My Code On Attrition

Ayush Jha

December 13, 2017

```r
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
library(ggplot2)
library(lattice)
library(ggthemes)
library(plyr)
```

```
## -------------------------------------------------------------------------

## You have loaded plyr after dplyr - this is likely to cause problems.
## If you need functions from both plyr and dplyr, please load plyr first, then dplyr:
## library(plyr); library(dplyr)

## -------------------------------------------------------------------------

##
## Attaching package: 'plyr'

## The following objects are masked from 'package:dplyr':
##
##     arrange, count, desc, failwith, id, mutate, rename, summarise,
##     summarize
```

```r
library(corrplot)
```

```
## corrplot 0.84 loaded
```

```r
library(forcats)
library(gridExtra)
```

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

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

```r
library(stringr)
library(caret)
library(formattable)
library(rpart)
library(rpart.plot)
library(Deducer)
```

```
## Loading required package: JGR

## Loading required package: rJava

## Loading required package: JavaGD

##
## Please type JGR() to launch console. Platform specific launchers (.exe and .app) can also be obtaine

## Loading required package: car

##
## Attaching package: 'car'

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

## Loading required package: MASS

##
## Attaching package: 'MASS'

## The following object is masked from 'package:formattable':
##
##     area

## The following object is masked from 'package:dplyr':
##
##     select

##
##
## Note Non-JGR console detected:
##  Deducer is best used from within JGR (http://jgr.markushelbig.org/).
##  To Bring up GUI dialogs, type deducer().
```

```r
library(Boruta)
```

```
## Loading required package: ranger
```

```r
library(DMwR)
```

```
## Loading required package: grid

##
## Attaching package: 'DMwR'

## The following object is masked from 'package:plyr':
##
##     join
```

```r
library(DT)
library(ROCR)
```

```
## Loading required package: gplots

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

```r
library(dummy)
```

```
## dummy 0.1.3
```

```
## dummyNews()
```

```r
library(caretEnsemble)
```

```
##
## Attaching package: 'caretEnsemble'
```

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

```r
library(caret)
```

```r
setwd("D:/My Datasets Library/ibm-hr-analytics-employee-attrition-performance")
```

We have been provided with the HR employee attrition data and build a model to predict the attrition.

1.Data Load

importing data using read_csv function

```r
library(readr)
```

```r
myds <- read.csv("D:/My Datasets Library/ibm-hr-analytics-employee-attrition-performance/WA_Fn-UseC_-HR
```

```r
View(myds)
```

2.Data Sanity Check

looking at summary

```r
summary.data.frame(myds)
```

```
##      ï..Age      Attrition            BusinessTravel   DailyRate
##  Min.   :18.00   No :1233   Non-Travel       : 150   Min.   : 102.0
##  1st Qu.:30.00   Yes: 237   Travel_Frequently: 277   1st Qu.: 465.0
##  Median :36.00              Travel_Rarely    :1043   Median : 802.0
##  Mean   :36.92                                       Mean   : 802.5
##  3rd Qu.:43.00                                       3rd Qu.:1157.0
##  Max.   :60.00                                       Max.   :1499.0
##
##                    Department  DistanceFromHome   Education
##  Human Resources       : 63   Min.   : 1.000   Min.   :1.000
##  Research & Development:961   1st Qu.: 2.000   1st Qu.:2.000
##  Sales                 :446   Median : 7.000   Median :3.000
##                               Mean   : 9.193   Mean   :2.913  
##                               3rd Qu.:14.000   3rd Qu.:4.000  
##                               Max.   :29.000   Max.   :5.000  
##                                                               
##            EducationField EmployeeCount EmployeeNumber
##  Human Resources : 27    Min.   :1     Min.   :   1.0  
##  Life Sciences   :606    1st Qu.:1     1st Qu.: 491.2  
##  Marketing       :159    Median :1     Median :1020.5  
##  Medical         :464    Mean   :1     Mean   :1024.9  
##  Other           : 82    3rd Qu.:1     3rd Qu.:1555.8  
##  Technical Degree:132    Max.   :1     Max.   :2068.0  
##                                                        
##  EnvironmentSatisfaction    Gender      HourlyRate     JobInvolvement
##  Min.   :1.000           Female:588   Min.   : 30.00   Min.   :1.00  
##  1st Qu.:2.000           Male  :882   1st Qu.: 48.00   1st Qu.:2.00  
##  Median :3.000                        Median : 66.00   Median :3.00  
##  Mean   :2.722                        Mean   : 65.89   Mean   :2.73  
##  3rd Qu.:4.000                        3rd Qu.: 83.75   3rd Qu.:3.00  
##  Max.   :4.000                        Max.   :100.00   Max.   :4.00  
##                                                                      
##     JobLevel                          JobRole    JobSatisfaction
##  Min.   :1.000   Sales Executive          :326   Min.   :1.000  
##  1st Qu.:1.000   Research Scientist       :292   1st Qu.:2.000  
##  Median :2.000   Laboratory Technician    :259   Median :3.000  
##  Mean   :2.064   Manufacturing Director   :145   Mean   :2.729  
##  3rd Qu.:3.000   Healthcare Representative:131   3rd Qu.:4.000  
##  Max.   :5.000   Manager                  :102   Max.   :4.000  
##                  (Other)                  :215                  
##   MaritalStatus MonthlyIncome    MonthlyRate    NumCompaniesWorked
##  Divorced:327   Min.   : 1009   Min.   : 2094   Min.   :0.000     
##  Married :673   1st Qu.: 2911   1st Qu.: 8047   1st Qu.:1.000     
##  Single  :470   Median : 4919   Median :14236   Median :2.000     
##                 Mean   : 6503   Mean   :14313   Mean   :2.693     
##                 3rd Qu.: 8379   3rd Qu.:20462   3rd Qu.:4.000     
##                 Max.   :19999   Max.   :26999   Max.   :9.000     
##                                                                   
##  Over18   OverTime   PercentSalaryHike PerformanceRating
##  Y:1470   No :1054   Min.   :11.00     Min.   :3.000    
##           Yes: 416   1st Qu.:12.00     1st Qu.:3.000    
##                      Median :14.00     Median :3.000    
##                      Mean   :15.21     Mean   :3.154    
##                      3rd Qu.:18.00     3rd Qu.:3.000    
##                      Max.   :25.00     Max.   :4.000    
##                                                         
##  RelationshipSatisfaction StandardHours StockOptionLevel TotalWorkingYears
##  Min.   :1.000            Min.   :80    Min.   :0.0000   Min.   : 0.00    
##  1st Qu.:2.000            1st Qu.:80    1st Qu.:0.0000   1st Qu.: 6.00    
##  Median :3.000            Median :80    Median :1.0000   Median :10.00    
##  Mean   :2.712            Mean   :80    Mean   :0.7939   Mean   :11.28    
##  3rd Qu.:4.000            3rd Qu.:80    3rd Qu.:1.0000   3rd Qu.:15.00    
##  Max.   :4.000            Max.   :80    Max.   :3.0000   Max.   :40.00    
##                                                                           
##  TrainingTimesLastYear WorkLifeBalance YearsAtCompany   YearsInCurrentRole
##  Min.   :0.000         Min.   :1.000   Min.   : 0.000   Min.   : 0.000    
##  1st Qu.:2.000         1st Qu.:2.000   1st Qu.: 3.000   1st Qu.: 2.000  
##  Median :3.000         Median :3.000   Median : 5.000   Median : 3.000  
##  Mean   :2.799         Mean   :2.761   Mean   : 7.008   Mean   : 4.229  
##  3rd Qu.:3.000         3rd Qu.:3.000   3rd Qu.: 9.000   3rd Qu.: 7.000  
##  Max.   :6.000         Max.   :4.000   Max.   :40.000   Max.   :18.000  
##                                                                         
##  YearsSinceLastPromotion YearsWithCurrManager
##  Min.   : 0.000          Min.   : 0.000      
##  1st Qu.: 0.000          1st Qu.: 2.000      
##  Median : 1.000          Median : 3.000      
##  Mean   : 2.188          Mean   : 4.123      
##  3rd Qu.: 3.000          3rd Qu.: 7.000      
##  Max.   :15.000          Max.   :17.000      
## 
```

another way to look into summary

```
summary(myds)
```

```
##      ï..Age      Attrition            BusinessTravel   DailyRate     
##  Min.   :18.00   No :1233   Non-Travel       : 150   Min.   : 102.0  
##  1st Qu.:30.00   Yes: 237   Travel_Frequently: 277   1st Qu.: 465.0  
##  Median :36.00              Travel_Rarely    :1043   Median : 802.0  
##  Mean   :36.92                                       Mean   : 802.5  
##  3rd Qu.:43.00                                       3rd Qu.:1157.0  
##  Max.   :60.00                                       Max.   :1499.0  
##                                                                      
##                   Department  DistanceFromHome   Education    
##  Human Resources       : 63   Min.   : 1.000   Min.   :1.000  
##  Research & Development:961   1st Qu.: 2.000   1st Qu.:2.000  
##  Sales                 :446   Median : 7.000   Median :3.000  
##                               Mean   : 9.193   Mean   :2.913  
##                               3rd Qu.:14.000   3rd Qu.:4.000  
##                               Max.   :29.000   Max.   :5.000  
##                                                               
##           EducationField EmployeeCount EmployeeNumber 
##  Human Resources : 27    Min.   :1     Min.   :   1.0  
##  Life Sciences   :606    1st Qu.:1     1st Qu.: 491.2  
##  Marketing       :159    Median :1     Median :1020.5  
##  Medical         :464    Mean   :1     Mean   :1024.9  
##  Other           : 82    3rd Qu.:1     3rd Qu.:1555.8  
##  Technical Degree:132    Max.   :1     Max.   :2068.0  
##                                                        
##  EnvironmentSatisfaction    Gender      HourlyRate     JobInvolvement
##  Min.   :1.000           Female:588   Min.   : 30.00   Min.   :1.00  
##  1st Qu.:2.000           Male  :882   1st Qu.: 48.00   1st Qu.:2.00  
##  Median :3.000                        Median : 66.00   Median :3.00  
##  Mean   :2.722                        Mean   : 65.89   Mean   :2.73  
##  3rd Qu.:4.000                        3rd Qu.: 83.75   3rd Qu.:3.00  
##  Max.   :4.000                        Max.   :100.00   Max.   :4.00  
##                                                                      
##     JobLevel                          JobRole    JobSatisfaction
##  Min.   :1.000   Sales Executive          :326   Min.   :1.000  
##  1st Qu.:1.000   Research Scientist       :292   1st Qu.:2.000  
##  Median :2.000   Laboratory Technician    :259   Median :3.000
##  Mean   :2.064   Manufacturing Director   :145   Mean   :2.729
##  3rd Qu.:3.000   Healthcare Representative:131   3rd Qu.:4.000
##  Max.   :5.000   Manager                  :102   Max.   :4.000
##                  (Other)                  :215
##   MaritalStatus MonthlyIncome    MonthlyRate    NumCompaniesWorked
##  Divorced:327   Min.   : 1009   Min.   : 2094   Min.   :0.000
##  Married :673   1st Qu.: 2911   1st Qu.: 8047   1st Qu.:1.000
##  Single  :470   Median : 4919   Median :14236   Median :2.000
##                 Mean   : 6503   Mean   :14313   Mean   :2.693
##                 3rd Qu.: 8379   3rd Qu.:20462   3rd Qu.:4.000
##                 Max.   :19999   Max.   :26999   Max.   :9.000
##
##  Over18   OverTime   PercentSalaryHike PerformanceRating
##  Y:1470   No :1054   Min.   :11.00     Min.   :3.000
##           Yes: 416   1st Qu.:12.00     1st Qu.:3.000
##                      Median :14.00     Median :3.000
##                      Mean   :15.21     Mean   :3.154
##                      3rd Qu.:18.00     3rd Qu.:3.000
##                      Max.   :25.00     Max.   :4.000
##
##  RelationshipSatisfaction StandardHours StockOptionLevel TotalWorkingYears
##  Min.   :1.000            Min.   :80    Min.   :0.0000   Min.   : 0.00
##  1st Qu.:2.000            1st Qu.:80    1st Qu.:0.0000   1st Qu.: 6.00
##  Median :3.000            Median :80    Median :1.0000   Median :10.00
##  Mean   :2.712            Mean   :80    Mean   :0.7939   Mean   :11.28
##  3rd Qu.:4.000            3rd Qu.:80    3rd Qu.:1.0000   3rd Qu.:15.00
##  Max.   :4.000            Max.   :80    Max.   :3.0000   Max.   :40.00
##
##  TrainingTimesLastYear WorkLifeBalance YearsAtCompany   YearsInCurrentRole
##  Min.   :0.000         Min.   :1.000   Min.   : 0.000   Min.   : 0.000
##  1st Qu.:2.000         1st Qu.:2.000   1st Qu.: 3.000   1st Qu.: 2.000
##  Median :3.000         Median :3.000   Median : 5.000   Median : 3.000
##  Mean   :2.799         Mean   :2.761   Mean   : 7.008   Mean   : 4.229
##  3rd Qu.:3.000         3rd Qu.:3.000   3rd Qu.: 9.000   3rd Qu.: 7.000
##  Max.   :6.000         Max.   :4.000   Max.   :40.000   Max.   :18.000
##
##  YearsSinceLastPromotion YearsWithCurrManager
##  Min.   : 0.000          Min.   : 0.000
##  1st Qu.: 0.000          1st Qu.: 2.000
##  Median : 1.000          Median : 3.000
##  Mean   : 2.188          Mean   : 4.123
##  3rd Qu.: 3.000          3rd Qu.: 7.000
##  Max.   :15.000          Max.   :17.000
##
```

```
dim(myds)
```

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

3. Check the missing value (if any)

```
sum(is.na(myds))
```

```
## [1] 0
```

we get no presence of missing value or NA value.

4. Chekcing Variable types

```r
str(myds)
```

```
## 'data.frame':    1470 obs. of  35 variables:
##  $ ï..Age                  : int  41 49 37 33 27 32 59 30 38 36 ...
##  $ Attrition               : Factor w/ 2 levels "No","Yes": 2 1 2 1 1 1 1 1 1 1 ...
##  $ BusinessTravel          : Factor w/ 3 levels "Non-Travel","Travel_Frequently",..: 3 2 3 2 3 2 3 3
##  $ DailyRate               : int  1102 279 1373 1392 591 1005 1324 1358 216 1299 ...
##  $ Department              : Factor w/ 3 levels "Human Resources",..: 3 2 2 2 2 2 2 2 2 2 ...
##  $ DistanceFromHome        : int  1 8 2 3 2 2 3 24 23 27 ...
##  $ Education               : int  2 1 2 4 1 2 3 1 3 3 ...
##  $ EducationField          : Factor w/ 6 levels "Human Resources",..: 2 2 5 2 4 2 4 2 2 4 ...
##  $ EmployeeCount           : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ EmployeeNumber          : int  1 2 4 5 7 8 10 11 12 13 ...
##  $ EnvironmentSatisfaction : int  2 3 4 4 1 4 3 4 4 3 ...
##  $ Gender                  : Factor w/ 2 levels "Female","Male": 1 2 2 1 2 2 1 2 2 2 ...
##  $ HourlyRate              : int  94 61 92 56 40 79 81 67 44 94 ...
##  $ JobInvolvement          : int  3 2 2 3 3 3 4 3 2 3 ...
##  $ JobLevel                : int  2 2 1 1 1 1 1 1 3 2 ...
##  $ JobRole                 : Factor w/ 9 levels "Healthcare Representative",..: 8 7 3 7 3 3 3 3 5 1
##  $ JobSatisfaction         : int  4 2 3 3 2 4 1 3 3 3 ...
##  $ MaritalStatus           : Factor w/ 3 levels "Divorced","Married",..: 3 2 3 2 2 3 2 1 3 2 ...
##  $ MonthlyIncome           : int  5993 5130 2090 2909 3468 3068 2670 2693 9526 5237 ...
##  $ MonthlyRate             : int  19479 24907 2396 23159 16632 11864 9964 13335 8787 16577 ...
##  $ NumCompaniesWorked      : int  8 1 6 1 9 0 4 1 0 6 ...
##  $ Over18                  : Factor w/ 1 level "Y": 1 1 1 1 1 1 1 1 1 1 ...
##  $ OverTime                : Factor w/ 2 levels "No","Yes": 2 1 2 2 1 1 2 1 1 1 ...
##  $ PercentSalaryHike       : int  11 23 15 11 12 13 20 22 21 13 ...
##  $ PerformanceRating       : int  3 4 3 3 3 3 4 4 4 3 ...
##  $ RelationshipSatisfaction: int  1 4 2 3 4 3 1 2 2 2 ...
##  $ StandardHours           : int  80 80 80 80 80 80 80 80 80 80 ...
##  $ StockOptionLevel        : int  0 1 0 0 1 0 3 1 0 2 ...
##  $ TotalWorkingYears       : int  8 10 7 8 6 8 12 1 10 17 ...
##  $ TrainingTimesLastYear   : int  0 3 3 3 3 2 3 2 2 3 ...
##  $ WorkLifeBalance         : int  1 3 3 3 3 2 2 3 3 2 ...
##  $ YearsAtCompany          : int  6 10 0 8 2 7 1 1 9 7 ...
##  $ YearsInCurrentRole      : int  4 7 0 7 2 7 0 0 7 7 ...
##  $ YearsSinceLastPromotion : int  0 1 0 3 2 3 0 0 1 7 ...
##  $ YearsWithCurrManager    : int  5 7 0 0 2 6 0 0 8 7 ...
```

we see how many rows and types are there. All are purely either intergers or character.

5.Removing Unwanted rows

From this overview we find that columns like **####over18,employeecount,standardhours* are not informative,so we removed it.**

we count number of rows

```
cat("No of Columns before removing:",ncol(myds),sep="/n")
```

```
## No of Columns before removing:/n35
```

Removed lines

```
myds1=myds[,!(names(myds) %in% c('Over18','EmployeeCount','StandardHours'))]
```

Count lines

```
cat("No of Columns after removing:",ncol(myds),sep="/n")
```

```
## No of Columns after removing:/n35
```

6.Removing rows with missing data (just in case we need to do)

removing the rows with missing values

```
nrow(data)
```

```
## NULL
```

```
data<- na.omit(data) ## removes the missing values
```

```
nrow(data)
```

```
## NULL
```

We dont have missing values

If we have to do missing value treatment:

mean imputation

median imputation

mode imputation

regression imputation

installing caret packï..Age

if only specific columns you want to keep .

```
myds3 <- data.frame(Attrition=rnorm(100)>0,OverTime=rnorm(100)>0)
head(myds)
```

```
##   ï..Age Attrition    BusinessTravel DailyRate            Department
## 1     41       Yes     Travel_Rarely      1102                  Sales
## 2     49        No Travel_Frequently       279 Research & Development
## 3     37       Yes     Travel_Rarely      1373 Research & Development
## 4     33        No Travel_Frequently      1392 Research & Development
## 5     27        No     Travel_Rarely       591 Research & Development
## 6     32        No Travel_Frequently      1005 Research & Development
##   DistanceFromHome Education EducationField EmployeeCount EmployeeNumber
## 1                1         2  Life Sciences             1              1
## 2                8         1  Life Sciences             1              2
## 3                2         2          Other             1              4
## 4                3         4  Life Sciences             1              5
## 5                2         1        Medical             1              7
## 6                2         2  Life Sciences             1              8
##   EnvironmentSatisfaction Gender HourlyRate JobInvolvement JobLevel
## 1                       2 Female         94              3        2
## 2                       3   Male         61              2        2
## 3                       4   Male         92              2        1
## 4                       4 Female         56              3        1
## 5                       1   Male         40              3        1
## 6                       4   Male         79              3        1
##                  JobRole JobSatisfaction MaritalStatus MonthlyIncome
## 1       Sales Executive               4        Single          5993
## 2    Research Scientist               2       Married          5130
## 3 Laboratory Technician               3        Single          2090
## 4    Research Scientist               3       Married          2909
## 5 Laboratory Technician               2       Married          3468
## 6 Laboratory Technician               4        Single          3068
##   MonthlyRate NumCompaniesWorked Over18 OverTime PercentSalaryHike
## 1       19479                  8      Y      Yes                11
## 2       24907                  1      Y       No                23
## 3        2396                  6      Y      Yes                15
## 4       23159                  1      Y      Yes                11
## 5       16632                  9      Y       No                12
## 6       11864                  0      Y       No                13
##   PerformanceRating RelationshipSatisfaction StandardHours
## 1                 3                        1            80
## 2                 4                        4            80
## 3                 3                        2            80
## 4                 3                        3            80
## 5                 3                        4            80
## 6                 3                        3            80
##   StockOptionLevel TotalWorkingYears TrainingTimesLastYear WorkLifeBalance
## 1                0                 8                     0               1
## 2                1                10                     3               3
## 3                0                 7                     3               3
## 4                0                 8                     3               3
## 5                1                 6                     3               3
## 6                0                 8                     2               2
##   YearsAtCompany YearsInCurrentRole YearsSinceLastPromotion
## 1              6                  4                       0
## 2             10                  7                       1
## 3              0                  0                       0
## 4              8                  7                       3
## 5              2                  2                       2
## 6              7                  7                       3
##   YearsWithCurrManager
## 1                    5
## 2                    7
## 3                    0
## 4                    0
## 5                    2
## 6                    6
```

testing true flase for variable being numeric

```r
sapply(myds, is.numeric)
```

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

7 changing values to numerica value from text to make it easy to use.

