Predicting Employee Attrition

Aaron Camacho

Introduction

"People are companies' most important assests" (Altman, 2017) and the cost to replace those assets is astonishing. "Josh Bersis of Deloitte believes the cost of losing an employee can range from tens of thousands of dollars to 1.5 – 2.0x the employee's annual salary" (Altman, 2017). If it is possible to predict which employees will leave the company and the potential reasons why, it can save companies hundreds of thousands of dollars. Also, by identifying the reasons why an employee might leave the company, the analysis can direct the company to make the necessary changes to increase employee morale and make a happier and more productive work place.

Research Question:

The purpose of this research paper is to answer the question: What are the most likely reasons for an employee to leave a company and can the turnover possibility be predicted before it happens?

Appearing simple at first, the question demands the consideration of various factors when attempting to determine whether one variable has a greater influence than another. There are 35 variables in the data set that could influence an employee to make a determination to leave the company: Age, Attrition, Business Travel, Daily Rate, Department, Distance From Home, Education, Education Field, Employee Count, Employee Number, Environment Satisfaction, Gender, Hourly Rate, Job involvement, Job Level, Job Role, Job Satisfaction, Marital Status, Monthly Income, Monthly Rate, Number of Companies Worked, Over 18, Over Time, Percent Salary Hike, Performance Rating, Relationship Satisfaction, Standard Hours, Stock Option Level, Total Working Years, Training Times Last Year, Work Life Balance, Years at company, Years in Current Role, Years Since Last Promotion, Years With Current Manager.

Hypothesis:

If an employee has more tenure with the company, he/she is less likely to leave. To answer the research question above and to approve or disprove the hypothesis, logistic regression will be used due to its ability to handle both categorical and numeric variables. However, other standard models will be compared to discover the best approach (i.e. decision tree, random forest, CART, etc.).

Data Collection

The first step is to collect and extract the data. The data itself comes from IBM through "Watson Analytics" and is a sample data set. It is a data set that can be used to explore the factors that may lead an employee to leave. The data was extracted with the following code:

```
import urllib.request
dls = "https://community.watsonanalytics.com/wp-content/uploads/2015/03/WA_Fn-UseC_-HR-Employee-Attrition.xlsx"
urllib.request.urlretrieve(dls, "WA_Fn-UseC_-HR-Employee-Attrition.xls")
```

Using Python to collect the data is quick and efficient. However, to extract, prepare and analyze the data R will be used for its ease of performing statistical calculations and visual representations.

Data Extraction and Preparation

The target variable is "Attrition" it is a Nominal categorical binary variable stated as a "Yes" or a "No" response. Attrition in the business world means the loss of employees through normal avenues (i.e. retirement, quitting, etc.). As we extract, prepare and analyze the data, possible predictor variables will also be identified.

First the data will be loaded into R with the following code:

```
# Load the Telco Churn Data
attrition <- read.csv("~/Desktop/r_intro/employee_attrition.csv")
```

As mentioned above there are 35 variables that IBM has collected on satisfaction, income, demographics and tenure. The data set incudes 1,470 rows each representing an employee. Below is the structure of the 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",..:
$ 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
$ 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
$ 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
$ 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
$ MonthlyIncome           : int  5993 5130 2090 2909 3468 3068 2670 2693 9526 5237 ...
$ MonthlyRate             : int  19479 24907 2396 23159 16632 11864 9964 13335 8787
$ 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 ...
```

One of the first things that stands out is the variable “ï..Age”. Such a titling will make data analysis difficult. We will change the variable “ï..Age” to “Age” with the following code segment.

```
#Rename Column"ï..Age" to "Age"
colnames(attrition)[colnames(attrition)=="ï..Age"] <- "Age"
```

Exploratory Data Analysis:

First, the summary() function is used to gain a brief yet deeper understanding of the variables.

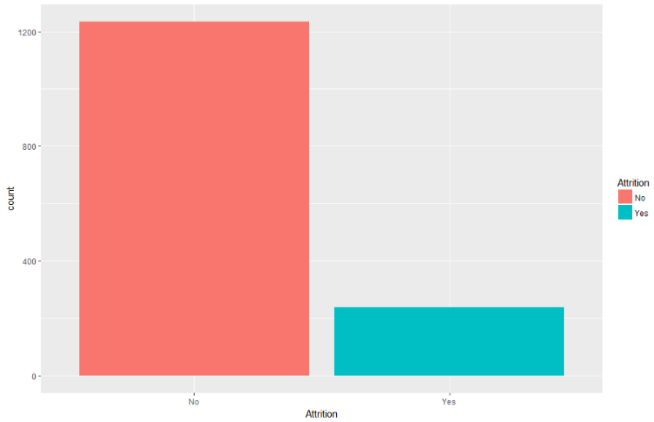
```
summary(attrition)
     i..Age       Attrition            BusinessTravel    DailyRate                        Department
 Min.   :18.00   No :1233   Non-Travel       : 150   Min.   : 102.0   Human Resources       : 63
 1st Qu.:30.00   Yes: 237   Travel_Frequently: 277   1st Qu.: 465.0   Research & Development:961
 Median :36.00              Travel_Rarely    :1043   Median : 802.0   Sales                 :446
 Mean   :36.92                                       Mean   : 802.5
 3rd Qu.:43.00                                       3rd Qu.:1157.0
 Max.   :60.00                                       Max.   :1499.0

 DistanceFromHome   Education            EducationField EmployeeCount EmployeeNumber
 Min.   : 1.000   Min.   :1.000   Human Resources : 27   Min.   :1     Min.   :   1.0
 1st Qu.: 2.000   1st Qu.:2.000   Life Sciences   :606   1st Qu.:1     1st Qu.: 491.2
 Median : 7.000   Median :3.000   Marketing       :159   Median :1     Median :1020.5
 Mean   : 9.193   Mean   :2.913   Medical         :464   Mean   :1     Mean   :1024.9
 3rd Qu.:14.000   3rd Qu.:4.000   Other           : 82   3rd Qu.:1     3rd Qu.:1555.8
 Max.   :29.000   Max.   :5.000   Technical Degree:132   Max.   :1     Max.   :2068.0

 EnvironmentSatisfaction    Gender      HourlyRate     JobInvolvement     JobLevel
 Min.   :1.000           Female:588   Min.   : 30.00   Min.   :1.00   Min.   :1.000
 1st Qu.:2.000           Male  :882   1st Qu.: 48.00   1st Qu.:2.00   1st Qu.:1.000
 Median :3.000                        Median : 66.00   Median :3.00   Median :2.000
 Mean   :2.722                        Mean   : 65.89   Mean   :2.73   Mean   :2.064
 3rd Qu.:4.000                        3rd Qu.: 83.75   3rd Qu.:3.00   3rd Qu.:3.000
 Max.   :4.000                        Max.   :100.00   Max.   :4.00   Max.   :5.000
```

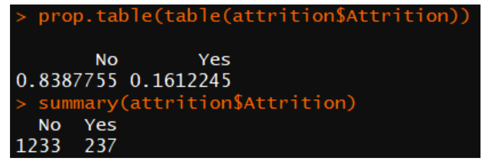
```
                      JobRole     JobSatisfaction  MaritalStatus MonthlyIncome    MonthlyRate
 Sales Executive          :326   Min.   :1.000   Divorced:327   Min.   : 1009   Min.   : 2094
 Research Scientist       :292   1st Qu.:2.000   Married :673   1st Qu.: 2911   1st Qu.: 8047
 Laboratory Technician    :259   Median :3.000   Single  :470   Median : 4919   Median :14236
 Manufacturing Director   :145   Mean   :2.729                  Mean   : 6503   Mean   :14313
 Healthcare Representative:131   3rd Qu.:4.000                  3rd Qu.: 8379   3rd Qu.:20462
 Manager                  :102   Max.   :4.000                  Max.   :19999   Max.   :26999
 (Other)                  :215
 NumCompaniesWorked Over18    OverTime   PercentSalaryHike PerformanceRating RelationshipSatisfaction
 Min.   :0.000      Y:1470   No :1054   Min.   :11.00     Min.   :3.000     Min.   :1.000
 1st Qu.:1.000               Yes: 416   1st Qu.:12.00     1st Qu.:3.000     1st Qu.:2.000
 Median :2.000                          Median :14.00     Median :3.000     Median :3.000
 Mean   :2.693                          Mean   :15.21     Mean   :3.154     Mean   :2.712
 3rd Qu.:4.000                          3rd Qu.:18.00     3rd Qu.:3.000     3rd Qu.:4.000
 Max.   :9.000                          Max.   :25.00     Max.   :4.000     Max.   :4.000

 StandardHours StockOptionLevel TotalWorkingYears TrainingTimesLastYear WorkLifeBalance YearsAtCompany
 Min.   :80    Min.   :0.0000   Min.   : 0.00     Min.   :0.000         Min.   :1.000   Min.   : 0.000
 1st Qu.:80    1st Qu.:0.0000   1st Qu.: 6.00     1st Qu.:2.000         1st Qu.:2.000   1st Qu.: 3.000
 Median :80    Median :1.0000   Median :10.00     Median :3.000         Median :3.000   Median : 5.000
 Mean   :80    Mean   :0.7939   Mean   :11.28     Mean   :2.799         Mean   :2.761   Mean   : 7.008
 3rd Qu.:80    3rd Qu.:1.0000   3rd Qu.:15.00     3rd Qu.:3.000         3rd Qu.:3.000   3rd Qu.: 9.000
 Max.   :80    Max.   :3.0000   Max.   :40.00     Max.   :6.000         Max.   :4.000   Max.   :40.000

 YearsInCurrentRole YearsSinceLastPromotion YearsWithCurrManager
 Min.   : 0.000     Min.   : 0.000          Min.   : 0.000
 1st Qu.: 2.000     1st Qu.: 0.000          1st Qu.: 2.000
 Median : 3.000     Median : 1.000          Median : 3.000
 Mean   : 4.229     Mean   : 2.188          Mean   : 4.123
 3rd Qu.: 7.000     3rd Qu.: 3.000          3rd Qu.: 7.000
 Max.   :18.000     Max.   :15.000          Max.   :17.000
```

Next we will take a look at our main variable of interest, Attrition. Approximately 16 percent of the work force in the data set quit.

Attrition:

```
# Attrition
ggplot(attrition,aes(Attrition,fill=Attrition))+geom_bar()
prop.table(table(attrition$Attrition))
summary(attrition$Attrition)
```



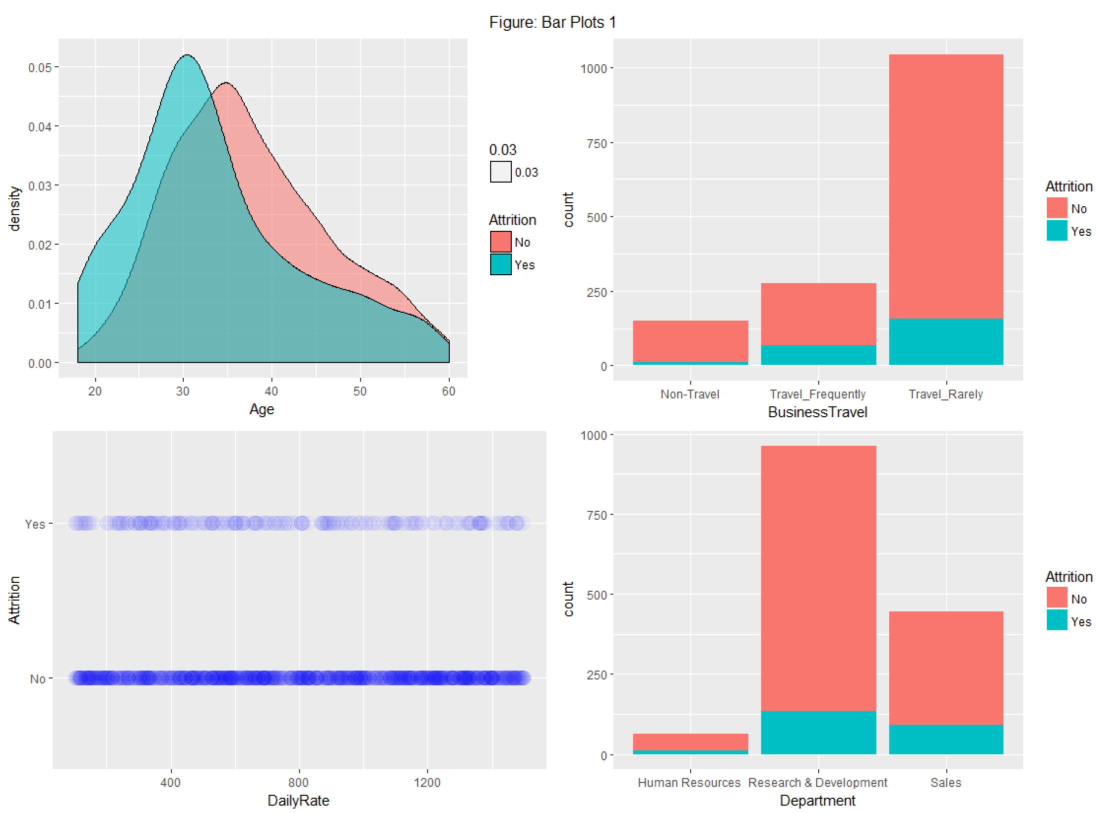


```
> prop.table(table(attrition$Attrition))

       No       Yes
0.8387755 0.1612245
> summary(attrition$Attrition)
  No  Yes
1233  237
```

Next we will evaluate each variable in the data set and how it relates to attrition in the data set.

Bar Plots 1: Age, BusinessTravel, DailyRate, Department:

```
# Bar Plots 1: Age, BusinessTravel, DailyRate, Department
p1 <- ggplot(attrition,aes(Age,fill=Attrition))+geom_density()+facet_grid(~Attrition)
p2 <- ggplot(attrition,aes(BusinessTravel,fill=Attrition))+geom_bar()
p3 <- ggplot(attrition,aes(DailyRate,Attrition))+geom_point(size=5,alpha = 0.03, col="b
lue")
p4 <- ggplot(attrition,aes(Department,fill = Attrition))+geom_bar()
grid.arrange(p1,p2,p3,p4,ncol=2,top = "Figure: Bar Plots 1")
```



Age: the majority of employees who leave approx. around 31 Years of age.

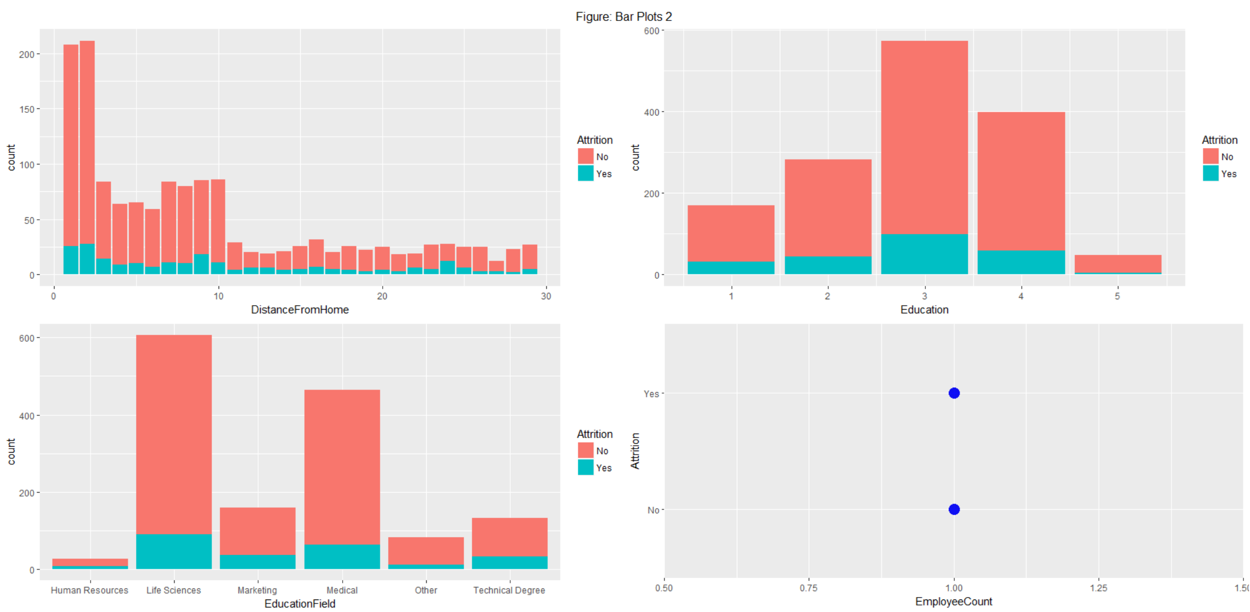
Business Travel: Employees who travel, are more likely to leave.

Daily Rate: There is no significant indications that can be found.

Department: R&D and Sales is where the most attrition occurred. However, it is important to note that the HR Department is proportionally smaller compared to the other departments.

Bar Plots 2: DistanceFromHome, Education, EducationField, EmployeeCount:

```
p5 <- ggplot(attrition,aes(DistanceFromHome,fill=Attrition))+geom_bar()
p6 <- ggplot(attrition,aes(Education,fill=Attrition))+geom_bar()
p7 <- ggplot(attrition,aes(EducationField,fill=Attrition))+geom_bar()
p8 <- ggplot(attrition,aes(EmployeeCount,Attrition))+geom_point(size=5,alpha = 0.03, col="blue")
grid.arrange(p5,p6,p7,p8,ncol=2,top = "Figure: Bar Plots 2")
```



Distance From Home: An unexpected result where employees who lived closer where more apt to leave.

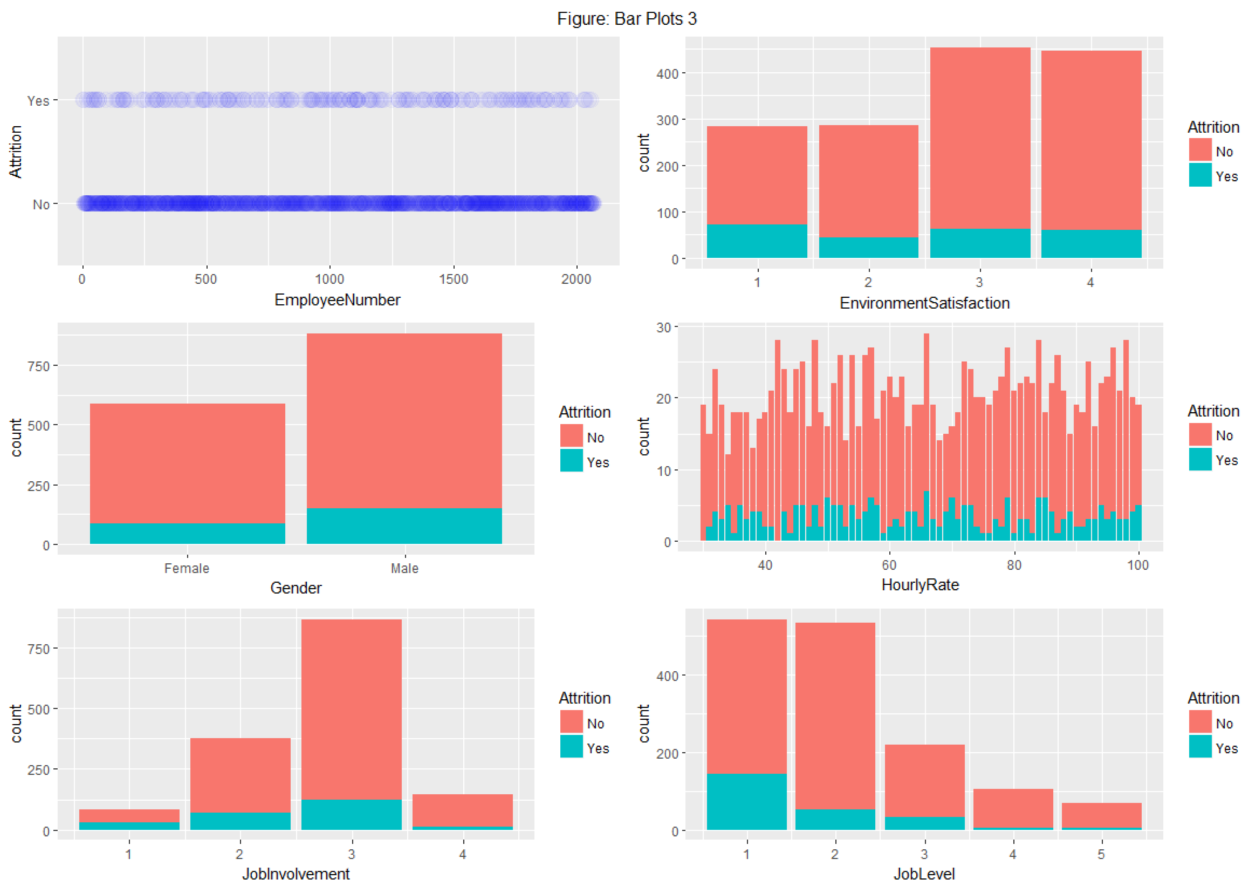
Education: 1 = "Below College", 2 = "College", 3 = "Bachelor", 4 = "Master", 5 = "Doctor" . Those with a bachelor's degree have the highest attrition. Important to note that there are very few employees with a doctorate degree. May have an impact on the amount that left in the Doctorate category.

Education Field: AS we saw in the Departments graph, those in an HR Field are less likely to leave. Again, this may be due to the low number of individuals in this group.

Employee Count: No significant findings. All numbers in variable are 1.

<u>Bar Plots 3: EmployeeNumber, EnvironmentSatisfaction, Gender, HourlyRate, JobInvolvement, JobLevel</u>

```
# Bar Plots 3: EmployeeNumber, EnvironmentSatisfaction, Gender, HourlyRate, JobInvolvement, JobLevel
p9 <- ggplot(attrition,aes(EmployeeNumber,Attrition))+geom_point(size=5,alpha = 0.03, col="blue")
p10 <- ggplot(attrition,aes(EnvironmentSatisfaction,fill=Attrition))+geom_bar()
p11 <- ggplot(attrition,aes(Gender,fill=Attrition))+geom_bar()
p12 <- ggplot(attrition,aes(HourlyRate,fill=Attrition))+geom_bar()
p13 <- ggplot(attrition,aes(JobInvolvement,fill=Attrition))+geom_bar()
p14 <- ggplot(attrition,aes(JobLevel,fill=Attrition))+geom_bar()
grid.arrange(p9,p10,p11,p12,p13,p14,ncol=2,top = "Figure: Bar Plots 3")
```



Employee Number: No significant findings.

Environment Satisfaction: 1 = "Low", 2 = "Medium", 3 = "High", 4 = "Very High". All levels are nearly the same. No significant findings.

Gender: Males are more likely to leave. However, there is 60% males and 40% female distribution which may be impacting the results.

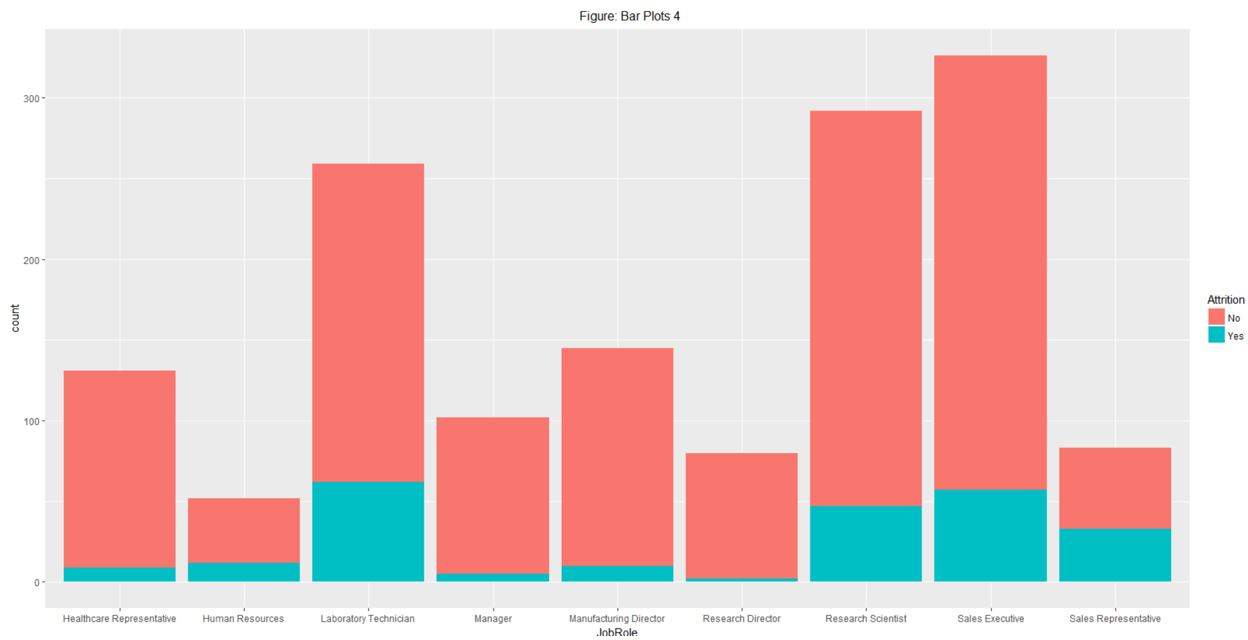
HourlyRate: No Significant findings. Also, there seems to be no direct relation to DailyRate.

Job Involvement: 1 = "Low", 2 = "Medium", 3 = "High", 4 = "Very High". It seems that the majority of employees who don't leave are either Very Highly involved or Low Involved in their Jobs. This may be correlated with the amount of pay they receive for the output of work performed.

JobLevel: An inferred meaning of ratings could be: 1 = "Entry level", 2 = "Junior Level", 3 = "Junior Manager", " 4 = "Senior level", 5 = "Senior Manger Level" but it is not sure. But, by looking at the graph it is clear that the high the job level the more unlikely an employee is to leave.

Bar Plots 4:JobRole

