Always Analytics Frisco Lay Employee Retention Consultation

Nicole Assenza

Setempber 9, 2022

#####YouTube video link: <https://youtu.be/i5hs0tR\_zCo>

#####RShiny App link: <https://seemantsrivastava.shinyapps.io/Attrition/>

#####GitHub repository link: <https://github.com/nassenza>

###Introduction:

####As a consultant for Always Analytics our team has worked hard to answer important questions: What the facot contributing to employee attrition? What factorsimpact the decisions of these employees to leave Frito Lay? Is salary too low? Job satisfaction leaving something to be desired? Is it a company culure issue where employees don't feel the work/life balance is unsatisfactory? The results might surprise you. We have completed the analysis of the dataset on employee attrition dataset you've provided. Hopefully this will help your HR department greatly in improving employee rentention rates. We've explored the data and built a model to predict what factors contribute to employees leaving the company or staying on. The datset provided has great wealth of info regarding 870 employees who've either chosen to quit or stay with the company. Thre are 36 features we've investigated to see which ones are most meaningful and impactful.

##Initial Exploration

###Summary: No missing data, which makes our job easier. One thing to note is that 84% of employees did choose to stay at Frito Lay. Which also means that fewer people quit, which is a bit tricky with lack of documentation of attrited employees but good for Frito which means, generally people are happy with their jobs.Next we will check for any missing/ not available values.

##EDA (Exploratory Data Analysis):

##Before we start modelling, we really need to explore the data and play around with the information to see what jumps out at us as possibly leading to some insights.What data correlates? What doesn't? We will answer those questions here.Some obvious ones come to mind: Older employees who keep the same job will have more working hours/years and income, for example, as well as having the same boss during that time. Let's see if that holds true.

##Interesting, Age, YearsatCompany, YearsinCurrentRole, YearswithCurrentManager, and TotalWorkingYears, Hourly Rate seem to be negatively correlated. Interesting! Now let's look at who's leaving, as you can see from the chart mostare happy staying at the company.

#### Income seems a likely choice for reason why someone would quit. Let's have a look-see if that's so.

#### Income in a major factor in levels of attrition, strangely enough the employees with higher salaries tend to stay. As we get into the 500-1000 montly income range, attrition sees a huge spike.

#### Next, we take a look at attrition levels in certrain departments.

#### Data doesn't lie: Sales and R%D have very high attrition rates: 53.57% and 41.14% respectively.HR has low attrition rates at around 4%.

