6.352e-06

            Sensitivity : 0.9120
            Specificity : 0.6714
         Pos Pred Value : 0.7872
         Neg Pred Value : 0.8512
             Prevalence : 0.5714
         Detection Rate : 0.5211
   Detection Prevalence : 0.6620
      Balanced Accuracy : 0.7917
```

```
'Positive' Class : No
```

Hide

```r
set.seed(1337)
Predictions_glm <- predict(model_glm, smote_test)
confusionMatrix(Predictions_glm, smote_test$Attrition)
```

```
Confusion Matrix and Statistics

          Reference
Prediction  No Yes
       No  271 107
       Yes  13 106

               Accuracy : 0.7586
                 95% CI : (0.7184, 0.7955)
    No Information Rate : 0.5714
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.4783
 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.9542
            Specificity : 0.4977
         Pos Pred Value : 0.7169
         Neg Pred Value : 0.8908
             Prevalence : 0.5714
         Detection Rate : 0.5453
   Detection Prevalence : 0.7606
      Balanced Accuracy : 0.7259

       'Positive' Class : No
```

Hide

```
set.seed(1337)
Predictions_svm <- predict(model_svm,smote_test)
confusionMatrix(Predictions_svm, smote_test$Attrition)
```

```
Confusion Matrix and Statistics

          Reference
Prediction  No Yes
       No  270 100
       Yes  14 113

               Accuracy : 0.7706
                 95% CI : (0.7311, 0.8069)
    No Information Rate : 0.5714
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.5068
 Mcnemar's Test P-Value : 1.707e-15

            Sensitivity : 0.9507
            Specificity : 0.5305
         Pos Pred Value : 0.7297
         Neg Pred Value : 0.8898
             Prevalence : 0.5714
         Detection Rate : 0.5433
   Detection Prevalence : 0.7445
      Balanced Accuracy : 0.7406

       'Positive' Class : No
```

Hide

```
Predictions_xgb <- predict(model_xgb, smote_test)
confusionMatrix(Predictions_xgb, smote_test$Attrition)
```

```
Confusion Matrix and Statistics
```

```
          Reference
Prediction  No Yes
       No  260  61
       Yes  24 152

               Accuracy : 0.829
                 95% CI : (0.7929, 0.861)
    No Information Rate : 0.5714
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.6431
 Mcnemar's Test P-Value : 9.432e-05

            Sensitivity : 0.9155
            Specificity : 0.7136
         Pos Pred Value : 0.8100
         Neg Pred Value : 0.8636
             Prevalence : 0.5714
         Detection Rate : 0.5231
   Detection Prevalence : 0.6459
      Balanced Accuracy : 0.8146

       'Positive' Class : No
```

Hide

```
set.seed(1337)
Predictions_lda <- predict(model_lda,smote_test)
confusionMatrix(Predictions_lda, smote_test$Attrition)
```

```
Confusion Matrix and Statistics

          Reference
Prediction  No Yes
       No  267  99
```

```
       Yes  17 114

               Accuracy : 0.7666
                 95% CI : (0.7269, 0.8031)
    No Information Rate : 0.5714
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.4994
 Mcnemar's Test P-Value : 5.45e-14

            Sensitivity : 0.9401
            Specificity : 0.5352
         Pos Pred Value : 0.7295
         Neg Pred Value : 0.8702
             Prevalence : 0.5714
         Detection Rate : 0.5372
   Detection Prevalence : 0.7364
      Balanced Accuracy : 0.7377

       'Positive' Class : No
```

Hide

```
roc_rf <- roc(as.numeric(smote_test$Attrition), as.numeric(Predictions_rf))
roc_rf$auc
```

```
Area under the curve: 0.7917
```

Hide

```
roc_svm <- roc(as.numeric(smote_test$Attrition), as.numeric(Predictions_svm))
roc_svm$auc
```

```
Area under the curve: 0.7406
```

Hide

```
roc_xgb <- roc(as.numeric(smote_test$Attrition), as.numeric(Predictions_xgb))
roc_xgb$auc
```