```r
myds$Attrition <- revalue(myds$Attrition, c("Yes"= 1))
```

```r
myds$Attrition <- revalue(myds$Attrition, c("No"= 0))
```

```r
head(myds$Attrition)
```

```
## [1] 1 0 1 0 0 0
## Levels: 0 1
```

```r
myds$OverTime <- revalue(myds$OverTime , c("Yes"= 1))
```

```r
myds$OverTime  <- revalue(myds$OverTime , c("No"= 0))
```

```
head(myds$OverTime )
```

```
## [1] 1 0 1 1 0 0
## Levels: 0 1
```

8.

very important, after turning value to numeric , change coloumn category also to numeric

```
myds$Attrition <- as.numeric(myds$Attrition)
```

```
myds$OverTime <- as.numeric(myds$OverTime)
```

some extra ways to convert values to numeric

```
#myds$Attrition [myds$Attrition == "Yes"] <- 1
#myds$Attrition [myds$Attrition == "No"] <- 0
```

8 lets first see attriation percentï..Age rate

```
round((prop.table(table(myds$Attrition)))*100,2)
```

```
##
##     1     2
## 83.88 16.12
```

this shows 16% attriation oocured yet.

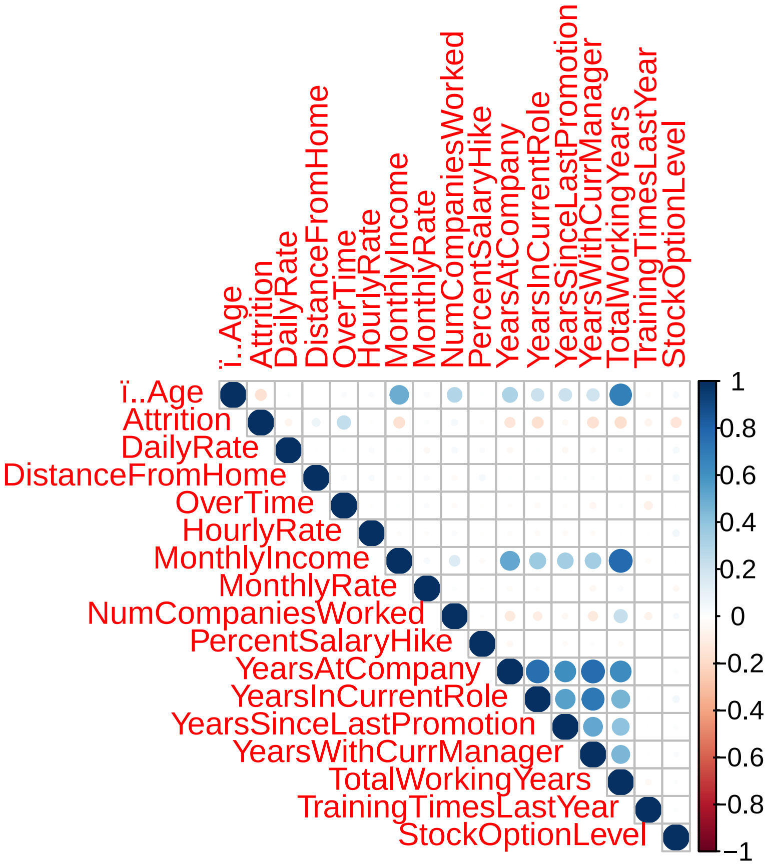
Exploratory Data Analysis

we will do bivariae and univariate analysis to see variables.

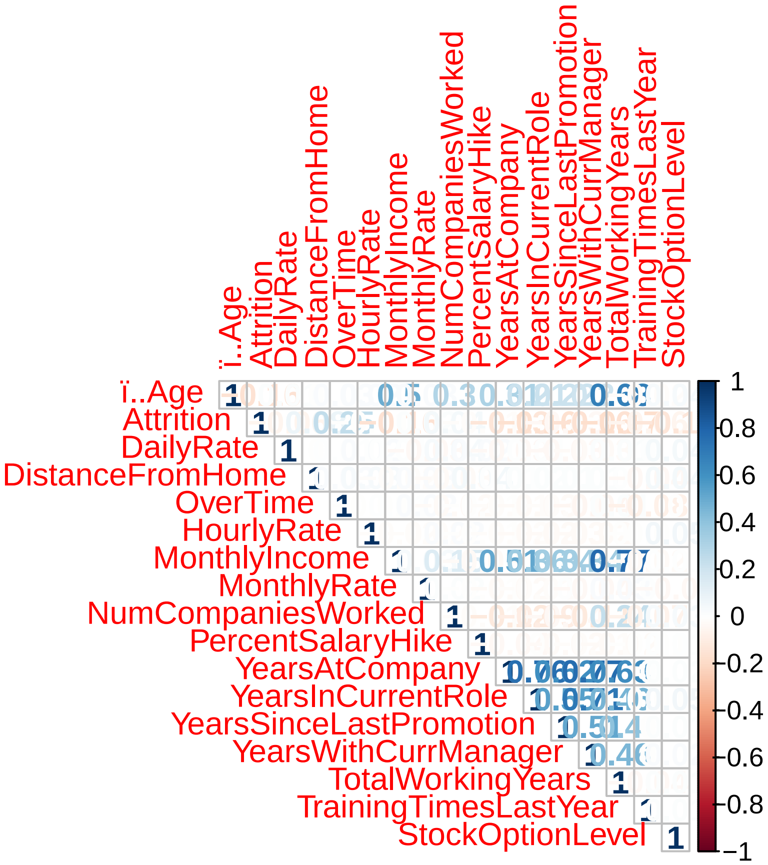
9

Correlation Plot

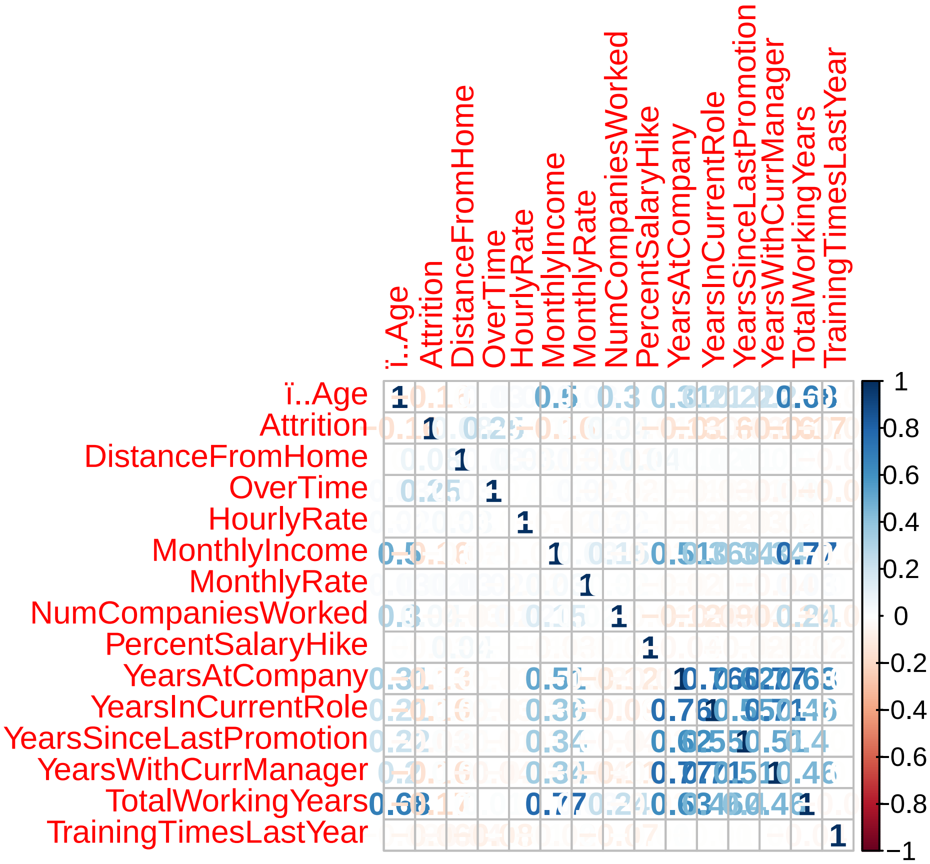
```
numeric=myds%>% dplyr::select(ï..Age,Attrition,DailyRate,DistanceFromHome,OverTime,HourlyRate,MonthlyIn
corrplot(cor(numeric),method="circle",type="upper")
```



```
numeric=myds%>% dplyr::select(ï..Age,Attrition,DailyRate,DistanceFromHome,OverTime,HourlyRate,MonthlyInc
corrplot(cor(numeric),method="number",type="upper")
```

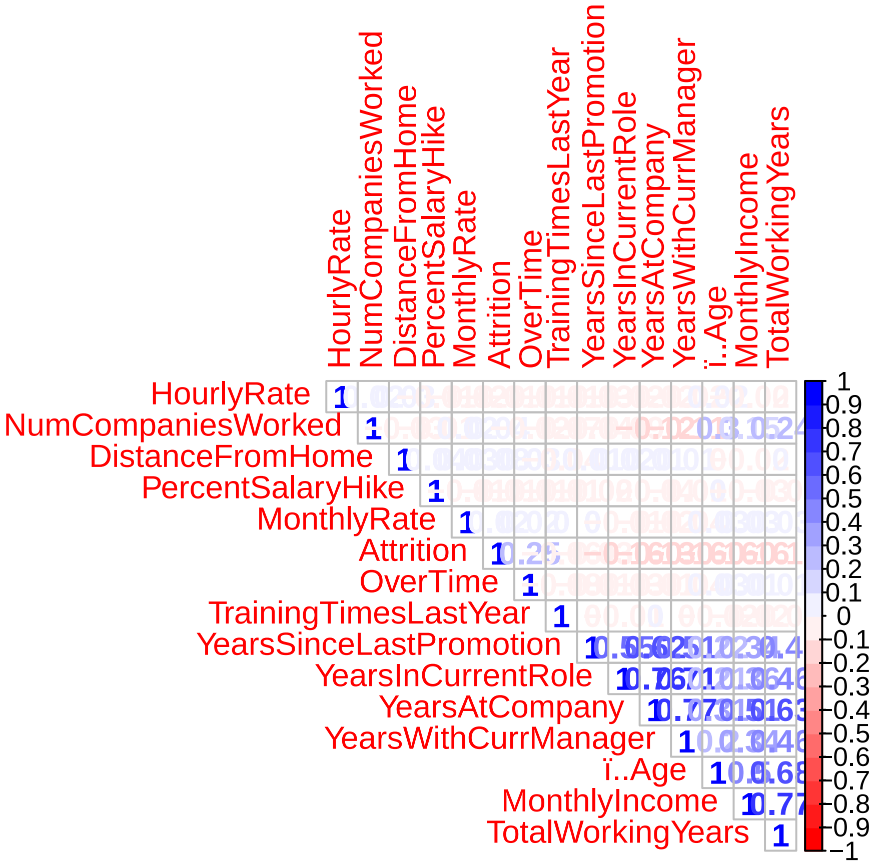


```
numeric=myds%>% dplyr::select(ï..Age,Attrition,DistanceFromHome,OverTime,HourlyRate,MonthlyIncome,Monthl
corrplot(cor(numeric),method="number",type="full")
```



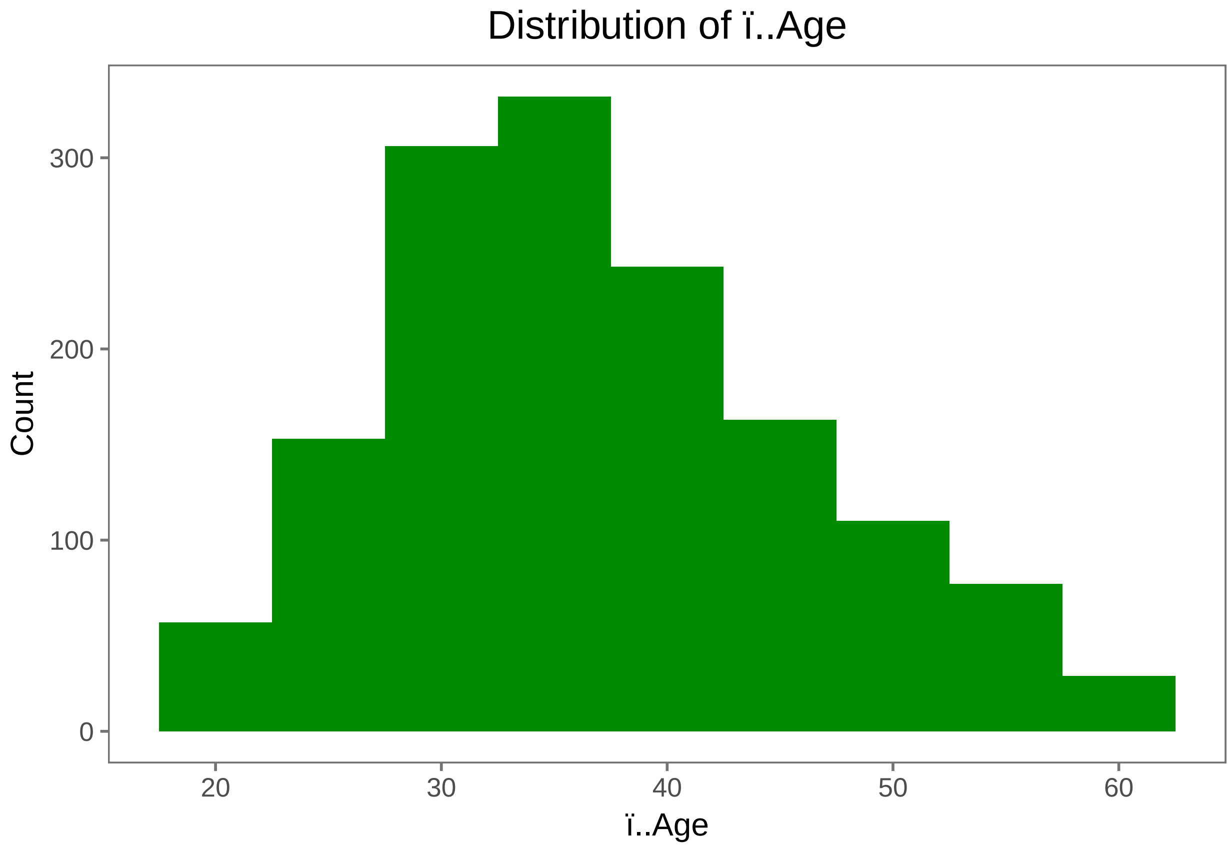
for different view

```
numeric=myds%>% dplyr::select(ï..Age,Attrition,DistanceFromHome,OverTime,HourlyRate,MonthlyIncome,Monthl
col<- colorRampPalette(c("red", "white", "blue"))(20)
corrplot(cor(numeric),method="number",type="upper", order="hclust",col=col)
```



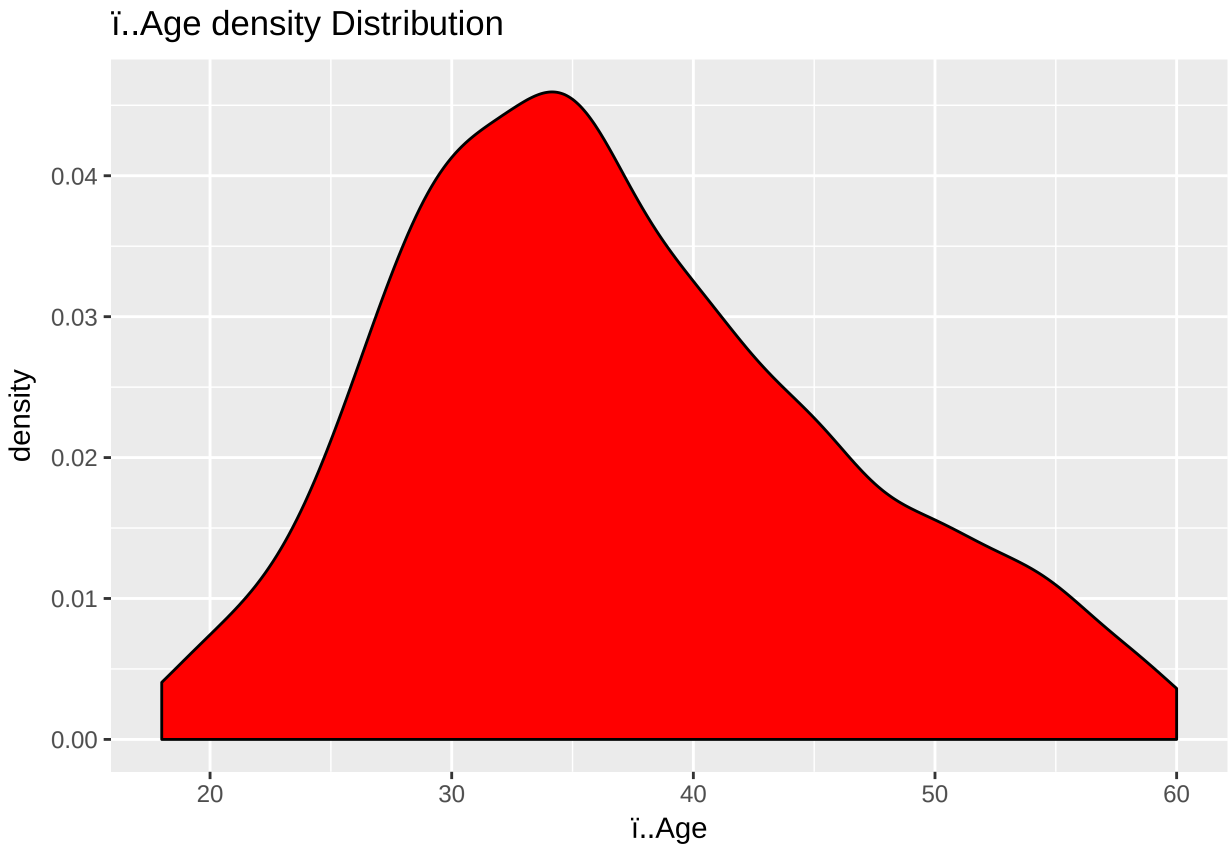
10. ggplotting - DISTRUBUTION OF ï..Age

```
ggplot(numeric,aes(ï..Age))+geom_histogram(binwidth=5,aes(y=..count..),fill="green4")+theme_few()+theme
```



11. Plotting for ï..Age distribution density

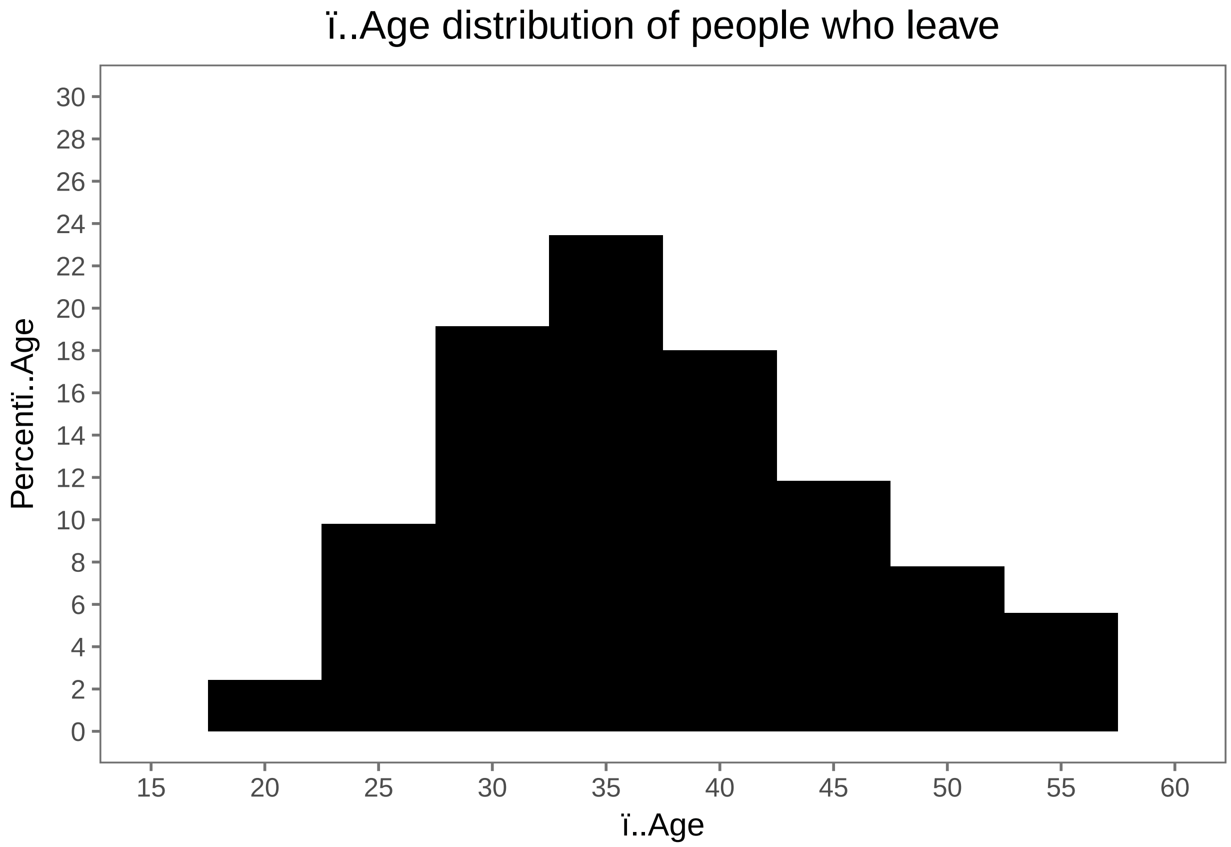
```
ggplot(myds, aes(x = ï..Age)) +
  geom_density(fill = "red") +
  ggtitle("ï..Age density Distribution")
```



####From the plot,we understand that median ï..Age is between 30 to 40 years and maximum is 60 years.

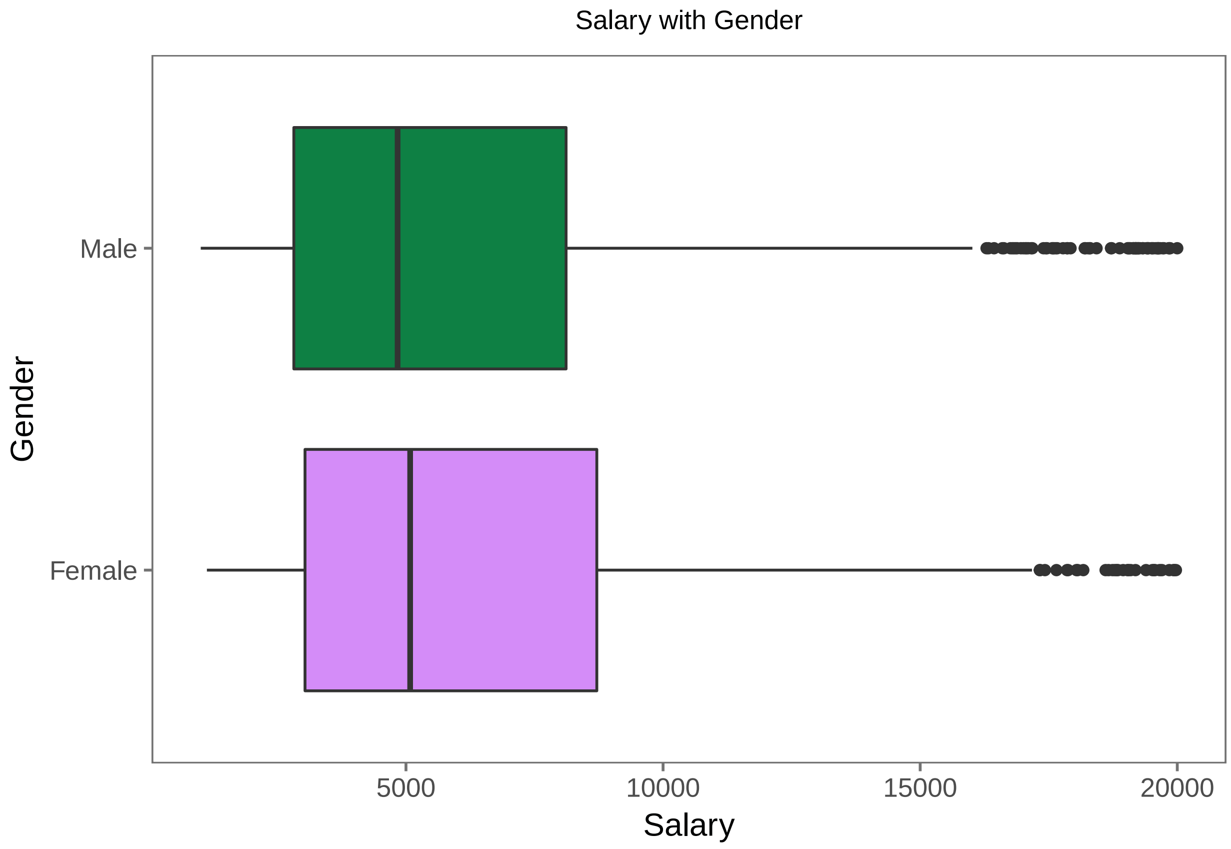
12 ï..Age distribution of attrition