##### In the next graph, HR, R&D and Sales have good retention rates.

scale\_x\_discrete(labels=function(x) str\_wrap(x,width=10))+scale\_fill\_brewer(palette="Set3")

```

#### Is this a relationship between gettng married and quitting a job? It would be interesting to see that broken down by gender. However, it doesn't seem very relavent to attrition rates.We can most likely to remove this as a key factor.

###Attrition/Marital Status/AgE/ monthly income:

#### Single people at any age seem to quit more, while the married regardless of age stay on and even divorced people tend to quit less.Salary is a large indicator for married people since they might (this is speculation) have families to support and require a higher paying salary or a more balanced relationship between work and home life. Single people are unattached so maybe they want to keep their options open.

####Attrition Vs Distance From Home:

##### Suprisingly, attrition rates decrease as the distance from home to office increase.

####Attrition Vs Payrates:

##### With the employees who quit, median monthly incomes are much less than those who stay employeed at Frito Lay. If you're making 5000 or less a much, I can somewhat understand looking for other opportunities. Although that's a very decent wage if you ask me.

####Attrition Vs Education:

#### Regardless of education levels, very few employees quit. Somewhat surprising results. The highest attrition rates are those with Bachelor's degrees, and also the highest rentention rates. This may have something to do with the fact that there are many, many people with Bachelor's degrees..

####Attrition Vs TotalWorkingYears:

##### Perhaps those with around 8 years or less of experience possibly are looking for a salary boost when they change companies.

####Attrition Vs Job Satisfaction:

##### The lower the satisfaction rate the greater impact the percentage of people who leave the company (27.1%) . It is unsual, however, that many people who leave still have a high or very high job satisfaction at nearly 31 percent. There must be some other indicators that make employees quit.

#####There is a visible trend in the category of people who do not leave whereas this is not so in the case of people who leave.

####Attrition Vs Worklife balance:

##### Another weird result: the differences between the employees who stay and thoes who leave have very similar satification in work/life balance. Let's try work environment to see what results that yields.

####Attrition Vs Environment Satisfaction:

#####H Puzzling. Those with low work environment satisfaction stay on almost identically percentage-wise as those who quit! Even the quitting employees have similarly high (those not quite as high as those happy with the environment) percentages of medium, high, and very high levels of satisfaction with the work environment.

####Attrition Vs OverTime:

##### Finally, a little less puzzling results! Apparently people love overtime! Nearly 80 percento of retained employees do not work overtime, with a small 24% of those who do. The employees who quit must really hate working overtime. If working overtime does keeps you away from resting, your family, and maybe even a lack of overtime pay may play a huge role in why people choose to quit.

####Attrition for years of experience vs monthly salary and their correlation:

##### My intuition about income has come to some fruition,the relationship between years of experience and monthly income is linear as indicated by said line. However, there is a crossover point where the retained employees wit the same number of years of experience earn higher income.

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

#### Attrition based on gender

Because attrition rates reported by Frito Lay are fairly low; male and females who quit are in the 15-17% range.

### Conclusion

The top variables that contribute to employee attritition are: Commute, Environmental Satisfaction, Overtime, Job Involvement, Work Life Balance, Business Travel and Years Since Last Promotion.

##Future Research

It may behoove Frito to get more data involving attrition since most of the data came from retained employees! As large company with many subsidiary companies, data collection could be obtained from those entities.

Thank you for your business. We at Always Analytics are excited to work with you team on future endeavors!

# Load Libraries  
library(dplyr) # for string functions  
library(ggplot2) # for plots  
library(kableExtra) # for table formatting  
library(knitr) # for presenting in html  
library(Hmisc) # for describing data  
library(caret) # for classification & regression training  
library(mlr) # for machine learing  
library(class) # for classification funtions  
library(corrplot) # for correlation matrix  
library(data.table) # for table creation for using one\_hot  
library(mltools) # for one\_hot dummy creation  
library(gridExtra) # for plots formatting

Abstract

Employees are valuable assets of an organization, and the unexpected departure of a key employee can be disruptive and costly. Our objective with this study is to identify factors that lead to employee attrition, enabling management to establish areas of focus to reduce employee churn.

Turnover is an expensive problem, both in terms of money and time so our data science team is both excited and well positioned to provide our clients with new insights to mitigate employee churn and its associated costs. <https://www.forbes.com/sites/billconerly/2018/08/12/companies-need-to-know-the-dollar-cost-of-employee-turnover/#591dbffd590a>

Here is information about how we used the data to build a employee attrition model.

Data Input

The training and validation data sets were read into data frames. We combined the data frames to one data frame to clean all the data at the same time if needed.