```
p15 <- ggplot(attrition,aes(JobRole,fill=Attrition))+geom_bar()
grid.arrange(p15,ncol=1,top = "Figure: Bar Plots 4")
prop.table(table(attrition$JobRole, attrition$Attrition))
```



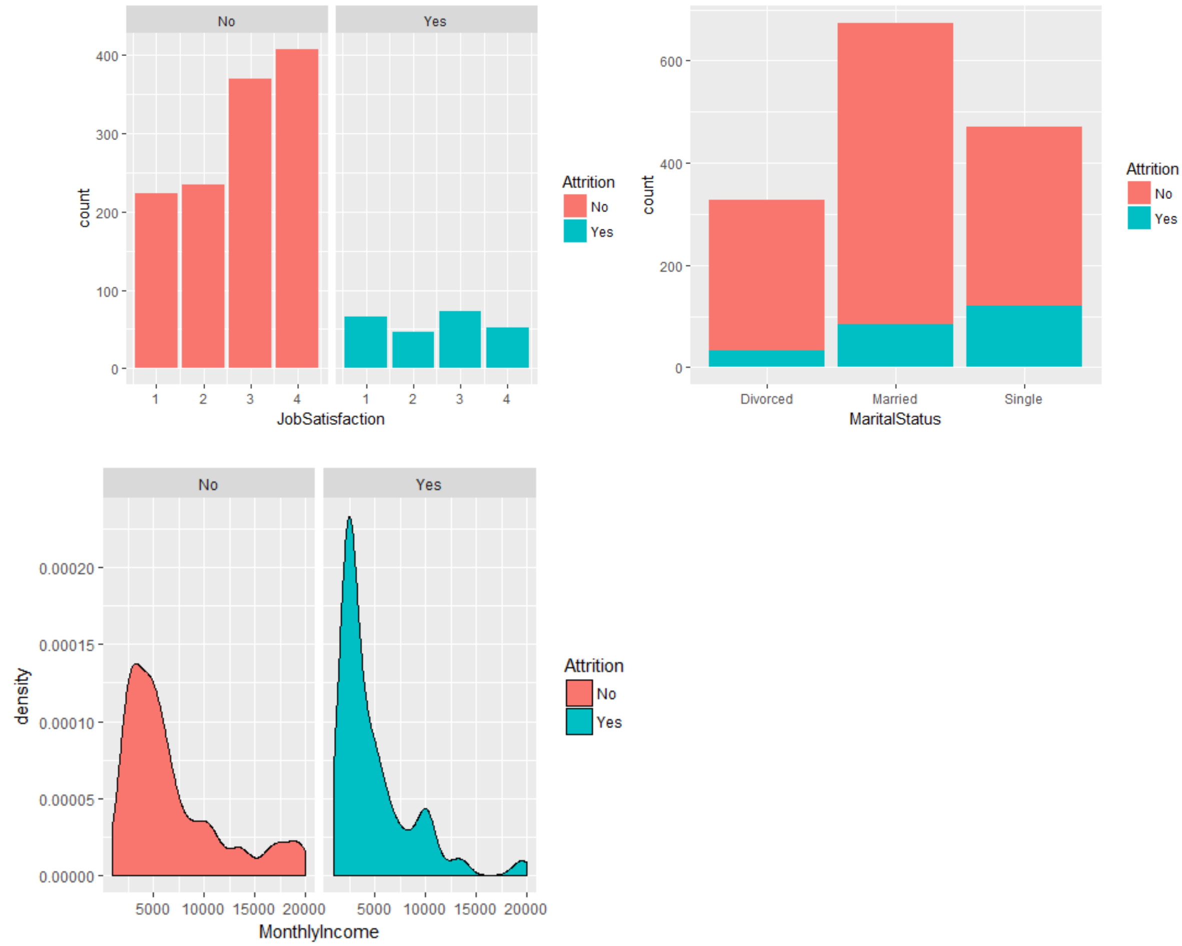

```
> prop.table(table(attrition$JobRole, attrition$Attrition))

                                     No         Yes
  Healthcare Representative 0.082993197 0.006122449
  Human Resources           0.027210884 0.008163265
  Laboratory Technician     0.134013605 0.042176871
  Manager                   0.065986395 0.003401361
  Manufacturing Director    0.091836735 0.006802721
  Research Director         0.053061224 0.001360544
  Research Scientist        0.166666667 0.031972789
  Sales Executive           0.182993197 0.038775510
  Sales Representative      0.034013605 0.022448980
```

Job Role: Proportions could be influenced to group size differences. However, the graph indicates that if an employee has one of the following job roles he/she is more likely to leave; Lab Tech, Research Scientist, Sales Executive, Sales Rep.

Bar Plots 5: JobSatisfaction, MaritalStatus, MonthlyIncome

```
p16 <- ggplot(attrition,aes(JobSatisfaction,fill=Attrition))+geom_bar()+facet_grid(~Attrition)
p17 <- ggplot(attrition,aes(MaritalStatus,fill=Attrition))+geom_bar()
p18 <- ggplot(attrition,aes(MonthlyIncome,fill=Attrition))+geom_density()+facet_grid(~Attrition)
grid.arrange(p16,p17,p18,ncol=2,top = "Figure: Bar Plots 5")
```

Figure: Bar Plots 5



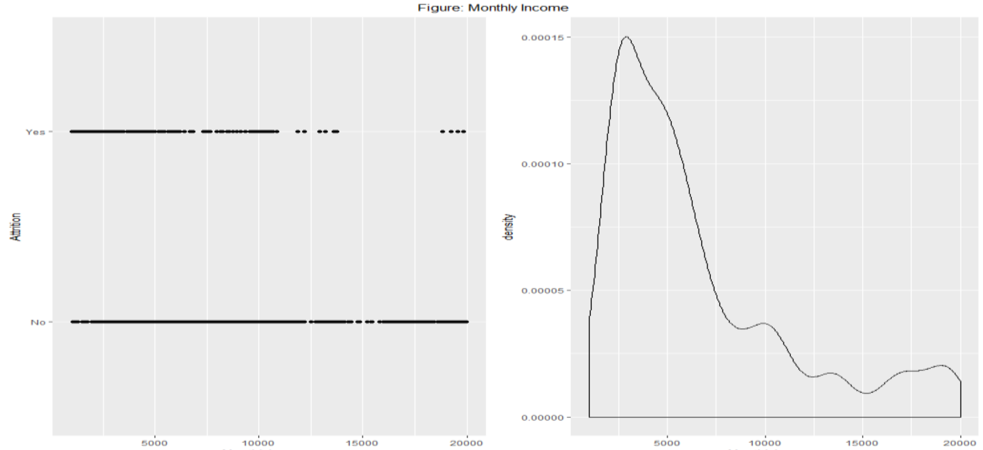
Job Satisfaction: 1 = "Low”, 2 = "Medium", 3 = "High", 4 = "Very High". Though attrition levels stay mostly the same, the number of employees who did not leave increases with job satisfaction.

Marital Status: Employees who are single are more likely to leave whereas, employees who are divorced are more likely to not leave.

Monthly Income: There are higher levels of attrition among the lower wage earners.

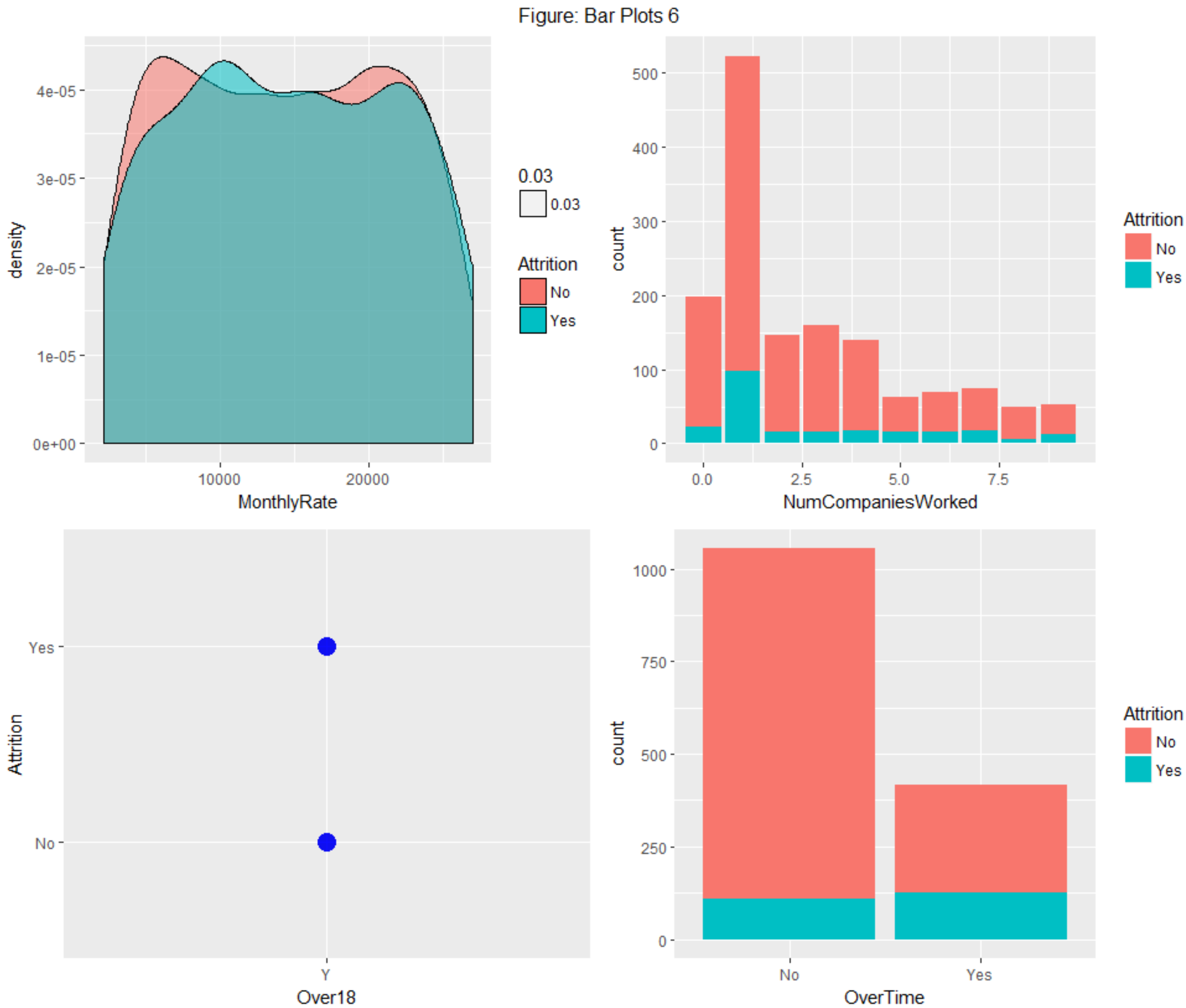
```
1. ggplot(attrition,aes(MonthlyIncome, Attrition))+geom_point()
2. ggplot(attrition,aes(MonthlyIncome))+geom_density()
```

Figure: Monthly Income



Bar Plots 6: MonthlyRate, NumCompaniesWorked, Over18, OverTime

```
p19 <- ggplot(attrition,aes(MonthlyRate,fill=Attrition))+geom_density()
p20 <- ggplot(attrition,aes(NumCompaniesWorked,fill=Attrition))+geom_bar()
p21 <- ggplot(attrition,aes(Over18,Attrition))+geom_point(size=5,alpha = 0.03, col="blue")
p22 <- ggplot(attrition,aes(OverTime,fill=Attrition))+geom_bar()
grid.arrange(p19,p20,p21,p22,ncol=2,top = "Figure: Bar Plots 6")
```



Monthly Rate: No Significant findings. Also, there seems to be little to no correlation to the Monthly Income variable.

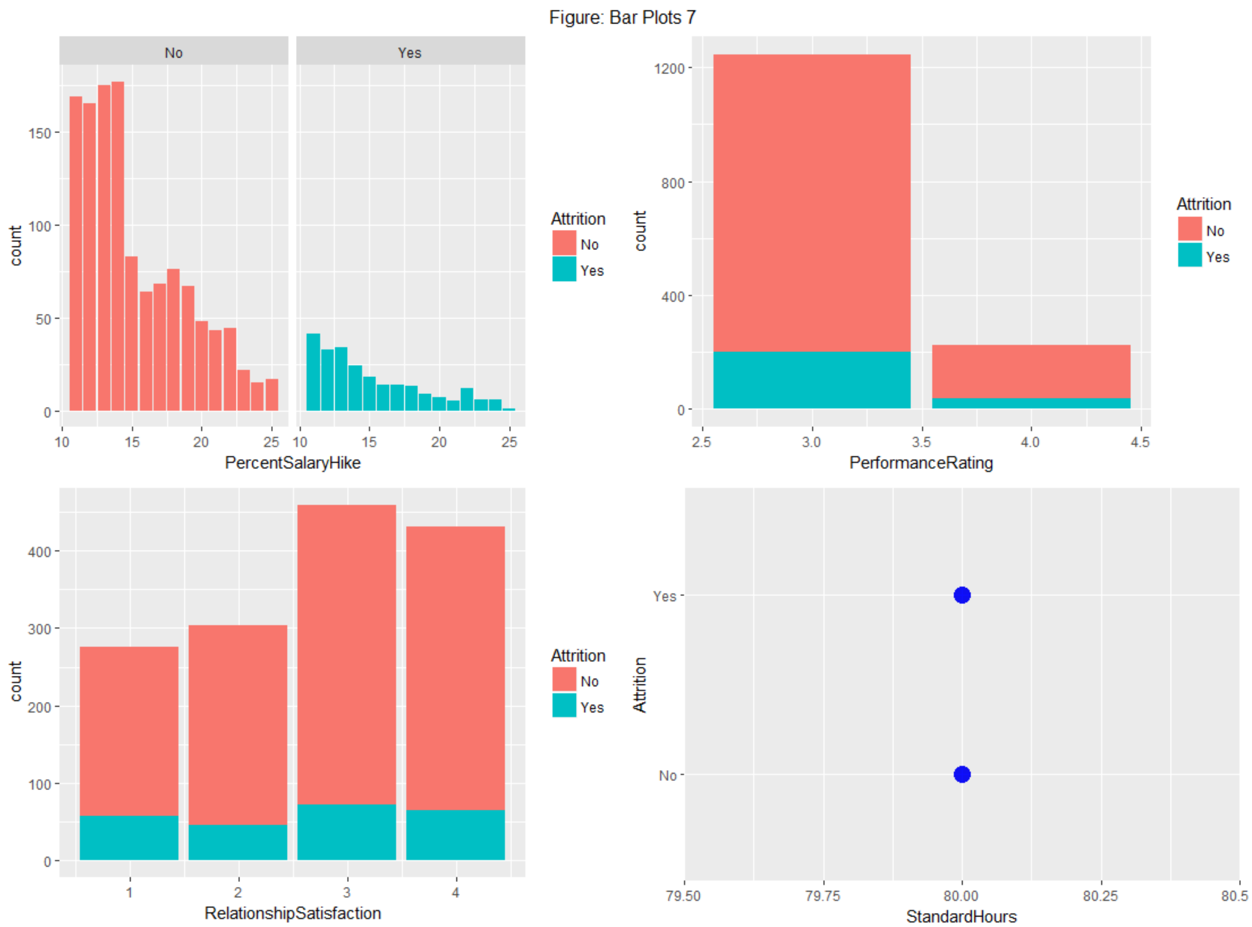
Number of Companies Worked: It is clear the if an employee has worked for only 1 company he/she is more likely to leave.

Over18: Not a significant variable. All employees are over 18 years old.

Over Time: Though attrition first appears to be nearly equal, a larger Proportion of employees working overtime are leaving.

<u>Bar Plots 7:PercentSalaryHike, PerformanceRating, RelationshipSatisfaction, StandardHours</u>

```
p23 <- ggplot(attrition,aes(PercentSalaryHike,fill=Attrition))+geom_bar()+facet_grid(~Attrition)
p24 <- ggplot(attrition,aes(PerformanceRating,fill = Attrition))+geom_bar()
p25 <- ggplot(attrition,aes(RelationshipSatisfaction,fill = Attrition))+geom_bar()
p26 <- ggplot(attrition,aes(StandardHours,Attrition))+geom_point(size=5,alpha = 0.03, col="blue")
grid.arrange(p23,p24,p25,p26,ncol=2,top = "Figure: Bar Plots 7")
```



Percent Salary Hike: Lower the percent salary hike equals more likely to leave.

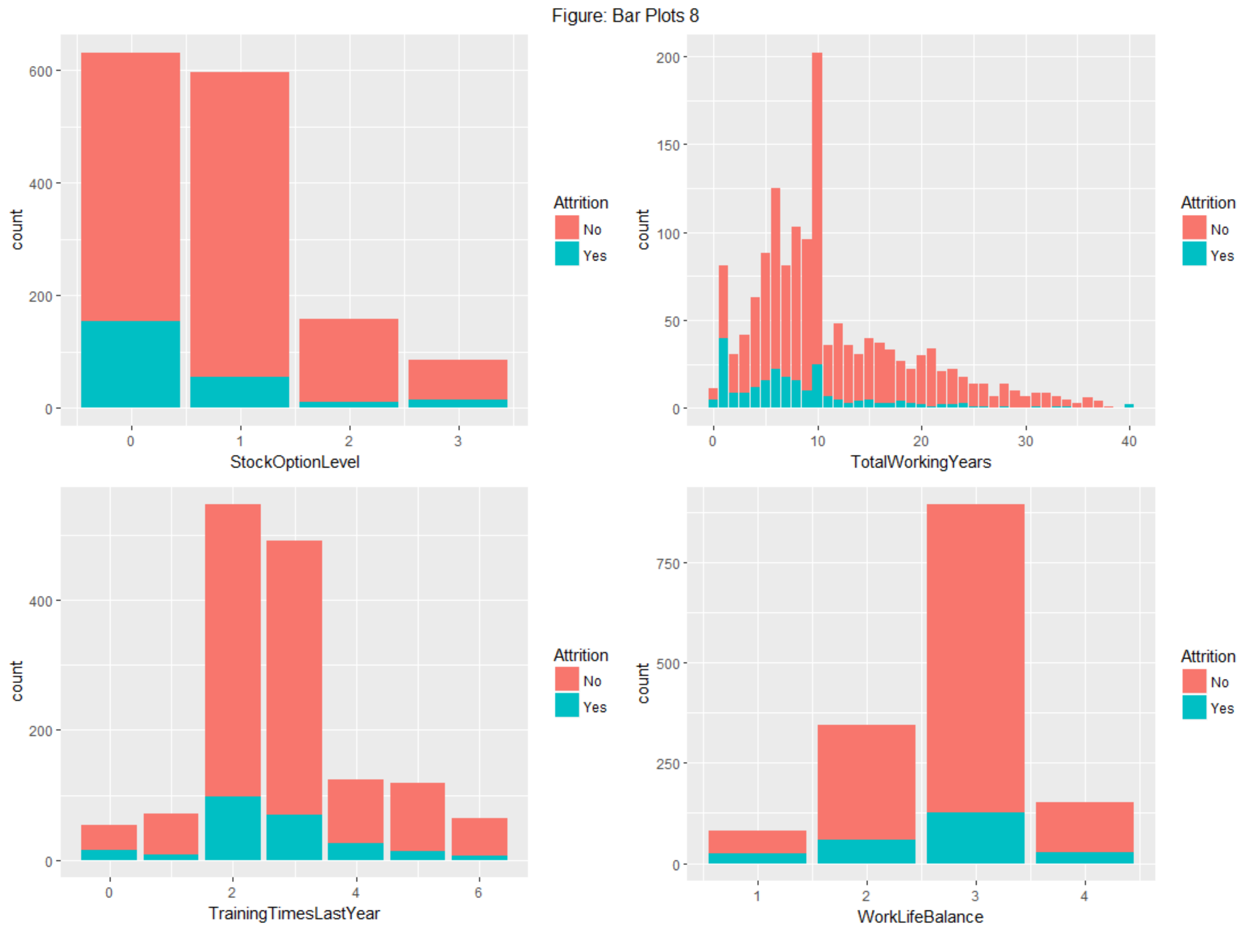
Performance Rating: 1 = "Low", 2 = "Good", 3 = "Excellent", 4 = "Outstanding". As expected, lower the performance rating more likely an employee is to leave.

Relationship Satisfaction: 1 = "Low", 2 = "Medium", 3 = "High", 4 = "Very High". Higher the relationship satisfaction the more employees don't leave.

Standard Hours: Not a significant variable. All employees have standard hours of 80.

Bar Plots 8:StockOptionLevel, TotalWorkingYears, TrainingTimesLastYear, WorkLifeBalance