```
myds %>% filter(Attrition == "1") %>% ggplot(aes(ï..Age))+
  geom_histogram(binwidth=5,aes(y=round(((..count..)/sum(..count..))*100,2)),fill="black")+
  theme_few()+theme(legend.position="none",plot.title = element_text(hjust=0.5,size=15))+
  labs(x="ï..Age",y="Percentï..Age",title="ï..Age distribution of people who leave")+scale_y_continuous
  scale_x_continuous(limits=c(15,60),breaks=seq(15,60,5))
```

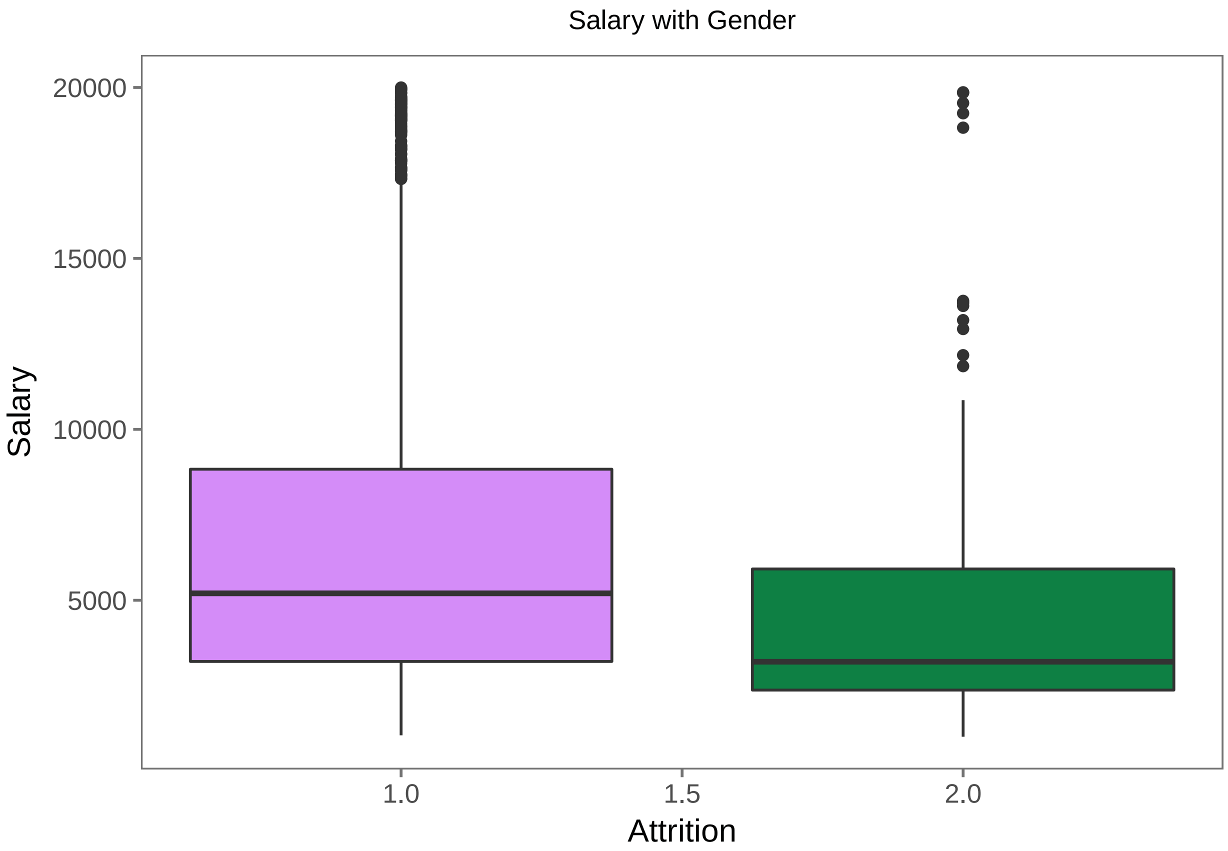


13 Boxplot for gender vs salary

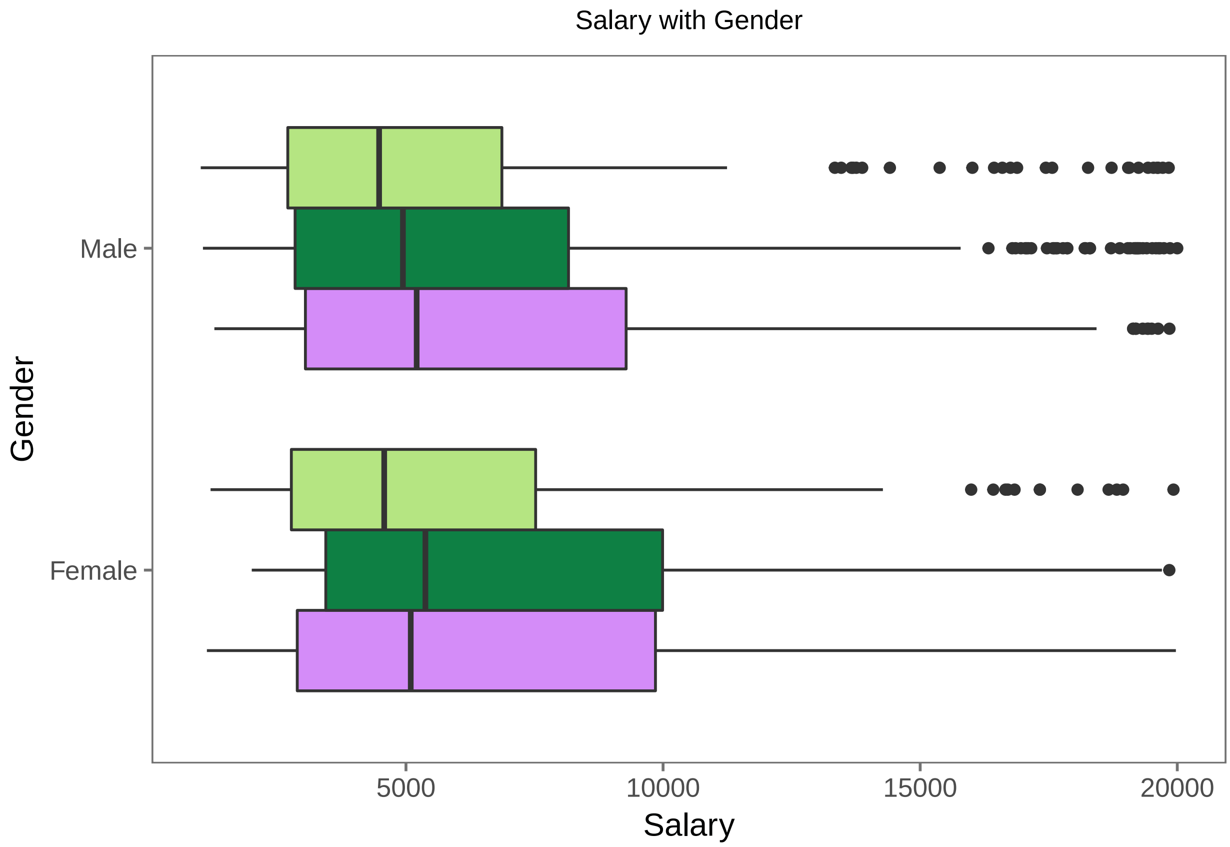
```
ggplot(myds,aes(Gender,MonthlyIncome,fill=Gender))+geom_boxplot()+theme_few()+theme(legend.position="no
```



```
ggplot(myds,aes(Attrition,MonthlyIncome,fill=factor(Attrition)))+geom_boxplot()+theme_few()+theme(legen
```



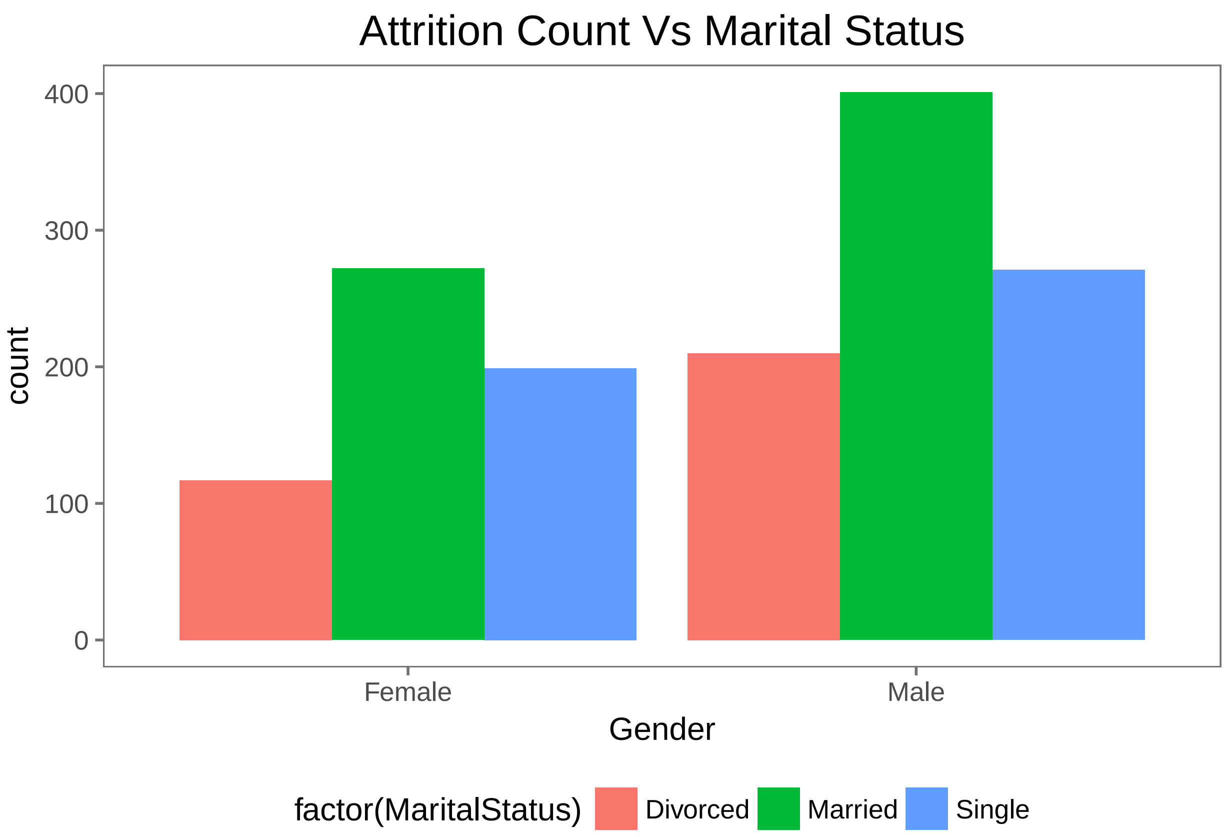
```
ggplot(myds,aes(Gender,MonthlyIncome,fill=factor(MaritalStatus)))+geom_boxplot()+theme_few()+theme(legen
```



We can see less salary is important factor

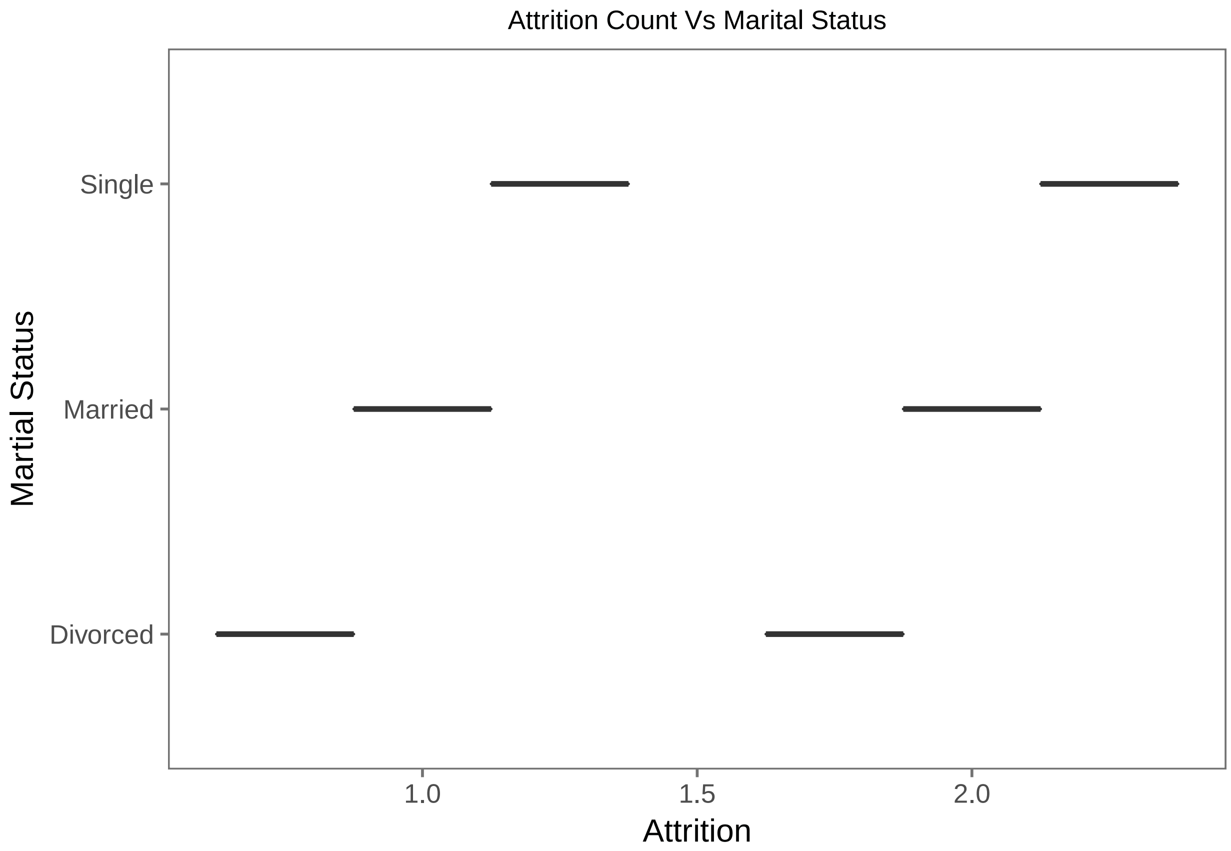
14 Attrition Vs Marital Status

```
ggplot(myds,aes(Gender,..count..,fill=factor(MaritalStatus)))+geom_bar(position=position_dodge())+theme_
```

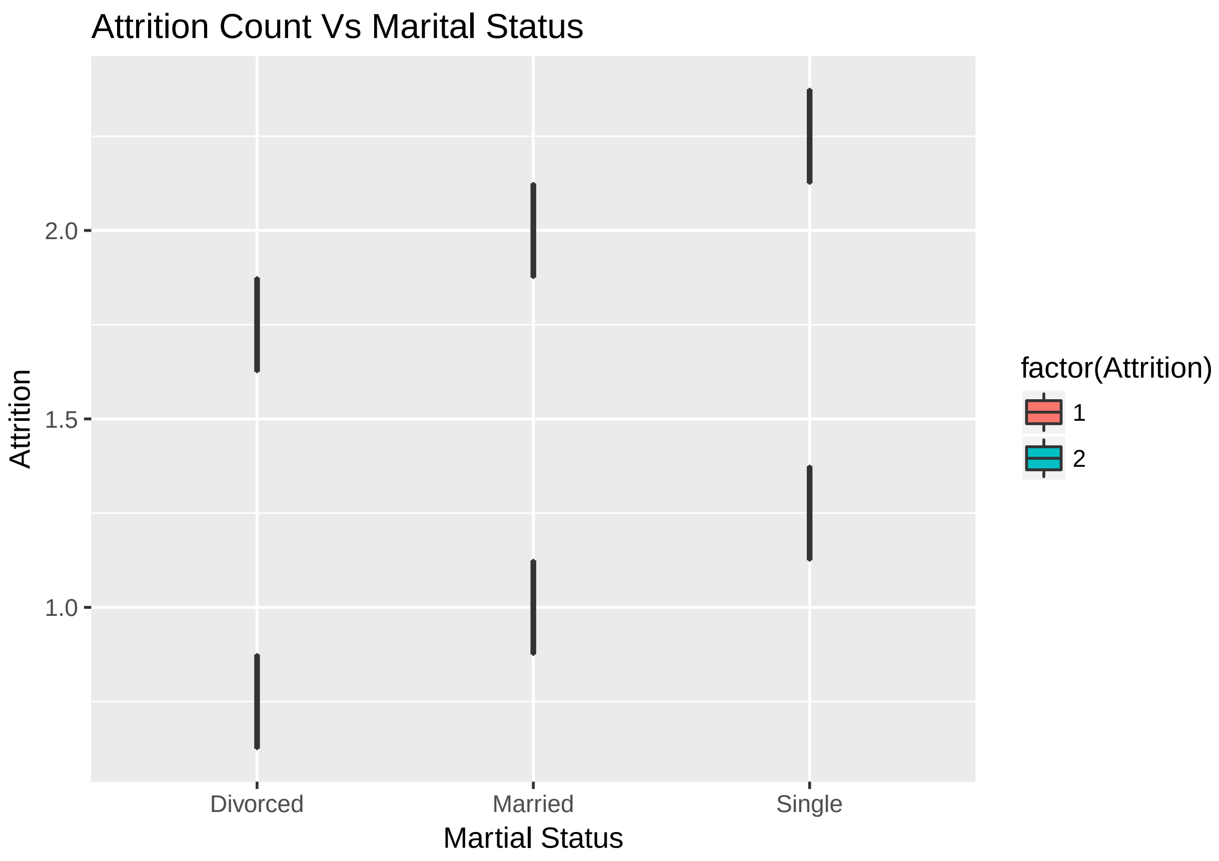


we can see easily married male have higher attriation number, and after that single.Overall ,its same for all.