# Read employee data from the training csv file and name the dataset train.  
train <- read.csv("C:/Users/ASUSL/OneDrive/Desktop/CASETSTUDY THE LASTONE/CaseStudy2-data.csv",header=TRUE,stringsAsFactors = TRUE,sep=",",encoding = "UTF-8")  
  
# Read employee data from the validation csv file and names the dataset train.  
validation <- read.csv("C:/Users/ASUSL/OneDrive/Desktop/CASETSTUDY THE LASTONE/CaseStudy2Validation.csv",header=TRUE,stringsAsFactors = TRUE,sep=",",encoding = "UTF-8")  
  
alldata = rbind(train,validation) # create dataframe with all data to clean all at same time

Data Exploration

There are 1170 observations with 37 variables (9 factors, 27 integers, 1 random number)

The following initial assessments/actions were completed: - Reviewed the data to confirm factor/categorical variable values exist in both. - Removed variables where all values are the same: StandardHours, EmployeeCount, Over18 - Removed unrelated variable Rand - Recoded Attrition to 0=No, 1=Yes

We classified the variables into 4 areas:

Demographics: Age, Gender, Marital Status, Education, Education Level and Distance from home (commute)

Attitudinal: Environmental Satisfaction, Job Satisfaction, Relationship Satisfaction, WorkLifeBalance,Job Involvement, Performance Rating

Work Demographics: Job ROle, Job Level, Business Travel, Department, Training

Financial: Hourly Rate (rates are collinear so showing only hourly rate), Overtime, Percent Salary Hike, Stock Option Levels

This document shows the process behind our analysis.

#] HMisc provides more details than str(alldata)  
Hmisc::describe(alldata, digits = 2, tabular = TRUE)

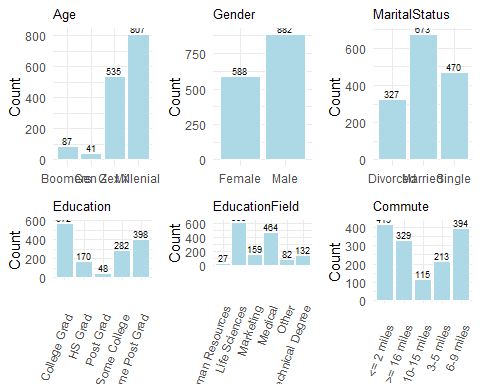
## alldata   
##   
## 37 Variables 1470 Observations  
## --------------------------------------------------------------------------------  
## ID   
## n missing distinct Info Mean Gmd .05 .10   
## 1470 0 1470 1 736 490 74 148   
## .25 .50 .75 .90 .95   
## 368 736 1103 1323 1397   
##   
## lowest : 1 2 3 4 5, highest: 1466 1467 1468 1469 1470  
## --------------------------------------------------------------------------------  
## Age   
## n missing distinct Info Mean Gmd .05 .10   
## 1470 0 43 0.999 37 10 24 26   
## .25 .50 .75 .90 .95   
## 30 36 43 50 54   
##   
## lowest : 18 19 20 21 22, highest: 56 57 58 59 60  
## --------------------------------------------------------------------------------  
## Attrition   
## n missing distinct   
## 1470 0 2   
##   
## Value No Yes  
## Frequency 1233 237  
## Proportion 0.839 0.161  
## --------------------------------------------------------------------------------  
## BusinessTravel   
## n missing distinct   
## 1470 0 3   
##   
## Value Non-Travel Travel\_Frequently Travel\_Rarely  
## Frequency 150 277 1043  
## Proportion 0.102 0.188 0.710  
## --------------------------------------------------------------------------------  
## DailyRate   
## n missing distinct Info Mean Gmd .05 .10   
## 1470 0 886 1 802 466 165 243   
## .25 .50 .75 .90 .95   
## 465 802 1157 1356 1424   
##   
## lowest : 102 103 104 105 106, highest: 1492 1495 1496 1498 1499  
## --------------------------------------------------------------------------------  
## Department   
## n missing distinct   
## 1470 0 3   
##   
## Value Human Resources Research & Development Sales  
## Frequency 63 961 446  
## Proportion 0.043 0.654 0.303  
## --------------------------------------------------------------------------------  
## DistanceFromHome   
## n missing distinct Info Mean Gmd .05 .10   
## 1470 0 29 0.993 9.2 8.8 1 1   
## .25 .50 .75 .90 .95   
## 2 7 14 23 26   
##   
## lowest : 1 2 3 4 5, highest: 25 26 27 28 29  
## --------------------------------------------------------------------------------  
## Education   
## n missing distinct Info Mean Gmd   
## 1470 0 5 0.913 2.9 1.1   
##   
## lowest : 1 2 3 4 5, highest: 1 2 3 4 5  
##   
## Value 1 2 3 4 5  
## Frequency 170 282 572 398 48  
## Proportion 0.116 0.192 0.389 0.271 0.033  
## --------------------------------------------------------------------------------  
## EducationField   
## n missing distinct   
## 1470 0 6   
##   
## lowest : Human Resources Life Sciences Marketing Medical Other   
## highest: Life Sciences Marketing Medical Other Technical Degree  
##   
## Value Human Resources Life Sciences Marketing Medical  
## Frequency 27 606 159 464  
## Proportion 0.018 0.412 0.108 0.316  
##   
## Value Other Technical Degree  
## Frequency 82 132  
## Proportion 0.056 0.090  
## --------------------------------------------------------------------------------  
## EmployeeCount   
## n missing distinct Info Mean Gmd   
## 1470 0 1 0 1 0   
##   
## Value 1  
## Frequency 1470  
## Proportion 1  
## --------------------------------------------------------------------------------  
## EmployeeNumber   
## n missing distinct Info Mean Gmd .05 .10   
## 1470 0 1470 1 1025 695 96 199   
## .25 .50 .75 .90 .95   
## 491 1020 1556 1857 1968   
##   
## lowest : 1 2 4 5 7, highest: 2061 2062 2064 2065 2068  
## --------------------------------------------------------------------------------  
## EnvironmentSatisfaction   
## n missing distinct Info Mean Gmd   
## 1470 0 4 0.928 2.7 1.2   
##   
## Value 1 2 3 4  
## Frequency 284 287 453 446  
## Proportion 0.193 0.195 0.308 0.303  
## --------------------------------------------------------------------------------  
## Gender   
## n missing distinct   
## 1470 0 2   
##   
## Value Female Male  
## Frequency 588 882  
## Proportion 0.4 0.6  
## --------------------------------------------------------------------------------  
## HourlyRate   
## n missing distinct Info Mean Gmd .05 .10   
## 1470 0 71 1 66 23 33 38   
## .25 .50 .75 .90 .95   
## 48 66 84 94 97   
##   
## lowest : 30 31 32 33 34, highest: 96 97 98 99 100  
## --------------------------------------------------------------------------------  
## JobInvolvement   
## n missing distinct Info Mean Gmd   
## 1470 0 