```
Area under the curve: 0.8146
```

Hide

```
roc_lda <- roc(as.numeric(smote_test$Attrition), as.numeric(Predictions_lda))
roc_lda$auc
```

```
Area under the curve: 0.7377
```

Hide

```
roc_glm <- roc(as.numeric(smote_test$Attrition), as.numeric(Predictions_glm))
roc_glm$auc
```

```
Area under the curve: 0.7259
```

Hide

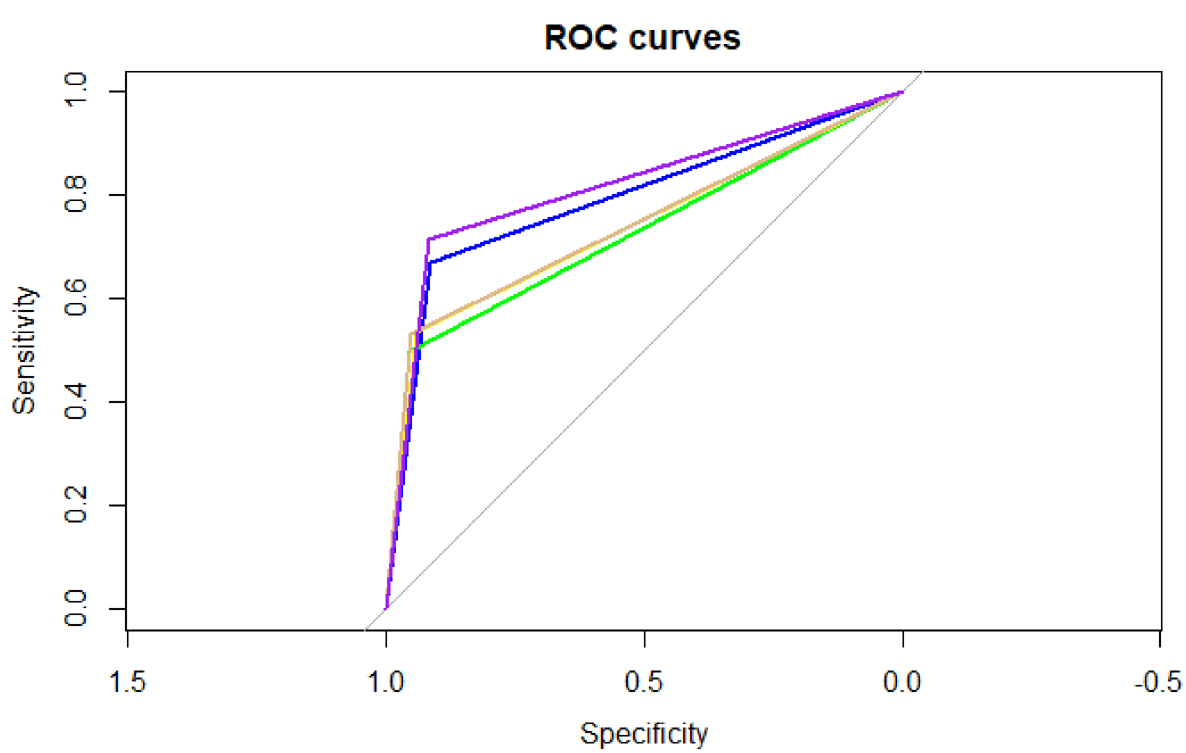
```
plot(roc_rf, ylim = c(0,1), main = "ROC curves", col = "blue")
plot(roc_glm, ylim = c(0,1),  col = "green", add = T)
```

Hide

```
plot(roc_lda, ylim = c(0,1),   col = "yellow", add = T)
plot(roc_svm, ylim = c(0,1),   col = "burlywood", add = T)
```

Hide

```
plot(roc_xgb, ylim = c(0,1),  col = "purple", add = T)
```



Hide

```
ggplot(smote_train,aes(Yearswithoutchange,fill=Attrition)) +
  geom_density(alpha=0.5)
```