```
ggplot(data=myds,mapping=aes(x=Attrition,y=MaritalStatus,fill=factor(Attrition)))+geom_boxplot()+theme_
```



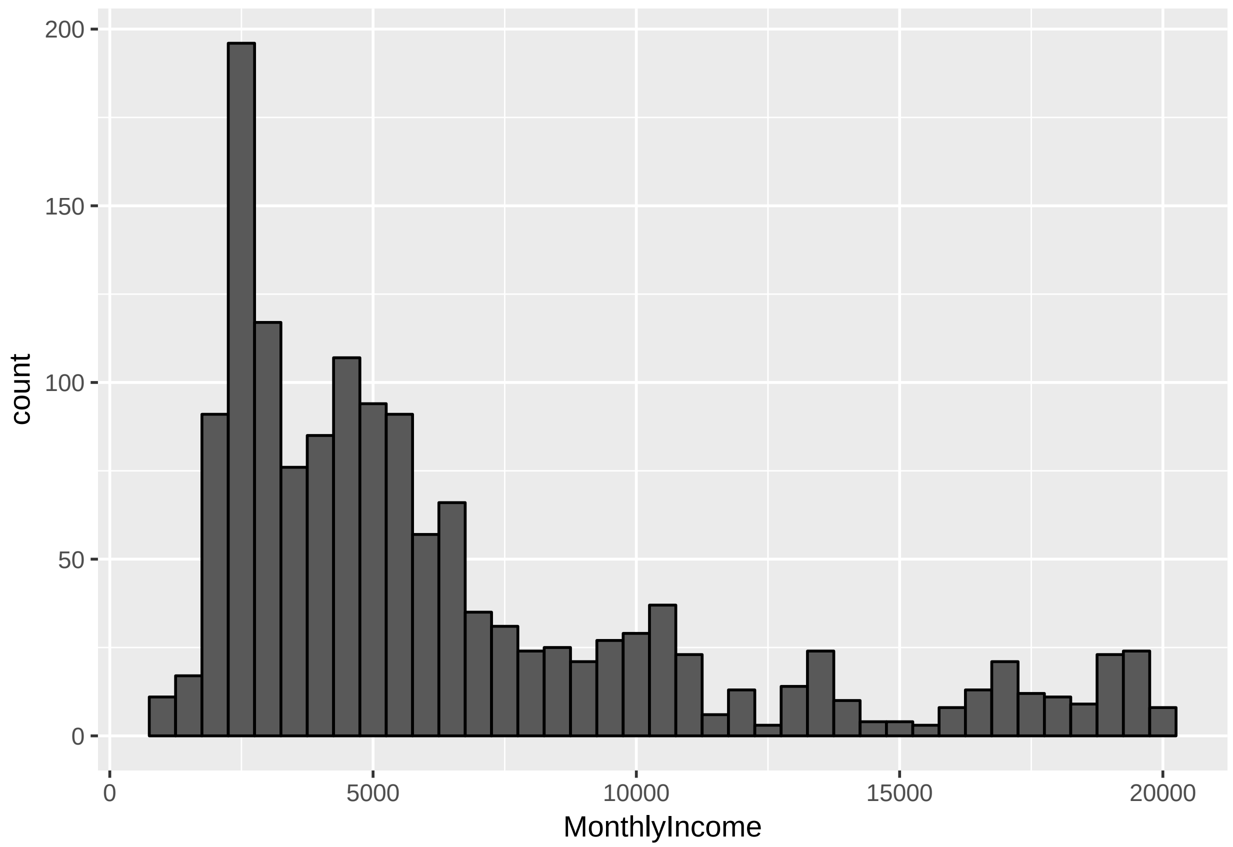
```
ggplot(data=myds,mapping=aes(x=Attrition,y=MaritalStatus,fill=factor(Attrition)))+geom_boxplot()+labs(x
```



this boxplot is not good but limits shows how martial status have difference.

15 Histogram - Monthlyincome vs Count of Employees

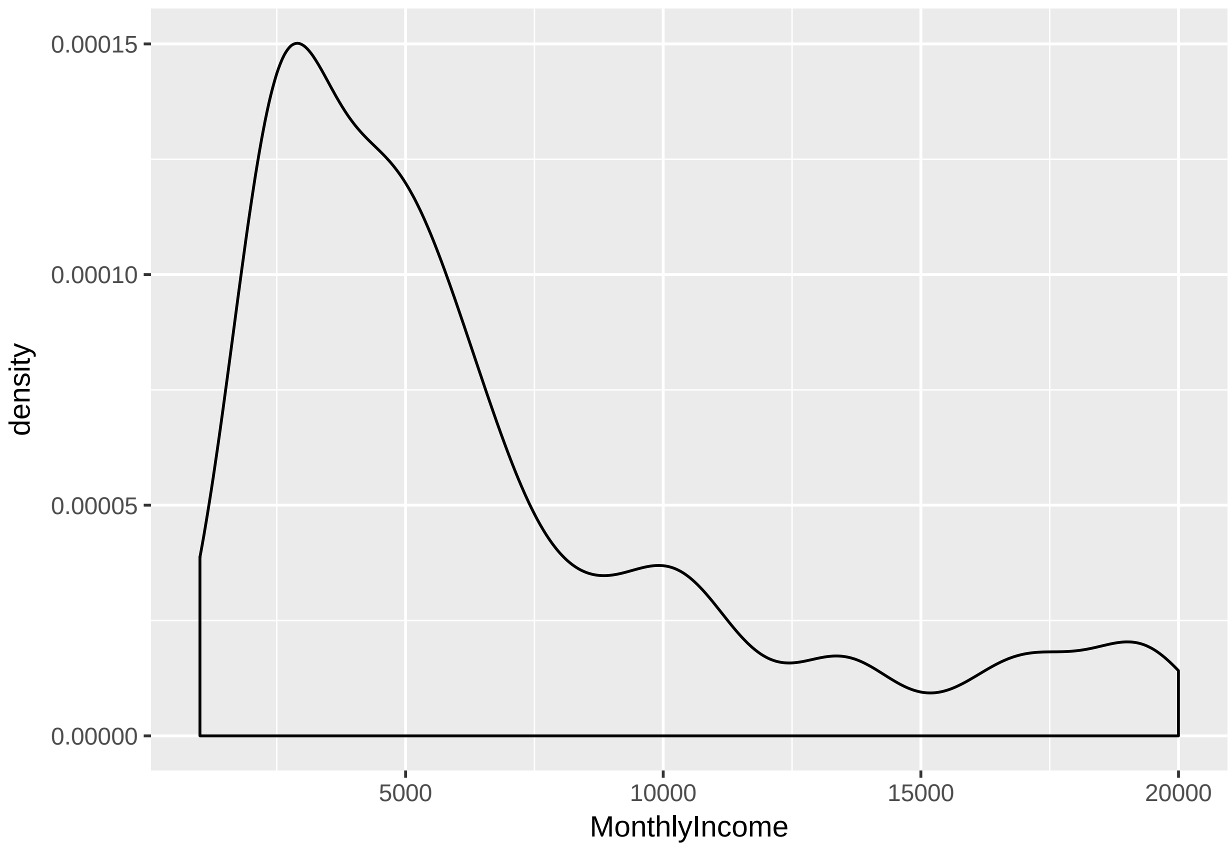
```
ggplot(myds, aes(MonthlyIncome) ) +
  geom_histogram(binwidth=500,color="Black")
```



we can see $2500 is the highest number of emplyees gettingsalary

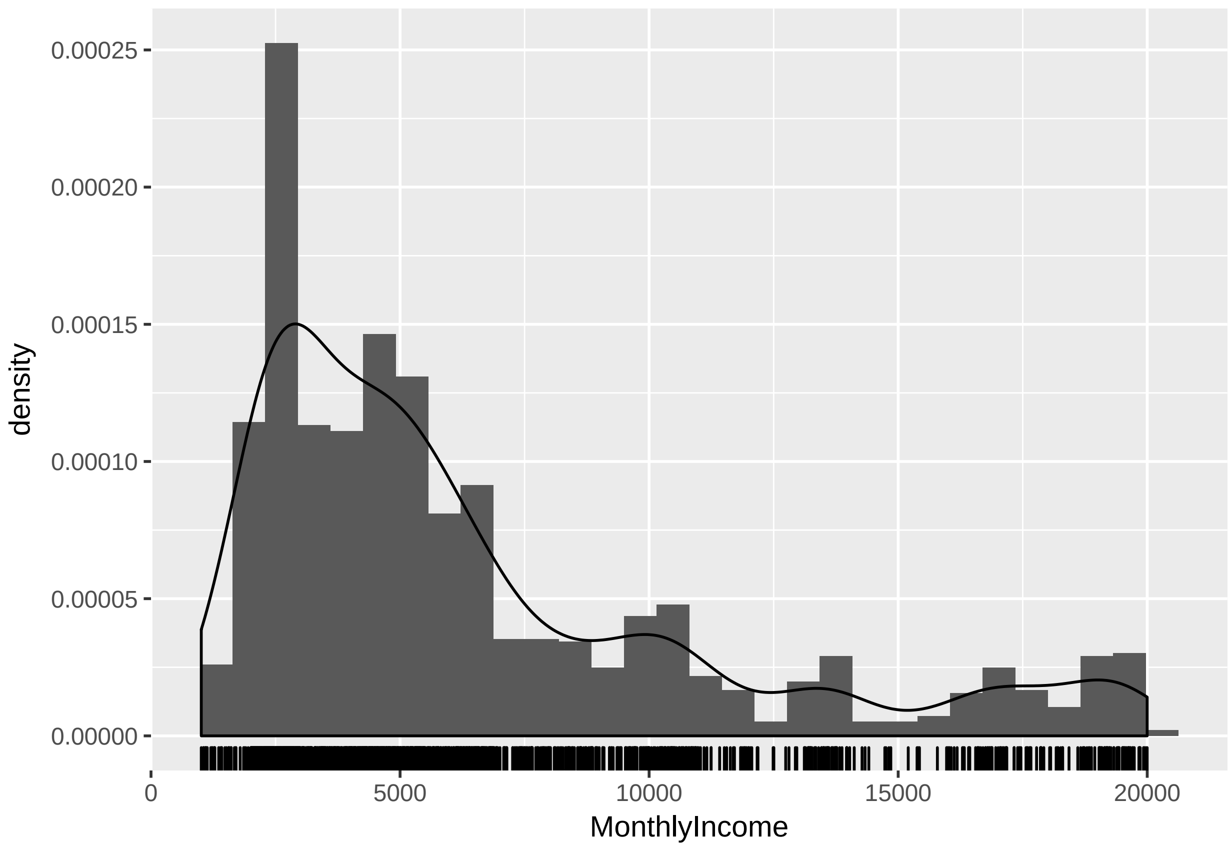
16 to see make it easy we can se density graph too

```
ggplot(myds, aes(MonthlyIncome)) +
  geom_density()
```



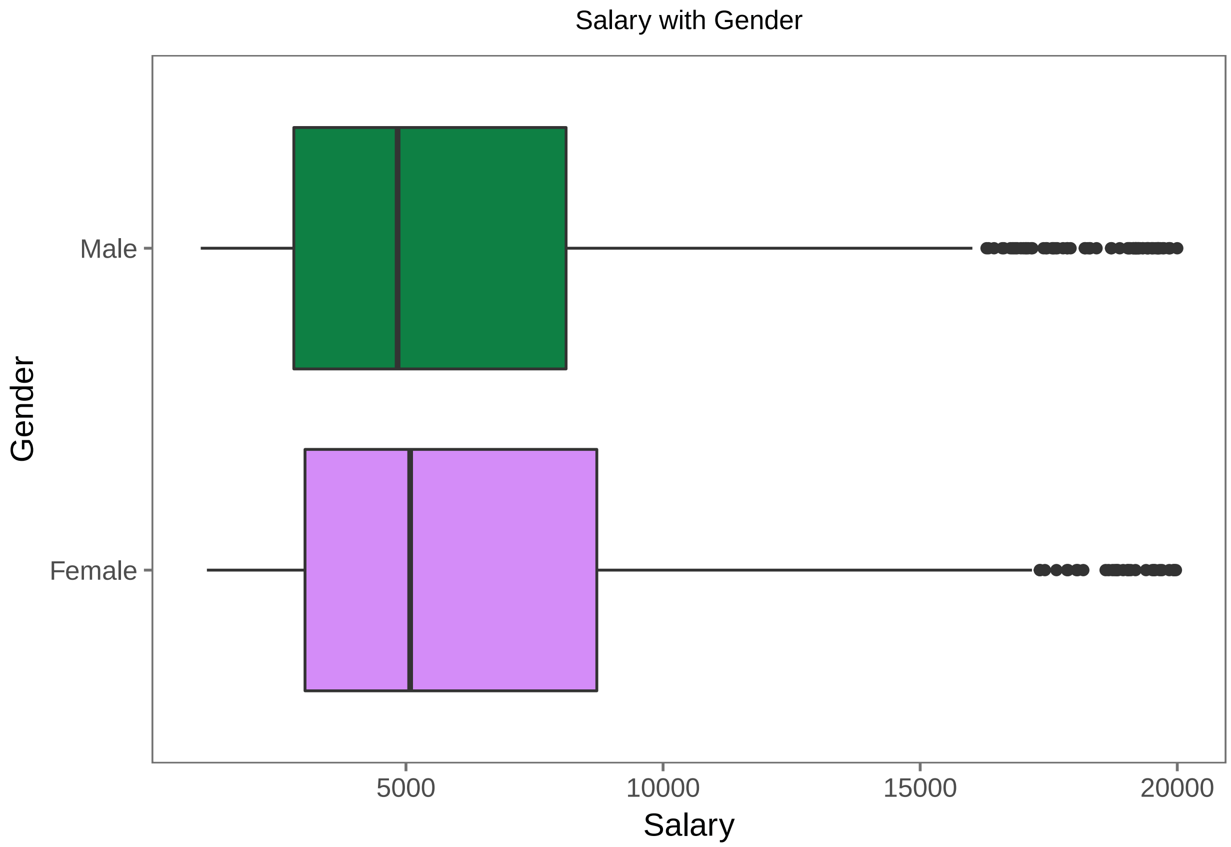
```
ggplot(data=myds) +
  geom_histogram( aes(MonthlyIncome, ..density..)) +
  geom_density( aes(MonthlyIncome, ..density..) ) +
  geom_rug( aes(MonthlyIncome) )
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



17 gender vs monthly income

```
ggplot(myds,aes(Gender,MonthlyIncome,fill=Gender))+geom_boxplot()+theme_few()+theme(legend.position="non
```



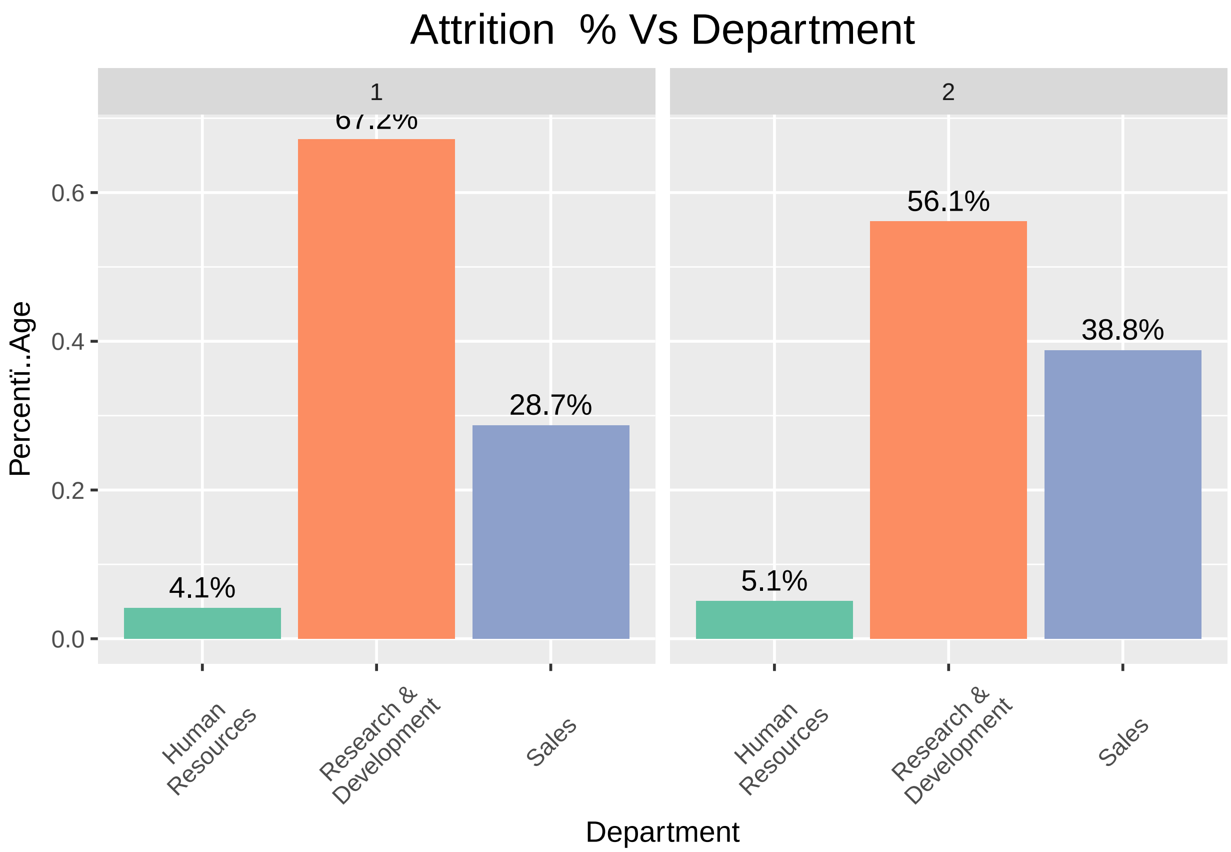
18 identifying number of departsments

```r
cat("There are",length(unique(myds$Department)),"unique departments in the dataset")
```

```
## There are 3 unique departments in the dataset
```

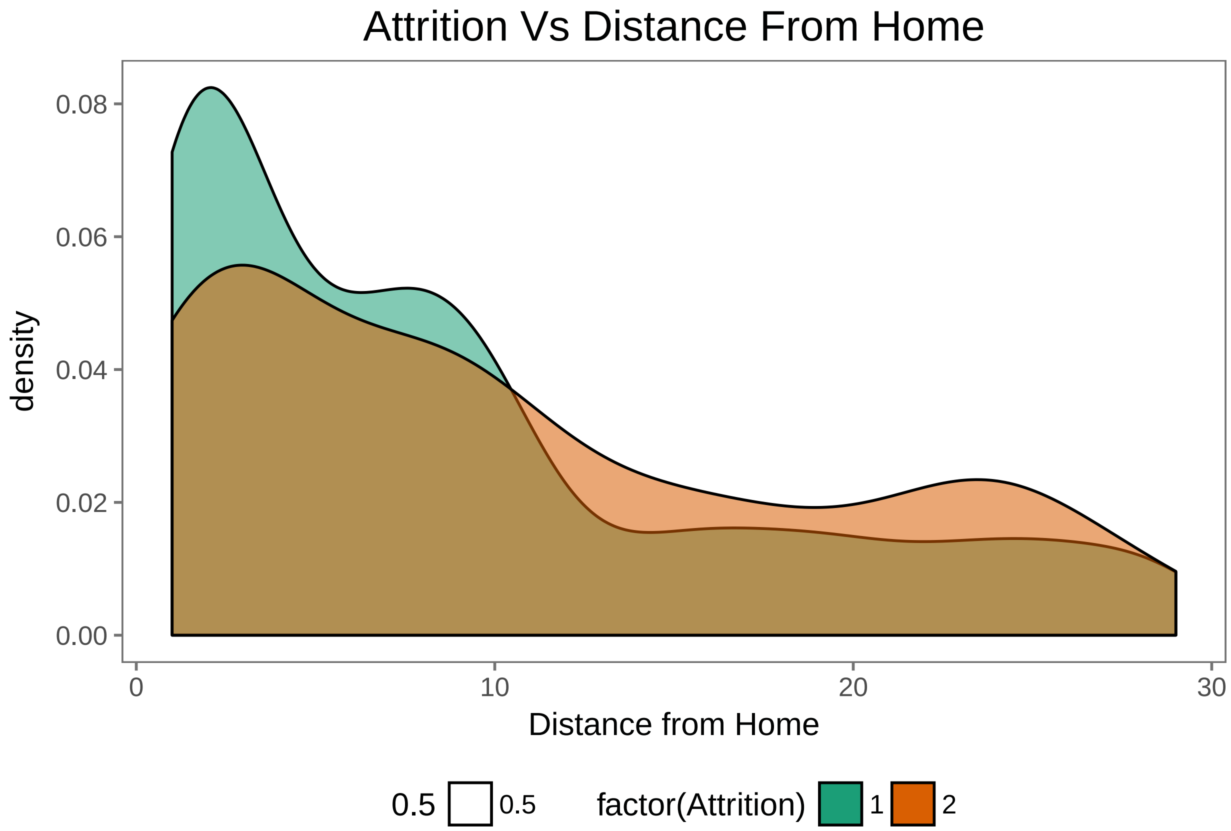
19.plotting ggplot with Dpeartment Vs Percentï..Age of Attrition

```r
ggplot(myds,aes(x=Department,group=Attrition))+geom_bar(aes(y=..prop..,fill=factor(..x..)),stat="count"
```



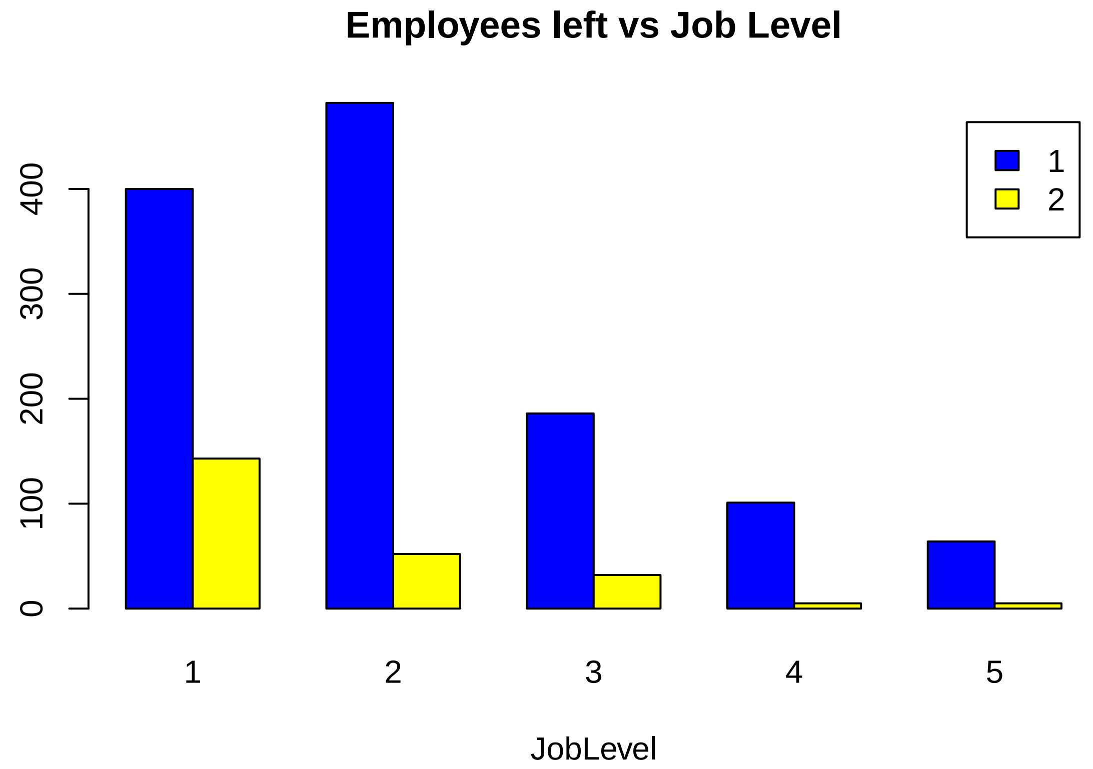
20 Attrition Vs Distance From Home

```
ggplot(myds,aes(x=DistanceFromHome,group=Attrition))+geom_density(aes(fill=factor(Attrition),alpha=0.5))
```



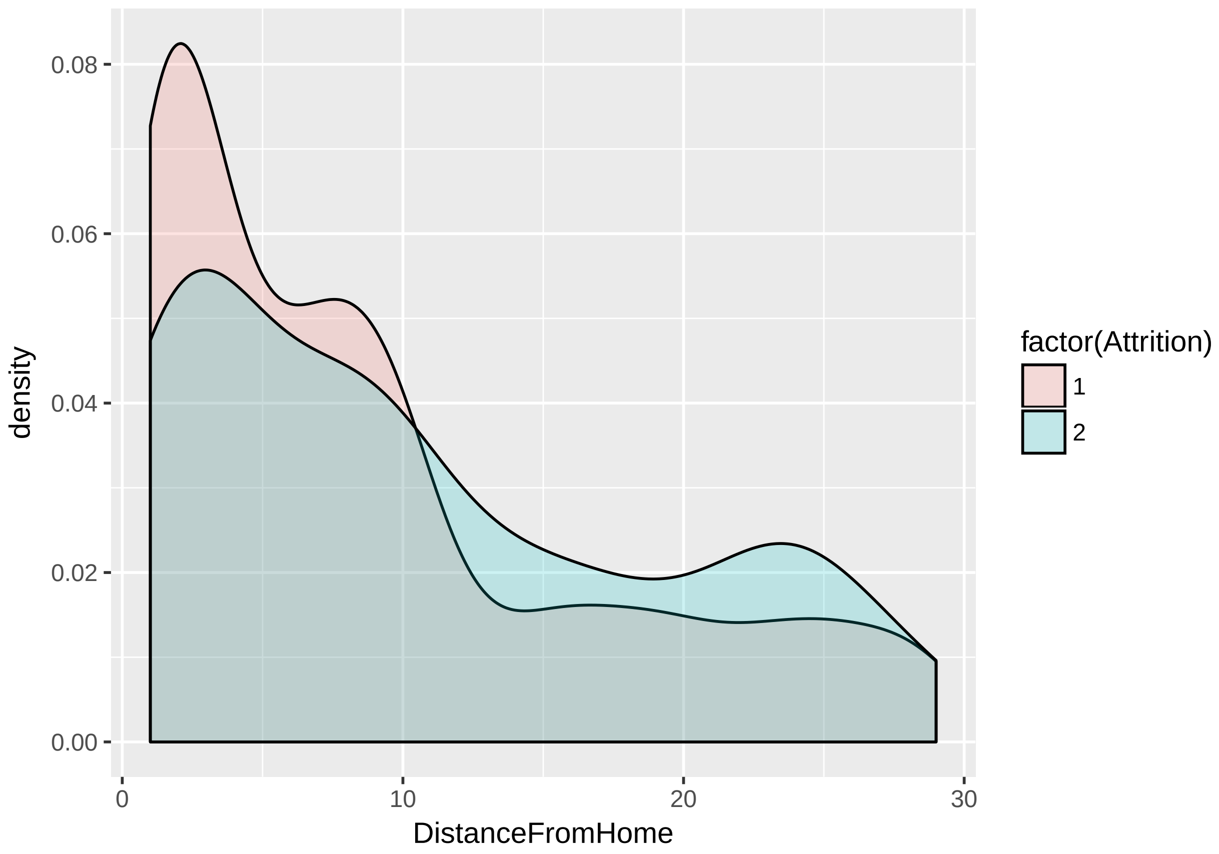
21 Plotting table for Joblevel vs Attrition

```
plottable1=table(myds$Attrition,myds$JobLevel)
barplot(plottable1, main="Employees left vs Job Level", xlab="JobLevel",col=c("Blue","Yellow"),legend=r
```



22 working line

```
ggplot(myds) + geom_density(aes(x = DistanceFromHome, fill = factor(Attrition)), alpha = 0.2)
```



23 #Attrition VS Marital Status

```
table_mar<-table(myds$MaritalStatus, myds$Attrition)
chisq.test(table_mar)
```

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

As p-value is less than alpha, attrition depends on the marital status of employees.