4 0.776 2.7 0.71   
##   
## Value 1 2 3 4  
## Frequency 83 375 868 144  
## Proportion 0.056 0.255 0.590 0.098  
## --------------------------------------------------------------------------------  
## JobLevel   
## n missing distinct Info Mean Gmd   
## 1470 0 5 0.898 2.1 1.2   
##   
## lowest : 1 2 3 4 5, highest: 1 2 3 4 5  
##   
## Value 1 2 3 4 5  
## Frequency 543 534 218 106 69  
## Proportion 0.369 0.363 0.148 0.072 0.047  
## --------------------------------------------------------------------------------  
## JobRole   
## n missing distinct   
## 1470 0 9   
##   
## lowest : Healthcare Representative Human Resources Laboratory Technician Manager Manufacturing Director   
## highest: Manufacturing Director Research Director Research Scientist Sales Executive Sales Representative   
## --------------------------------------------------------------------------------  
## JobSatisfaction   
## n missing distinct Info Mean Gmd   
## 1470 0 4 0.928 2.7 1.2   
##   
## Value 1 2 3 4  
## Frequency 289 280 442 459  
## Proportion 0.197 0.190 0.301 0.312  
## --------------------------------------------------------------------------------  
## MaritalStatus   
## n missing distinct   
## 1470 0 3   
##   
## Value Divorced Married Single  
## Frequency 327 673 470  
## Proportion 0.222 0.458 0.320  
## --------------------------------------------------------------------------------  
## MonthlyIncome   
## n missing distinct Info Mean Gmd .05 .10   
## 1470 0 1349 1 6503 4868 2098 2318   
## .25 .50 .75 .90 .95   
## 2911 4919 8379 13776 17821   
##   
## lowest : 1009 1051 1052 1081 1091, highest: 19859 19926 19943 19973 19999  
## --------------------------------------------------------------------------------  
## MonthlyRate   
## n missing distinct Info Mean Gmd .05 .10   
## 1470 0 1427 1 14313 8221 3385 4603   
## .25 .50 .75 .90 .95   
## 8047 14236 20462 24002 25432   
##   
## lowest : 2094 2097 2104 2112 2122, highest: 26956 26959 26968 26997 26999  
## --------------------------------------------------------------------------------  
## NumCompaniesWorked   
## n missing distinct Info Mean Gmd .05 .10   
## 1470 0 10 0.95 2.7 2.7 0 0   
## .25 .50 .75 .90 .95   
## 1 2 4 7 8   
##   
## lowest : 0 1 2 3 4, highest: 5 6 7 8 9  
##   
## Value 0 1 2 3 4 5 6 7 8 9  
## Frequency 197 521 146 159 139 63 70 74 49 52  
## Proportion 0.134 0.354 0.099 0.108 0.095 0.043 0.048 0.050 0.033 0.035  
## --------------------------------------------------------------------------------  
## Over18   
## n missing distinct value   
## 1470 0 1 Y   
##   
## Value Y  
## Frequency 1470  
## Proportion 1  
## --------------------------------------------------------------------------------  
## OverTime   
## n missing distinct   
## 1470 0 2   
##   
## Value No Yes  
## Frequency 1054 416  
## Proportion 0.717 0.283  
## --------------------------------------------------------------------------------  
## PercentSalaryHike   
## n missing distinct Info Mean Gmd .05 .10   
## 1470 0 15 0.988 15 4 11 11   
## .25 .50 .75 .90 .95   
## 12 14 18 21 22   
##   
## lowest : 11 12 13 14 15, highest: 21 22 23 24 25  
##   
## Value 11 12 13 14 15 16 17 18 19 20 21  
## Frequency 210 198 209 201 101 78 82 89 76 55 48  
## Proportion 0.143 0.135 0.142 0.137 0.069 0.053 0.056 0.061 0.052 0.037 0.033  
##   
## Value 22 23 24 25  
## Frequency 56 28 21 18  
## Proportion 0.038 0.019 0.014 0.012  
## --------------------------------------------------------------------------------  
## PerformanceRating   
## n missing distinct Info Mean Gmd   
## 1470 0 2 0.39 3.2 0.26   
##   
## Value 3 4  
## Frequency 1244 226  
## Proportion 0.846 0.154  
## --------------------------------------------------------------------------------  
## RelationshipSatisfaction   
## n missing distinct Info Mean Gmd   
## 1470 0 4 0.929 2.7 1.2   
##   
## Value 1 2 3 4  
## Frequency 276 303 459 432  
## Proportion 0.188 0.206 0.312 0.294  
## --------------------------------------------------------------------------------  
## StandardHours   
## n missing distinct Info Mean Gmd   
## 1470 0 1 0 80 0   
##   
## Value 80  
## Frequency 1470  
## Proportion 1  
## --------------------------------------------------------------------------------  
## StockOptionLevel   
## n missing distinct Info Mean Gmd   
## 1470 0 4 0.853 0.79 0.88   
##   
## Value 0 1 2 3  
## Frequency 631 596 158 85  
## Proportion 0.429 0.405 0.107 0.058  
## --------------------------------------------------------------------------------  
## TotalWorkingYears   
## n missing distinct Info Mean Gmd .05 .10   
## 1470 0 40 0.995 11 8.3 1 3   
## .25 .50 .75 .90 .95   
## 6 10 15 23 28   
##   
## lowest : 0 1 2 3 4, highest: 35 36 37 38 40  
## --------------------------------------------------------------------------------  
## TrainingTimesLastYear   
## n missing distinct Info Mean Gmd   
## 1470 0 7 0.91 2.8 1.4   
##   
## lowest : 0 1 2 3 4, highest: 2 3 4 5 6  
##   
## Value 0 1 2 3 4 5 6  
## Frequency 54 71 547 491 123 119 65  
## Proportion 0.037 0.048 0.372 0.334 0.084 0.081 0.044  
## --------------------------------------------------------------------------------  
## WorkLifeBalance   
## n missing distinct Info Mean Gmd   
## 1470 0 4 0.762 2.8 0.7   
##   
## Value 1 2 3 4  
## Frequency 80 344 893 153  
## Proportion 0.054 0.234 0.607 0.104  
## --------------------------------------------------------------------------------  
## YearsAtCompany   
## n missing distinct Info Mean Gmd .05 .10   
## 1470 0 37 0.993 7 6.2 1 1   
## .25 .50 .75 .90 .95   
## 3 5 9 15 20   
##   
## lowest : 0 1 2 3 4, highest: 33 34 36 37 40  
## --------------------------------------------------------------------------------  
## YearsInCurrentRole   
## n missing distinct Info Mean Gmd .05 .10   
## 1470 0 19 0.974 4.2 3.9 0 0   
## .25 .50 .75 .90 .95   
## 2 3 7 9 11   
##   
## lowest : 0 