```
p27 <- ggplot(attrition,aes(StockOptionLevel,fill = Attrition))+geom_bar()
p28 <- ggplot(attrition,aes(TotalWorkingYears,fill = Attrition))+geom_bar()
p29 <- ggplot(attrition,aes(TrainingTimesLastYear,fill = Attrition))+geom_bar()
p30 <- ggplot(attrition,aes(WorkLifeBalance,fill = Attrition))+geom_bar()
grid.arrange(p27,p28,p29,p30,ncol=2,top = "Figure: Bar Plots 8")
```



Stock Option Level: Larger the stock option level less likely an employee is to leave. It is expected that there would be more 0 and 1 levels because most employees would have very little to no stock options.

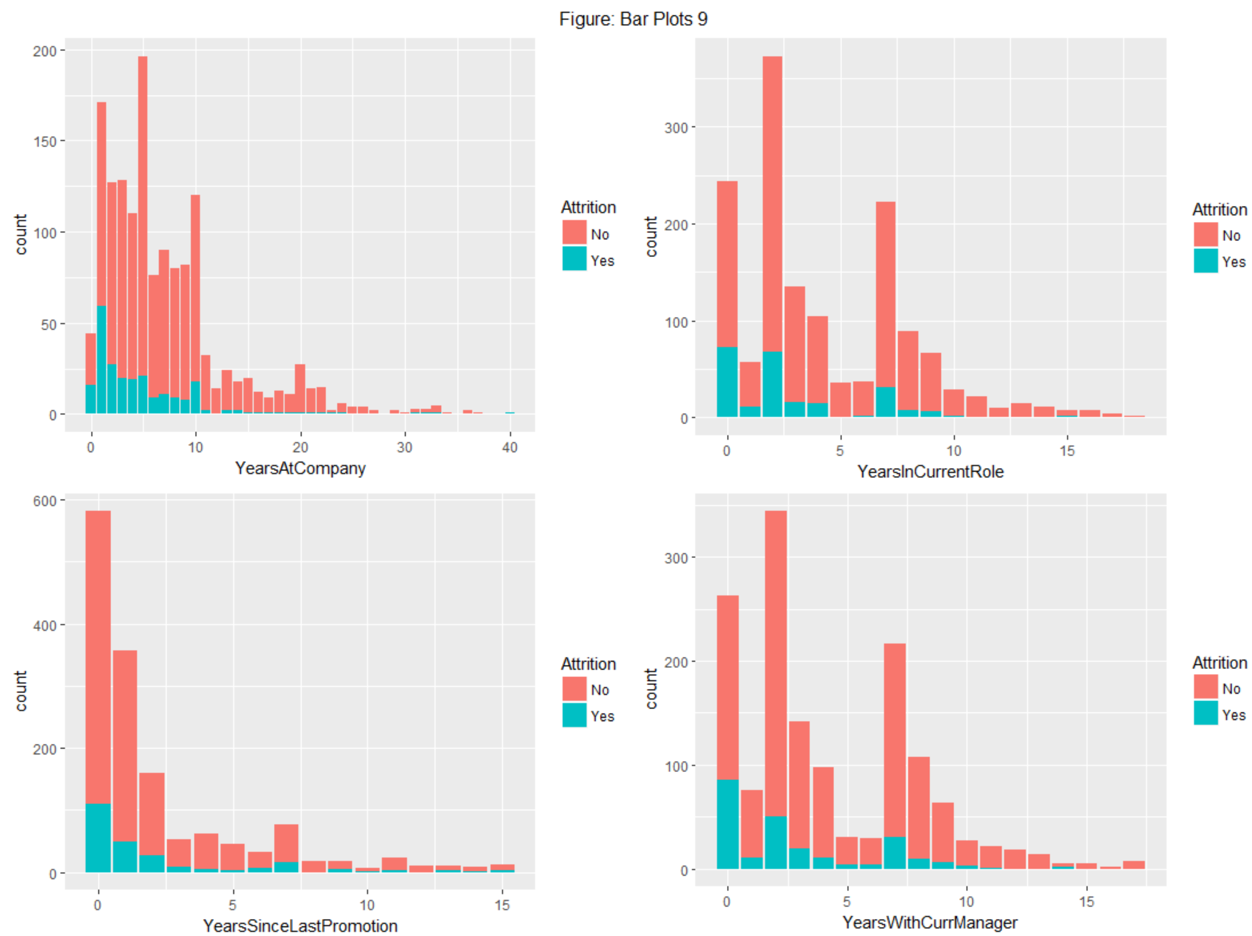
Total Working Years: The more years of working the less likely you are to leave. 1 year highly likely to leave. It appears years 0 to 12 have a high chance of attrition.

Training Times Last Year: 2 to 3 trainings seem to indicate a higher chance of attrition. Though the majority of employees seem to have 2 or 3 trainings.

Work Life Balance: 1 = "Bad", 2 = "Good", 3 = "Better", 4 = "Best". Those that have a higher work life balance are more likely to not leave.

Bar Plots 9: YearsAtCompany, YearsInCurrentRole, YearsSinceLastPromotion, YearsWithCurrentManager

```
p31 <- ggplot(attrition,aes(YearsAtCompany,fill = Attrition))+geom_bar()
p32 <- ggplot(attrition,aes(YearsInCurrentRole,fill = Attrition))+geom_bar()
p33 <- ggplot(attrition,aes(YearsSinceLastPromotion,fill = Attrition))+geom_bar()
p34 <- ggplot(attrition,aes(YearsWithCurrManager,fill = Attrition))+geom_bar()
grid.arrange(p31,p32,p33,p34,ncol=2,top = "Figure: Bar Plots 9")
```



Years at Company: Employees with less tenure are leaving more. However, that is also where the majority of employee tenure is, 0 to 10 years.

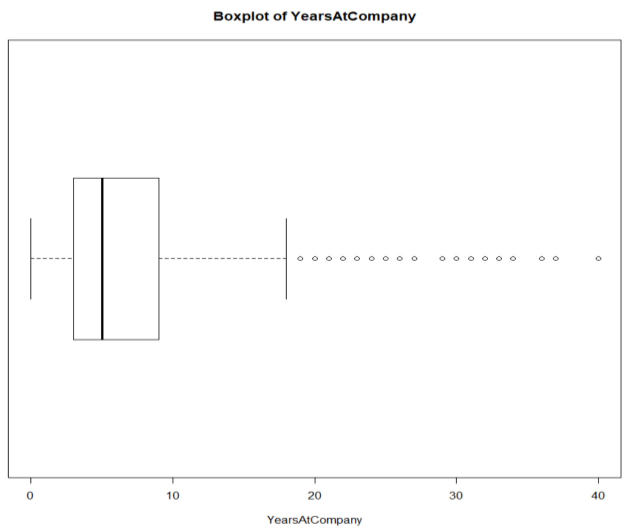
Years In Current Role: Employees with less years in role are leaving. However, we do not know if they just left for another position within the same company.

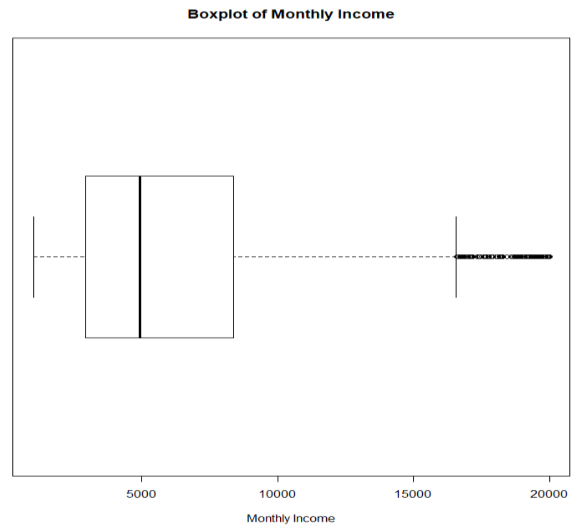
Years Since Last Promotion: It appears that those that have recently got a new promotion, 0 to 3 years, are more likely to leave.

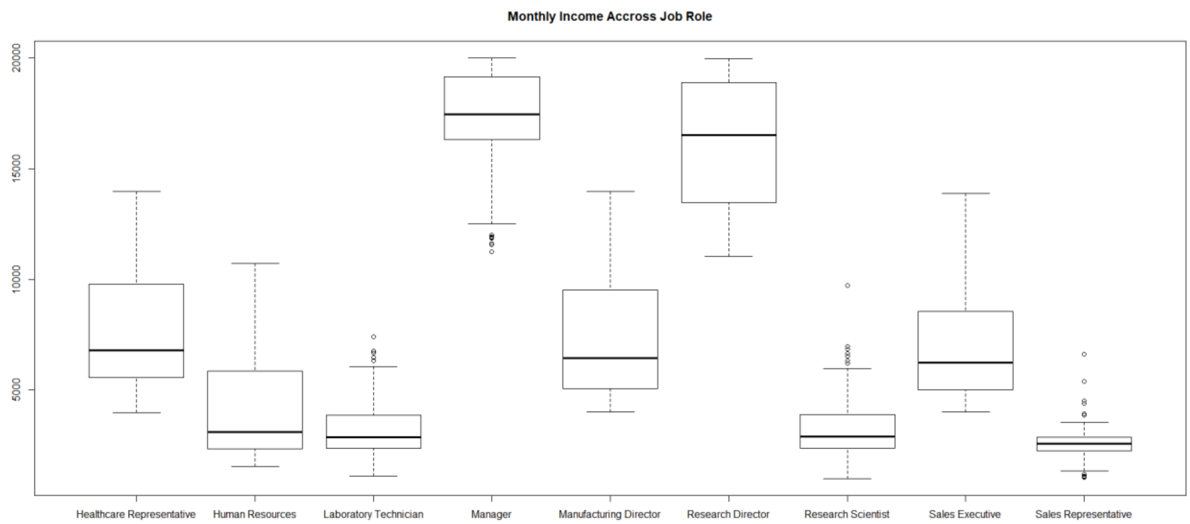
Years With Current Manager: Mangers play a large role in retention. Increased years with manager decreases chances of attrition.

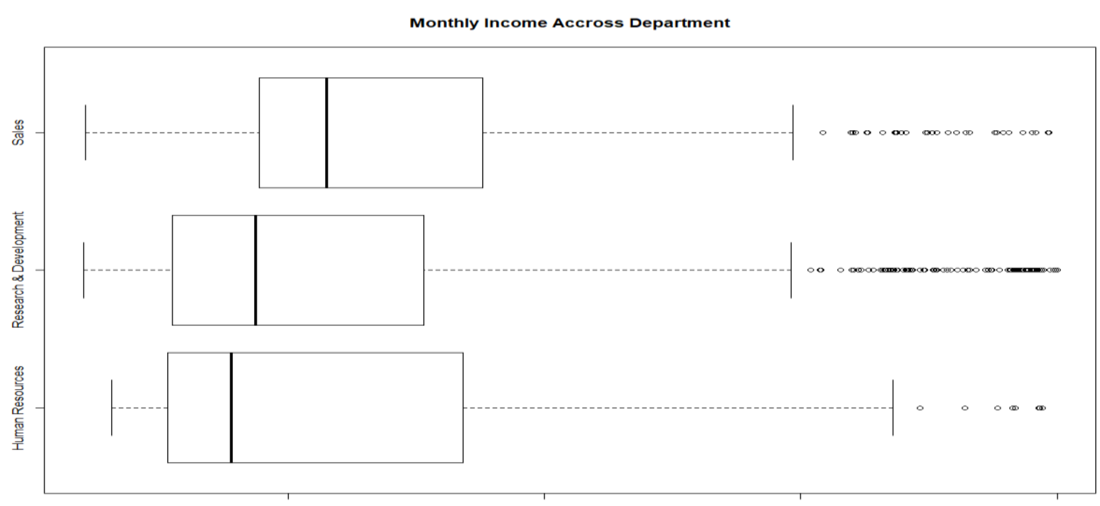
Significant Outlier Variables: YearsAtComapny, Monthly Income

Monthly Income and Years At Company variables appear to have significant outliers. The boxplots below the extent of the outliers.









Both variables, YearAtCompany and MonthlyIncome, will be removed from the data set.

Unique Variable Creation

Though the original data set gives a good feel for why an employee may leave the company, we can create some new variables based off of conventions of the given variables. The Hypothesis states, If an employee has more tenure with the company, he/she is less likely to leave. As we saw from the "YearsAtCompany" graph above, employees with more tenure were in fact less likely to leave. Though this answers the hypothesis we can take it a step further and deeper. Consider the following three variables:

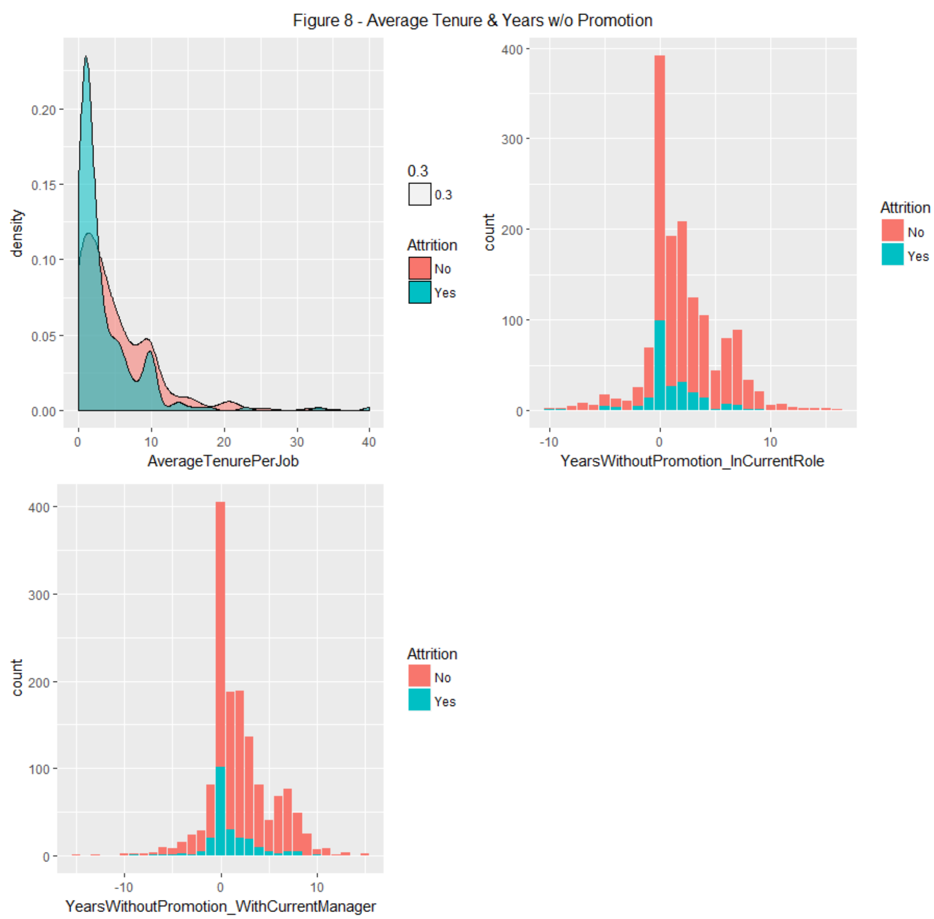
Average Tenure per Job = TotalWorkingYears/NumCompaniesWroked. For this group we will make the assumption that they are motivated by change. These individuals work for a company for a few years but end moving to a new company within a few years. We will see that those with a lower average tenure per a job are more apt to leave.

Years without Promotion in Current Role = YearsInCurrentRole – YearsSinceLastPromotion. The assumption here is that employees that are seeking growth through a promotion are more likely to leave if a promotion is not gained within a reasonable amount of time. It is impotant to note that it is unclear whether the current role was a promotion or not. This may be the reason there is negative values found.

Years without Promotion with Current Manager: YearsWithCurrentManager – YearsSinceLastPromotion. Like the variable above, the assumption here is that employees that are seeking growth through a promotion are more likely to leave if a promotion is not gained within a reasonable amount of time. However, it focuses on time with a manager vs current time in a role. It is important to note that it is possible that an employee might have a new manager but has not received a promotion which would cause a negative value.

<u>Bar Plots 10: AveragetenurePerJob_Plot, Years without Promotion_InCurrent Role, Years without Promotion with Current Manager.</u>

```r
# Unique Variable Creation
attrition$AverageTenurePerJob <- ifelse(attrition$NumCompaniesWorked!=0, attrition$TotalWorkingYears/attrition$NumCompaniesWorked,0)
attrition$YearsWithoutPromotion_InCurrentRole <- attrition$YearsInCurrentRole - attrition$YearsSinceLastPromotion
attrition$YearsWithoutPromotion_WithCurrentManager <- attrition$YearsWithCurrManager - attrition$YearsSinceLastPromotion

averagetenurePerJob_Plot <- ggplot(attrition,aes(AverageTenurePerJob, fill=Attrition, alpha = 0.3))+geom_density()
ywopcurrole_Plot <- ggplot(attrition,aes(YearsWithoutPromotion_InCurrentRole, fill=Attrition))+geom_bar()
ywopcurmanager_Plot <- ggplot(attrition,aes(YearsWithoutPromotion_WithCurrentManager, fill=Attrition))+geom_bar()
grid.arrange(averagetenurePerJob_Plot, ywopcurrole_Plot, ywopcurmanager_Plot, ncol=2,top = "Figure 8 - Average Tenure & Years w/o Promotion")
```

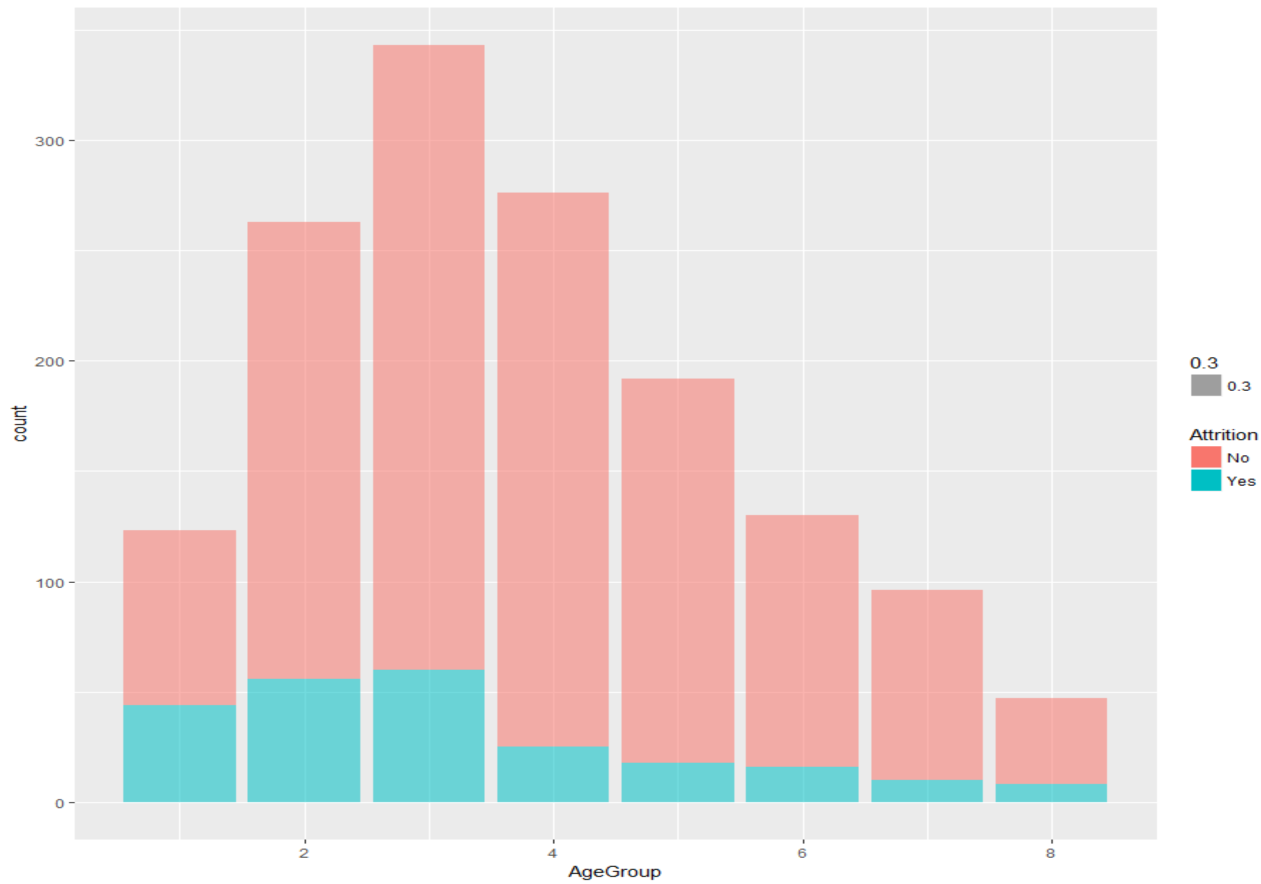


Binning

To assist in the analysis of the data set the following variables will be put into groups to gain a better understanding of how they affect attrition.

Age Group:

1 = 25 or less years	5 = 41 to 45
2 = 26 to 30	6 = 46 to 50
3 = 31 to 35	7 = 51 to 55
4 = 36 to 40	8 = 56 or greater




```
> prop.table(table(attrition$AgeGroup, attrition$Attrition))

           No         Yes
  1 0.053741497 0.029931973
  2 0.140816327 0.038095238
  3 0.192517007 0.040816327
  4 0.170748299 0.017006803
  5 0.118367347 0.012244898
  6 0.077551020 0.010884354
  7 0.058503401 0.006802721
  8 0.026530612 0.005442177
```

Age Group: Quite noticeably group 1 through 3 or ages 18 to 35 have the highest attrition rates. The chance of attrition as it were, seems to drop once an individual reaches the age of 36 years old.

Distance Group:

1 = 5 or less miles

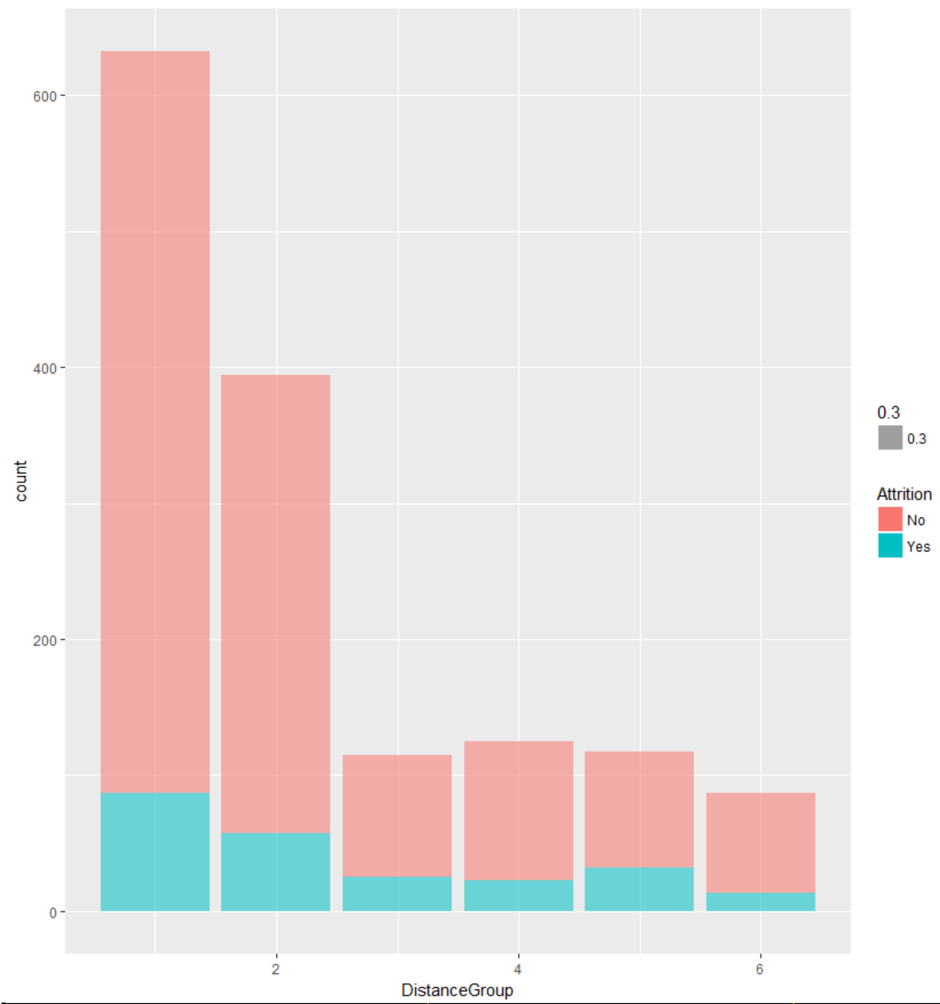
2 = 6 to 10

3 = 11 to 15

4 = 16 to 20

5 = 21 to 25

6 = Greater than 25




```
> prop.table(table(attrition$DistanceGroup, attrition$Attrition))