24 identifying Travel Frequency

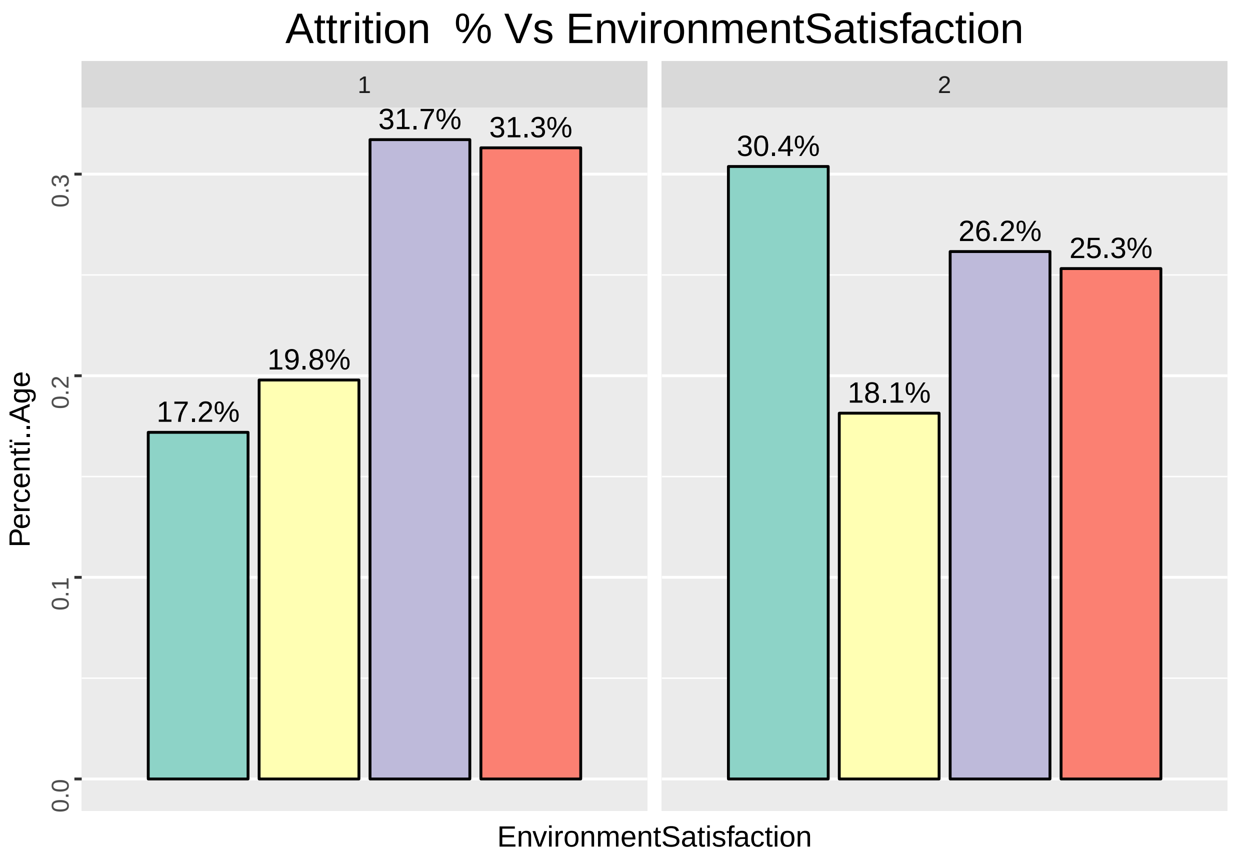
```
cat("There are",length(unique(myds$Department)),"unique departments in the dataset")
```

```
## There are 3 unique departments in the dataset
```

25 plotting ggplot with Dpeartment Vs Percentï..Age of Attrition

```
ggplot(myds, aes(x = EnvironmentSatisfaction, group = Attrition)) + geom_bar(
  aes(y = ..prop.., fill = factor(..x..)),
  stat = "count",
  position = position_dodge(),
  colour = "black"
) + scale_fill_manual(values = c("#999999", "#E69F00")) + facet_grid( ~
                                                                        Attrition) + theme(
                                                                          axis.text = element_text(
                                                                            angle = 90,
                                                                            vjust = 10,
                                                                            hjust = 10
                                                                          ),
                                                                          legend.position = "Bottom",
                                                                          plot.title = element_text(siz
                                                                        ) + labs(x = "EnvironmentSatisf
```

```
## Scale for 'fill' is already present. Adding another scale for 'fill',
## which will replace the existing scale.
```



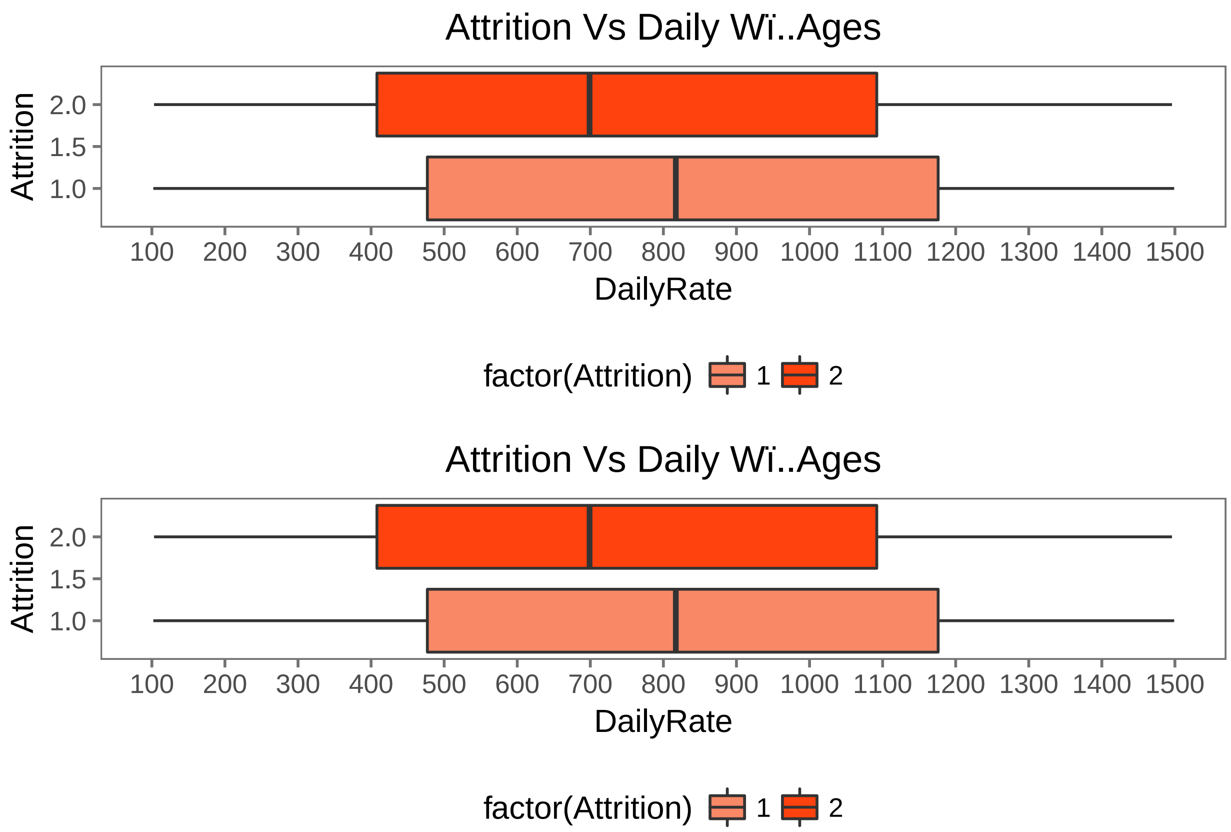
26 Attrition Vs Payrates

this is because those who paid less might leave early

```
g1=ggplot(myds,aes(Attrition,DailyRate,fill=factor(Attrition)))+geom_boxplot()+theme_few()+theme(plot.t
```

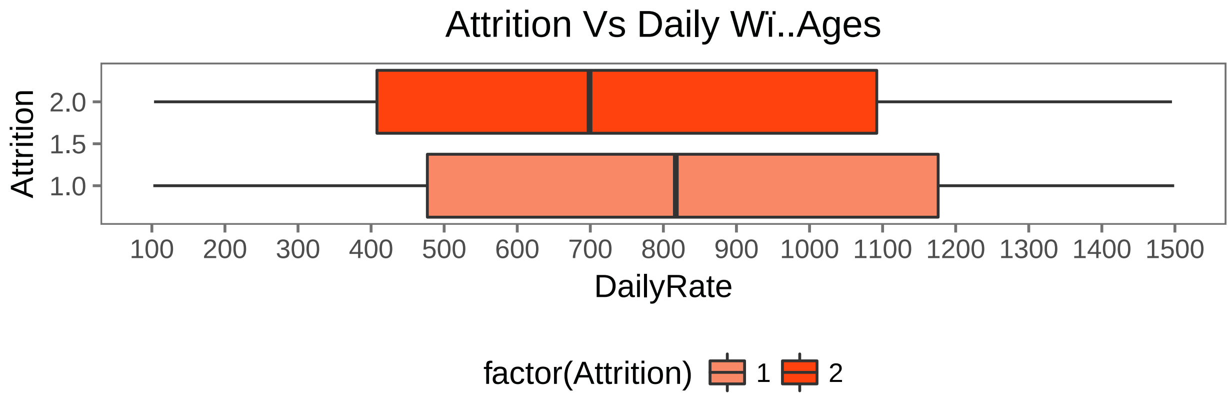
```
g2=ggplot(myds,aes(Attrition,DailyRate,fill=factor(Attrition)))+geom_boxplot()+theme_few()+theme(plot.t
```

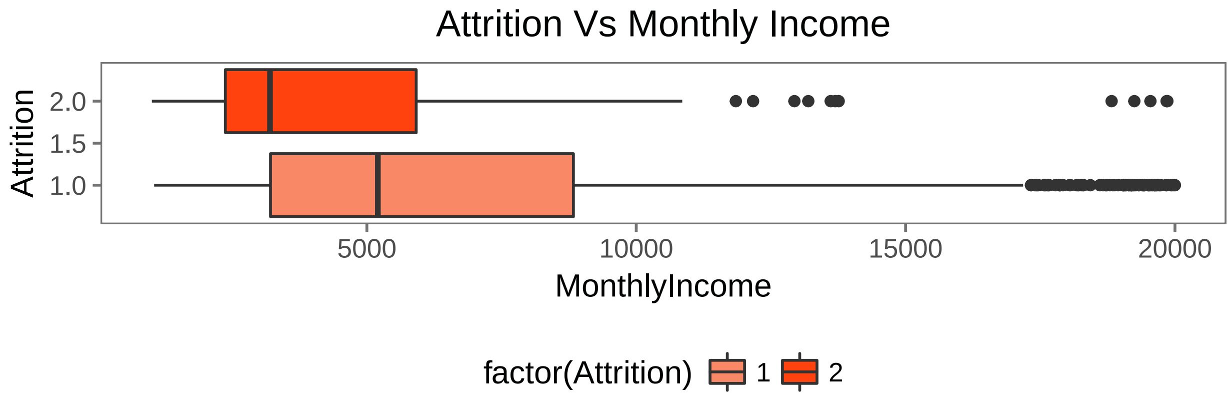
```
grid.arrange(g1,g2,nrow=2)
```



```
g2=ggplot(myds,aes(Attrition,MonthlyIncome,fill=factor(Attrition)))+geom_boxplot()+theme_few()+theme(pl
```

```
grid.arrange(g1,g2,nrow=2)
```





27 Boxplotting for Attrition vs dailyrate

boxplot(myds*Attrition myds*DailyRate,col=rainbow(3),notch=FALSE)

ggplot(myds,aes(Attrition,HourlyRate,fill=factor(Attrition)))+geom_boxplot()+theme_few()+theme(plot.title=element_text
Vs Hourly Wï..Ages”)

28 Attrition VS Monthly Income

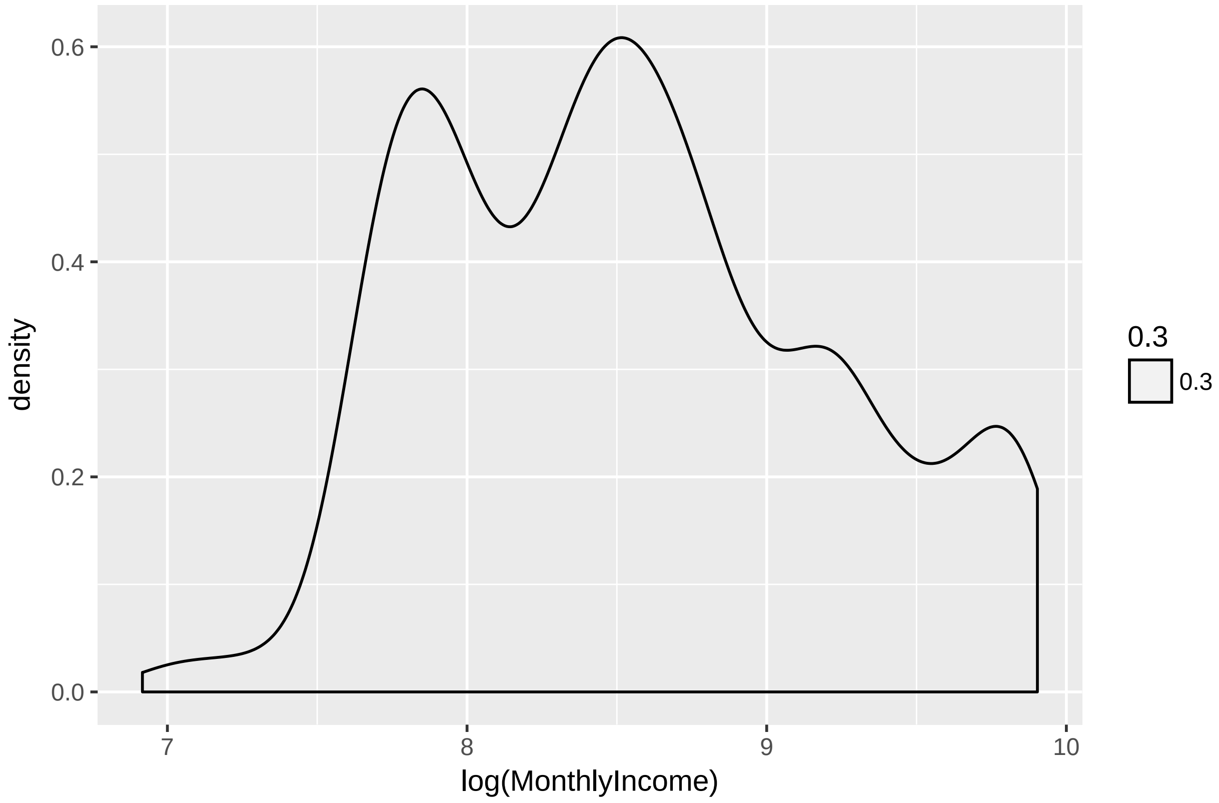
```
t.test(myds$MonthlyIncome~myds$Attrition)
```

```
##
##  Welch Two Sample t-test
##
## data:  myds$MonthlyIncome by myds$Attrition
## t = 7.4826, df = 412.74, p-value = 4.434e-13
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  1508.244 2583.050
## sample estimates:
## mean in group 1 mean in group 2
##        6832.740        4787.093
```

As t.test shows, attrition is highly dependent on monthly income.

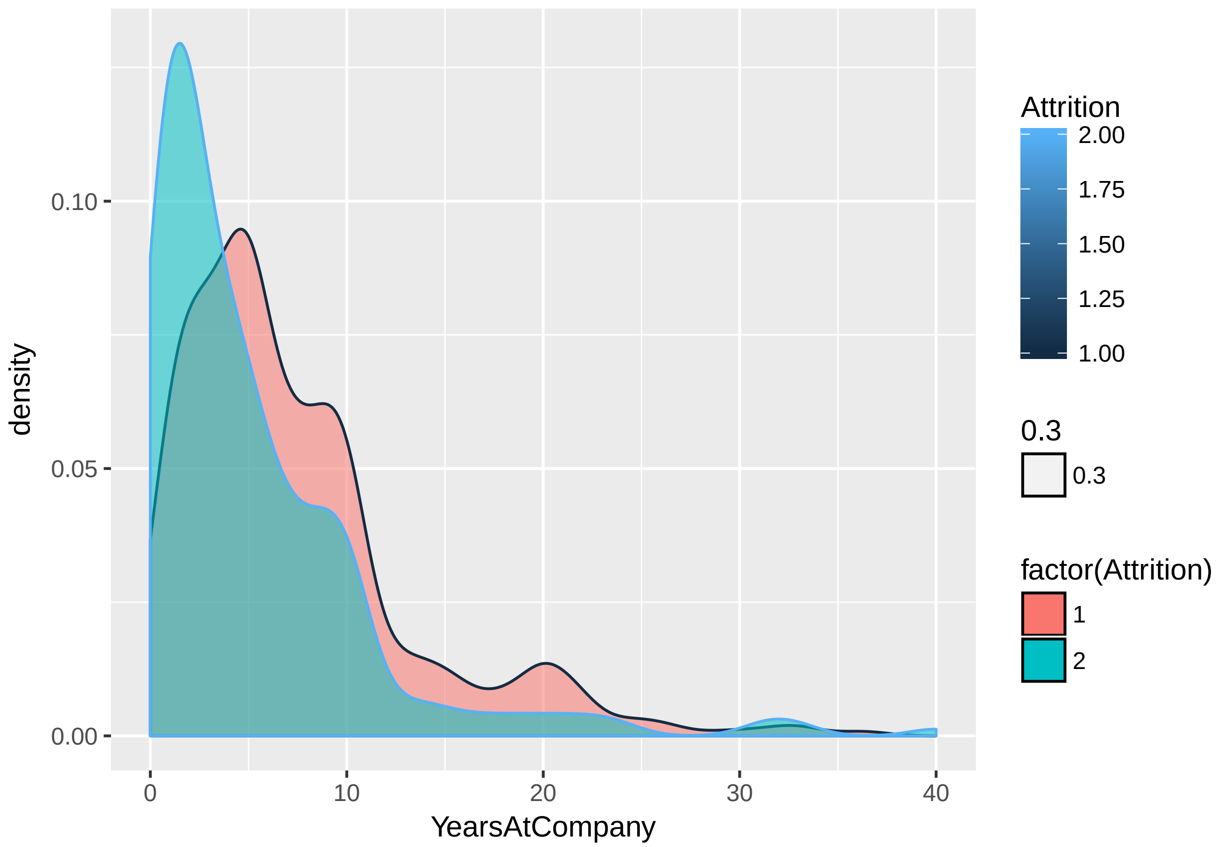
29 Attrition - log(Monthly Income)

```
ggplot(myds, aes(x =  log(MonthlyIncome), fill =Attrition,
                     colour = Attrition, alpha = .3)) +
  geom_density() + ggtitle("")
```



30 YearsAtCompany - Attrition

```
ggplot(myds, aes(x = YearsAtCompany, fill = factor(Attrition),
                     colour = Attrition, alpha = .3)) +
  geom_density()
```