1 2 3 4, highest: 14 15 16 17 18  
##   
## Value 0 1 2 3 4 5 6 7 8 9 10  
## Frequency 244 57 372 135 104 36 37 222 89 67 29  
## Proportion 0.166 0.039 0.253 0.092 0.071 0.024 0.025 0.151 0.061 0.046 0.020  
##   
## Value 11 12 13 14 15 16 17 18  
## Frequency 22 10 14 11 8 7 4 2  
## Proportion 0.015 0.007 0.010 0.007 0.005 0.005 0.003 0.001  
## --------------------------------------------------------------------------------  
## YearsSinceLastPromotion   
## n missing distinct Info Mean Gmd .05 .10   
## 1470 0 16 0.922 2.2 3 0 0   
## .25 .50 .75 .90 .95   
## 0 1 3 7 9   
##   
## lowest : 0 1 2 3 4, highest: 11 12 13 14 15  
##   
## Value 0 1 2 3 4 5 6 7 8 9 10  
## Frequency 581 357 159 52 61 45 32 76 18 17 6  
## Proportion 0.395 0.243 0.108 0.035 0.041 0.031 0.022 0.052 0.012 0.012 0.004  
##   
## Value 11 12 13 14 15  
## Frequency 24 10 10 9 13  
## Proportion 0.016 0.007 0.007 0.006 0.009  
## --------------------------------------------------------------------------------  
## YearsWithCurrManager   
## n missing distinct Info Mean Gmd .05 .10   
## 1470 0 18 0.976 4.1 3.9 0 0   
## .25 .50 .75 .90 .95   
## 2 3 7 9 10   
##   
## lowest : 0 1 2 3 4, highest: 13 14 15 16 17  
##   
## Value 0 1 2 3 4 5 6 7 8 9 10  
## Frequency 263 76 344 142 98 31 29 216 107 64 27  
## Proportion 0.179 0.052 0.234 0.097 0.067 0.021 0.020 0.147 0.073 0.044 0.018  
##   
## Value 11 12 13 14 15 16 17  
## Frequency 22 18 14 5 5 2 7  
## Proportion 0.015 0.012 0.010 0.003 0.003 0.001 0.005  
## --------------------------------------------------------------------------------  
## Rand   
## n missing distinct Info Mean Gmd .05 .10   
## 1470 0 1470 1 -0.0055 1.1 -1.716 -1.287   
## .25 .50 .75 .90 .95   
## -0.674 0.019 0.647 1.268 1.611   
##   
## lowest : -3.019377 -2.722264 -2.631626 -2.586027 -2.585977  
## highest: 2.879533 3.024861 3.060540 3.084939 3.183079  
## --------------------------------------------------------------------------------

# There are 1170 observations with 37 variables (9 factors, 27 integers, 1 random number)  
# Reviewed train and validate data to confirm factor/categorical variable values exist in both  
# Remove variables where all values are the same: StandardHours, EmployeeCount, Over18  
# Remove unrelated variable  
alldata <- select(alldata,-c(StandardHours, EmployeeCount, Over18, Rand))  
##Manually created data dictionary based on results below with a notes field  
#DataDict <- read.csv("C:/Users/ASUSL/OneDrive/Desktop/CASETSTUDY THE LASTONE/Datadictionary.csv,header=TRUE,sep=",",encoding = "UTF-8")

# Mutate for EDA  
  
plotalldata <- alldata %>% mutate(   
Age = as.factor   
(ifelse(Age %in% 0:21,"Gen Z", ifelse(Age %in% 22:37,"Millenial",   
ifelse(Age %in% 38:53,"Gex X", ifelse(Age %in% 54:72,"Boomers",   
 "Silent"))))),  
#Generation Cutoff Ref:   
#http://www.pewresearch.org/fact-tank/2018/03/01/defining-generations-where-millennials-end-and-post-millennials-begin/  
  
DistanceFromHome = as.factor   
(ifelse(DistanceFromHome %in% 0:2,"<= 2 miles", ifelse(DistanceFromHome %in% 3:5,"3-5 miles",   
 ifelse(DistanceFromHome %in% 6:10,"6-9 miles", ifelse(DistanceFromHome %in% 10:15,"10-15 miles",  
">= 16 miles"))))),  
  
Education = as.factor #Source lastresearch project, best estimate of values  
(ifelse(Education == 1,"HS Grad", ifelse(Education == 2, "Some College",   
ifelse(Education == 3, "College Grad", ifelse(Education == 4, "Some Post Grad",  
"Post Grad"))))),  
   
EnvironmentSatisfaction = as.factor #4 point Likert, best estimate of values  
(ifelse(EnvironmentSatisfaction == 1, "Low",ifelse(EnvironmentSatisfaction == 2, "Med",   
ifelse(EnvironmentSatisfaction == 3, "High", "Very High")))),  
  
JobInvolvement = as.factor #4 point Likert, best estimate of values  
(ifelse(JobInvolvement == 1, "Low",ifelse(JobInvolvement == 2, "Med",  
ifelse(JobInvolvement == 3, "High","Very High")))),  
  
JobLevel = as.factor #5 point Scale estimate of values  
(ifelse(JobLevel == 1, "Contributor",ifelse(JobLevel == 2, "Supervisor",  
ifelse(JobLevel == 3, "Manager",ifelse(JobLevel == 4, "Director",  
"Co Leadership"))))),  
  
JobSatisfaction = as.factor #4 point Likert, best estimate of values  
(ifelse(JobSatisfaction == 1, "Low",ifelse(JobSatisfaction == 2, "Med",  
ifelse(JobSatisfaction == 3, "High","Very High")))),  
  
PerformanceRating = as.factor #4 point Likert, best estimate of values  
(ifelse(PerformanceRating == 1, "Low",ifelse(PerformanceRating == 2, "Med",  
ifelse(PerformanceRating == 3, "High", "Very High")))),  
  
RelationshipSatisfaction = as.factor #4 point Likert, best estimate of values  
(ifelse(RelationshipSatisfaction == 1, "Low",ifelse(RelationshipSatisfaction == 2, "Med",  
ifelse(RelationshipSatisfaction == 3, "High", "Very High")))),  
  
WorkLifeBalance = as.factor #4 point Likert, best estimate of values  
(ifelse(WorkLifeBalance == 1, "Low",ifelse(WorkLifeBalance == 2, "Med",  
ifelse(WorkLifeBalance == 3, "High", "Very High"))))  
)

In the demographic data, it is as expected

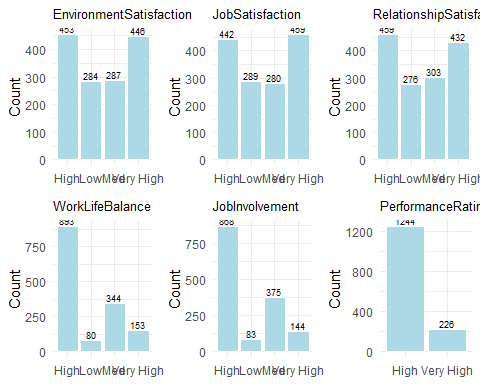
#Demographics  
  
#Plot Counts for High Level View  
#Demographics: Age, Gender, Marital Status, Education, Education Level and Distance from home (commute)  
  
AgePlot <- plotalldata %>% group\_by(Age) %>% summarise(counts = n()) %>%  
 ggplot(aes(x = as.factor(Age), y = counts)) +  
 geom\_bar(stat = 'identity', fill = "light blue",col = "white") +  
 ggtitle("Age") + ylab("Count") + theme\_minimal() +   
 geom\_text(aes(label=counts), size = 2.5, position=position\_dodge(width=0.3), vjust=-0.25) +  
 theme(plot.title = element\_text(size =10),axis.title.x=element\_blank())   
  
GenderPlot <- plotalldata %>% group\_by(Gender) %>% summarise(counts = n()) %>%  
 ggplot(aes(x = as.factor(Gender), y = counts)) +  
 geom\_bar(stat = 'identity', fill = "light blue",col = "white") +  
 ggtitle("Gender") + ylab("Count") + theme\_minimal() +   