             No         Yes
  1 0.370748299 0.059183673
  2 0.229251701 0.038775510
  3 0.061224490 0.017006803
  4 0.069387755 0.015646259
  5 0.057823129 0.021768707
  6 0.050340136 0.008843537
```

Distance Group: The majority of employees live within 10 miles or less from work which is representative of group 1 & 2. Attrition is also greatest in these groups which may be due to the fact that they are the largest groups.

Years with Manager Group:

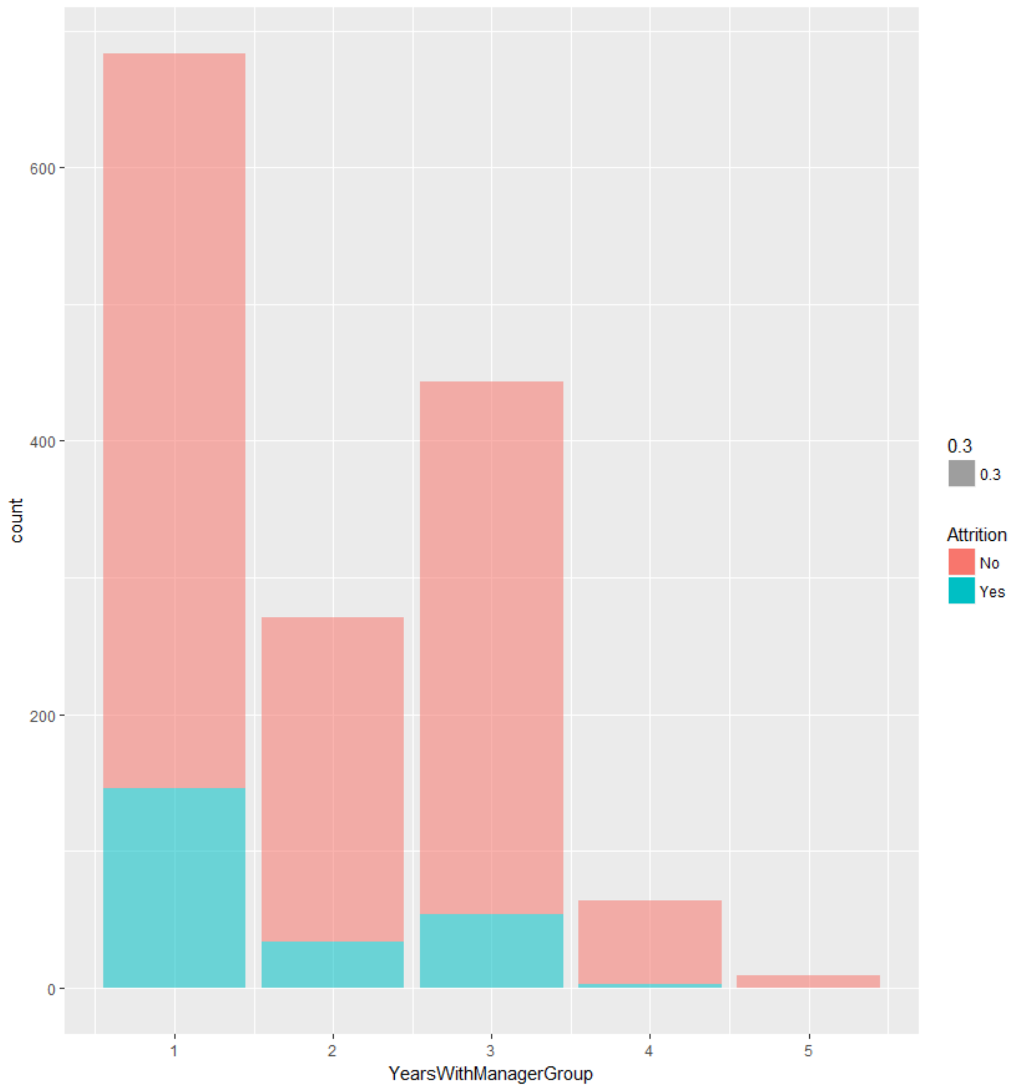
1 = 2 or less years

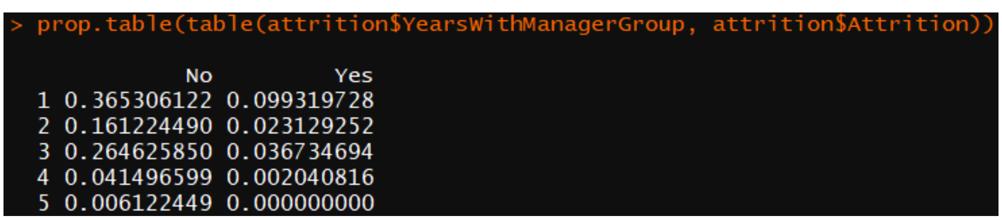
2 = 3 to 5

3 = 6 to 10

4 = 11 to 15

5 = Greater than 15 years



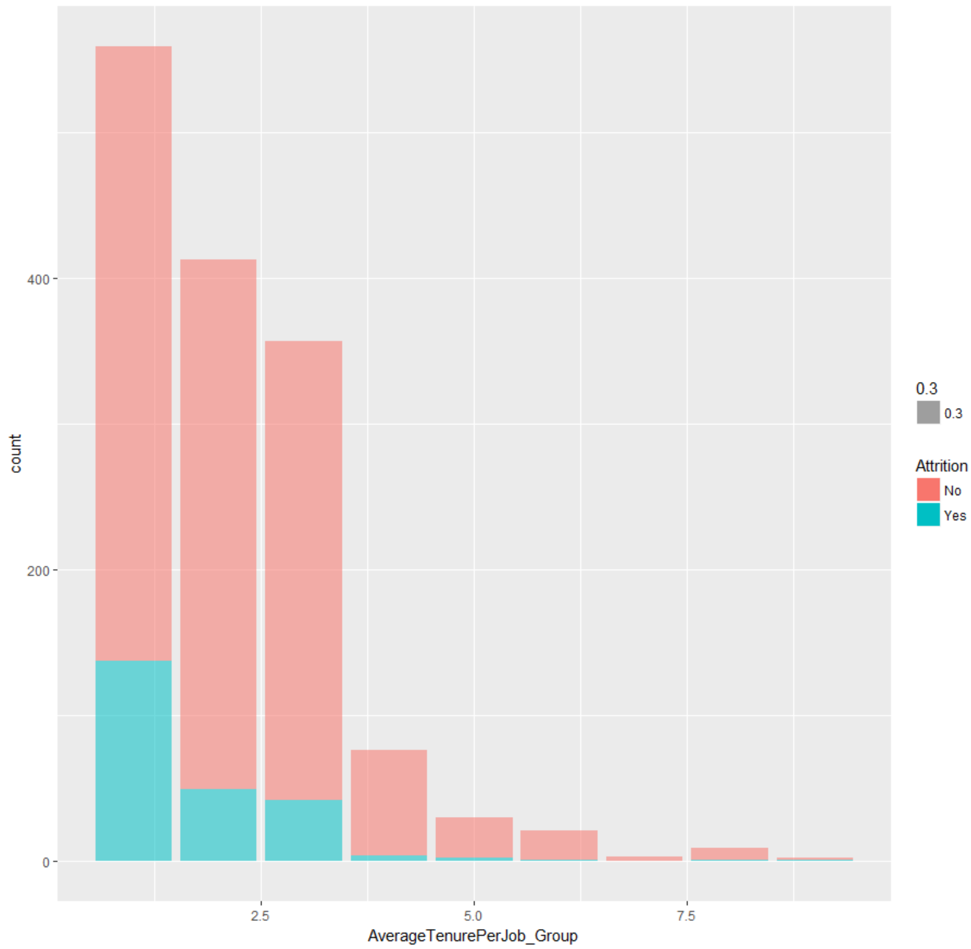


```
> prop.table(table(attrition$YearsWithManagerGroup, attrition$Attrition))

             No         Yes
  1 0.365306122 0.099319728
  2 0.161224490 0.023129252
  3 0.264625850 0.036734694
  4 0.041496599 0.002040816
  5 0.006122449 0.000000000
```

Years with Manager Group: Employees that have a tenure of 2 or less year with their current manager are more likely to leave.

Average Tenure per Job Group:

1 = 2 or less years
2 = 3 to 5
3 = 6 to 10
4 = 11 to 15
5 = 16 to 20
6 = 21 to 25
7 = 26 to 30
8 = 31 to 35
9 = Greater than 35

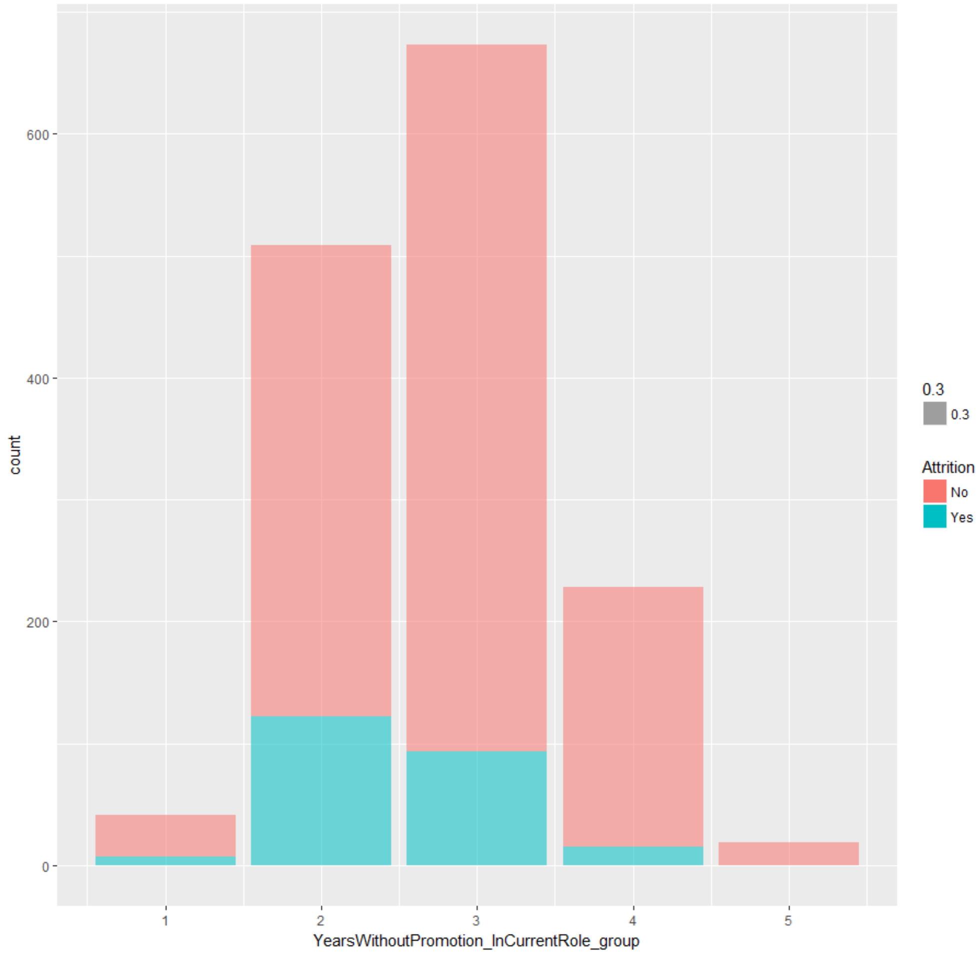



```
> prop.table(table(attrition$AverageTe

           No          Yes
 1 0.2870748299 0.0931972789
 2 0.2476190476 0.0333333333
 3 0.2142857143 0.0285714286
 4 0.0489795918 0.0027210884
 5 0.0190476190 0.0013605442
 6 0.0136054422 0.0006802721
 7 0.0020408163 0.0000000000
 8 0.0054421769 0.0006802721
 9 0.0006802721 0.0006802721
```

Average Tenure per Job Group: Higher the average tenure per a job, the less likely an employee is to leave.

Years without Promotion in Current Role Group:

1 = -5 or less years
2 = -4 to 0
3 = 1 to 5
4 = 6 to 10
5 = Greater than 10




```
> prop.table(table(attrition$Years

                No          Yes
  1 0.023129252 0.004761905
  2 0.263265306 0.082993197
  3 0.394557823 0.063265306
  4 0.144897959 0.010204082
  5 0.012925170 0.000000000
```

Years without Promotion in Current Role Group: Employees who have not received a promotion within a 5 year period are more likely to leave.

Years without Promotion with Current Manager Group:

1 = -10 or less years

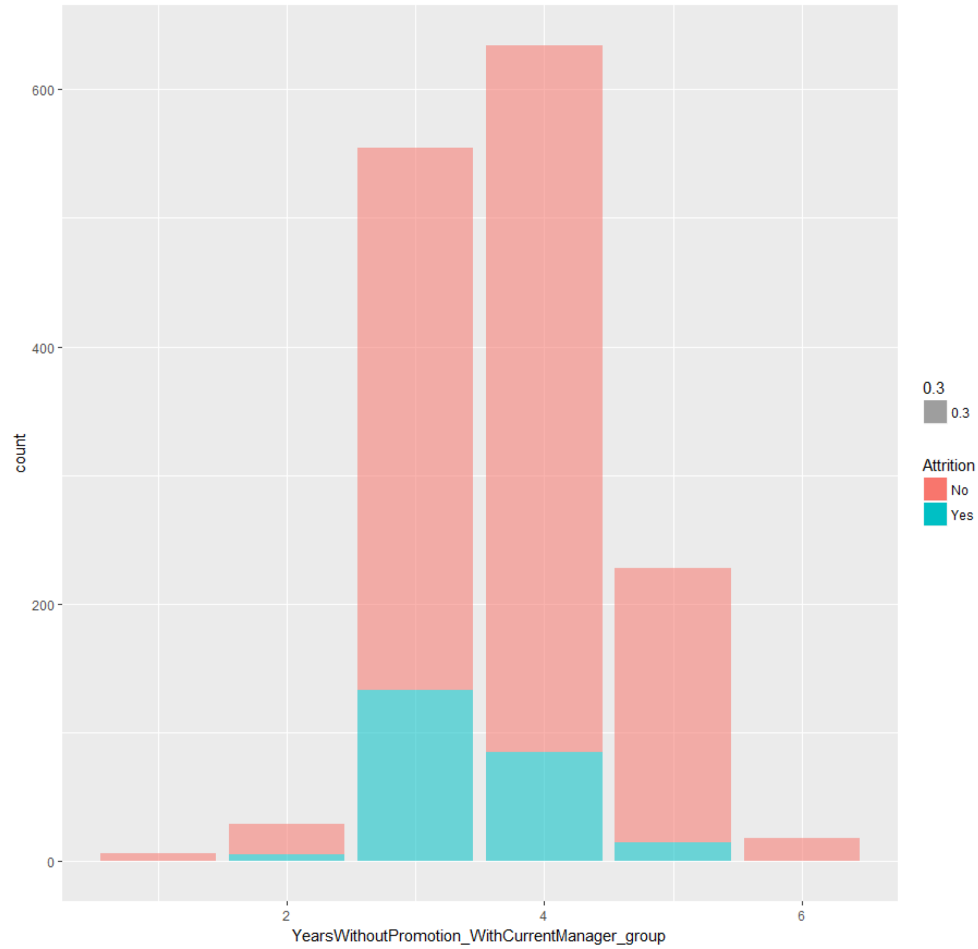
2 = -9 to -5

3 = -4 to 0

4 = 1 to 5

5 = 6 to 10

6 = Greater than 10




```
> prop.table(table(attrition$YearsW
))

          No         Yes
  1 0.004081633 0.000000000
  2 0.016326531 0.003401361
  3 0.287074830 0.090476190
  4 0.373469388 0.057823129
  5 0.145578231 0.009523810
  6 0.012244898 0.000000000
```

Years without Promotion with Current Manager Group: Employees that have not received a promotion in the last 5 years with their current manager whether new or not, are more likely to leave.

Total Working Years Group:

1 = 2 or less years

2 = 3 to 5

3 = 6 to 10

4 = 11 to 15

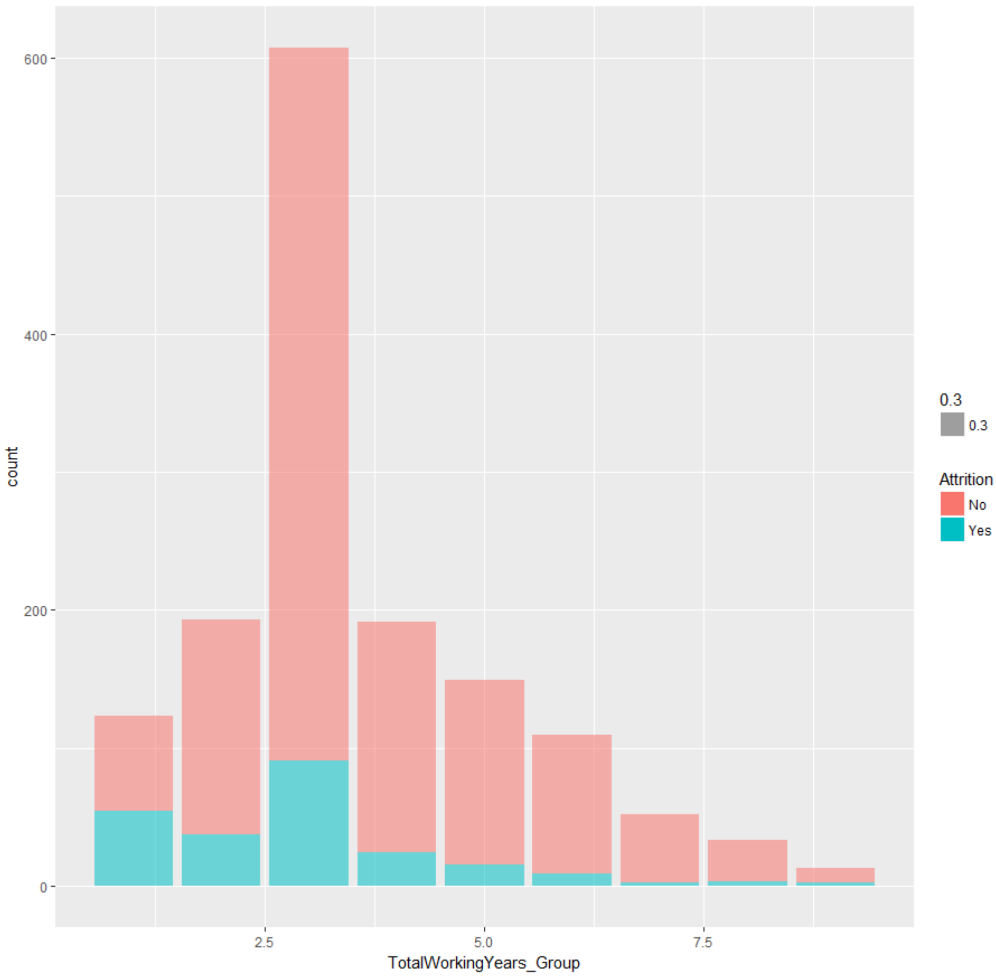
5 = 16 to 20

6 = 21 to 25

7 = 26 to 30

8 = 31 to 35

9 = Greater than 35 years




```
> prop.table(table(attrition$TotalWor

             No         Yes
  1 0.046938776 0.036734694
  2 0.106122449 0.025170068
  3 0.351020408 0.061904762
  4 0.113605442 0.016326531
  5 0.091156463 0.010204082
  6 0.068027211 0.006122449
  7 0.034013605 0.001360544
  8 0.020408163 0.002040816
  9 0.007482993 0.001360544
```

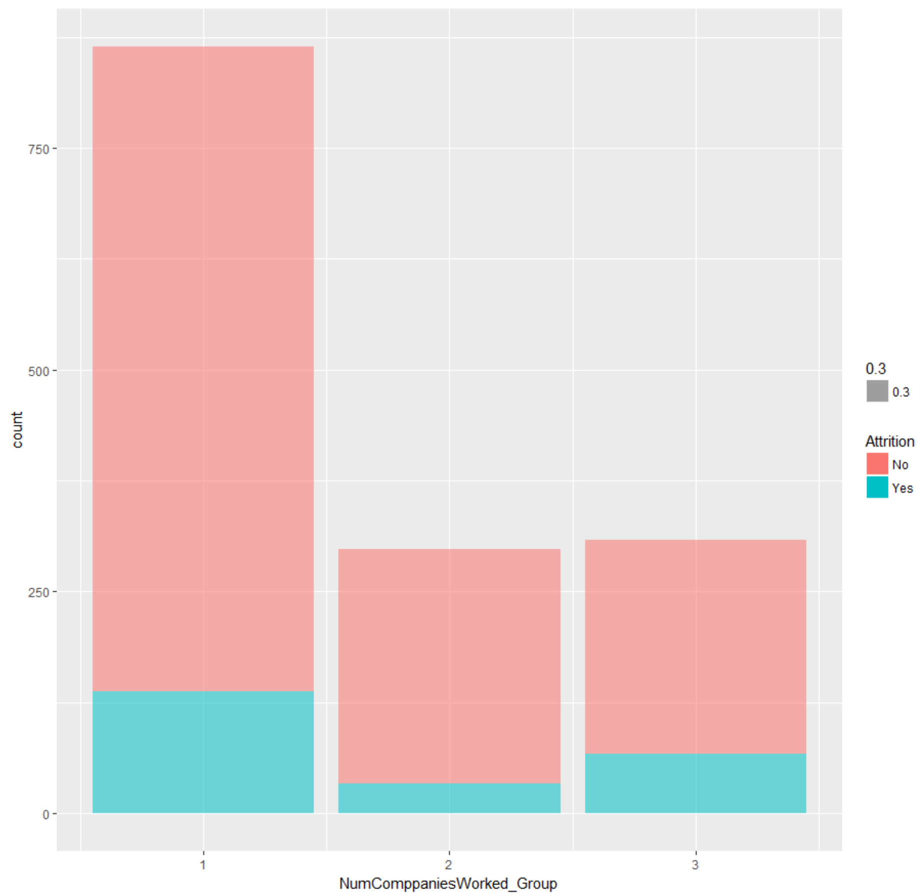
Total Working Years Group: Greater the working years decreases the chance of attrition.

Number of Companies Worked Group:

1 = 2 or less Companies

2 = 3 companies

3 = Greater than 4 Companies




```
> prop.table(table(attrition$NumCon

             No         Yes
  1 0.49455782 0.09319728
  2 0.18027211 0.02244898
  3 0.16394558 0.04557823
```

Number of Companies Worked Group: There is more attrition in those that have work with 2 or less companies.

Data Cleaning

Find Missing Values:

Goals of Data Cleaning are to [1] find and remove missing values and [2] and address any anomalies in the data. Missing values in the data were found in with the following code:

```
sapply(attrition, function(x) sum(is.na(x))) # No missing values
```

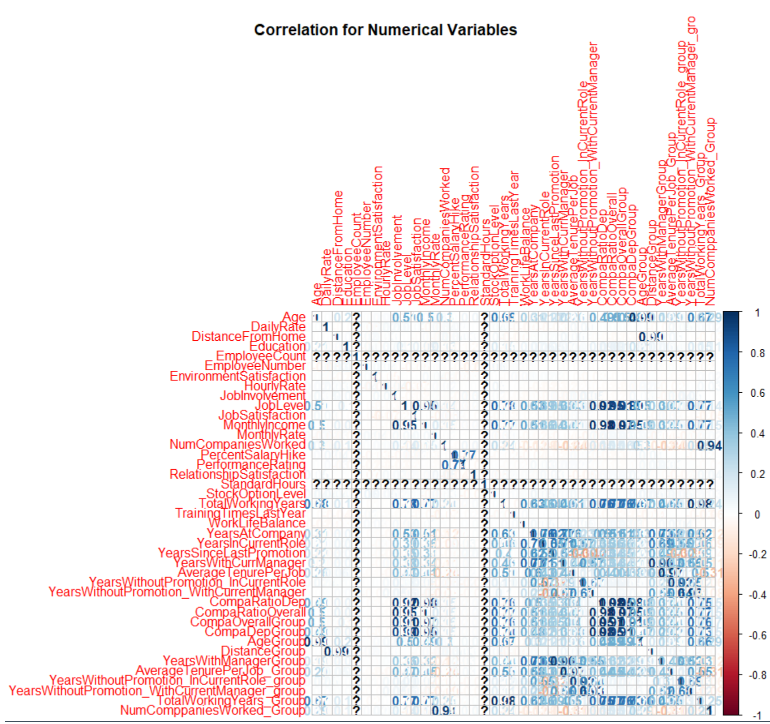
```
> sapply(attrition, function(x) sum(is.na(x))) # No missing values
                                     Age                                         Attrition                         BusinessTravel
                                       0                                                 0                                      0
                               DailyRate                                        Department                       DistanceFromHome
                                       0                                                 0                                      0
                               Education                                    EducationField                          EmployeeCount
                                       0                                                 0                                      0
                          EmployeeNumber                           EnvironmentSatisfaction                                 Gender
                                       0                                                 0                                      0
                              HourlyRate                                    JobInvolvement                               JobLevel
                                       0                                                 0                                      0
                                 JobRole                                   JobSatisfaction                          MaritalStatus
                                       0                                                 0                                      0
                           MonthlyIncome                                       MonthlyRate                     NumCompaniesWorked
                                       0                                                 0                                      0
                                  Over18                                          OverTime                      PercentSalaryHike
                                       0                                                 0                                      0
                       PerformanceRating                          RelationshipSatisfaction                          StandardHours
                                       0                                                 0                                      0
                        StockOptionLevel                                 TotalWorkingYears                  TrainingTimesLastYear
                                       0                                                 0                                      0
                         WorkLifeBalance                                    YearsAtCompany                     YearsInCurrentRole
                                       0                                                 0                                      0
                 YearsSinceLastPromotion                              YearsWithCurrManager                 AverageTenurePerJob
                                       0                                                 0                                      0
     YearsWithoutPromotion_InCurrentRole           YearsWithoutPromotion_WithCurrentManager                               AgeGroup
                                       0                                                 0                                      0
                           DistanceGroup                                YearsWithManagerGroup            AverageTenurePerJob_Group
                                       0                                                 0                                      0
YearsWithoutPromotion_InCurrentRole_group YearsWithoutPromotion_WithCurrentManager_group               TotalWorkingYears_Group
                                       0                                                 0                                      0
                 NumCompaniesWorked_Group
                                       0
```

No missing values are found in the data set.

Correlation:

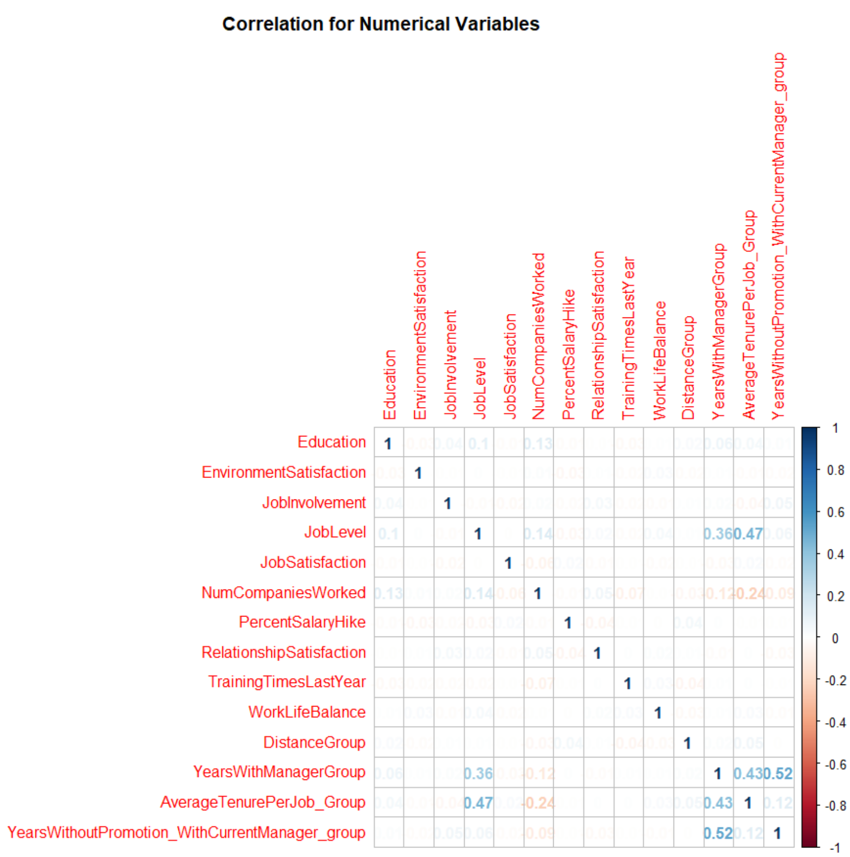
Discover correlation between numeric variables.

```
numeric_variables <- sapply(attrition, is.numeric)
matrix <- cor(attrition[,numeric_variables])
corrplot(matrix, main="\n\nCorrelation for Numerical Variables", method="number")
```



In the Correlation Plot created above, we can see highly correlated variables, which are variables that are approaching +/- 1, and variables that have a near zero variance which is represented with a "?". Variables that are at greater than or less than 0.60/-0.60 respectively, will be removed from the data set. Also, variables that have a near zero variance will also be removed. Beyond, this variables that have been grouped will be removed in favor of the new grouped variable. However, NumCompaniesWorked will be kept instead of NumCompaniesWorked_Group because the p-value is lower in the original variable. We will keep the following variables and remove the rest.

```
# Variables to Keep
"
Attrition, BusinessTravel, Department, Education, EducationField,
EnvironmentSatisfaction, Gender, JobInvolvement, JobLevel, JobRole,
JobSatisfaction, MaritalStatus, NumCompaniesWorked, OverTime,
PercentSalaryHike, RelationshipSatisfaction, TrainingTimesLastYear,
WorkLifeBalance, DistanceGroup, YearsWithManagerGroup,
AverageTenurePerJob_Group, YearsWithoutPromotion_InCurrentRole_group,
YearsWithoutPromotion_WithCurrentManager_group, TotalWorkingYears_Group

"
attrition <- attrition[,c(2,3,5,7,8,11,12,14,15,16,17,18,21,23,24,26,30,31,43:46,48)]
```



Now the data set is clean and correlation diminished. Now to perform further analysis.

Feature Analysis

The analytic method that will be applied is hierarchal clustering with k-modes. The evaluative method that will be applied to the data is logistic regression.

<u>Hierarchal Clustering K-modes:</u>

Hierarchal Clustering K – Modes was chosen for the analytic method because of its ability to handle categorical data and represent which values occur most in the data set. PCA and other options lacked the ability to cope with the categorical data contained within the data.

```
############################################## K-Modes Algorithm ##########################################
library(klaR)
data.to.cluster <- attrition
cluster.results <- kmodes(data.to.cluster, 3, iter.max = 10, weighted = FALSE)
cluster.results
summary(cluster.results)
```

```
> cluster.results
K-modes clustering with 3 clusters of sizes 741, 431, 298

Cluster modes:
  Attrition BusinessTravel             Department Education EducationField EnvironmentSatisfaction Gender JobInvolvement JobLevel
1        No  Travel_Rarely Research & Development         3        Medical                       3   Male              3        2
2        No  Travel_Rarely Research & Development         3  Life Sciences                       3 Female              3        1
3        No  Travel_Rarely Research & Development         2  Life Sciences                       4   Male              2        2
             JobRole JobSatisfaction MaritalStatus NumCompaniesWorked OverTime PercentSalaryHike RelationshipSatisfaction
1    Sales Executive               3       Married                  1       No                12                        3
2 Research Scientist               4       Married                  1       No                13                        2
3    Sales Executive               4        Single                  1       No                14                        4
  TrainingTimesLastYear WorkLifeBalance DistanceGroup YearsWithManagerGroup AverageTenurePerJob_Group
1                     2               3             1                     1                         1
2                     3               3             1                     1                         2
3                     3               3             1                     1                         2
  YearsWithoutPromotion_WithCurrentManager_group
1                                              3
2                                              4
3                                              4
```

```
Clustering vector:
   [1] 3 3 3 2 1 3 2 1 1 1 1 3 2 1 2 1 3 1 2 2 3 2 1 3 1 2 2 1 3 1 1 3 3 1 1 2 3 2 2 2 1 2 1 3 3 1 1 1 3 3 3 1 1 2 1 2 3 1 3 1 2 3 1
  [64] 2 1 1 3 1 1 1 3 2 1 2 2 2 1 2 2 1 1 2 3 2 1 3 1 2 1 3 3 1 1 1 1 1 1 1 1 1 2 1 2 1 3 2 2 3 1 1 1 3 1 1 1 2 1 1 2 2 1 1 2 1 1 2
 [127] 2 1 1 1 2 2 1 1 2 1 3 3 3 1 1 2 1 2 3 2 3 1 2 1 1 1 3 1 3 1 1 2 1 2 1 1 2 2 1 2 1 1 3 2 1 1 3 1 2 2 2 3 1 2 2 2 2 1 3 2 1 3 1
 [190] 1 1 2 3 3 3 2 1 1 1 2 1 3 1 1 1 2 2 2 3 3 1 1 3 3 2 2 1 1 3 3 3 2 1 1 1 2 2 1 1 2 2 2 1 2 1 2 2 3 2 3 1 1 1 3 1 1 2 3 2 3 1 1
 [253] 2 3 3 1 1 1 2 1 2 3 3 1 1 1 1 3 1 2 1 1 2 1 3 2 2 1 2 3 1 1 3 3 1 2 1 3 2 2 2 1 1 3 2 2 1 1 1 3 3 1 1 1 1 1 3 2 3 1 1 1 2 1 1
 [316] 2 3 1 2 1 3 3 1 1 2 2 1 1 3 1 2 1 3 1 1 1 1 1 1 1 1 3 1 3 1 2 3 1 2 3 1 2 2 1 1 1 3 2 1 1 1 2 3 1 2 1 1 3 1 3 2 2 1 2 3 1 1 2
 [379] 3 2 1 1 1 2 1 1 2 1 2 1 3 1 1 1 1 1 1 3 1 3 3 2 1 1 1 1 1 2 2 2 2 2 1 3 2 3 1 1 2 2 1 2 3 1 1 1 1 1 2 1 1 1 3 2 1 1 1 3 1 1 1
 [442] 1 3 1 1 1 3 1 2 2 3 1 1 1 1 1 2 1 1 3 1 3 3 2 2 2 2 1 3 1 1 1 1 3 2 3 1 1 1 2 1 2 1 3 3 1 1 2 2 3 2 2 2 3 2 2 1 1 1 1 1 2 1 3 2
 [505] 3 2 1 1 2 1 1 1 3 1 2 1 2 1 3 2 1 3 3 1 3 3 3 1 1 2 2 2 1 3 1 3 1 3 1 1 2 2 2 1 1 1 1 1 3 1 1 1 1 3 1 1 3 1 3 2 2 1 2 2 1 3 3
 [568] 3 1 3 2 2 1 1 2 1 1 2 3 2 2 1 1 2 2 1 1 2 1 2 1 1 2 1 2 3 2 3 1 2 1 1 2 2 2 3 2 2 1 1 2 2 1 1 1 1 1 1 2 1 2 3 3 2 3 1 1 1 3 3
 [631] 3 2 1 2 3 2 2 2 1 2 3 3 2 2 2 1 1 1 2 2 2 1 1 3 1 3 1 2 2 3 3 1 2 2 1 3 2 2 1 1 1 2 1 2 2 1 2 2 1 1 2 2 2 1 1 1 1 1 2 1 1 1 1
 [694] 3 3 1 1 2 1 2 1 1 1 2 1 3 1 1 3 3 1 2 2 2 3 2 1 1 1 3 2 1 1 2 1 3 2 3 1 1 1 1 2 1 2 3 3 1 2 1 2 2 2 2 1 1 2 1 1 2 1 1 2 1 3 2
 [757] 1 1 3 2 1 1 2 1 1 1 1 2 1 2 3 2 2 3 3 1 2 2 3 1 3 1 1 2 1 2 2 2 2 3 3 1 2 3 3 3 2 1 1 1 1 1 1 1 1 1 1 3 1 2 2 1 1 2 2 1 1 3 1
 [820] 2 3 3 3 2 3 1 1 2 1 3 1 1 1 2 2 1 2 1 1 3 1 1 2 1 1 2 1 3 1 1 1 1 2 1 1 2 1 2 2 2 2 2 2 2 1 1 1 1 1 3 1 2 1 1 1 3 1 3 1 3 1 3
 [883] 1 2 1 1 1 1 2 2 2 2 1 2 1 1 1 3 2 1 1 1 2 1 3 2 1 1 3 2 1 3 2 1 3 1 3 1 1 3 1 1 3 3 1 1 3 1 2 1 1 1 2 2 2 1 1 2 1 1 1 1 1 2 1
 [946] 2 1 1 1 1 1 1 2 2 1 3 3 1 3 1 1 1 1 3 1 2 2 2 1 1 2 3 2 1 3 1 1 1 1 1 2 1 2 1 1 1 3 1 2 1 1 1 2 3 2 3 3 2 1 1
 [ reached getOption("max.print") -- omitted 470 entries ]

Within cluster simple-matching distance by cluster:
[1] 8244 4512 3220

Available components:
[1] "cluster"    "size"       "modes"      "withindiff" "iterations" "weighted"
```

```
> summary(cluster.results)
           Length Class      Mode
cluster    1470   -none-     numeric
size          3   table      numeric
modes        22   data.frame list
withindiff    3   -none-     numeric
iterations    1   -none-     numeric
weighted      1   -none-     logical
```

Business Travel: “Travel_Rarely” reoccurs the most in all three clusters.

Department: Research & Development reoccurs the most in all three clusters.

AverageTentурePerJob_Group: In two of three clusters group “2” or tenure of “2 to 3 years” is most reoccurring.

YearsWithManagerGroup: Group 1 or “2 or less years” reoccurs the most in all three clusters.

OverTime: “No” occurs most in all three clusters.

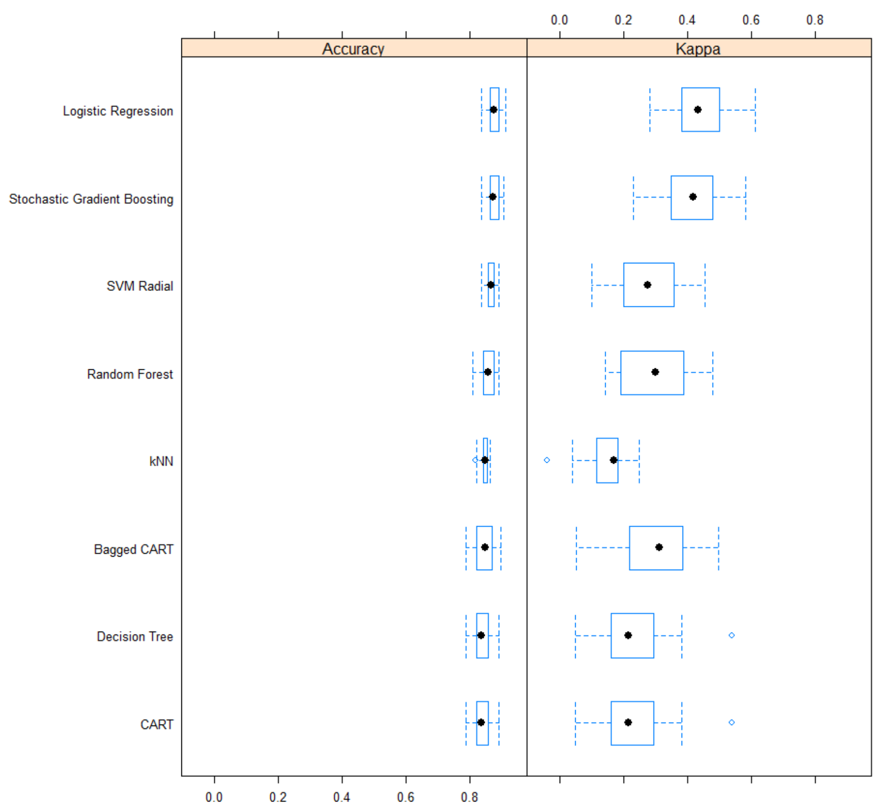
Comparative Model Testing:

In the data analysis we want to identify the strong predictor variables and which methods will produce the most accurate results. Now that our data set is cleaned and prepared, different methods will be compared against one another to see which methods we should consider further.

```
# rename dataset to keep code below generic
dataset_test <- attrition

control <- trainControl(method="repeatedcv", number=10, repeats=3)
seed <- 7

metric <- "Accuracy"
preProcess=c("center", "scale")
```

```
# Linear Discriminant Analysis
set.seed(seed)
fit.lda <- train(Attrition~., data=dataset_test, method="lda", metric=metric, preProc=c("center", "scale"), trControl=control)
# Logistic Regression
set.seed(seed)
fit.glm <- train(Attrition~., data=dataset_test, method="glm", metric=metric, trControl=control)
# GLMNET
set.seed(seed)
fit.glmnet <- train(Attrition~., data=dataset_test, method="glmnet", metric=metric, preProc=c("center", "scale"), trControl=contr
# SVM Radial
set.seed(seed)
fit.svmRadial <- train(Attrition~., data=dataset_test, method="svmRadial", metric=metric, preProc=c("center", "scale"), trControl
# kNN
set.seed(seed)
fit.knn <- train(Attrition~., data=dataset_test, method="knn", metric=metric, preProc=c("center", "scale"), trControl=control)
# Naive Bayes
# set.seed(seed)
# fit.nb <- train(Attrition~., data=dataset_test, method="nb", metric=metric, trControl=control)
# CART
set.seed(seed)
fit.cart <- train(Attrition~., data=dataset_test, method="rpart", metric=metric, trControl=control)
# C5.0
# set.seed(seed)
# fit.c50 <- train(Attrition~., data=dataset_test, method="C5.0", metric=metric, trControl=control)
# Bagged CART
set.seed(seed)
fit.treebag <- train(Attrition~., data=dataset_test, method="treebag", metric=metric, trControl=control)
# Random Forest
set.seed(seed)
fit.rf <- train(Attrition~., data=dataset_test, method="rf", metric=metric, trControl=control)
# Stochastic Gradient Boosting (Generalized Boosted Modeling)
set.seed(seed)
fit.gbm <- train(Attrition~., data=dataset_test, method="gbm", metric=metric, trControl=control, verbose=FALSE)
# Decision Tree
set.seed(seed)
fit.dt <- train(Attrition~., data=dataset_test, method="rpart", metric=metric, trControl=control)

results <- resamples(list("Logistic Regression"=fit.glm,"SVM Radial"=fit.svmRadial, kNN=fit.knn, CART=fit.cart,
                          "Bagged CART"=fit.treebag, "Random Forest"=fit.rf, "Stochastic Gradient Boosting"=fit.gbm,
                          "Decision Tree" =fit.dt ))
```

```
# Table comparison
summary(results)
results

# boxplot comparison
bwplot(results)
# Dot-plot comparison
dotplot(results)
```



As you can see, out of all the options for an evaluative method, logistic regression was the most accurate of each of the models. Therefore, it was chosen to analyze the data set. Likewise, logistic regression is able to handle all of the categorical variables with ease.

Logistic Regression:

```
# ------------------------------LOGISITIC REGRESSION-------------------
nrow(attrition)
# 1st Split data into training and testing sets:

train <- createDataPartition(attrition$Attrition,p=0.7,list=FALSE)
set.seed(2017)
training <- attrition[train,]
testing <- attrition[-train,]
# Check Spliting Results
dim(training); dim(testing)

# Fitting the LOg Regresssion Model
mod_fit <- glm(Attrition ~ .,family=binomial(link="logit"),data=training)
mod_fit
summary(mod_fit)
```

```
> summary(mod_fit)

Call:
glm(formula = Attrition ~ ., family = binomial(link = "logit"),
    data = training)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9869  -0.4742  -0.2378  -0.0676   3.6523
```

```
Coefficients:
                                               Estimate Std. Error z value Pr(>|z|)
(Intercept)                                  -8.407e+00  1.615e+03  -0.005 0.995845
BusinessTravelTravel_Frequently               2.051e+00  5.424e-01   3.782 0.000156 ***
BusinessTravelTravel_Rarely                   1.195e+00  5.043e-01   2.369 0.017853 *
DepartmentResearch & Development              1.380e+01  1.615e+03   0.009 0.993182
DepartmentSales                              -8.055e-01  1.723e+03   0.000 0.999627
Education                                     3.803e-03  1.082e-01   0.035 0.971970
EducationFieldLife Sciences                  -9.740e-01  1.095e+00  -0.890 0.373712
EducationFieldMarketing                      -2.419e-01  1.159e+00  -0.209 0.834598
EducationFieldMedical                        -1.130e+00  1.091e+00  -1.036 0.300256
EducationFieldOther                          -1.376e+00  1.175e+00  -1.171 0.241654
EducationFieldTechnical Degree                7.863e-02  1.130e+00   0.070 0.944529
EnvironmentSatisfaction                      -5.218e-01  1.028e-01  -5.077 3.83e-07 ***
GenderMale                                    1.367e-01  2.262e-01   0.604 0.545634
JobInvolvement                               -5.254e-01  1.520e-01  -3.457 0.000545 ***
JobLevel                                     -1.383e-01  2.687e-01  -0.515 0.606853
JobRoleHuman Resources                        1.506e+01  1.615e+03   0.009 0.992556
JobRoleLaboratory Technician                  1.288e+00  5.782e-01   2.228 0.025861 *
JobRoleManager                                1.344e-01  9.792e-01   0.137 0.890793
JobRoleManufacturing Director                 3.697e-01  6.276e-01   0.589 0.555834
JobRoleResearch Director                     -1.172e+00  1.189e+00  -0.986 0.324261
JobRoleResearch Scientist                     7.077e-01  5.865e-01   1.207 0.227508
JobRoleSales Executive                        1.515e+01  6.005e+02   0.025 0.979870
JobRoleSales Representative                   1.663e+01  6.005e+02   0.028 0.977911
JobSatisfaction                              -3.116e-01  9.868e-02  -3.158 0.001590 **
MaritalStatusMarried                          4.129e-01  3.052e-01   1.353 0.176132
MaritalStatusSingle                           1.334e+00  3.152e-01   4.233 2.30e-05 ***
NumCompaniesWorked                            1.979e-01  5.321e-02   3.719 0.000200 ***
OverTimeYes                                   2.155e+00  2.399e-01   8.982  < 2e-16 ***
PercentSalaryHike                            -1.438e-02  2.990e-02  -0.481 0.630509
RelationshipSatisfaction                     -2.765e-01  1.026e-01  -2.694 0.007062 **
TrainingTimesLastYear                        -2.360e-01  8.849e-02  -2.667 0.007650 **
WorkLifeBalance                              -4.322e-01  1.547e-01  -2.793 0.005216 **
AgeGroup                                     -1.691e-01  8.368e-02  -2.021 0.043260 *
DistanceGroup                                 1.979e-01  6.751e-02   2.932 0.003371 **
YearsWithManagerGroup                         3.838e-01  1.608e-01   2.387 0.017004 *
AverageTenurePerJob_Group                    -3.602e-02  1.556e-01  -0.231 0.816955
YearsWithoutPromotion_WithCurrentManager_group -9.787e-01  1.815e-01  -5.393 6.95e-08 ***
TotalWorkingYears_Group                      -2.583e-01  1.487e-01  -1.737 0.082315 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 909.69  on 1029  degrees of freedom
Residual deviance: 575.87  on  992  degrees of freedom
AIC: 651.87

Number of Fisher Scoring iterations: 16
```

```
> qchisq(0.95, 992)
[1] 1066.385
```

The critical value at 95 percent confidence and 992 degrees of freedom is 1,066.385. Since the residual deviance of 575.87 is less than the critical value the null model is not rejected. In other words, we have a reliable model at 95 percent confidence level. Also, we can see that the most most significant variables are “Business Travel, Environmental Satisfaction, Job Involvement, Marital Status, Num of Companies Worked, Overtime, Years Without Promotion With Current Manager group”.

<u>Deviance Analysis Table:</u>

```
> anova(mod_fit, test="Chisq")
Analysis of Deviance Table

Model: binomial, link: logit

Response: Attrition

Terms added sequentially (first to last)


                                             Df Deviance Resid. Df Resid. Dev  Pr(>Chi)    
NULL                                                          1029     909.69              
BusinessTravel                                2   20.042      1027     889.65 4.446e-05 ***
Department                                    2    6.262      1025     883.39 0.0436803 *  
Education                                     1    0.955      1024     882.43 0.3284005    
EducationField                                5    7.411      1019     875.02 0.1918219    
EnvironmentSatisfaction                       1   12.845      1018     862.17 0.0003383 ***
Gender                                        1    0.084      1017     862.09 0.7713941    
JobInvolvement                                1   13.973      1016     848.12 0.0001855 ***
JobLevel                                      1   61.379      1015     786.74 4.709e-15 ***
JobRole                                       8   13.768      1007     772.97 0.0880148 .  
JobSatisfaction                               1    7.764      1006     765.21 0.0053308 ** 
MaritalStatus                                 2   15.634      1004     749.57 0.0004028 ***
NumCompaniesWorked                            1   11.592      1003     737.98 0.0006625 ***
OverTime                                      1   86.254      1002     651.73 < 2.2e-16 ***
PercentSalaryHike                             1    0.077      1001     651.65 0.7810048    
RelationshipSatisfaction                      1    6.615      1000     645.03 0.0101106 *  
TrainingTimesLastYear                         1    7.924       999     637.11 0.0048790 ** 
WorkLifeBalance                               1    7.062       998     630.05 0.0078757 ** 
AgeGroup                                      1    8.287       997     621.76 0.0039934 ** 
DistanceGroup                                 1    8.734       996     613.03 0.0031241 ** 
YearsWithManagerGroup                         1    1.630       995     611.40 0.2017557    
AverageTenurePerJob_Group                     1    1.993       994     609.41 0.1580754    
YearsWithoutPromotion_WithCurrentManager_group 1   30.452       993     578.95 3.421e-08 ***
TotalWorkingYears_Group                       1    3.081       992     575.87 0.0792097 .  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The deviance table shows that as you add a variable there is a reduction in the deviance. However, some variables have a greater impact on reduction of deviance than others. The variables “OverTime”, “JobLevel”, “YearsWithoutPromotion_WithCurrentManager_group” and “BusinessTravel” have some the greatest reduction impact and all have low p-values. On the other hand, the variables “Education”, "YearsWithManagerGroup", "AverageTenurePerjob_group", and "TotalWorkingYears_Group" have low low p-values and have a much smaller impact on the reduction of the residual deviance.