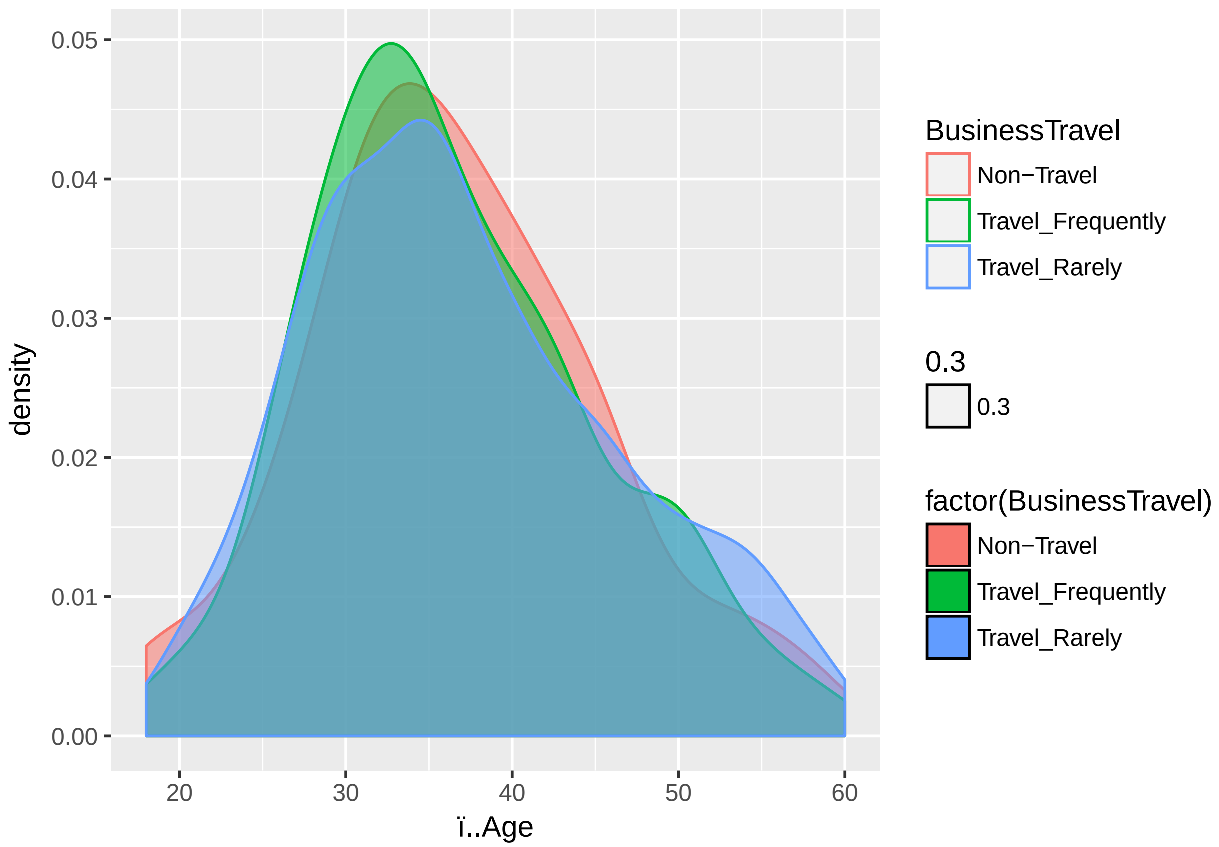
```
t.test(myds$YearsAtCompany~myds$Attrition)
```

```
##
##  Welch Two Sample t-test
##
## data:  myds$YearsAtCompany by myds$Attrition
## t = 5.2826, df = 338.21, p-value = 2.286e-07
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  1.404805 3.071629
## sample estimates:
## mean in group 1 mean in group 2
##        7.369019        5.130802
```

T.test shwows .attrition is dependent on Years at company

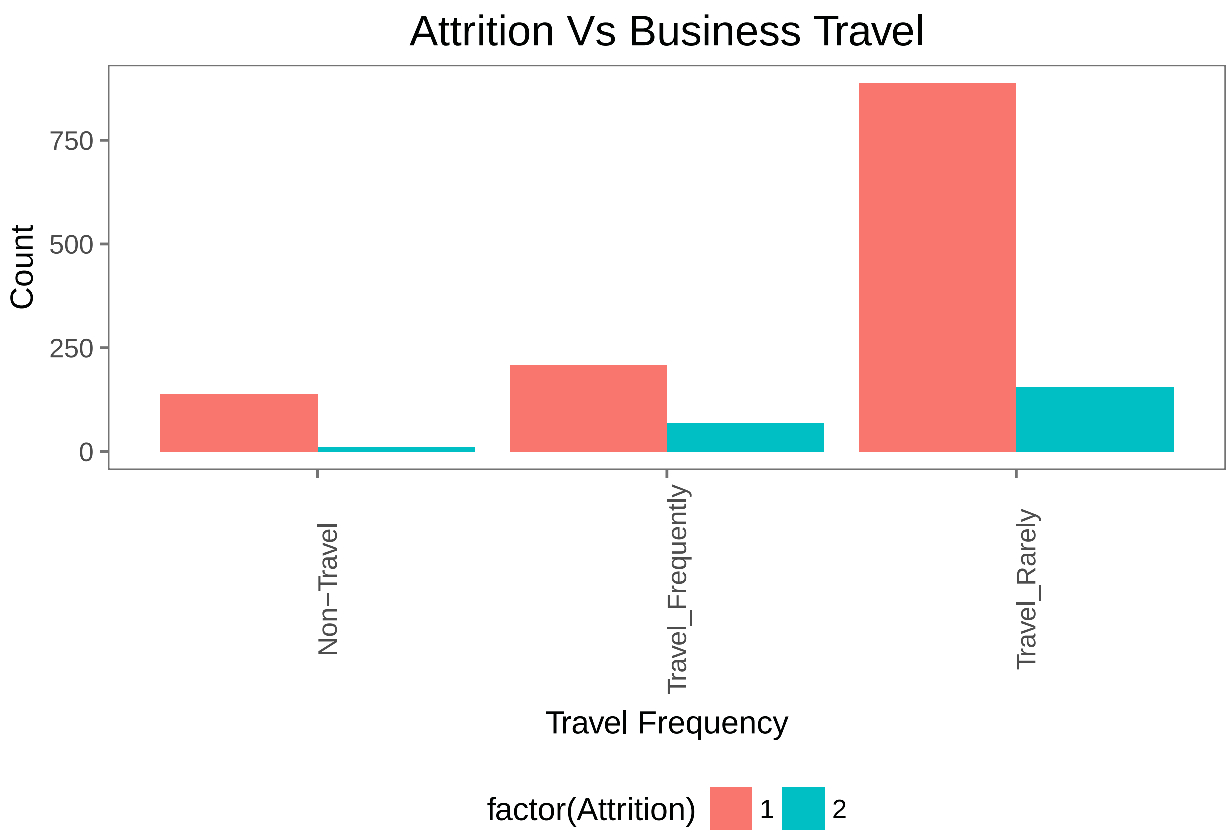
34plotting of distance travel vs Attrition

```
ggplot(myds, aes(x = ï..Age,
                    fill = factor(BusinessTravel),
                    colour = BusinessTravel, alpha = .3)) +
  geom_density()
```



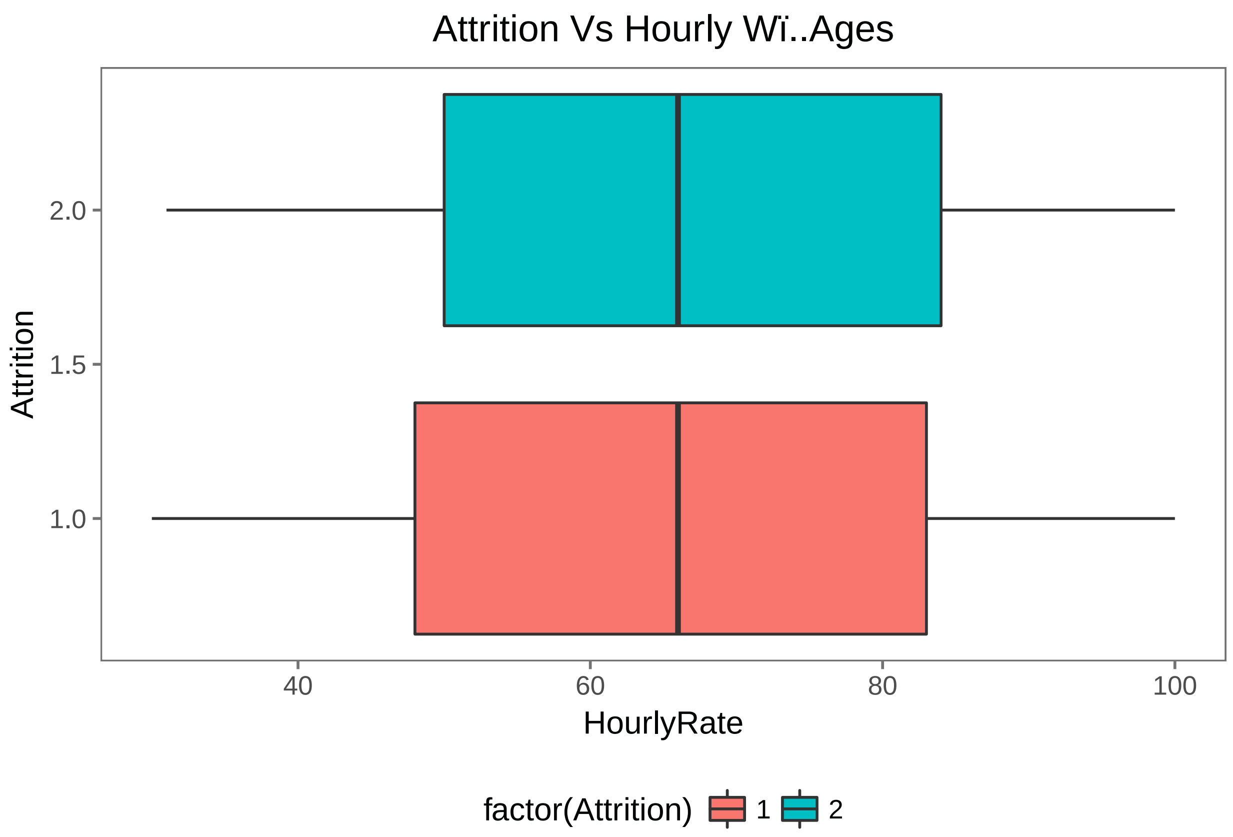
35 again plotting ofattrition vs distance travel

```
ggplot(myds,aes(BusinessTravel,fill=factor(Attrition)))+geom_bar(stat="count",aes(y=..count..),position=
```



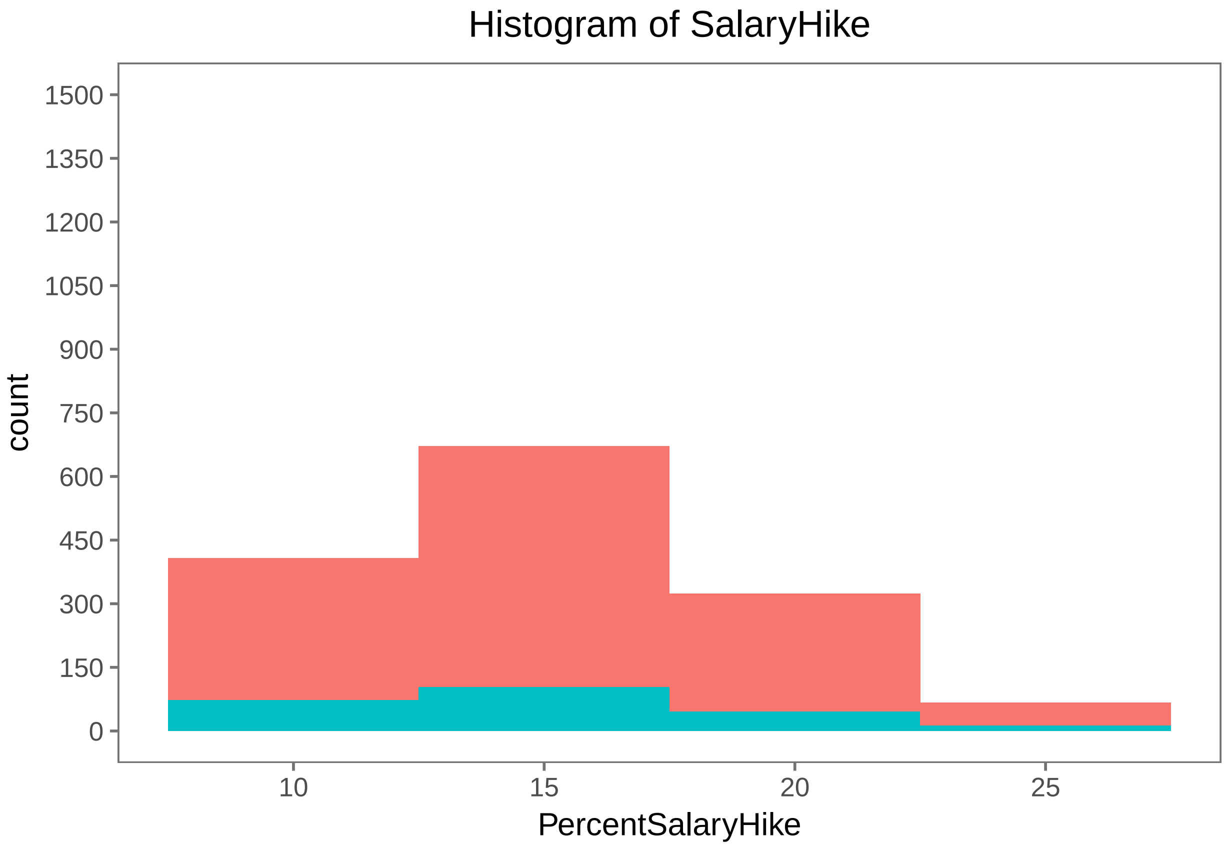
36 Attrition Vs Hourly Rate

```
ggplot(myds,aes(Attrition,HourlyRate,fill=factor(Attrition)))+geom_boxplot()+theme_few()+theme(plot.titl
```



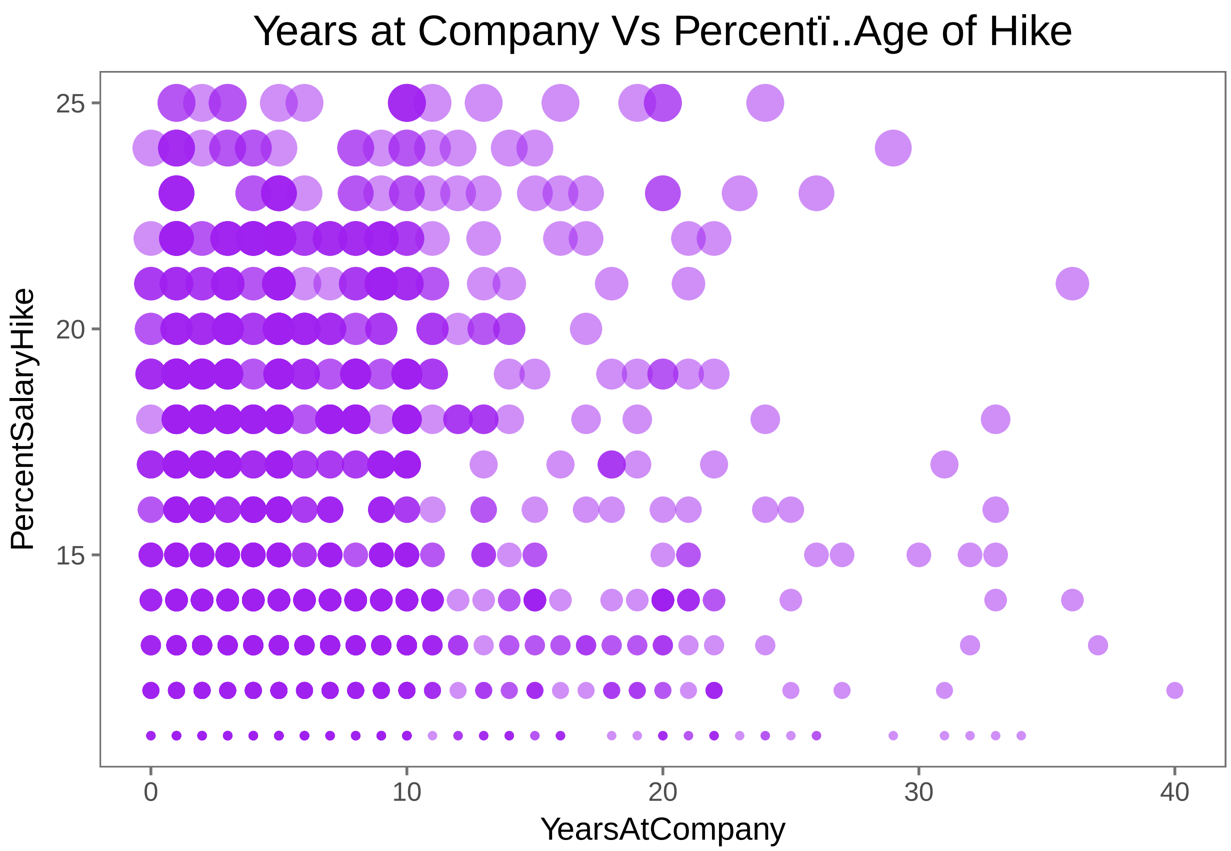
37 Percentï..Age of salary hike

```
ggplot(myds,aes(PercentSalaryHike,..count..,fill=factor(Attrition)))+geom_histogram(binwidth=5)+theme_f
```



38 plotting again for years at company vs percentï..Age of hike employees recieve

```
myds %>%
ggplot(aes(YearsAtCompany,PercentSalaryHike,size=PercentSalaryHike))+geom_point(color="purple",alpha=0.5
```



Observation : Here too we see no relation between the two factors.Even People who have lesser year of stint at the company have received maximum hike.

39 Which role is paid more?

Precaution : load Stringr if str__wrap error comes up

```
temp=myds %>% group_by(JobRole) %>% summarise(salary=median(MonthlyIncome)) %>% arrange(desc(salary))
ggplot(temp,aes(factor(JobRole,levels=(JobRole)),salary))+geom_bar(stat="identity",fill="gold4")+coord_
```

```
## Error in factor(JobRole, levels = (JobRole)): object 'JobRole' not found
```

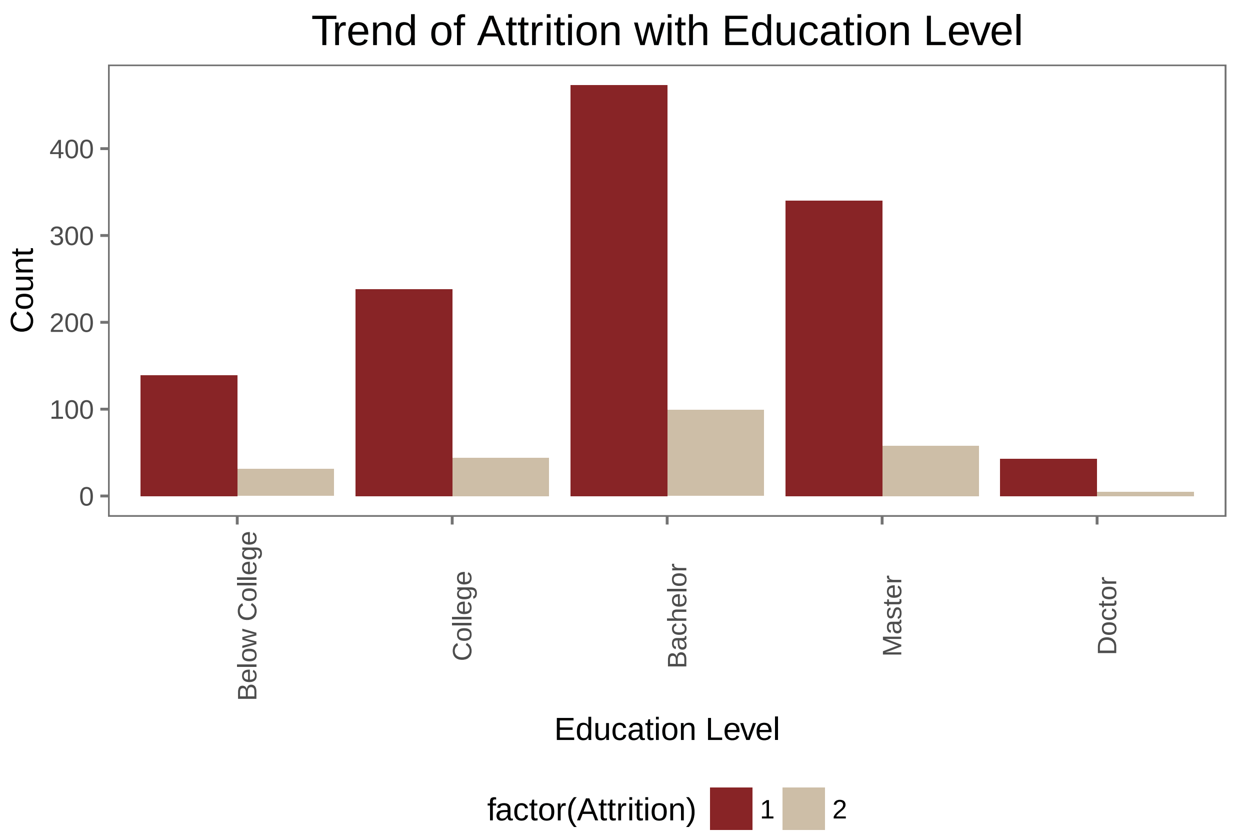
*** Manï..Ager,Research director,Healthcare representative have higher median salary whereas HR,Sales rep have been paid a lower salary**

40 Education,EducationField:

load forcat

```
temp= myds %>% mutate(Education=factor(Education)) %>% mutate(Education=fct_recode(Education,'Below Col
```

```
ggplot(temp,aes(Education,fill=factor(Attrition)))+geom_bar(stat="count",aes(y=..count..),position=posi
```



Observation : Mostly bachelors education holder and least by Doctor but cant draw clear consluion, so we will look at education field too.

ggplot(temp,aes(Education,fill=factor(Attrition)))+geom_bar(stat=“count”,aes(y=..count..),position=position_dodge())+the
= element_text(angle=90))+labs(x=“Education Level”,y=“Count”,title=“Education levels and field of education”)+scale_fill_canva(palette=“Unique and striking”)+facet_grid(~EducationField)

Observation: Life science and medical contribute much to datasets and least by Hr.