 geom\_text(aes(label=counts), size = 2.5, position=position\_dodge(width=0.3), vjust=-0.25) +  
 theme(plot.title = element\_text(size =10),axis.title.x=element\_blank())   
  
MaritalStatusPlot <- plotalldata %>% group\_by(MaritalStatus) %>% summarise(counts = n()) %>%  
 ggplot(aes(x = as.factor(MaritalStatus), y = counts)) +  
 geom\_bar(stat = 'identity', fill = "light blue",col = "white") +  
 ggtitle("MaritalStatus") + ylab("Count") + theme\_minimal() +   
 geom\_text(aes(label=counts), size = 2.5, position=position\_dodge(width=0.3), vjust=-0.25) +  
 theme(plot.title = element\_text(size =10),axis.title.x=element\_blank())   
  
EducationPlot <- plotalldata %>% group\_by(Education) %>% summarise(counts = n()) %>%  
 ggplot(aes(x = as.factor(Education), y = counts)) +  
 geom\_bar(stat = 'identity', fill = "light blue",col = "white") +  
 ggtitle("Education") + ylab("Count") + theme\_minimal() +   
 geom\_text(aes(label=counts), size = 2.5, position=position\_dodge(width=0.3), vjust=-0.25) +  
 theme(plot.title = element\_text(size =10),axis.title.x=element\_blank()) +  
 theme(axis.text.x=element\_text(angle=70,vjust=.7,hjust=1))  
  
EducationFieldPlot <- plotalldata %>% group\_by(EducationField) %>% summarise(counts = n()) %>%  
 ggplot(aes(x = as.factor(EducationField), y = counts)) +  
 geom\_bar(stat = 'identity', fill = "light blue",col = "white") +  
 ggtitle("EducationField") + ylab("Count") + theme\_minimal() +   
 geom\_text(aes(label=counts), size = 2.5, position=position\_dodge(width=0.3), vjust=-0.25) +  
 theme(plot.title = element\_text(size =10),axis.title.x=element\_blank()) +  
 theme(axis.text.x=element\_text(angle=70,vjust=.7,hjust=1))  
  
CommutePlot <- plotalldata %>% group\_by(DistanceFromHome) %>% summarise(counts = n()) %>%  
 ggplot(aes(x = as.factor(DistanceFromHome), y = counts)) +  
 geom\_bar(stat = 'identity', fill = "light blue",col = "white") +  
 ggtitle("Commute") + ylab("Count") + theme\_minimal() +   
 geom\_text(aes(label=counts), size = 2.5, position=position\_dodge(width=0.3), vjust=-0.25) +  
 theme(plot.title = element\_text(size =10),axis.title.x=element\_blank()) +  
 theme(axis.text.x=element\_text(angle=70,vjust=.8,hjust=1))  
  
grid.arrange(AgePlot, GenderPlot, MaritalStatusPlot, EducationPlot, EducationFieldPlot,CommutePlot, nrow = 2, ncol = 3)



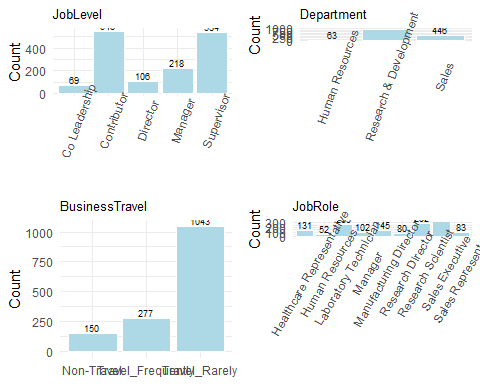
In the demographic data, the JobPerformanceRating is unusual, there are only employees marked “high” and “very high”. This warrants follow up, as there is an expected relationship between low performers and attrition, the absences of low performers makes the model less reliable.

The Satisfaction ratings are also left skewed, which means there may be an issue with how the satisfaction numbers are measured, there could be evidence that employees may feel pressured to score high, or again the low performing employees could be not in the sample.

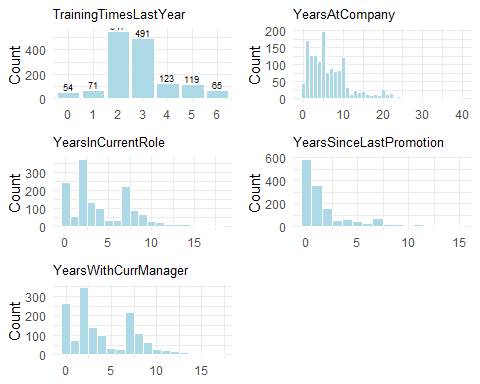
#Attitudinal  
  
#Plot Counts for High Level View  
#Attitudinal: Environmental Satisfaction, Job Satisfaction, Relationship Satisfaction, WorkLifeBalance,Job Involvement, Performance Rating   
  
EnvironSatPlot <- plotalldata %>% group\_by(EnvironmentSatisfaction) %>% summarise(counts = n()) %>%  
 ggplot(aes(x = as.factor(EnvironmentSatisfaction), y = counts)) +  
 geom\_bar(stat = 'identity', fill = "light blue",col = "white") +  
 ggtitle("EnvironmentSatisfaction") + ylab("Count") + theme\_minimal() +   
 geom\_text(aes(label=counts), size = 2.5, position=position\_dodge(width=0.3), vjust=-0.25) +  
 theme(plot.title = element\_text(size =10),axis.title.x=element\_blank())   
  
JobSatPlot <- plotalldata %>% group\_by(JobSatisfaction) %>% summarise(counts = n()) %>%  
 ggplot(aes(x = as.factor(JobSatisfaction), y = counts)) +  
 geom\_bar(stat = 'identity', fill = "light blue",col = "white") +  
 ggtitle("JobSatisfaction") + ylab("Count") + theme\_minimal() +   
 geom\_text(aes(label=counts), size = 2.5, position=position\_dodge(width=0.3), vjust=-0.25) +  
 theme(plot.title = element\_text(size =10),axis.title.x=element\_blank())   
  
RelatSatPlot <- plotalldata %>% group\_by(RelationshipSatisfaction) %>% summarise(counts = n()) %>%  
 ggplot(aes(x = as.factor(RelationshipSatisfaction), y = counts)) +  
 geom\_bar(stat = 'identity', fill = "light blue",col = "white") +  
 ggtitle("RelationshipSatisfaction") + ylab("Count") + theme\_minimal() +   
 geom\_text(aes(label=counts), size = 2.5, position=position\_dodge(width=0.3), vjust=-0.25) +  
 theme(plot.title = element\_text(size =10),axis.title.x=element\_blank())   
  
WoLiBaPlot <- plotalldata %>% group\_by(WorkLifeBalance) %>% summarise(counts = n()) %>%  
 ggplot(aes(x = as.factor(WorkLifeBalance), y = counts)) +  
 geom\_bar(stat = 'identity', fill = "light blue",col = "white") +  
 ggtitle("WorkLifeBalance") + ylab("Count") + theme\_minimal() +   
 geom\_text(aes(label=counts), size = 2.5, position=position\_dodge(width=0.3), vjust=-0.25) +  
 theme(plot.title = element\_text(size =10),axis.title.x=element\_blank())   
  
JobInvPlot <- plotalldata %>% group\_by(JobInvolvement) %>% summarise(counts = n()) %>%  
 ggplot(aes(x = as.factor(JobInvolvement), y = counts)) +  
 geom\_bar(stat = 'identity', fill = "light blue",col = "white") +  
 ggtitle("JobInvolvement") + ylab("Count") + theme\_minimal() +   
 geom\_text(aes(label=counts), size = 2.5, position=position\_dodge(width=0.3), vjust=-0.25) +  
 theme(plot.title = element\_text(size =10),axis.title.x=element\_blank())   
   
PerfRatingPlot <- plotalldata %>% group\_by(PerformanceRating) %>% summarise(counts = n()) %>%  
 ggplot(aes(x = as.factor(PerformanceRating), y = counts)) +  
 geom\_bar(stat = 'identity', fill = "light blue",col = "white") +  
 ggtitle("PerformanceRating") + ylab("Count") + theme\_minimal() +   
 geom\_text(aes(label=counts), size = 2.5, position=position\_dodge(width=0.3), vjust=-0.25) +  