Logistic Regression Accuracy:

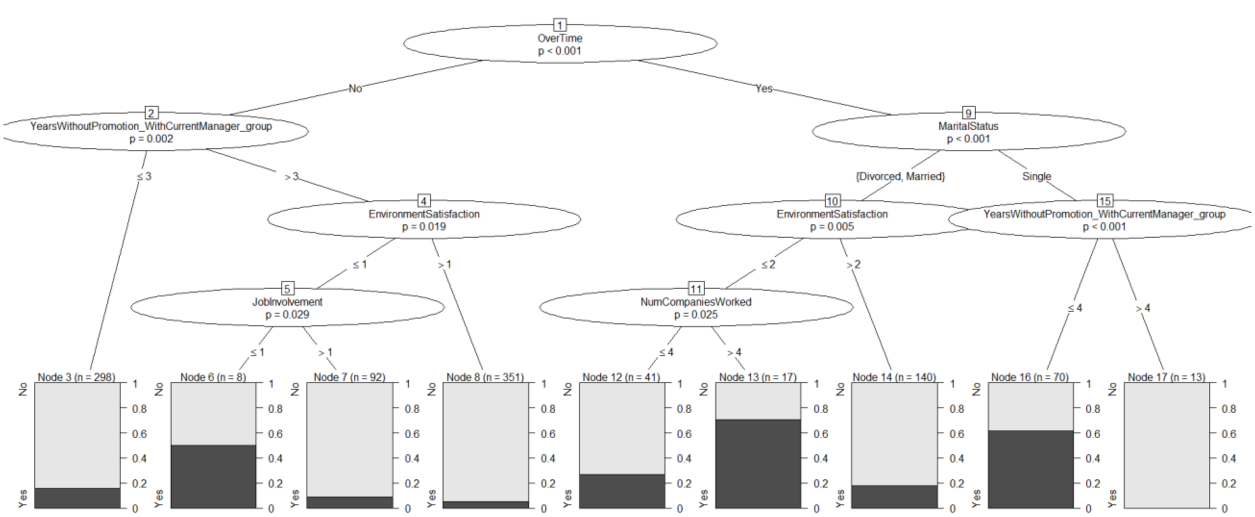
```
# Logistic Regression Accuracy or the predictive ability of the mod_fit
testing$Attrition <- as.character(testing$Attrition)
testing$Attrition[testing$Attrition=="No"] <- "0"
testing$Attrition[testing$Attrition=="Yes"] <- "1"
fitted.results <- predict(mod_fit,newdata=testing,type='response')
fitted.results <- ifelse(fitted.results > 0.5,1,0)
misClasificError <- mean(fitted.results != testing$Attrition)
print(paste('Logistic Regression Accuracy',1-misClasificError))
```

```
[1] "Logistic Regression Accuracy 0.877272727272727"
```

Decision Tree:

Though Logistic regression was found to be more accurate it is often hard to interpret unless you are trained. Decision trees, on the other hand, are naturally easy to interpret. For this purpose, the decision tree model will also be used.

```
# ----------------Decision Tree Selective 2---------------
tree_selective_2 <- ctree(Attrition ~ BusinessTravel + EnvironmentSatisfaction + JobInvolvement
                          + MaritalStatus + NumCompaniesWorked + OverTime
                          + YearsWithoutPromotion_WithCurrentManager_group, training)
plot(tree_selective_2)

# Decision Tree Confusion Matrix
pred_tree_selective_2 <- predict(tree_selective_2, testing)
print("Confusion Matrix for Decision Tree"); table(Predicted = pred_tree_selective_2, Actual = testing$Attrition)

# Decision Tree Accuracy
p1_selective_2 <- predict(tree_selective_2, training)
tab1_selective_2 <- table(Predicted = p1_selective_2, Actual = training$Attrition)
tab2_selective_2 <- table(Predicted = pred_tree_selective_2, Actual = testing$Attrition)
print(paste('Decision Tree Accuracy',sum(diag(tab2_selective_2))/sum(tab2_selective_2)))
"Accuracy is not imporved over log regression"
```



```
> print("Confusion Matrix for Decision Tr
[1] "Confusion Matrix for Decision Tree"
         Actual
Predicted   0   1
      No  344  51
      Yes  25  20
```

```
[1] "Decision Tree Accuracy 0.827272727272727"
```

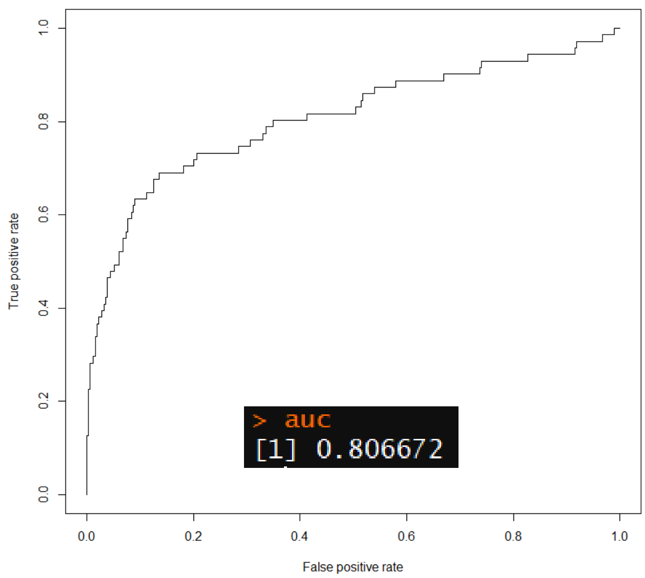
The seven most significant variables from the logistic regression model were used to create the above decision tree. We can see that the accuracy is slightly lower than the logistic regression. Of the seven variables, the most significant is Overtime. Meaning the variable Overtime and whether or not an employee has overtime is the best variable to indicate whether or not a customer will leave. Three other things we can pull from the data are (1) Employees that are single and in group 4 or less (have 5 years or less) without a promotion with their current manager are more likely to leave. (2) If an employee has worked for more than four companies he/she is more like to leave than if they have worked for four or less companies. (3) If an employee has a job involvement level less than or equal to one he/she is more likely to leave. (4) Environmental Satisfaction seems to have and influence on whether an employee will leave a company especially if their environmental satisfaction level is less than or equal to one.

Discriminate Analysis:

To check if the data is discriminating two methods are used, Chi-squared and ROC Curve. The Chi-Squared was calculated above in the logistic regression section. As we saw, with a 95 percent confidence level, that the critical value was lager than the residual deviance. Thus we saw that the logistic regression model was discriminating and good fit. As for the ROC curve, we will need to calculate the area under the curve. A possible area value under the curve is

```
# Compute AUC for predicting Class with the model
prob <- predict(mod_fit_one, newdata=testing, type="response")
pred <- prediction(prob, testing$Attrition)
perf <- performance(pred, measure = "tpr", x.measure = "fpr")
plot(perf)

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

between 0.50 and 1.00. If the curve for the selected model is 0.80 or greater, it can be determined that model is discriminate. The code to the left is used to create and calculate the ROC Curve. The calculated area under the curve for the logistic regression model is 0.806672. We find that the logistic regression model is discriminate.



Conclusion and Implications

<u>Variable Importance – Logistic Regression Model:</u>

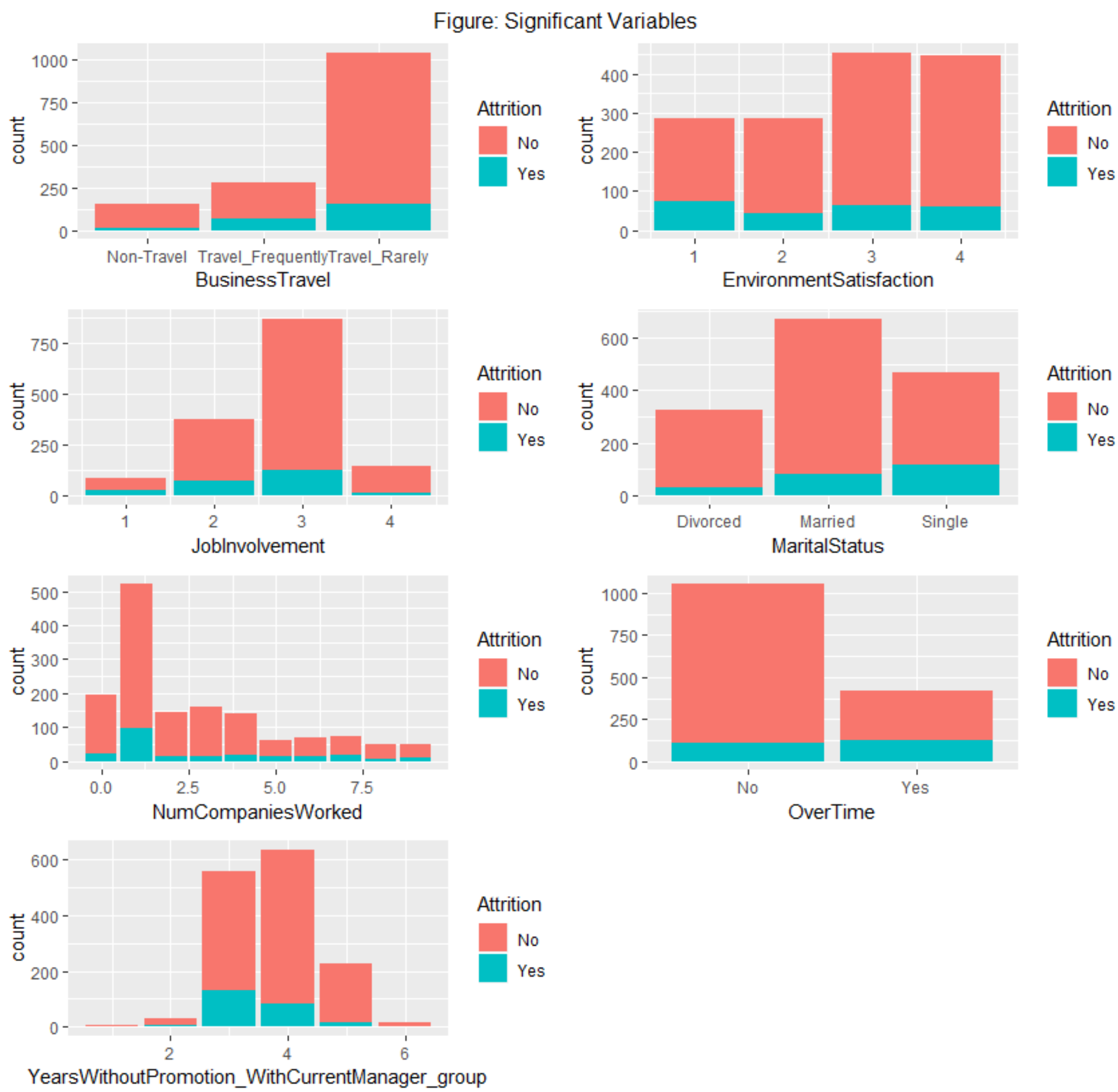
From the caret package the varImp method was used to calculate/rank each variable in terms of importance/significance while predicting the variable "Attrition". Below is the code used and the results:

```
sig_var <- train(Attrition ~ ., data=attrition, method ="glm", family="binomial")
varImp(sig_var)
```

```
> varImp(sig_var)
glm variable importance

  only 20 most important variables shown (out of 35)

                                              Overall
OverTimeYes                                    100.00
YearsWithoutPromotion_WithCurrentManager_group  59.14
MaritalStatusSingle                             56.35
EnvironmentSatisfaction                         51.60
JobSatisfaction                                 50.22
JobInvolvement                                  45.03
BusinessTravelTravel_Frequently                 44.09
DistanceGroup                                   38.72
RelationshipSatisfaction                        29.36
WorkLifeBalance                                 28.93
NumCompaniesWorked                              28.76
`JobRoleLaboratory Technician`                  28.38
TrainingTimesLastYear                           26.54
BusinessTravelTravel_Rarely                     23.81
GenderMale                                      21.64
`JobRoleSales Representative`                   20.05
YearsWithManagerGroup                           18.33
JobLevel                                        17.34
AverageTenurePerJob_Group                       14.40
EducationFieldMedical                           14.07
```

We can see that when an employee has overtime it has the greatest impact on the variable attrition. Overtime has a 40 percent greater impact on Attrition that the next most significant variable, YearsWithoutPromotion_withCurrent Manager_group. In fact, the variable OverTime has consistently been the most significant variable throughout the entire study independent of a particular model.

We started this research paper with the question: "What are the most likely reasons for an employee to leave a company and can the turnover possibility be predicted before it happens?" From above we know that employees with overtime are more likely to leave. Yet, throughout the analysis process we found that six other variables also have significance when predicting whether an employee will leave: BusinessTravel, EnvironmentSatisfaction, JobInvolvement, MaritalStatus, NumCompaniesWorked, YearsWithoutPromotion_WithCurrentManager_group. Each of these have p-values equal to or less than 0.001. Below is a recap analysis



There are many reasons why an employee might leave, but through these variables we start to have some insight. We being to understand that employees that work overtime are more apt to leave, and employees that have worked at two or less companies are also more likely to leave. Furthermore, an employee that has not received a promotion within five years will leave for another opportunity. It is also important to note that employees that are single will leave more often that someone that is married or divorced. The amount of business travel also seems to have an impact on attrition. Excellent working conditions or outstanding environmental satisfaction are important to employees. Employees that have a good environment clean environment to work in are more likely to stay.

The question now is, can attrition be predicted before it happens? Yes, I believe it can. Though this data set is an example, periodically employees are asked to perform surveys and end of year evaluations. These surveys and evaluations, coupled with HR records one could compile a data set that could be analyzed like the sample data set. You could then use the predictor variables to identify employees that are at risk for leaving. However, I would recommend not going to employees saying, “It has come to our attention that you may consider leaving.” This would surely induce mistrust and discontent within the workplace. However, HR could use this information to make recommendations on policy changes, reassignments, promotions and other positive changes that would encourage employees to stay with the company thus decreasing the cost of attrition.

The first action I would recommend to take is to identify the reasons for overtime and attempt to reduce if not eliminate the need for employees to work excessive hours. As this has one of the greatest impacts attrition it would prudent to start here. A second recommendation would be to identify employees that have not received a promotion within the three to ten years. These employees should be evaluated against their performance record and their managers recommendations for whether or not a promotion should be given.

It is important to note a major a limitation to the analysis. Though we can take a data set and perform analysis and attempt to discover why an employee may leave, the reasons could be endless. For example, though we know that those that work overtime are more likely to leave, we do not know if it is merely the extra hours or whether it is something else entirely. Perhaps the environmental satisfaction is low and the employee would be happy to work a few extra hours if their environment was so not so poor. To predict human emotion is very difficult if not impossible.

To overcome this limitation, I purpose two approaches to further study the data set. First, surveys can be sent out to employees that have previously left the company. The surveys could ask for further information and understanding on the reasons they left. Questions could include but are not limited to the following: Did you feel valued at work? If there was another position that interested you at (company name) would you return? Out of a scale from 1 to 10, one being the unsatisfactory and ten being Outstandingly satisfactory, rate your experience with (company name). Out of on a scale of 1 through 5 rate your satisfaction with your manager. etc. There are many options for questions but they should seek to gain further understanding. Also, though selective answer questions are good open-ended questions should also be used to give a chance for the former employee to express their options without being confined to a predetermined set of answers.

The second approach as mentioned above, HR should use the analysis to direct decisions. The purpose of the analysis is not to make decisions but to help in the guiding of decisions. Furthermore, the HR department should focus on a follow-up survey for current employees that focuses on deepening the understanding of the analysis. The goal being that it would shine a light on policies and procedures that could be changed to better the workplace experience and environment.

References

Altman, J. (2017, 01 18). *How Much Does Employee Turnover Really Cost?* Retrieved from HUFFPOST: https://www.huffingtonpost.com/entry/how-much-does-employee-turnover-really-cost_us_587fbaf9e4b0474ad4874fb7

EXHIBIT A: Entire R Code

```r
getwd()


# Packages installed
install.packages("ggpubr")
install.packages("rmarkdown")
install.packages("corrplot")
install.packages("ggplot2")
install.packages("gridExtra")
install.packages("ggthemes")
install.packages("caret")
install.packages("randomForest")
install.packages("party")
install.packages("stringi")
install.packages("Hmisc")
install.packages("pastecs")
install.packages("dplyr")
install.packages("olsrr")
install.packages("devtools")
# install.packages("lmtest")

# Library List
library(plyr)
library(corrplot)
library(ggplot2)
library(gridExtra)
library(ggthemes)
library(caret)
library(MASS)
library(randomForest)
library(party)
library(Hmisc)
library(pastecs)
library(psych)
library(dplyr)
library(grid)