41 Number of companies worked:

```
temp2 = myds %>% group_by(Attrition,NumCompaniesWorked) %>% tally(sort=TRUE)
```

```
ggplot(temp,aes(NumCompaniesWorked,n,fill=factor(Attrition),label=n))+geom_bar(stat="identity",position
```

```
## Don't know how to automatically pick scale for object of type function. Defaulting to continuous.
## Don't know how to automatically pick scale for object of type function. Defaulting to continuous.
```

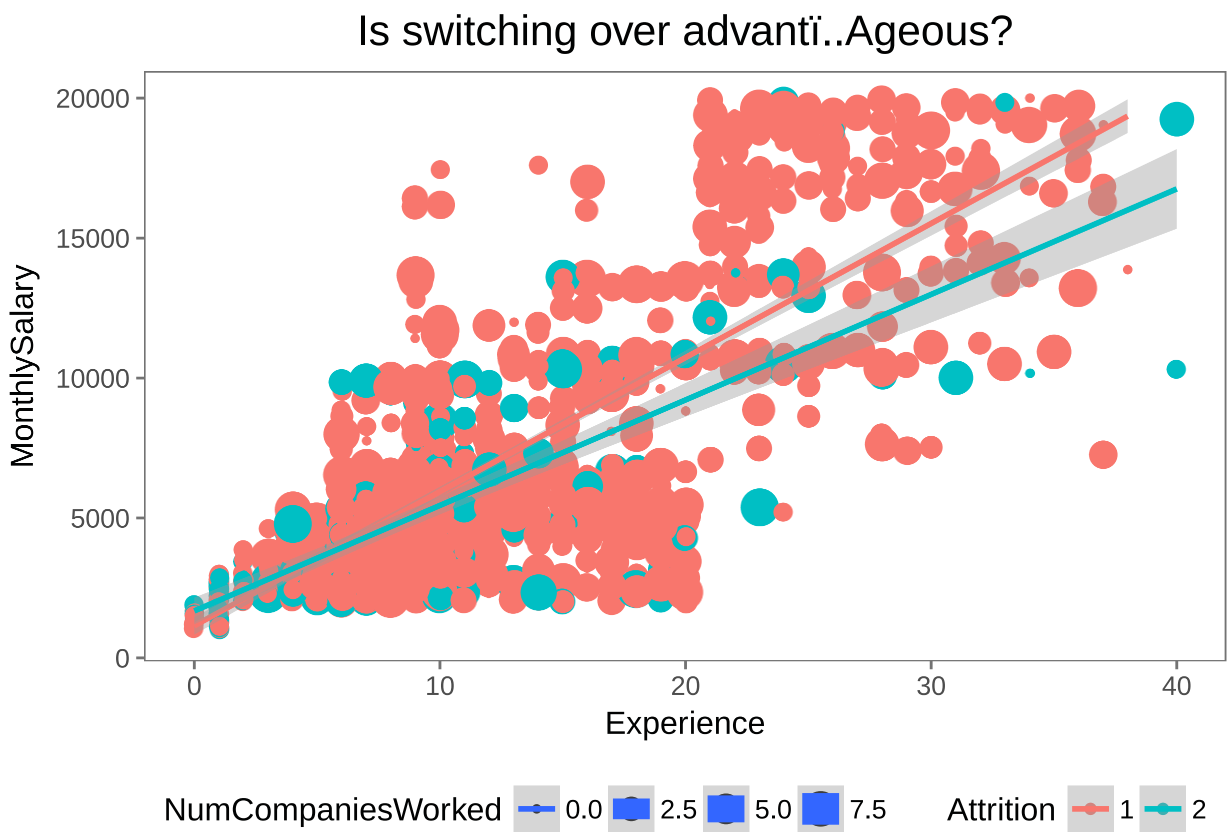
```
## Error in (function (..., row.names = NULL, check.rows = FALSE, check.names = TRUE, : arguments imply
```

Observation : We see people worked at least 1 company switch mostly and equal ratio for rest with low rates

Conclusion: Higher experince or switch lesser ittration rate

42 Swtiching is over adventegious or not ?

```
ggplot(myds,aes(TotalWorkingYears,MonthlyIncome,size=NumCompaniesWorked,col=factor(Attrition)))+geom_po
```



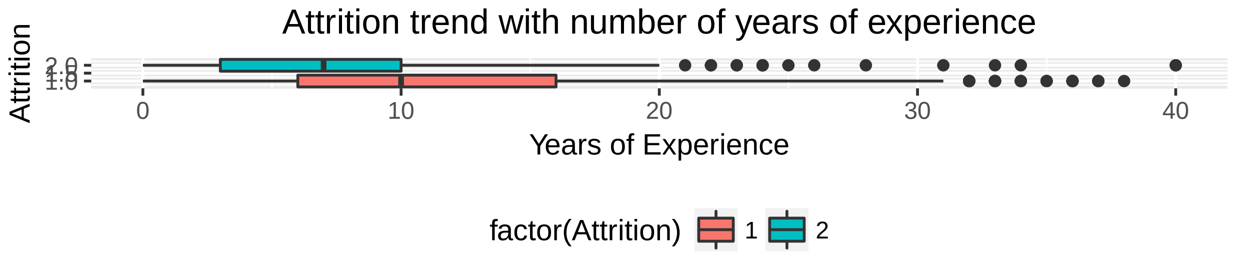
Is switching over advantï..Ageous?

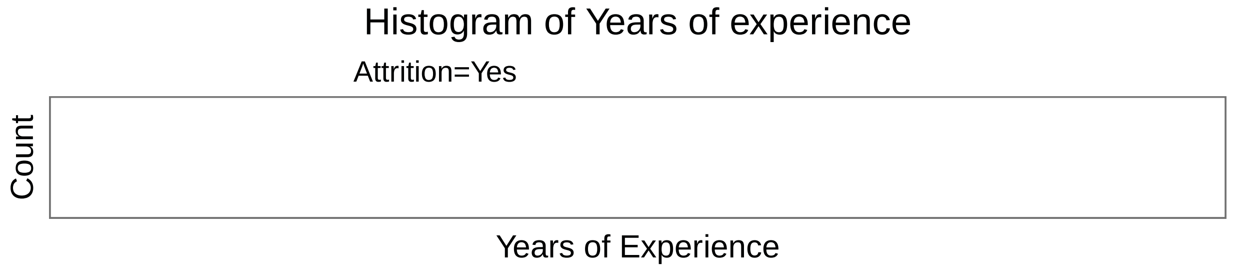
```
g1=ggplot(myds,aes(Attrition,TotalWorkingYears,fill=factor(Attrition)))+geom_boxplot()+theme(legend.pos
```

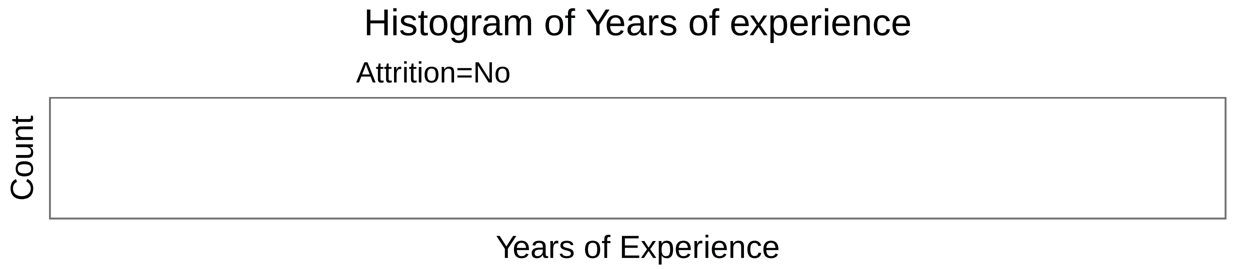
```
g2=myds %>% filter(Attrition=="Yes") %>% ggplot(aes(TotalWorkingYears,..count..,fill=factor(Attrition))
```

```
g3=myds %>% filter(Attrition=="No") %>% ggplot(aes(TotalWorkingYears,..count..,fill=factor(Attrition))
```

```
grid.arrange(g1,g2,g3,nrow=3)
```







Boxplot and histogram shows that there is a siginificant difference between the number of experience with attrition levels.

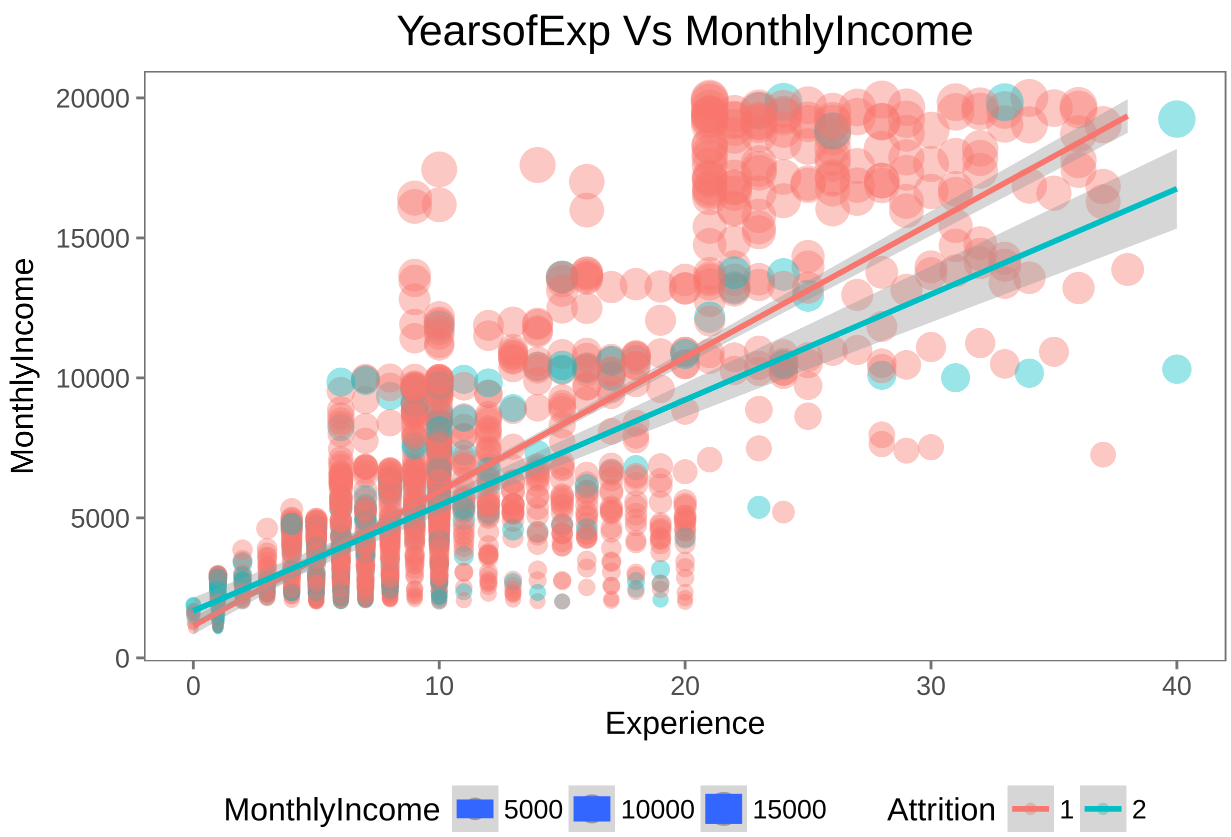
It is noted that people with less than 10 years of experience prefer to jump to another company whereas after that the jump drops.

The histgram for both the attrition levels is right skewed.

43 plot a scatter plot for years of experience vs monthly salary and see the correlation

geom_smooth,geom_point

```
ggplot(myds,aes(TotalWorkingYears,MonthlyIncome,size=MonthlyIncome,col=factor(Attrition)))+geom_point(a
```

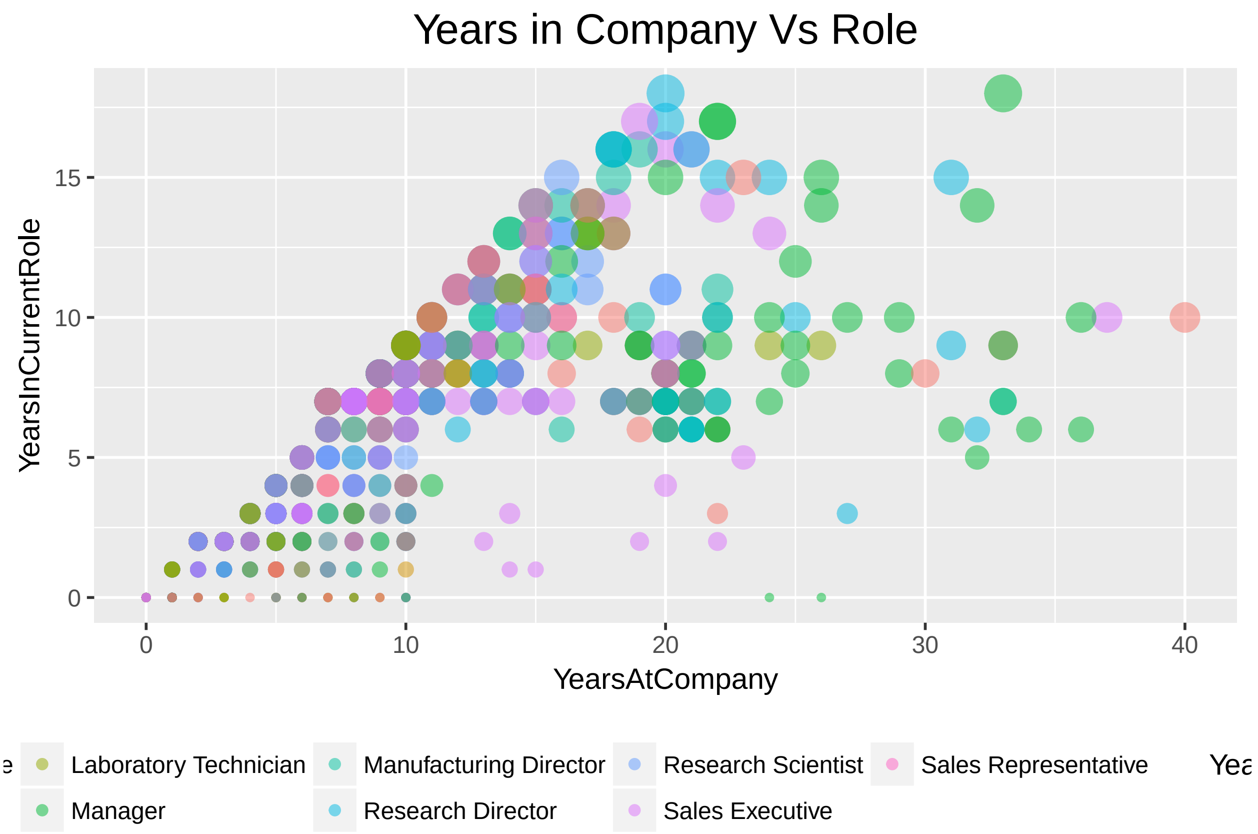


As expected,there exists a linear relationship between years of experience and monthly income as shown by the line.

There is a point in the graph,where the lines seems to intersect after which the no attrition line has higher monthly income compared to yes attrition line.

44 Analsysis on Specific role based tenure duration

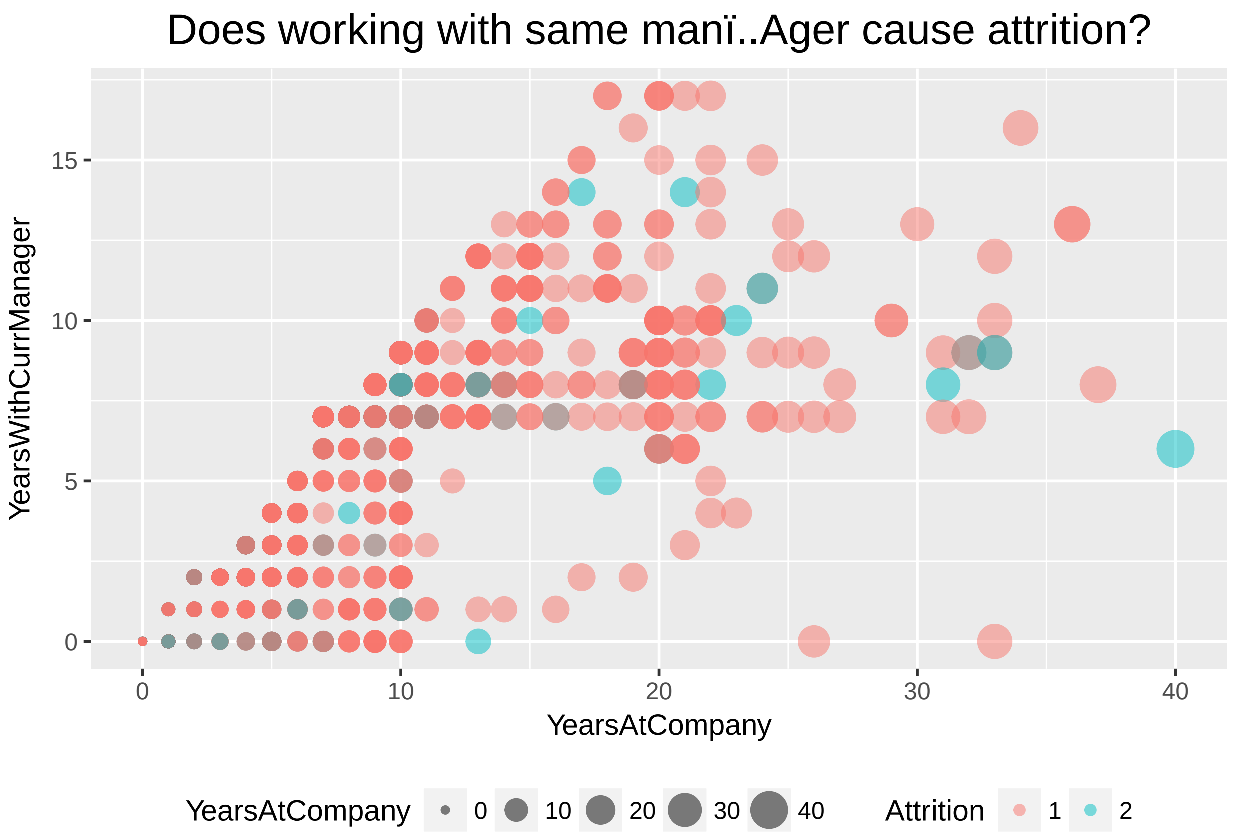
```
ggplot(myds,aes(YearsAtCompany,YearsInCurrentRole,col=factor(JobRole),size=YearsInCurrentRole))+geom_poi
```



45 we see direct relation in these tw variables

Working under same manï..Ager causes attrition

```
ggplot(myds,aes(YearsAtCompany,YearsWithCurrManager,col=factor(Attrition),size=YearsAtCompany))+geom_po
```



Observation: We get no clear relation as its scattered

46 Attrition Vs Categorical Variables:

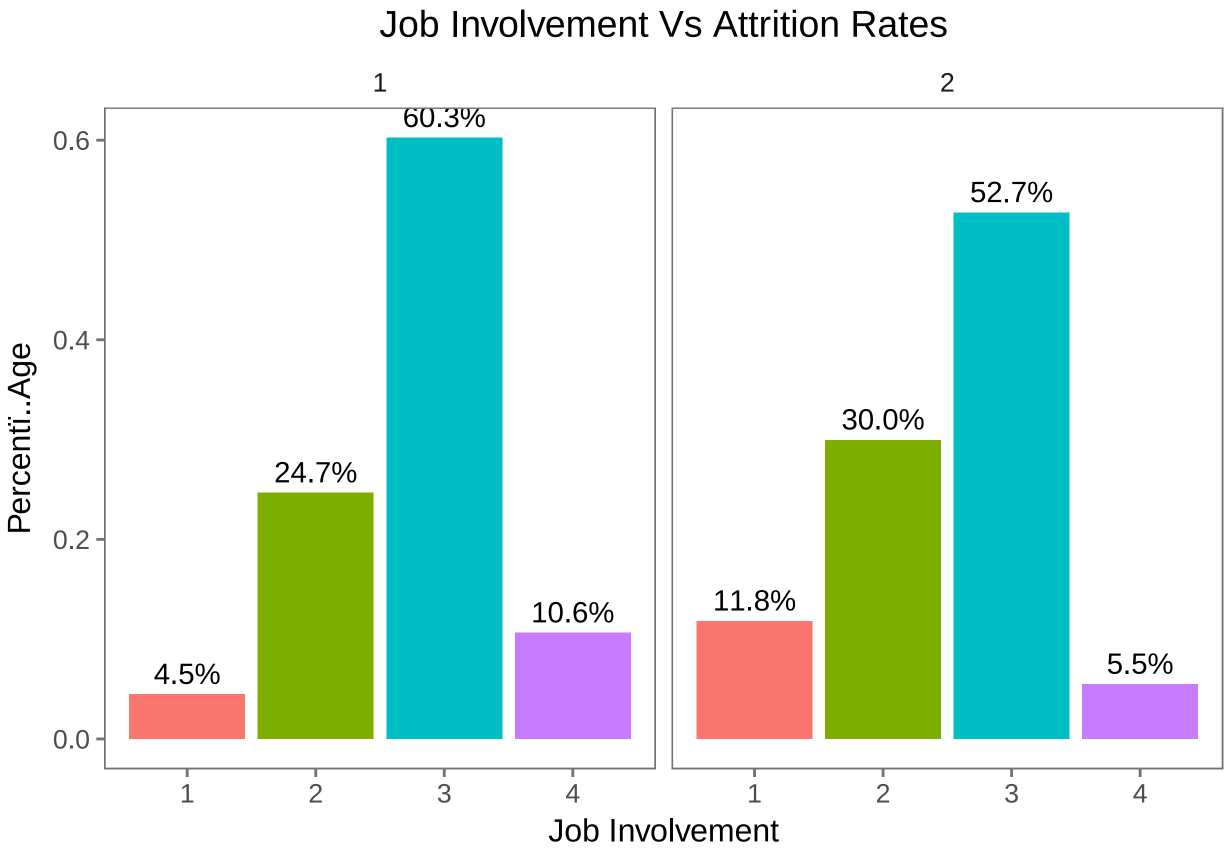
```
temp4 = myds %>% mutate(JobInvolvement=factor(JobInvolvement)) %>% mutate(JobInvolvement=fct_recode(JobI
round((prop.table(table(temp$JobInvolvement)))*100,2)
```

```
##
##     1     2     3     4
##  5.65 25.51 59.05  9.80
```

59 % have high job involvement whereas 25 % have medium involvement in the job.Let us check how this relates to attrition.