 theme(plot.title = element\_text(size =10),axis.title.x=element\_blank())   
  
grid.arrange(EnvironSatPlot, JobSatPlot, RelatSatPlot, WoLiBaPlot, JobInvPlot, PerfRatingPlot, nrow = 2, ncol = 3)



#Plot Counts for High Level View  
#Work Demographics: Job ROle, Job Level, Business Travel, Department, Training   
#distribution of work demo details. Expect Right tail.  
  
JobLevelPlot <- plotalldata %>% group\_by(JobLevel) %>% summarise(counts = n()) %>%  
 ggplot(aes(x = as.factor(JobLevel), y = counts)) +  
 geom\_bar(stat = 'identity', fill = "light blue",col = "white") +  
 ggtitle("JobLevel") + ylab("Count") + theme\_minimal() +   
 geom\_text(aes(label=counts), size = 2.5, position=position\_dodge(width=0.3), vjust=-0.25) +  
 theme(plot.title = element\_text(size =10),axis.title.x=element\_blank()) +  
 theme(axis.text.x=element\_text(angle=70,vjust=.8))  
  
JobRolePlot <- plotalldata %>% group\_by(JobRole) %>% summarise(counts = n()) %>%  
 ggplot(aes(x = as.factor(JobRole), y = counts)) +  
 geom\_bar(stat = 'identity', fill = "light blue",col = "white") +  
 ggtitle("JobRole") + ylab("Count") + theme\_minimal() +   
 geom\_text(aes(label=counts), size = 2.5, position=position\_dodge(width=0.3), vjust=-0.25) +  
 theme(plot.title = element\_text(size =10),axis.title.x=element\_blank()) +  
 theme(axis.text.x=element\_text(angle=60,vjust=.8,hjust=.5))   
  
BusinessTravelPlot <- plotalldata %>% group\_by(BusinessTravel) %>% summarise(counts = n()) %>%  
 ggplot(aes(x = as.factor(BusinessTravel), y = counts)) +  
 geom\_bar(stat = 'identity', fill = "light blue",col = "white") +  
 ggtitle("BusinessTravel") + ylab("Count") + theme\_minimal() +   
 geom\_text(aes(label=counts), size = 2.5, position=position\_dodge(width=0.3), vjust=-0.25) +  
 theme(plot.title = element\_text(size =10),axis.title.x=element\_blank())   
  
DepartmentPlot <- plotalldata %>% group\_by(Department) %>% summarise(counts = n()) %>%  
 ggplot(aes(x = as.factor(Department), y = counts)) +  
 geom\_bar(stat = 'identity', fill = "light blue",col = "white") +  
 ggtitle("Department") + ylab("Count") + theme\_minimal() +   
 geom\_text(aes(label=counts), size = 2.5, position=position\_dodge(width=0.3), vjust=-0.25) +  
 theme(plot.title = element\_text(size =10),axis.title.x=element\_blank()) +  
 theme(axis.text.x=element\_text(angle=70,vjust=.8,hjust=.5))  
  
TrainingPlot <- plotalldata %>% group\_by(TrainingTimesLastYear) %>% summarise(counts = n()) %>%  
 ggplot(aes(x = as.factor(TrainingTimesLastYear), y = counts)) +  
 geom\_bar(stat = 'identity', fill = "light blue",col = "white") +  
 ggtitle("TrainingTimesLastYear") + ylab("Count") + theme\_minimal() +   
 geom\_text(aes(label=counts), size = 2.5, position=position\_dodge(width=0.3), vjust=-0.25) +  
 theme(plot.title = element\_text(size =10),axis.title.x=element\_blank())  
  
YearsAtCompanyPlot <- ggplot(plotalldata) +   
 geom\_histogram(aes(YearsAtCompany), binwidth = 1, fill = "light blue",col = "white") +   
 ggtitle("YearsAtCompany") + ylab("Count") + theme\_minimal()+   
 theme(plot.title = element\_text(size =10),axis.title.x=element\_blank()) # add main & axis titles  
  
YearsInCurrentRolePlot <- ggplot(plotalldata) +   
 geom\_histogram(aes(YearsInCurrentRole), binwidth = 1, fill = "light blue",col = "white") +   
 ggtitle("YearsInCurrentRole") + ylab("Count") + theme\_minimal() + # add main & axis titles  
 theme(plot.title = element\_text(size =10),axis.title.x=element\_blank())   
   
YearsSinceLastPromotionPlot <- ggplot(plotalldata) +   
 geom\_histogram(aes(YearsSinceLastPromotion), binwidth = 1, fill = "light blue",col = "white") +   
 ggtitle("YearsSinceLastPromotion")+ ylab("Count") + theme\_minimal() + # add main & axis titles  
 theme(plot.title = element\_text(size =10),axis.title.x=element\_blank())   
  
YearsWithCurrManagerPlot<- ggplot(plotalldata) +   
 geom\_histogram(aes(YearsWithCurrManager), binwidth = 1, fill = "light blue",col = "white") +   
 ggtitle("YearsWithCurrManager") + ylab("Count") + theme\_minimal() + # add main & axis titles  
 theme(plot.title = element\_text(size =10),axis.title.x=element\_blank())   
  
grid.arrange(JobLevelPlot, DepartmentPlot, BusinessTravelPlot, JobRolePlot, nrow = 2, ncol = 2)

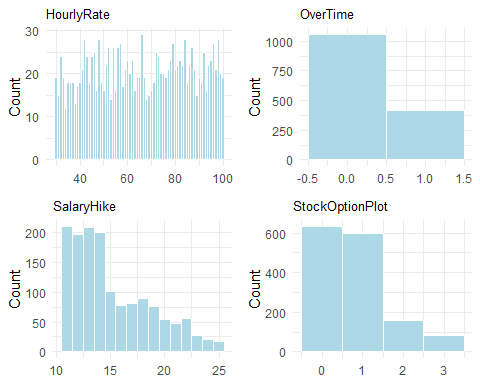


grid.arrange(TrainingPlot, YearsAtCompanyPlot, YearsInCurrentRolePlot, YearsSinceLastPromotionPlot,YearsWithCurrManagerPlot, nrow = 3, ncol = 2)



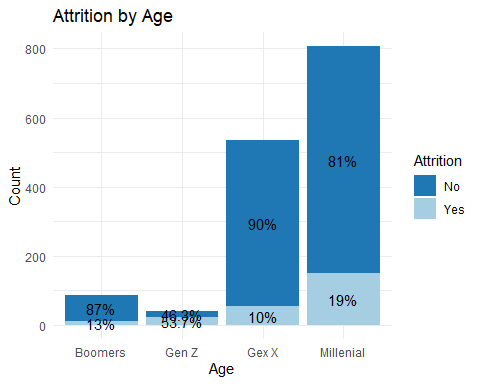
Financial Plots

HourlyRatePlot <- ggplot(plotalldata) +   
 geom\_histogram(aes(HourlyRate), binwidth = 1, fill = "light blue",col = "white") +   
 ggtitle("HourlyRate") + ylab("Count") + theme\_minimal()+   
 theme(plot.title = element\_text(size =10),axis.title.x=element\_blank())   
  
OT <- as.data.frame(ifelse(plotalldata [,22] == "Yes", 1, ifelse(plotalldata[,22] == "No", 0, 99))) # Recode OT with 0s and 1s  
names(OT)<-c("OverTime")  
Overtimeplot <- ggplot(OT) +   
 geom\_histogram(aes(OverTime), binwidth = 1, fill = "light blue",col = "white") +   
 ggtitle("OverTime") + ylab("Count") + theme\_minimal()+   
 theme(plot.title = element\_text(size =10),axis.title.x=element\_blank())   
  
PercentSalaryHikePlot <- ggplot(plotalldata) +   
 geom\_histogram(aes(PercentSalaryHike), binwidth = 1, fill = "light blue",col = "white") +   
 ggtitle("SalaryHike") + ylab("Count") + theme\_minimal()+   
 theme(plot.title = element\_text(size =10),axis.title.x=element\_blank())  
  
StockOptionPlot <- ggplot(plotalldata) +   
 geom\_histogram(aes(StockOptionLevel), binwidth = 1, fill = "light blue",col = "white") +   
 ggtitle("StockOptionPlot") + ylab("Count") + theme\_minimal() +   
 theme(plot.title = element\_text(size =10),axis.title.x=element\_blank())   
grid.arrange( HourlyRatePlot, Overtimeplot, PercentSalaryHikePlot, StockOptionPlot, nrow = 2, ncol = 2)

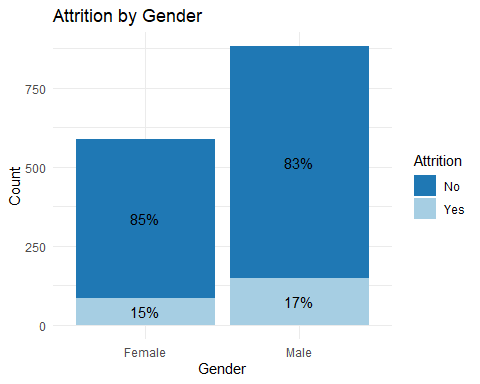