# Load the Telco Churn Data
attrition <- read.csv("~/Desktop/r_intro/employee_attrition.csv")


############################### EDA ###################################################
####

# Variables
colnames(attrition)

"
```

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

## Display basic distribution of variables and view data
str(attrition)
summary(attrition)
class(attrition)
head(attrition)
View(attrition)

#Rename Column"ï..Age" to "Age"
colnames(attrition)[colnames(attrition)=="ï..Age"] <- "Age"
# colnames(attrition)[1] <- "Age" # Renaming the column

# Attrition
ggplot(attrition,aes(Attrition,fill=Attrition))+geom_bar()
prop.table(table(attrition$Attrition))
summary(attrition$Attrition)

# Bar Plots 1: Age, BusinessTravel, DailyRate, Department
p1 <- ggplot(attrition,aes(Age,fill=Attrition, alpha = 0.03))+geom_density()
p2 <- ggplot(attrition,aes(BusinessTravel,fill=Attrition))+geom_bar()
p3 <- ggplot(attrition,aes(DailyRate,Attrition))+geom_point(size=5,alpha = 0.03, col="blue")
p4 <- ggplot(attrition,aes(Department,fill = Attrition))+geom_bar()
grid.arrange(p1,p2,p3,p4,ncol=2,top = "Figure: Bar Plots 1")

# Age: the majority of employees who leave approx. around 31 Years of age.
# Business Travel: Employees who travel, are more likely to leave.
# Daily Rate: There is no significant indications that can be found.
# Department: R&D and Sales is where the most attrition occurred. However, it is important to note that the HR Department is proportionally smaller compared to the other departments.

# Bar Plots 2: DistanceFromHome, Education, EducationField, EmployeeCount
p5 <- ggplot(attrition,aes(DistanceFromHome,fill=Attrition))+geom_bar()
p6 <- ggplot(attrition,aes(Education,fill=Attrition))+geom_bar()
p7 <- ggplot(attrition,aes(EducationField,fill=Attrition))+geom_bar()
p8 <- ggplot(attrition,aes(EmployeeCount,Attrition))+geom_point(size=5,alpha = 0.03, col="blue")
grid.arrange(p5,p6,p7,p8,ncol=2,top = "Figure: Bar Plots 2")

# Distance From Home: An unexpected result where employees who lived closer where more apt to leave.
# Education:1 = "Below College", 2 = "College", 3 = "Bachelor", 4 = "Master", 5 = "Doctor" . Those with a bachelors degree have the highest attrition.  Important to note that there are very few employees with a doctorate degree.  May have an impact on the amount that left in the Doctorate category.
# Education Field: AS we saw in the Departments graph, those in an HR Field are less likely to leave. Again, this may be due to the low number of individuals in this group.
# Employee Count: No significant findings. All numbers in variable are 1.

# Bar Plots 3: EmployeeNumber, EnvironmentSatisfaction, Gender, HourlyRate, JobInvolvement, JobLevel
```

```
p9 <- ggplot(attrition,aes(EmployeeNumber,Attrition))+geom_point(size=5,alpha = 0.03, col="blue")
p10 <- ggplot(attrition,aes(EnvironmentSatisfaction,fill=Attrition))+geom_bar()
p11 <- ggplot(attrition,aes(Gender,fill=Attrition))+geom_bar()
p12 <- ggplot(attrition,aes(HourlyRate,fill=Attrition))+geom_bar()
p13 <- ggplot(attrition,aes(JobInvolvement,fill=Attrition))+geom_bar()
p14 <- ggplot(attrition,aes(JobLevel,fill=Attrition))+geom_bar()
grid.arrange(p9,p10,p11,p12,p13,p14,ncol=2,top = "Figure: Bar Plots 3")

# Employee Number: No significant findings.
# Environment Satisfaction:  1 = "Low",  2 = "Medium",  3 = "High", 4 = "Very High". All levels are nearly the same. No significnat findings from graph.
  prop.table(table(attrition$EnvironmentSatisfaction, attrition$Attrition) ) # Again no significant findings.
# Gender: Males are more likey to leave. However, there is 60% males and 40% female distribution which may be impacting the results.
  prop.table(table(attrition$Gender, attrition$Attrition) )
  table(attrition$Gender)/length(attrition$Gender)
# HourlyRate : No Significant findings. Also, there seems to be no direct relation to DailyRate.
# Job Involvement: 1 = "Low",  2 = "Medium",  3 = "High", 4 = "Very High". It seems that the majority of employees who don't leave are either Very Highly involved or Low Involved in their Jobs. This may be coorelated with the amount of pay they recieve for the output of work performed.
# JobLevel: An infered meaning of ratings could be: 1 = "Entry level", 2 = "Junior Level", 3 = "Junior Manager", " 4 = "Senior level", 5 = "Senior Manger Level" but it is not sure. But, by looking at the graph it is clear that the high the job level the more unlikely an employee is to leave.

# Bar Plots 4:JobRole
p15 <- ggplot(attrition,aes(JobRole,fill=Attrition))+geom_bar()
grid.arrange(p15,ncol=1,top = "Figure: Bar Plots 4")
prop.table(table(attrition$JobRole, attrition$Attrition))

# Job Role: Proportions could be influenced to group size differences. However, the graph indicates that if an employee has one of the  following job roles he/she is more likely to leave; Lab Tech, Research Scientist, Sales Executive, Sales Rep.
  prop.table(table(attrition$JobRole, attrition$Attrition)) #Cooberates above statement.

# Bar Plots 5: JobSatisfaction, MaritalStatus, MonthlyIncome
p16 <- ggplot(attrition,aes(JobSatisfaction,fill=Attrition))+geom_bar()+facet_grid(~Attrition)
p17 <- ggplot(attrition,aes(MaritalStatus,fill=Attrition))+geom_bar()
p18 <- ggplot(attrition,aes(MonthlyIncome,fill=Attrition))+geom_density()+facet_grid(~Attrition)
grid.arrange(p16,p17,p18,ncol=2,top = "Figure: Bar Plots 5")

# Job Satisfaction: 1 = "Low",  2 = "Medium",  3 = "High", 4 = "Very High". Though attrition levels stay mostly the same, the amount of employees who did not leave increases with Job Satisfaction.
  prop.table(table(attrition$JobSatisfaction, attrition$Attrition)) #Cooberates above statement.
# Marital Status:Employees who are single are more likely to leave whereas, employees who are divorced are more likely to not leave.
# Monthly Income: There are higher levels of attrition among the lower wage earners.
  mi1 <- ggplot(attrition,aes(MonthlyIncome, Attrition))+geom_point()
  mi2 <- ggplot(attrition,aes(MonthlyIncome))+geom_density()
  grid.arrange(mi1,mi2,ncol=2,top = "Figure: Monthly Income")

# Bar Plots 6: MonthlyRate, NumCompaniesWorked, Over18, OverTime
p19 <- ggplot(attrition,aes(MonthlyRate,fill=Attrition, alpha = 0.03))+geom_density()
p20 <- ggplot(attrition,aes(NumCompaniesWorked,fill=Attrition))+geom_bar()
```

```
        p21 <- ggplot(attrition,aes(Over18,Attrition))+geom_point(size=5,alpha = 0.03, col="blue")
        p22 <- ggplot(attrition,aes(OverTime,fill=Attrition))+geom_bar()
        grid.arrange(p19,p20,p21,p22,ncol=2,top = "Figure: Bar Plots 6")

        # Monthly Rate: No Significant findings. Also, there seems to be little to no correlation to the Monthly Income variable.
        # Number of Companies Worked: It is clear the if an employee has worked for only 1 company he/she is more likely to leave.
        # Over18: Not a significant variable.  All employees are over 18 years old.
        # Over Time: Though attrition first appears to be nearly equal, a larger Proportion of employees working overtime are leaving.
          prop.table(table(attrition$OverTime, attrition$Attrition)) #Cooberates above statement.

        # Bar Plots 7:PercentSalaryHike, PerformanceRating, RelationshipSatisfaction, StandardHours
        p23 <- ggplot(attrition,aes(PercentSalaryHike,fill=Attrition))+geom_bar()+facet_grid(~Attrition)
        p24 <- ggplot(attrition,aes(PerformanceRating,fill = Attrition))+geom_bar()
        p25 <- ggplot(attrition,aes(RelationshipSatisfaction,fill = Attrition))+geom_bar()
        p26 <- ggplot(attrition,aes(StandardHours,Attrition))+geom_point(size=5,alpha = 0.03, col="blue")
        grid.arrange(p23,p24,p25,p26,ncol=2,top = "Figure: Bar Plots 7")

        # Percent Salary Hike: Lower the percent salary hike equals more likely to leave.
        # Performance Rating: 1 = "Low", 2 = "Good", 3 = "Excellent", 4 = "Outstanding". As expected, lower the performance rating more likely an employee is to leave.
        # Relationship Satisfaction: 1 = "Low",  2 = "Medium",  3 = "High", 4 = "Very High". Higher the relationship satisfaction the more employees don't leave.
        # Standard Hours: Not a significant variable. All employees have standard hours of 80.

        # Bar Plots 8:StockOptionLevel, TotalWorkingYears, TrainingTimesLastYear, WorkLifeBalance
        p27 <- ggplot(attrition,aes(StockOptionLevel,fill = Attrition))+geom_bar()
        p28 <- ggplot(attrition,aes(TotalWorkingYears,fill = Attrition))+geom_bar()
        p29 <- ggplot(attrition,aes(TrainingTimesLastYear,fill = Attrition))+geom_bar()
        p30 <- ggplot(attrition,aes(WorkLifeBalance,fill = Attrition))+geom_bar()
        grid.arrange(p27,p28,p29,p30,ncol=2,top = "Figure: Bar Plots 8")

        # Stock Option Level: Larger the stock option level less likely an employee is to leave.  It is expected that there would be more 0 and 1 levels because most employees would have very little to no stock options.
        # Total Working Years: The more years of working the less likely you are to leave.  1 year highly likely to leave. It appears years 0 to 12 have a high chance of attrition.
        # Training Times Last Year: 2 to 3 trainings seem to indicate a higher chance of attrition. Though the majority of employees seem to have 2 or 3 trainings.
        # Work Life Balance: 1 = "Bad", 2 = "Good", 3 = "Better", 4 = "Best". Those that have a higher work life balance are more likely to not leave.
          prop.table(table(attrition$WorkLifeBalance, attrition$Attrition)) #Cooberates above statement.

        # Bar Plots 9: YearsAtCompany, YearsInCurrentRole, YearsSinceLastPromotion, YearsWithCurrentManager
        p31 <- ggplot(attrition,aes(YearsAtCompany,fill = Attrition))+geom_bar()
        p32 <- ggplot(attrition,aes(YearsInCurrentRole,fill = Attrition))+geom_bar()
        p33 <- ggplot(attrition,aes(YearsSinceLastPromotion,fill = Attrition))+geom_bar()
        p34 <- ggplot(attrition,aes(YearsWithCurrManager,fill = Attrition))+geom_bar()
        grid.arrange(p31,p32,p33,p34,ncol=2,top = "Figure: Bar Plots 9")
```

```
# Years at Company: Employees with less tenure are leaving more.  However, that is also where the majority of employee tenure is, 0 to 10 years.
# Years In Current Role: Employees with less years in role are leaving.  However, we do not know if they just left for another position within the same company.
# Years Since Last Promotion: It appears that those that have recently got a new promotion, 0 to 3 years, are more likely to leave.
# Years With Current Manager: Mangers play a large role in retention. Increased years with manager decreases chances of attrition.



######New Variables
# Unique Variable Creation
attrition$AverageTenurePerJob <- ifelse(attrition$NumCompaniesWorked!=0, attrition$TotalWorkingYears/attrition$NumCompaniesWorked,0)
attrition$YearsWithoutPromotion_InCurrentRole <- attrition$YearsInCurrentRole - attrition$YearsSinceLastPromotion
attrition$YearsWithoutPromotion_WithCurrentManager <- attrition$YearsWithCurrManager - attrition$YearsSinceLastPromotion

averagetenurePerJob_Plot <- ggplot(attrition,aes(AverageTenurePerJob, fill=Attrition, alpha = 0.3))+geom_density()
ywopcurrole_Plot <- ggplot(attrition,aes(YearsWithoutPromotion_InCurrentRole, fill=Attrition))+geom_bar()
ywopcurmanager_Plot <- ggplot(attrition,aes(YearsWithoutPromotion_WithCurrentManager, fill=Attrition))+geom_bar()
grid.arrange(averagetenurePerJob_Plot, ywopcurrole_Plot, ywopcurmanager_Plot, ncol=2,top = "Figure 8 - Average Tenure & Years w/o Promotion")



################### Binning ##################

attrition$AgeGroup <- with(attrition,
                           ifelse(Age>55,8,
                               ifelse(Age>50,7,
                                   ifelse(Age>45,6,
                                       ifelse(Age>40,5,
                                           ifelse(Age>35,4,
                                               ifelse(Age>30,3,
                                                   ifelse(Age>25,2,1))))))))

attrition$DistanceGroup <- with(attrition,
                           ifelse(DistanceFromHome>25,6,
                               ifelse(DistanceFromHome>20,5,
                                   ifelse(DistanceFromHome>15,4,
                                       ifelse(DistanceFromHome>10,3,
                                           ifelse(DistanceFromHome>5,2,1))))))

attrition$YearsWithManagerGroup <- with(attrition,
                           ifelse(YearsWithCurrManager>15,5,
                               ifelse(YearsWithCurrManager>10,4,
                                   ifelse(YearsWithCurrManager>5,3,
                                       ifelse(YearsWithCurrManager>2,2,1)))))

attrition$AverageTenurePerJob_Group <- with(attrition,
                           ifelse(AverageTenurePerJob>35,9,
                               ifelse(AverageTenurePerJob>30,8,
```

```
                                                    ifelse(AverageTenurePerJob>25,7,
                                                        ifelse(AverageTenurePerJob>20,6,
                                                            ifelse(AverageTenurePerJob>15,5,
                                                                ifelse(AverageTenurePerJob>10,4,
        ifelse(AverageTenurePerJob>5,3,
            ifelse(AverageTenurePerJob>2,2,1))))))))

      attrition$YearsWithoutPromotion_InCurrentRole_group <- with(attrition,
                                                    ifelse(YearsWithoutPromotion_InCurrentRole>10,5,
                                                        ifelse(YearsWithoutPromotion_InCurrentRole>5,4,
                                                            ifelse(YearsWithoutPromotion_InCurrentRole>0,3,
                                                                ifelse(YearsWithoutPromotion_InCurrentRole>-5,2,1)))))


      attrition$YearsWithoutPromotion_WithCurrentManager_group <- with(attrition,
                                                    ifelse(YearsWithoutPromotion_WithCurrentManager>10,6,
                                                        ifelse(YearsWithoutPromotion_WithCurrentManager>5,5,
                                                            ifelse(YearsWithoutPromotion_WithCurrentManager>0,4,
        ifelse(YearsWithoutPromotion_WithCurrentManager>-5,3,
            ifelse(YearsWithoutPromotion_WithCurrentManager>-10,2,1))))))

      attrition$TotalWorkingYears_Group <- with(attrition,ifelse(TotalWorkingYears>35,9,
                                                    ifelse(TotalWorkingYears>30,8,
                                                        ifelse(TotalWorkingYears>25,7,
                                                            ifelse(TotalWorkingYears>20,6,
    ifelse(TotalWorkingYears>15,5,
        ifelse(TotalWorkingYears>10,4,
            ifelse(TotalWorkingYears>5,3,
                ifelse(TotalWorkingYears>2,2,1)))))))))

      attrition$NumComppaniesWorked_Group <- with(attrition,
                                            ifelse(NumCompaniesWorked>4,3,
                                                ifelse(NumCompaniesWorked>2,2,1)))


      attrition$YearsAtCompany_Group <- with(attrition,ifelse(YearsAtCompany>35,9,
                                                    ifelse(YearsAtCompany>30,8,
                                                        ifelse(YearsAtCompany>25,7,
                                                            ifelse(YearsAtCompany>20,6,
```

```
ifelse(YearsAtCompany>15,5,
       ifelse(YearsAtCompany>10,4,
              ifelse(YearsAtCompany>5,3,
                     ifelse(YearsAtCompany>2,2,1))))))))



######------Grouped/Binned Variables
# Bar Plots 10: AgeGroup, DistanceGroup, YearsWithManagerGroup, AverageTenurePerJob_Group
p35 <- ggplot(attrition,aes(AgeGroup, fill = Attrition, alpha = 0.3))+geom_bar()
p36 <- ggplot(attrition,aes(DistanceGroup, fill = Attrition, alpha = 0.3))+geom_bar()
p37 <- ggplot(attrition,aes(YearsWithManagerGroup, fill = Attrition, alpha = 0.3))+geom_bar()
p38 <- ggplot(attrition,aes(AverageTenurePerJob_Group, fill = Attrition, alpha = 0.3))+geom_bar()
grid.arrange(p35,p36,p37,p38,ncol=2,top = "Figure: Bar Plots 10")

# Bar Plots 11: YearsWithoutPromotion_InCurrentRole_group, YearsWithoutPromotion_WithCurrentManager_group, TotalWorkingYears_Group, NumComppaniesWorked_Group
p39 <- ggplot(attrition,aes(YearsWithoutPromotion_InCurrentRole_group, fill = Attrition, alpha = 0.3))+geom_bar()
p40 <- ggplot(attrition,aes(YearsWithoutPromotion_WithCurrentManager_group, fill = Attrition, alpha = 0.3))+geom_bar()
p41 <- ggplot(attrition,aes(TotalWorkingYears_Group, fill = Attrition, alpha = 0.3))+geom_bar()
p42 <- ggplot(attrition,aes(NumComppaniesWorked_Group, fill = Attrition, alpha = 0.3))+geom_bar()
grid.arrange(p39,p40,p41,p42,ncol=2,top = "Figure: Bar Plots 11")

# Bar Plots 12: YearsWithCompany_Group
p43 <- ggplot(attrition,aes(YearsAtCompany_Group, fill = Attrition, alpha = 0.3))+geom_bar()
grid.arrange(p43,ncol=1,top = "Figure: Bar Plots 12")



################################### Data Cleaning #########################################
## Find missing values
sapply(attrition, function(x) sum(is.na(x))) # No missing values


#-------------NUmeric Variables---------------
# -----Correlation-----
# Discover Correlation between Numneric Variables
numeric_variables <- sapply(attrition, is.numeric)
matrix <- cor(attrition[,numeric_variables])
corrplot(matrix, main="\n\nCorrelation for Numerical Variables", method="number")

#-------------------------------OUTLIERS---------------------------
boxplot(attrition$YearsAtCompany, horizontal = TRUE,
        main = "Boxplot of YearsAtCompany", xlab = "YearsAtCompany")
boxplot(attrition$MonthlyIncome, horizontal = TRUE,
        main = "Boxplot of MonthlyIncome", xlab = "MonthlyIncome")

# Variables to Keep
"
Attrition, BusinessTravel, Department, Education, EducationField,
```

```
EnvironmentSatisfaction, Gender, JobInvolvement, JobLevel, JobRole,
JobSatisfaction, MaritalStatus, NumCompaniesWorked, OverTime,
PercentSalaryHike, RelationshipSatisfaction, TrainingTimesLastYear,
WorkLifeBalance, DistanceGroup, YearsWithManagerGroup,
AverageTenurePerJob_Group, YearsWithoutPromotion_InCurrentRole_group,
YearsWithoutPromotion_WithCurrentManager_group, NumCompaniesWorked_Group
YearsAtCompany_Group
"
colnames(attrition)
attrition <- attrition[,c(2,3,5,7,8,11,12,14,15,16,17,18,21,23,24,26,30,31,40, 41,42,44)]


################################ Analysis & Statistics ####################################

##GROUPED VARIABLES
# AgeGroup
str(attrition$AgeGroup)
summary(attrition$AgeGroup)
var(attrition$AgeGroup)
sd(attrition$AgeGroup)
hist(attrition$AgeGroup, main = "Histogram of AgeGroup", xlab = "YearsAtCompany", col = "blue")
boxplot(attrition$AgeGroup, horizontal = TRUE,
        main = "Boxplot of AgeGroup", xlab = "AgeGroup")
table(attrition$Attrition, attrition$AgeGroup)
prop.table(table(attrition$AgeGroup, attrition$Attrition))

# DistanceGroup
str(attrition$DistanceGroup)
summary(attrition$DistanceGroup)
var(attrition$DistanceGroup)
sd(attrition$DistanceGroup)
hist(attrition$DistanceGroup, main = "Histogram of DistanceGroup", xlab = "YearsAtCompany", col = "blue")
boxplot(attrition$DistanceGroup, horizontal = TRUE,
        main = "Boxplot of DistanceGroup", xlab = "DistanceGroup")
table(attrition$Attrition, attrition$DistanceGroup)
prop.table(table(attrition$DistanceGroup, attrition$Attrition))

# YearsWithManagerGroup
str(attrition$YearsWithManagerGroup)
summary(attrition$YearsWithManagerGroup)
var(attrition$YearsWithManagerGroup)
sd(attrition$YearsWithManagerGroup)
hist(attrition$YearsWithManagerGroup, main = "Histogram of YearsWithManagerGroup", xlab = "YearsAtCompany", col = "blue")
boxplot(attrition$YearsWithManagerGroup, horizontal = TRUE,
        main = "Boxplot of YearsWithManagerGroup", xlab = "YearsWithManagerGroup")
table(attrition$Attrition, attrition$YearsWithManagerGroup)
prop.table(table(attrition$YearsWithManagerGroup, attrition$Attrition))

# AverageTenurePerJob_Group
str(attrition$AverageTenurePerJob_Group)
summary(attrition$AverageTenurePerJob_Group)
var(attrition$AverageTenurePerJob_Group)
sd(attrition$AverageTenurePerJob_Group)
hist(attrition$AverageTenurePerJob_Group, main = "Histogram of AverageTenurePerJob_Group", xlab = "YearsAtCompany", col = "blue")
boxplot(attrition$AverageTenurePerJob_Group, horizontal = TRUE,
        main = "Boxplot of AverageTenurePerJob_Group", xlab = "AverageTenurePerJob_Group")
table(attrition$Attrition, attrition$AverageTenurePerJob_Group)
prop.table(table(attrition$AverageTenurePerJob_Group, attrition$Attrition))
```

```r

        # YearsWithoutPromotion_InCurrentRole_group
        str(attrition$YearsWithoutPromotion_InCurrentRole_group)
        summary(attrition$YearsWithoutPromotion_InCurrentRole_group)
        var(attrition$YearsWithoutPromotion_InCurrentRole_group)
        sd(attrition$YearsWithoutPromotion_InCurrentRole_group)
        hist(attrition$YearsWithoutPromotion_InCurrentRole_group, main = "Histogram of YearsWithoutPromotion_InCurrentRole_group", xlab = "YearsAtCompany", col = "blue")
        boxplot(attrition$YearsWithoutPromotion_InCurrentRole_group, horizontal = TRUE,
                main = "Boxplot of YearsWithoutPromotion_InCurrentRole_group", xlab = "YearsWithoutPromotion_InCurrentRole_group")
        table(attrition$Attrition, attrition$YearsWithoutPromotion_InCurrentRole_group)
        prop.table(table(attrition$YearsWithoutPromotion_InCurrentRole_group, attrition$Attrition))

        # YearsWithoutPromotion_WithCurrentManager_group
        str(attrition$YearsWithoutPromotion_WithCurrentManager_group)
        summary(attrition$YearsWithoutPromotion_WithCurrentManager_group)
        var(attrition$YearsWithoutPromotion_WithCurrentManager_group)
        sd(attrition$YearsWithoutPromotion_WithCurrentManager_group)
        hist(attrition$YearsWithoutPromotion_WithCurrentManager_group, main = "Histogram of YearsWithoutPromotion_WithCurrentManager_group", xlab = "YearsAtCompany", col = "blue")
        boxplot(attrition$YearsWithoutPromotion_WithCurrentManager_group, horizontal = TRUE,
                main = "Boxplot of YearsWithoutPromotion_WithCurrentManager_group", xlab = "YearsWithoutPromotion_WithCurrentManager_group")
        table(attrition$Attrition, attrition$YearsWithoutPromotion_WithCurrentManager_group)
        prop.table(table(attrition$YearsWithoutPromotion_WithCurrentManager_group, attrition$Attrition))


        # TotalWorkingYears_Group
        summary(attrition$TotalWorkingYears_Group)
        var(attrition$TotalWorkingYears_Group)
        sd(attrition$TotalWorkingYears_Group)
        hist(attrition$TotalWorkingYears_Group, main = "Histogram of TotalWorkingYears_Group", xlab = "YearsAtCompany", col = "blue")
        boxplot(attrition$TotalWorkingYears_Group, horizontal = TRUE,
                main = "Boxplot of TotalWorkingYears_Group", xlab = "TotalWorkingYears_Group")
        table(attrition$Attrition, attrition$TotalWorkingYears_Group)
        prop.table(table(attrition$TotalWorkingYears_Group, attrition$Attrition))

        # NumComppaniesWorked_Group
        summary(attrition$NumComppaniesWorked_Group)
        var(attrition$NumComppaniesWorked_Group)
        sd(attrition$NumComppaniesWorked_Group)
        hist(attrition$NumComppaniesWorked_Group, main = "Histogram of NumComppaniesWorked_Group", xlab = "YearsAtCompany", col = "blue")