47 ggplotting for Job Involvement vs Attrition Rates

```
ggplot(temp,aes(x=JobInvolvement,group=Attrition))+geom_bar(stat="count",aes(y=..prop..,fill=factor(..x
```



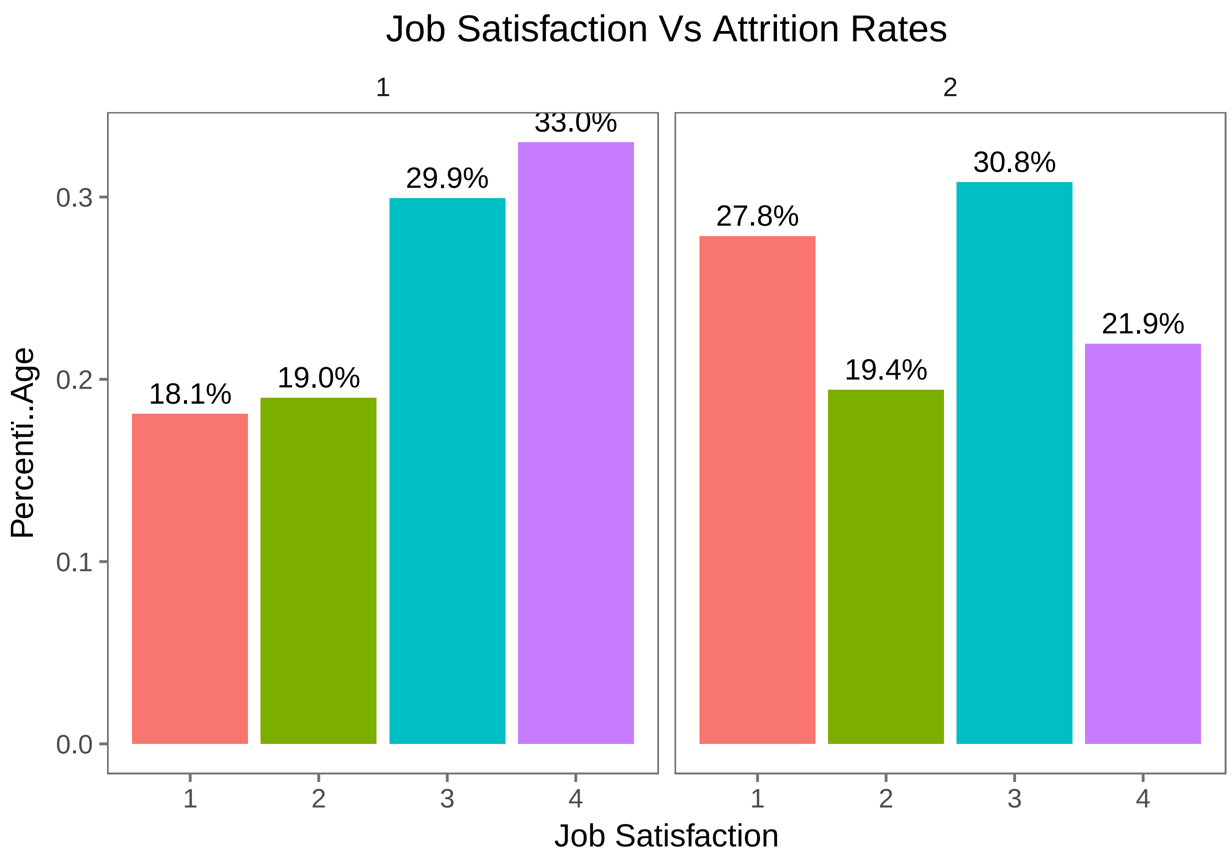
People with high job involvement have higher attrition rates followed by medium involvement people.But,equal number of percentï..Age of people have also shown no attrition rates.

48Job Satisfaction

Creating subsets with temp name

```
temp5 = myds %>% mutate(JobSatisfaction=factor(JobSatisfaction)) %>% mutate(JobSatisfaction=fct_recode(
```

```
ggplot(temp,aes(x=JobSatisfaction,group=Attrition))+geom_bar(stat="count",aes(y=..prop..,fill=factor(..
```



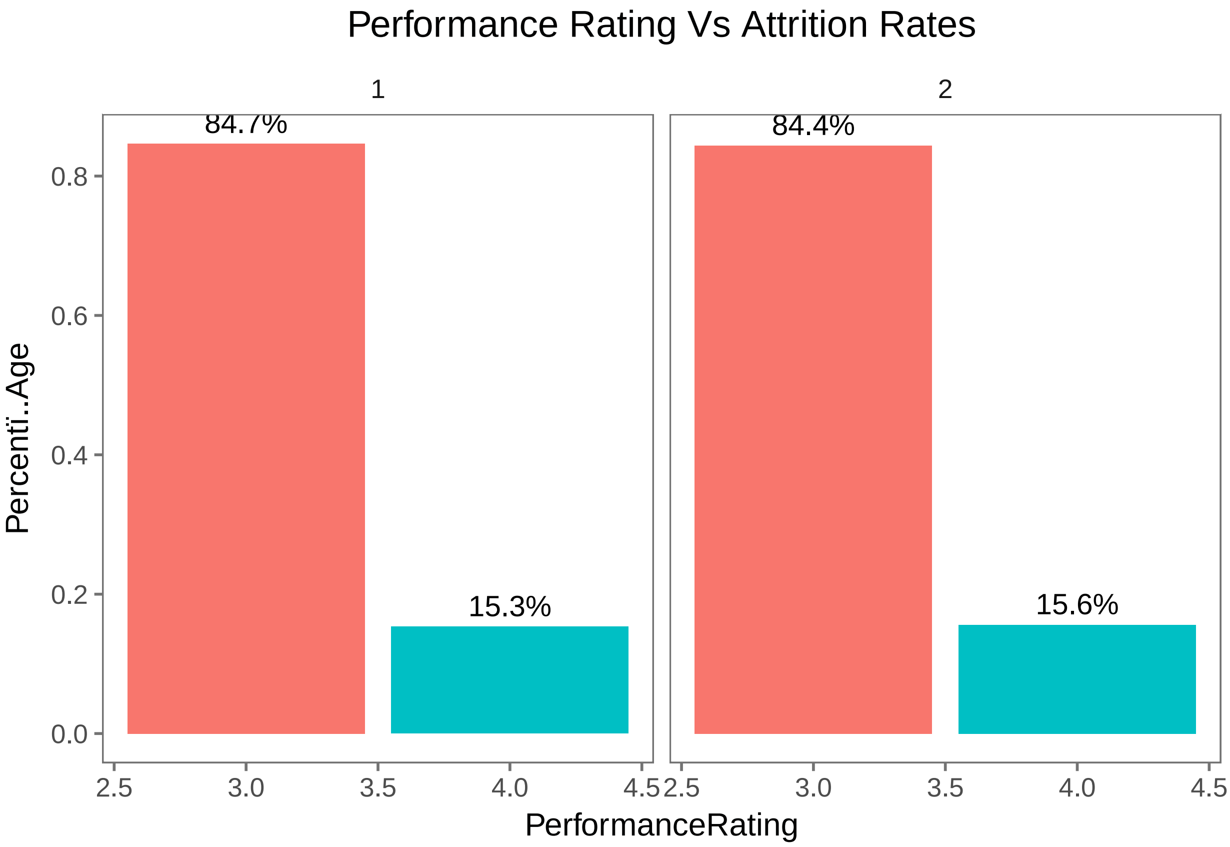
Clearly its visible Job Satisfaction play important vaiarable role in understanding

out of those who leave about 30.8 % have experience high job satisfaction.Therefore,there should be some other factor which triggers their exit from the present company.

49 Performance Rating:

```
temp6 = myds %>% mutate(PerformanceRating=factor(PerformanceRating)) %>% mutate(PerformanceRating=fct_r
```

```
ggplot(temp,aes(x=PerformanceRating,group=Attrition))+geom_bar(stat="count",aes(y=..prop..,fill=factor(
```

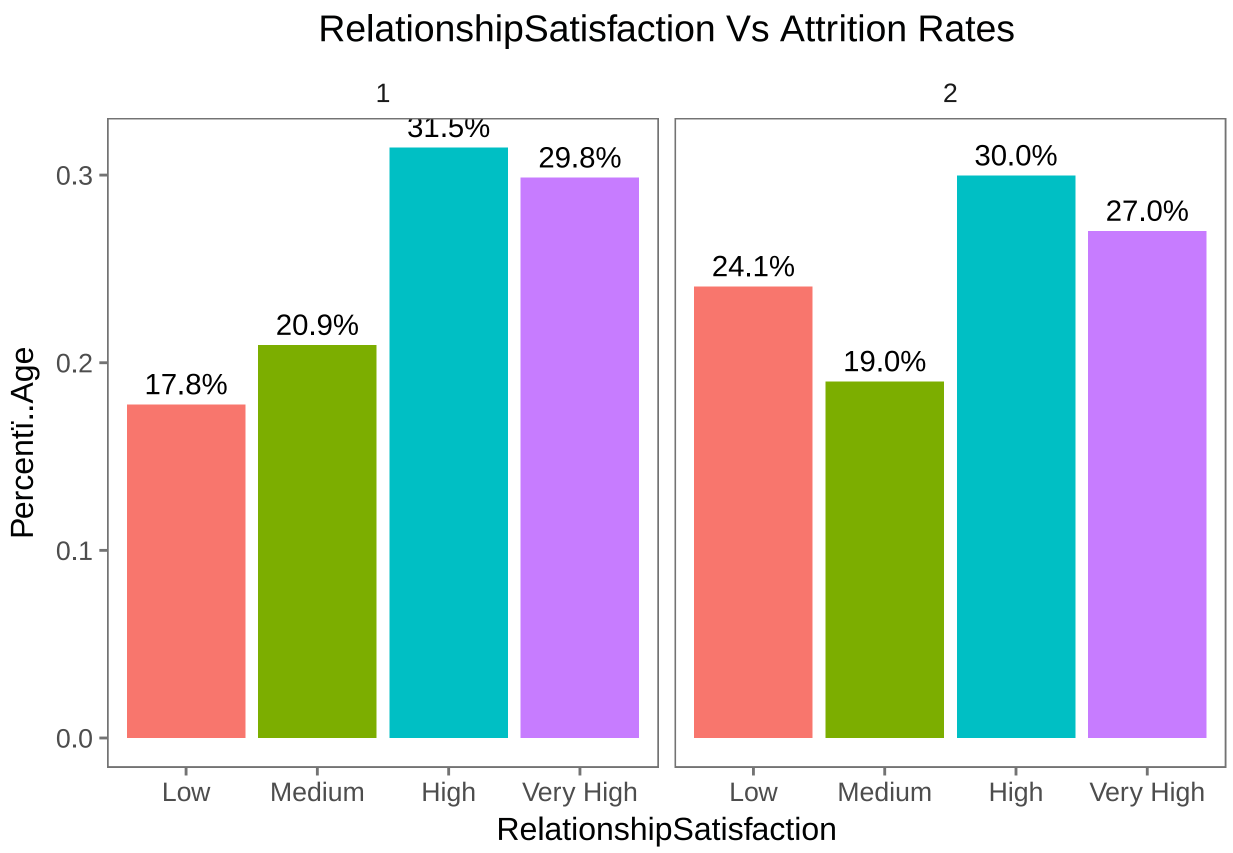


Variable rating for low = 1, gppd = 2 is not aviable in set

Observation: Same percentï..Age which shows no impact of it.

50 Relationship Satisfaction:

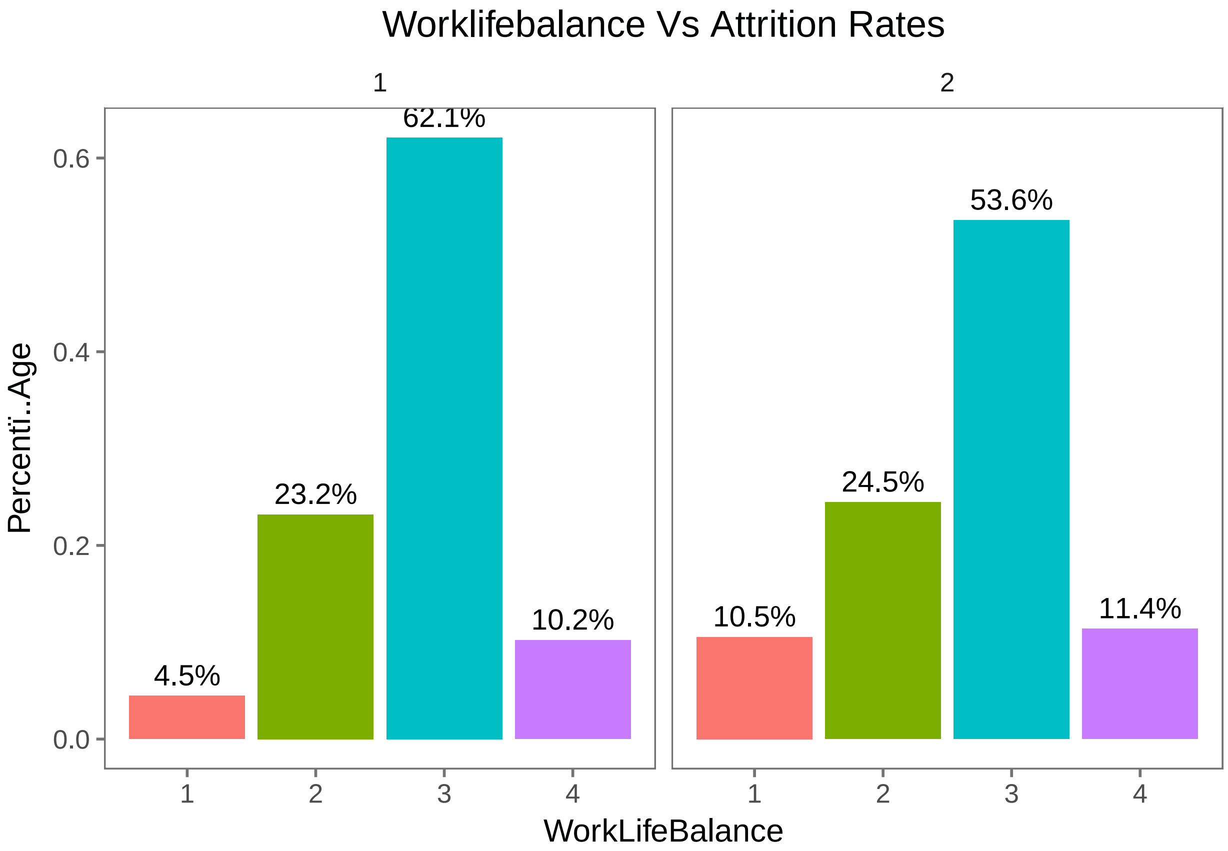
```
temp= myds%>% mutate(RelationshipSatisfaction=factor(RelationshipSatisfaction)) %>% mutate(Relationship
ggplot(temp,aes(x=RelationshipSatisfaction,group=Attrition))+geom_bar(stat="count",aes(y=..prop..,fill=
```



In this too,we find that almost 57 % (combining high and very high) experience attrition whereas similar percentï..Age have also stayed within the company.

51 Worklife balance:

```
temp7 = myds %>% mutate(WorkLifeBalance=factor(WorkLifeBalance)) %>% mutate(WorkLifeBalance=fct_recode(
ggplot(temp,aes(x=WorkLifeBalance,group=Attrition))+geom_bar(stat="count",aes(y=..prop..,fill=factor(..
```

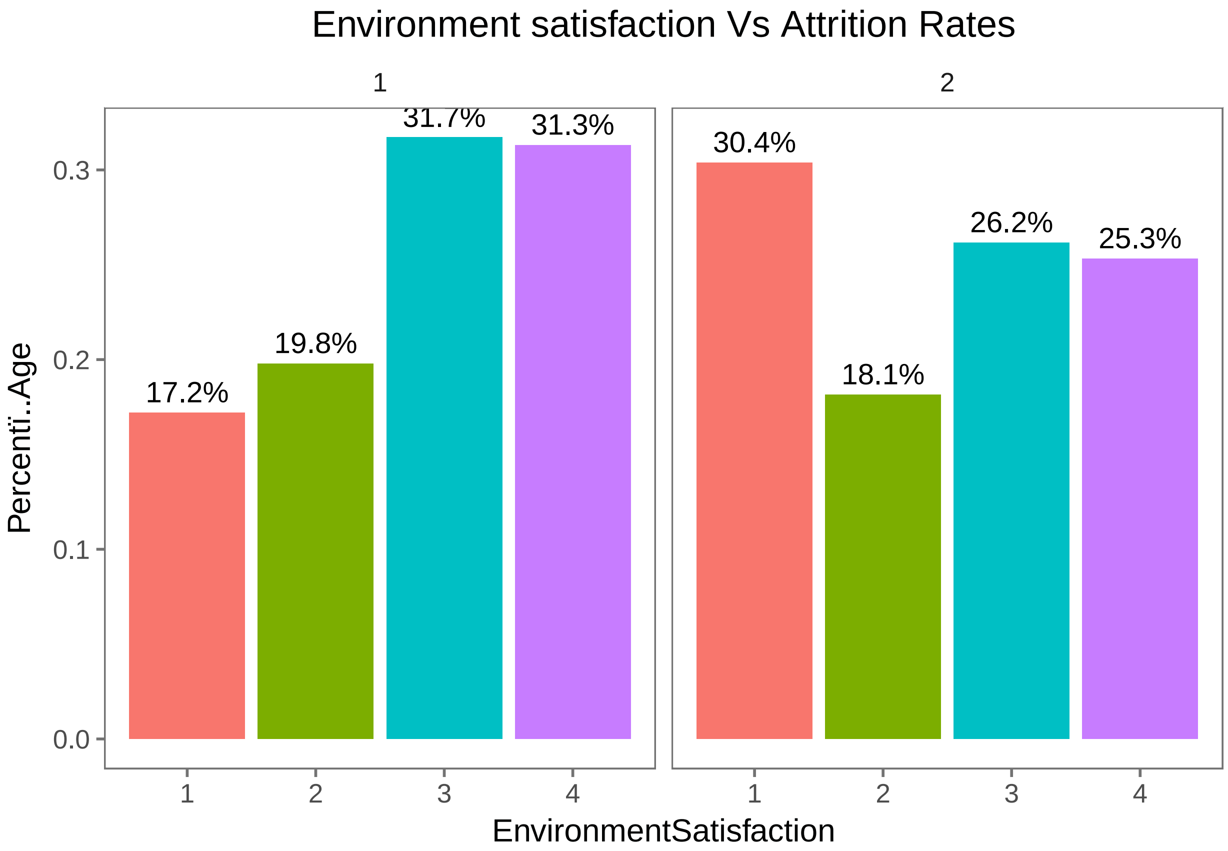


in this also we do not find any major conlusion

42 Environment Satisfaction:

```
temp8 = myds %>% mutate(EnvironmentSatisfaction=factor(EnvironmentSatisfaction)) %>% mutate(Environment
```

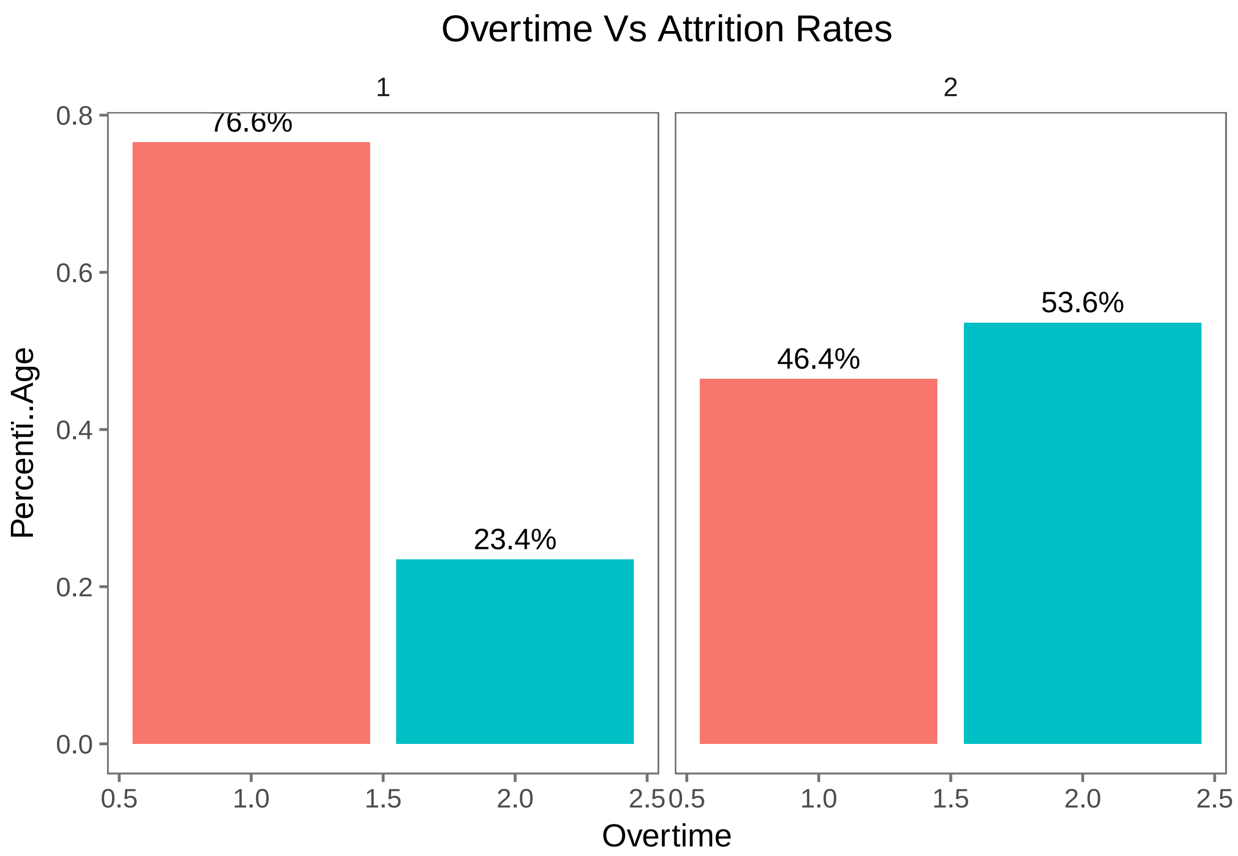
```
ggplot(temp,aes(x=EnvironmentSatisfaction,group=Attrition))+geom_bar(stat="count",aes(y=..prop..,fill=f
```



Here we see that people having low environment satisfaction (30.4%) leave the company.

53. Attrition Vs OverTime:

```
ggplot(myds,aes(x=OverTime,group=Attrition))+geom_bar(stat="count",aes(y=..prop..,fill=factor(..x..)))+
```



53 % of those who experience attrition have worked overtime whereas 76 % of those who have not experienced overtime have not left the company.Therefore overtime is a strong indicator of attrition.

54. Attrition VS Training times last year

```
t.test(myds$TrainingTimesLastYear~myds$Attrition)
```

```
##
##  Welch Two Sample t-test
##
## data:  myds$TrainingTimesLastYear by myds$Attrition
## t = 2.3305, df = 339.56, p-value = 0.02036
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.03251776 0.38439273
## sample estimates:
## mean in group 1 mean in group 2
##        2.832928        2.624473
```

As p-value is less than alpha, attrition rate depends on trainings.

55. Attrition VS Work/Life Balance

```
table_balance<-table(myds$WorkLifeBalance, myds$Attrition)
chisq.test(table_balance)
```

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

Attrition is dependent on Work/Life balance because p-value is less than alpha.