        boxplot(attrition$NumComppaniesWorked_Group, horizontal = TRUE,
                main = "Boxplot of NumComppaniesWorked_Group", xlab = "NumComppaniesWorked_Group")
        table(attrition$Attrition, attrition$NumComppaniesWorked_Group)
        prop.table(table(attrition$NumComppaniesWorked_Group, attrition$Attrition))


        ##CATEGORICAL Variables
        # Gender
        summary(attrition$Gender)
        prop.table(table(attrition$Gender))
        table(attrition$Attrition, attrition$Gender)
        prop.table(table(attrition$Gender, attrition$Attrition))
```

```
ggplot(attrition,aes(Gender,fill=Attrition))+geom_bar()

# Attrition
summary(attrition$Attrition)
prop.table(table(attrition$Attrition))
ggplot(attrition,aes(Attrition,fill=Attrition))+geom_bar()

# Business Travel
summary(attrition$BusinessTravel)
prop.table(table(attrition$BusinessTravel))
table(attrition$Attrition, attrition$BusinessTravel)
prop.table(table(attrition$BusinessTravel, attrition$Attrition))
ggplot(attrition,aes(BusinessTravel,fill=Attrition))+geom_bar()

# Department
summary(attrition$Department)
prop.table(table(attrition$Department))
table(attrition$Attrition, attrition$Department)
prop.table(table(attrition$Department, attrition$Attrition))
ggplot(attrition,aes(Department,fill=Attrition))+geom_bar()

# Education Field
summary(attrition$EducationField)
prop.table(table(attrition$EducationField))
table(attrition$Attrition, attrition$EducationField)
prop.table(table(attrition$EducationField, attrition$Attrition))
ggplot(attrition,aes(EducationField,fill=Attrition))+geom_bar()

# Job Role
summary(attrition$JobRole)
prop.table(table(attrition$JobRole))
table(attrition$Attrition, attrition$JobRole)
prop.table(table(attrition$JobRole, attrition$Attrition))
ggplot(attrition,aes(JobRole,fill=Attrition))+geom_bar()

# Marital Status
summary(attrition$MaritalStatus)
prop.table(table(attrition$MaritalStatus))
table(attrition$Attrition, attrition$MaritalStatus)
prop.table(table(attrition$MaritalStatus, attrition$Attrition))
ggplot(attrition,aes(MaritalStatus,fill=Attrition))+geom_bar()

# Over Time
summary(attrition$OverTime)
prop.table(table(attrition$OverTime))
table(attrition$Attrition, attrition$OverTime)
prop.table(table(attrition$OverTime, attrition$Attrition))
ggplot(attrition,aes(OverTime,fill=Attrition))+geom_bar()



##NUMERIC VARIABLES
# YearsWithoutPromotion_WithCurrentManager
summary(attrition$YearsWithoutPromotion_WithCurrentManager)
var(attrition$YearsWithoutPromotion_WithCurrentManager)
sd(attrition$YearsWithoutPromotion_WithCurrentManager)
hist(attrition$YearsWithoutPromotion_WithCurrentManager, main = "Histogram of YearsWithoutPromotion_WithCurrentManager", xlab = "YearsWithCurrManager", col = "blue")
boxplot(attrition$YearsWithoutPromotion_WithCurrentManager, horizontal = TRUE,

        main = "Boxplot of YearsWithoutPromotion_WithCurrentManager", xlab = "YearsWithoutPromotion_WithCurrentManager")
ggplot(attrition,aes(YearsWithoutPromotion_WithCurrentManager,fill=Attrition, alpha = 0.03))+geom_density()

# YearsWithCurrManager
```

```r
summary(attrition$YearsWithCurrManager)
var(attrition$YearsWithCurrManager)
sd(attrition$YearsWithCurrManager)
hist(attrition$YearsWithCurrManager, main = "Histogram of YearsWithCurrManager", xlab = "YearsWithCurrManager", col = "blue")
boxplot(attrition$YearsWithCurrManager, horizontal = TRUE,
        main = "Boxplot of YearsWithCurrManager", xlab = "YearsWithCurrManager")

ggplot(attrition,aes(YearsWithCurrManager,fill=Attrition, alpha = 0.03))+geom_density()

# HourlyRate
summary(attrition$HourlyRate)
var(attrition$HourlyRate)
sd(attrition$HourlyRate)
hist(attrition$HourlyRate, main = "Histogram of HourlyRate", xlab = "HourlyRate", col = "blue")
boxplot(attrition$HourlyRate, horizontal = TRUE,
        main = "Boxplot of HourlyRate", xlab = "HourlyRate")
ggplot(attrition,aes(HourlyRate,fill=Attrition, alpha = 0.03))+geom_density()

# Age
summary(attrition$Age)
var(attrition$Age)
sd(attrition$Age)
hist(attrition$Age, main = "Histogram of Age", xlab = "Age", col = "blue")
boxplot(attrition$Age, horizontal = TRUE,
        main = "Boxplot of Age", xlab = "Age")
ggplot(attrition,aes(Age,fill=Attrition, alpha = 0.03))+geom_density()

# Distance from Home
summary(attrition$DistanceFromHome)
var(attrition$DistanceFromHome)
sd(attrition$DistanceFromHome)
hist(attrition$DistanceFromHome, main = "Histogram of DistanceFromHome", xlab = "DistanceFromHome", col = "blue")
boxplot(attrition$DistanceFromHome, horizontal = TRUE,
        main = "Boxplot of DistanceFromHome", xlab = "DistanceFromHome")
ggplot(attrition,aes(DistanceFromHome,fill=Attrition, alpha = 0.03))+geom_density()

# Education
summary(attrition$Education)
var(attrition$Education)
sd(attrition$Education)
hist(attrition$Education, main = "Histogram of Education", xlab = "Education", col = "blue")
boxplot(attrition$Education, horizontal = TRUE,
        main = "Boxplot of Education", xlab = "Education")
ggplot(attrition,aes(Education,fill=Attrition, alpha = 0.03))+geom_bar()


# Enviroment Satisfaction
summary(attrition$EnvironmentSatisfaction)
var(attrition$EnvironmentSatisfaction)
sd(attrition$EnvironmentSatisfaction)
hist(attrition$EnvironmentSatisfaction, main = "Histogram of EnvironmentSatisfaction", xlab = "EnvironmentSatisfaction", col = "blue")
boxplot(attrition$EnvironmentSatisfaction, horizontal = TRUE,
        main = "Boxplot of EnvironmentSatisfaction", xlab = "EnvironmentSatisfaction")
ggplot(attrition,aes(EnvironmentSatisfaction,fill=Attrition, alpha = 0.03))+geom_bar()


# JobInvolvement
```

```
    summary(attrition$JobInvolvement)
    var(attrition$JobInvolvement)
    sd(attrition$JobInvolvement)
    hist(attrition$JobInvolvement, main = "Histogram of JobInvolvement", xlab = "Job Involvement", col = "blue")
    boxplot(attrition$JobInvolvement, horizontal = TRUE,
            main = "Boxplot of JobInvolvement", xlab = "JobInvolvement")
    ggplot(attrition,aes(JobInvolvement,fill=Attrition, alpha = 0.03))+geom_bar()


    # JobSatisfaction
    summary(attrition$JobSatisfaction)
    var(attrition$JobSatisfaction)
    sd(attrition$JobSatisfaction)
    hist(attrition$JobSatisfaction, main = "Histogram of JobSatisfaction", xlab = "JobSatisfaction", col = "blue")
    boxplot(attrition$JobSatisfaction, horizontal = TRUE,
            main = "Boxplot of JobSatisfaction", xlab = "JobSatisfaction")
    ggplot(attrition,aes(JobSatisfaction,fill=Attrition, alpha = 0.03))+geom_bar()


    # Monthly Income
    summary(attrition$MonthlyIncome)
    var(attrition$MonthlyIncome)
    sd(attrition$MonthlyIncome)
    hist(attrition$MonthlyIncome, main = "Histogram of Monthy Income", xlab = "MonthlyIncome", col = "blue")
    boxplot(attrition$MonthlyIncome, horizontal = TRUE,
            main = "Boxplot of Monthly Income", xlab = "Monthly Income")
    ggplot(attrition,aes(MonthlyIncome,fill=Attrition, alpha = 0.03))+geom_density()

      # Outliers Univariate - Monthly Income
        outlier_values <- boxplot.stats(attrition$MonthlyIncome)$out # outlier values
        boxplot(attrition$MonthlyIncome, main="Monthly Income", boxwex=0.1, horizontal = TRUE,
                xlab = "Monthly Income")
        mtext(paste("Outliers: ", paste(outlier_values, collapse = ", ")), cex = 0.6)
      # Outliers Bivariate - Monthly Income
        # For categorical variable
          boxplot(MonthlyIncome ~ JobRole, data=attrition, main="Monthly Income Accross Job Role",
                  horizontal = FALSE)  # clear pattern is noticeable.
          boxplot(MonthlyIncome ~ Department, data=attrition, main="Monthly Income Accross Department",
                  horizontal = FALSE)  # this may not be significant, as day of week variable is a subset of the month var.
      # Plot of data with Outliers
        par(mfrow=c(1,1))
        plot(attrition$YearsAtCompany, attrition$MonthlyIncome, xlim=c(0, 40), ylim=c(1000, 20000),
             main="With Outliers", xlab="YearsAtCompany", ylab="MonthlyIncome", pch="*", col="red", cex=2)
        abline(lm(MonthlyIncome ~ YearsAtCompany, data = attrition), col="blue", lwd = 3, lty=2)


    # NumCompaniesWorked
    summary(attrition$NumCompaniesWorked)
    var(attrition$NumCompaniesWorked)
    sd(attrition$NumCompaniesWorked)
    hist(attrition$NumCompaniesWorked, main = "Histogram of NumCompaniesWorked", xlab = "NumCompaniesWorked", col = "blue")
    boxplot(attrition$NumCompaniesWorked, horizontal = TRUE,
            main = "Boxplot of NumCompaniesWorked", xlab = "NumCompaniesWorked")
```

```
    ggplot(attrition,aes(NumCompaniesWorked,fill=Attrition, alpha = 0.03))+geom_density()
    table(attrition$NumCompaniesWorked)
    prop.table(table(attrition$NumCompaniesWorked))
    prop.table(table(attrition$NumCompaniesWorked, attrition$Attrition))


    # PercentSalaryHike
    summary(attrition$PercentSalaryHike)
    var(attrition$PercentSalaryHike)
    sd(attrition$PercentSalaryHike)
    hist(attrition$PercentSalaryHike, main = "Histogram of PercentSalaryHike", xlab = "PercentSalaryHike", col = "blue")
    boxplot(attrition$PercentSalaryHike, horizontal = TRUE,
            main = "Boxplot of PercentSalaryHike", xlab = "PercentSalaryHike")
    ggplot(attrition,aes(PercentSalaryHike,fill=Attrition, alpha = 0.03))+geom_density()


    # RelationshipSatisfaction
    summary(attrition$RelationshipSatisfaction)
    var(attrition$RelationshipSatisfaction)
    sd(attrition$RelationshipSatisfaction)
    hist(attrition$RelationshipSatisfaction, main = "Histogram of RelationshipSatisfaction", xlab = "RelationshipSatisfaction", col = "blue")
    boxplot(attrition$RelationshipSatisfaction, horizontal = TRUE,
            main = "Boxplot of RelationshipSatisfaction", xlab = "RelationshipSatisfaction")
    ggplot(attrition,aes(RelationshipSatisfaction,fill=Attrition, alpha = 0.03))+geom_bar()


    # StockOptionLevel
    summary(attrition$StockOptionLevel)
    var(attrition$StockOptionLevel)
    sd(attrition$StockOptionLevel)
    hist(attrition$StockOptionLevel, main = "Histogram of StockOptionLevel", xlab = "StockOptionLevel", col = "blue")
    boxplot(attrition$StockOptionLevel, horizontal = TRUE,
            main = "Boxplot of StockOptionLevel", xlab = "StockOptionLevel")
    ggplot(attrition,aes(StockOptionLevel,fill=Attrition, alpha = 0.03))+geom_bar()


    # TrainingTimesLastYear
    summary(attrition$TrainingTimesLastYear)
    var(attrition$TrainingTimesLastYear)
    sd(attrition$TrainingTimesLastYear)
    hist(attrition$TrainingTimesLastYear, main = "Histogram of TrainingTimesLastYear", xlab = "TrainingTimesLastYear", col = "blue")
    boxplot(attrition$TrainingTimesLastYear, horizontal = TRUE,
            main = "Boxplot of TrainingTimesLastYear", xlab = "TrainingTimesLastYear")
    ggplot(attrition,aes(TrainingTimesLastYear,fill=Attrition, alpha = 0.03))+geom_bar()

    # WorkLifeBalance
    summary(attrition$WorkLifeBalance)
    var(attrition$WorkLifeBalance)
    sd(attrition$WorkLifeBalance)
    hist(attrition$WorkLifeBalance, main = "Histogram of WorkLifeBalance", xlab = "WorkLifeBalance", col = "blue")
    boxplot(attrition$WorkLifeBalance, horizontal = TRUE,
            main = "Boxplot of WorkLifeBalance", xlab = "WorkLifeBalance")
    ggplot(attrition,aes(WorkLifeBalance,fill=Attrition, alpha = 0.03))+geom_bar()

    # YearsAtCompany
```

```
summary(attrition$YearsAtCompany)
var(attrition$YearsAtCompany)
sd(attrition$YearsAtCompany)
hist(attrition$YearsAtCompany, main = "Histogram of YearsAtCompany", xlab = "YearsAtCompany", col = "blue")
boxplot(attrition$YearsAtCompany, horizontal = TRUE,
        main = "Boxplot of YearsAtCompany", xlab = "YearsAtCompany")
  # Outliers Univariate - Monthly Income
  outlier_values <- boxplot.stats(attrition$YearsAtCompany)$out # outlier values

  boxplot(attrition$YearsAtCompany, main="Monthly Income", boxwex=0.1, horizontal = TRUE,
        xlab = "YearsAtCompany")
  mtext(paste("Outliers: ", paste(outlier_values, collapse = ", ")), cex = 0.6)

  # Outliers Bivariate - YearsAtCompany
  # For categorical variable
  boxplot(YearsAtCompany ~ JobRole, data=attrition, main="YearsAtCompany Accross Job
        Role", ylab = "YearsAtComapny", xlab = "JobRole")  # clear pattern is noticeable.
  boxplot(YearsAtCompany ~ Department, data=attrition, main="YearsAtCompany Accross
        Department", ylab = "YearsAtComapny", xlab = "Department")  # this may not be significant, as day of week variable is a subset of the month var.
  # Plot of data with Outliers
  par(mfrow=c(1,1))
  plot(attrition$YearsAtCompany, attrition$MonthlyIncome, xlim=c(0, 40), ylim=c(1000, 20000),
     main="With Outliers", xlab="MonthlyIncome", ylab="YearsAtCompany", pch="*", col="red", cex=2)
  abline(lm(MonthlyIncome ~ YearsAtCompany, data = attrition), col="blue", lwd = 3, lty=2)
ggplot(attrition,aes(YearsAtCompany,fill=Attrition, alpha = 0.03))+geom_bar()

# CompaOverallGroup
str(attrition$CompaOverallGroup)
summary(attrition$CompaOverallGroup)
var(attrition$CompaOverallGroup)
sd(attrition$CompaOverallGroup)
hist(attrition$CompaOverallGroup, main = "Histogram of CompaOverallGroup", xlab = "CompaOverallGroup", col = "blue")
boxplot(attrition$CompaOverallGroup, horizontal = TRUE,
        main = "Boxplot of CompaOverallGroup", xlab = "CompaOverallGroup")
table(attrition$Attrition, attrition$CompaOverallGroup)
prop.table(table(attrition$CompaOverallGroup, attrition$Attrition))
ggplot(attrition,aes(CompaOverallGroup, fill = Attrition, alpha = 0.3))+geom_bar()

########################################### K-Modes Algorithm ####################################################
library(klaR)
data.to.cluster <- attrition
cluster.results <- kmodes(data.to.cluster, 3, iter.max = 10, weighted = FALSE)
cluster.results
summary(cluster.results)

# ---------------------Comparative "TEST" Testing-------------------------------

# install.packages("klaR")
# install.packages("caret")

# load libraries
library(mlbench)
library(caret)
```

```
library(klaR)

# rename dataset to keep code below generic
dataset_test <- attrition

control <- trainControl(method="repeatedcv", number=10, repeats=3)
seed <- 7

metric <- "Accuracy"
preProcess=c("center", "scale")


# Linear Discriminant Analysis
set.seed(seed)
fit.lda <- train(Attrition~., data=dataset_test, method="lda", metric=metric, preProc=c("center", "scale"), trControl=control)
# Logistic Regression
set.seed(seed)
fit.glm <- train(Attrition~., data=dataset_test, method="glm", metric=metric, trControl=control)
# GLMNET
set.seed(seed)
fit.glmnet <- train(Attrition~., data=dataset_test, method="glmnet", metric=metric, preProc=c("center", "scale"), trControl=control)
# SVM Radial
set.seed(seed)
fit.svmRadial <- train(Attrition~., data=dataset_test, method="svmRadial", metric=metric, preProc=c("center", "scale"), trControl=control, fit=FALSE)
# kNN
set.seed(seed)
fit.knn <- train(Attrition~., data=dataset_test, method="knn", metric=metric, preProc=c("center", "scale"), trControl=control)
# Naive Bayes
# set.seed(seed)
# fit.nb <- train(Attrition~., data=dataset_test, method="nb", metric=metric, trControl=control)
# CART
set.seed(seed)
fit.cart <- train(Attrition~., data=dataset_test, method="rpart", metric=metric, trControl=control)
# C5.0
# set.seed(seed)
# fit.c50 <- train(Attrition~., data=dataset_test, method="C5.0", metric=metric, trControl=control)
# Bagged CART
set.seed(seed)
fit.treebag <- train(Attrition~., data=dataset_test, method="treebag", metric=metric, trControl=control)
# Random Forest
set.seed(seed)
fit.rf <- train(Attrition~., data=dataset_test, method="rf", metric=metric, trControl=control)
# Stochastic Gradient Boosting (Generalized Boosted Modeling)
set.seed(seed)
fit.gbm <- train(Attrition~., data=dataset_test, method="gbm", metric=metric, trControl=control, verbose=FALSE)
# Decision Tree
set.seed(seed)
fit.dt <- train(Attrition~., data=dataset_test, method="rpart", metric=metric, trControl=control)

results <- resamples(list("Logistic Regression"=fit.glm,"SVM Radial"=fit.svmRadial, kNN=fit.knn, CART=fit.cart,
                          "Bagged CART"=fit.treebag, "Random Forest"=fit.rf, "Stochastic Gradient Boosting"=fit.gbm,
                          "Decision Tree" =fit.dt ))
```

```
        # Table comparison
        summary(results)
        results


        # boxplot comparison
        bwplot(results)
        # Dot-plot comparison
        dotplot(results)


        # -------------------------------LOGISITIC REGRESSION---------------------------------------------
        nrow(attrition)
        # 1st Split data into training and testing sets:

        train <- createDataPartition(attrition$Attrition,p=0.7,list=FALSE)
        set.seed(2017)
        training <- attrition[train,]
        testing <- attrition[-train,]
        # Check Spliting Results
        dim(training); dim(testing)

        # Fitting the LOg Regresssion Model
        mod_fit <- glm(Attrition ~ .,family=binomial(link="logit"),data=training)
        mod_fit
        summary(mod_fit)

        qchisq(0.95, 992)

        # Predictive Model for Attrition - Most Significant Variables
        "Age, BusinessTravel, DistanceFromHome, EnviromentSatisfaction, Gender,
        JobInvolvement, JobRole, JobSatisfaction, NumCompaniesWorked, OverTime,
        RelationshipSatisfaction"

        # ANOVA of model_log
        anova(mod_fit, test="Chisq")

        # Logistic Regression Accuracy or the predictive ability of the mod_fit
        testing$Attrition <- as.character(testing$Attrition)
        testing$Attrition[testing$Attrition=="No"] <- "0"
        testing$Attrition[testing$Attrition=="Yes"] <- "1"
        fitted.results <- predict(mod_fit,newdata=testing,type='response')
        fitted.results <- ifelse(fitted.results > 0.5,1,0)
        misClasificError <- mean(fitted.results != testing$Attrition)
        print(paste('Logistic Regression Accuracy',1-misClasificError))

        # Log Reg Confusion Matrix
        print("Confusion Matrix for Logistic Regression"); table(testing$Attrition, fitted.results > 0.5)


        mod_fit_selective <- glm(Attrition ~ Age + BusinessTravel + DistanceFromHome + EnvironmentSatisfaction + Gender +
                                 JobInvolvement + JobRole + JobSatisfaction + NumCompaniesWorked + OverTime +
                                 RelationshipSatisfaction
                                 , family=binomial(link="logit"),data=training)
        summary(mod_fit_selective)

        # ANOVA of model_log
        anova(mod_fit_selective, test="Chisq")

        # Logistic Regression Accuracy or the predictive ability of the mod_fit_selective
        testing$Attrition <- as.character(testing$Attrition)
```

```r
testing$Attrition[testing$Attrition=="No"] <- "0"
testing$Attrition[testing$Attrition=="Yes"] <- "1"
fitted.results <- predict(mod_fit_selective,newdata=testing,type='response')
fitted.results <- ifelse(fitted.results > 0.5,1,0)
misClasificError <- mean(fitted.results != testing$Attrition)
print(paste('Logistic Regression Accuracy',1-misClasificError))

# Log Reg Confusion Matrix
print("Confusion Matrix for Logistic Regression"); table(testing$Attrition, fitted.results > 0.5)



# ---------------------Odds Ratio--------------------
library(MASS)
exp(cbind(OR=coef(mod_fit), confint(mod_fit)))




# ====================================Decision Trees====================================

# ---------------------Decision Tree------------------
tree <- ctree(Attrition~., training)
plot(tree)

# Decision Tree Confusion Matrix
pred_tree <- predict(tree, testing)
print("Confusion Matrix for Decision Tree"); table(Predicted = pred_tree, Actual = testing$Attrition)

# Decision Tree Accuracy
p1 <- predict(tree, training)
tab1 <- table(Predicted = p1, Actual = training$Attrition)
tab2 <- table(Predicted = pred_tree, Actual = testing$Attrition)
print(paste('Decision Tree Accuracy',sum(diag(tab2))/sum(tab2)))
"Accuracy is not imporved over log regression"


# ----------------Decision Tree Selective---------------
tree_selective <- ctree(Attrition ~ ., training)
plot(tree_selective)

# Decision Tree Confusion Matrix
pred_tree_selective <- predict(tree_selective, testing)
print("Confusion Matrix for Decision Tree"); table(Predicted = pred_tree_selective, Actual = testing$Attrition)

# Decision Tree Accuracy
p1_selective <- predict(tree_selective, training)
tab1_selective <- table(Predicted = p1_selective, Actual = training$Attrition)
tab2_selective <- table(Predicted = pred_tree_selective, Actual = testing$Attrition)
print(paste('Decision Tree Accuracy',sum(diag(tab2_selective))/sum(tab2_selective)))
"Accuracy is not imporved over log regression"

summary(attrition$YearsWithoutPromotion_WithCurrentManager_group)

# ----------------Decision Tree Selective 2---------------
tree_selective_2 <- ctree(Attrition ~ BusinessTravel + EnvironmentSatisfaction + JobInvolvement
                          + MaritalStatus + NumCompaniesWorked + OverTime
                          + YearsWithoutPromotion_WithCurrentManager_group, training)
```

```r
plot(tree_selective_2)

# Decision Tree Confusion Matrix
pred_tree_selective_2 <- predict(tree_selective_2, testing)
print("Confusion Matrix for Decision Tree"); table(Predicted = pred_tree_selective_2, Actual = testing$Attrition)

# Decision Tree Accuracy
p1_selective_2 <- predict(tree_selective_2, training)
tab1_selective_2 <- table(Predicted = p1_selective_2, Actual = training$Attrition)
tab2_selective_2 <- table(Predicted = pred_tree_selective_2, Actual = testing$Attrition)
print(paste('Decision Tree Accuracy',sum(diag(tab2_selective_2))/sum(tab2_selective_2)))
"Accuracy is not imporved over log regression"


#### Discriminate Analysis ################

library(lmtest)
lrtest(mod_fit_one, mod_fit_two)

# Discriminate Analysis
mod_fit_one <- glm(Attrition ~ ., data=training, family="binomial")
mod_fit_two <- glm(Attrition ~ BusinessTravel + EnvironmentSatisfaction + JobInvolvement
                   + MaritalStatus + NumCompaniesWorked + OverTime
                   + YearsWithoutPromotion_WithCurrentManager_group, data=training, family="binomial")

library(lmtest)
lrtest(mod_fit_one, mod_fit_two)

#### RocK Curve ####
# install.packages("ROCR")
library(ROCR)
# Compute AUC for predicting Class with the model
prob <- predict(mod_fit_one, newdata=testing, type="response")
pred <- prediction(prob, testing$Attrition)
perf <- performance(pred, measure = "tpr", x.measure = "fpr")
plot(perf)

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

# VarImp
sig_var <- train(Attrition ~ ., data=attrition, method ="glm", family="binomial")
varImp(sig_var)

#Significant Variables: BusinessTravel, EnvironmentalSatisfaction, JobInvolvement, MaritalStatus,  NumCompaniesWorked, OverTime, YearsWithoutPromotion_WithCurrentManager_group
sv1 <- ggplot(attrition,aes(BusinessTravel,fill = Attrition))+geom_bar()
sv2 <- ggplot(attrition,aes(EnvironmentSatisfaction,fill = Attrition))+geom_bar()
sv3 <- ggplot(attrition,aes(JobInvolvement,fill = Attrition))+geom_bar()
sv4 <- ggplot(attrition,aes(MaritalStatus,fill = Attrition))+geom_bar()
sv5 <- ggplot(attrition,aes(NumCompaniesWorked,fill = Attrition))+geom_bar()
sv6 <- ggplot(attrition,aes(OverTime,fill = Attrition))+geom_bar()
sv7 <- ggplot(attrition,aes(YearsWithoutPromotion_WithCurrentManager_group,fill = Attrition))+geom_bar()
grid.arrange(sv1,sv2,sv3,sv4,sv5,sv6,sv7,ncol=2,top = "Figure: Significant Variables")
```

```
# Cleaned Data Output attrition
write.csv(attrition, file = "cleaned_attrition.csv", row.names = FALSE)
```