Demographic Attrition Plots

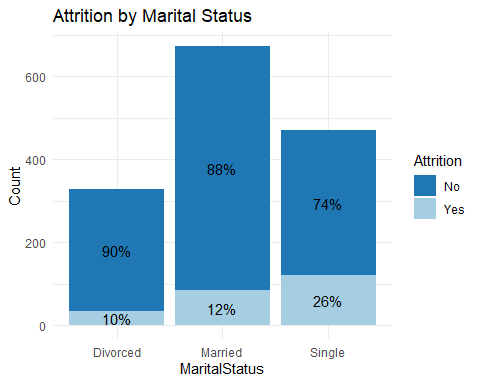
# Attrition by Age  
percentData <- plotalldata %>% group\_by(Age) %>% count(Attrition) %>%  
 mutate(ratio=scales::percent(n/sum(n)))  
ggplot(plotalldata,aes(x=Age,fill=Attrition)) +  
 ggtitle("Attrition by Age") + ylab("Count") + # add main & axis titles  
 scale\_fill\_brewer(palette="Paired",direction=-1)+ theme\_minimal() + # change color palette  
 geom\_bar(aes(y = (..count..)),position=position\_dodge()) + # side-by-side bars  
 geom\_bar(aes(y = (..count..))) +   
 geom\_text(data=percentData,aes(y=n,label=ratio),position=position\_stack(vjust=.5)) # % labels



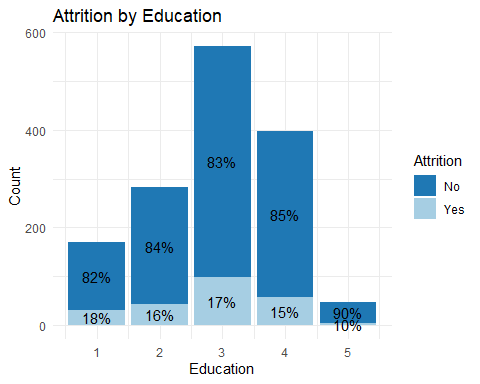
# Attrition by Gender  
percentData <- alldata %>% group\_by(Gender) %>% count(Attrition) %>%  
 mutate(ratio=scales::percent(n/sum(n)))  
ggplot(alldata,aes(x=Gender,fill=Attrition)) +  
 ggtitle("Attrition by Gender") + ylab("Count") + # add main & axis titles  
 scale\_fill\_brewer(palette="Paired",direction=-1)+ theme\_minimal() + # change color palette  
 geom\_bar(aes(y = (..count..)),position=position\_dodge()) + # side-by-side bars  
 geom\_bar(aes(y = (..count..))) +   
 geom\_text(data=percentData,aes(y=n,label=ratio),position=position\_stack(vjust=.5)) # % labels



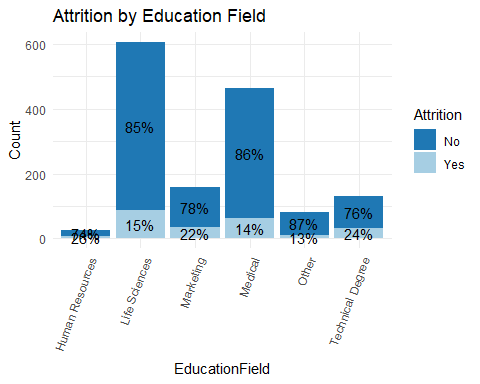
# Attrition by Marital Status  
percentData <- alldata %>% group\_by(MaritalStatus) %>% count(Attrition) %>%  
 mutate(ratio=scales::percent(n/sum(n)))  
ggplot(alldata,aes(x=MaritalStatus,fill=Attrition)) +  
 ggtitle("Attrition by Marital Status") + ylab("Count") + # add main & axis titles  
 scale\_fill\_brewer(palette="Paired",direction=-1)+ theme\_minimal() + # change color palette  
 geom\_bar(aes(y = (..count..)),position=position\_dodge()) + # side-by-side bars  
 geom\_bar(aes(y = (..count..))) +   
 geom\_text(data=percentData,aes(y=n,label=ratio),position=position\_stack(vjust=.5)) # % labels



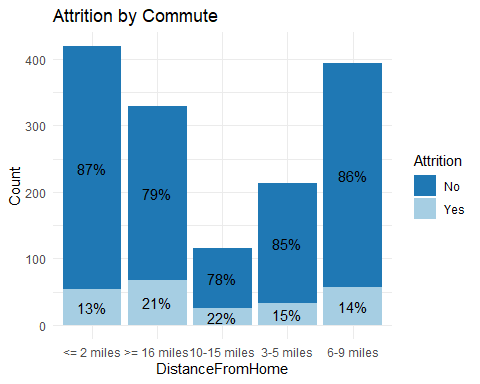
# Attrition by Education  
percentData <- alldata %>% group\_by(Education) %>% count(Attrition) %>%  
 mutate(ratio=scales::percent(n/sum(n)))  
ggplot(alldata,aes(x=Education,fill=Attrition)) +  
 ggtitle("Attrition by Education") + ylab("Count") + # add main & axis titles  
 scale\_fill\_brewer(palette="Paired",direction=-1)+ theme\_minimal() + # change color palette  
 geom\_bar(aes(y = (..count..)),position=position\_dodge()) + # side-by-side bars  
 geom\_bar(aes(y = (..count..))) +   
 geom\_text(data=percentData,aes(y=n,label=ratio),position=position\_stack(vjust=.5)) # % labels



# Attrition by Education Field  
percentData <- alldata %>% group\_by(EducationField) %>% count(Attrition) %>%  
 mutate(ratio=scales::percent(n/sum(n)))  
ggplot(alldata,aes(x=EducationField,fill=Attrition)) +  
 ggtitle("Attrition by Education Field") + ylab("Count") + # add main & axis titles  
 scale\_fill\_brewer(palette="Paired",direction=-1)+ theme\_minimal() + # change color palette  
 geom\_bar(aes(y = (..count..)),position=position\_dodge()) + # side-by-side bars  
 geom\_bar(aes(y = (..count..))) +   
 geom\_text(data=percentData,aes(y=n,label=ratio),position=position\_stack(vjust=.5)) + # % labels  
 theme(axis.text.x=element\_text(angle=70,hjust=1))

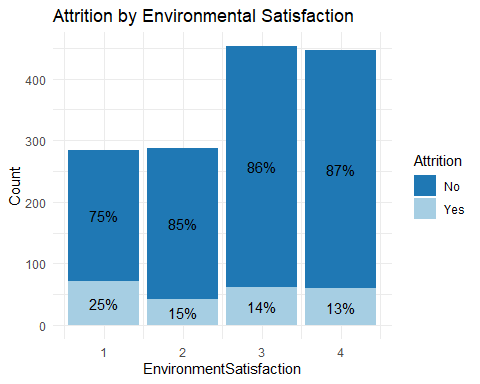


# Attrition by Commute  
percentData <- plotalldata %>% group\_by(DistanceFromHome) %>% count(Attrition) %>%  
 mutate(ratio=scales::percent(n/sum(n)))  
ggplot(plotalldata,aes(x=DistanceFromHome,fill=Attrition)) +  
 ggtitle("Attrition by Commute") + ylab("Count") + # add main & axis titles  
 scale\_fill\_brewer(palette="Paired",direction=-1)+ theme\_minimal() + # change color palette  
 geom\_bar(aes(y = (..count..)),position=position\_dodge()) + # side-by-side bars  
 geom\_bar(aes(y = (..count..))) +   
 geom\_text(data=percentData,aes(y=n,label=ratio),position=position\_stack(vjust=.5)) # % labels

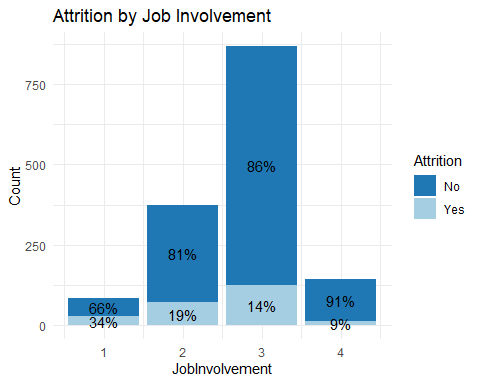


Attitudinal Attrition Plots

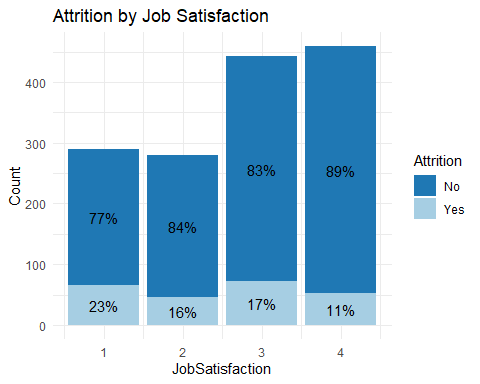
# Attrition by Environmental Satisfaction  
percentData <- alldata %>% group\_by(EnvironmentSatisfaction) %>% count(Attrition) %>%  
 mutate(ratio=scales::percent(n/sum(n)))  
ggplot(alldata,aes(x=EnvironmentSatisfaction,fill=Attrition)) +  
 ggtitle("Attrition by Environmental Satisfaction") + ylab("Count") + # add main & axis titles  
 scale\_fill\_brewer(palette="Paired",direction=-1)+ theme\_minimal() + # change color palette  
 geom\_bar(aes(y = (..count..)),position=position\_dodge()) + # side-by-side bars  
 geom\_bar(aes(y = (..count..))) +   
 geom\_text(data=percentData,aes(y=n,label=ratio),position=position\_stack(vjust=.5)) # % labels



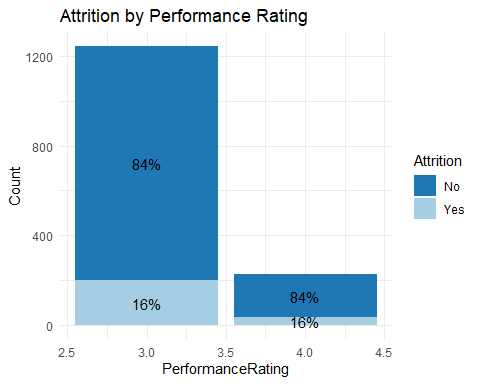
# Attrition by Job Involvement  
percentData <- alldata %>% group\_by(JobInvolvement) %>% count(Attrition) %>%  
 mutate(ratio=scales::percent(n/sum(n)))  
ggplot(alldata,aes(x=JobInvolvement,fill=Attrition)) +  
 ggtitle("Attrition by Job Involvement") + ylab("Count") + # add main & axis titles  
 scale\_fill\_brewer(palette="Paired",direction=-1)+ theme\_minimal() + # change color palette  
 geom\_bar(aes(y = (..count..)),position=position\_dodge()) + # side-by-side bars  
 geom\_bar(aes(y = (..count..))) +   
 geom\_text(data=percentData,aes(y=n,label=ratio),position=position\_stack(vjust=.5)) # % labels



# Attrition by Job Satisfaction  
percentData <- alldata %>% group\_by(JobSatisfaction) %>% count(Attrition) %>%  
 mutate(ratio=scales::percent(n/sum(n)))  
ggplot(alldata,aes(x=JobSatisfaction,fill=Attrition)) +  
 ggtitle("Attrition by Job Satisfaction") + ylab("Count") + # add main & axis titles  
 scale\_fill\_brewer(palette="Paired",direction=-1)+ theme\_minimal() + # change color palette  
 geom\_bar(aes(y = (..count..)),position=position\_dodge()) + # side-by-side bars  
 geom\_bar(aes(y = (..count..))) +   
 geom\_text(data=percentData,aes(y=n,label=ratio),position=position\_stack(vjust=.5)) # % labels

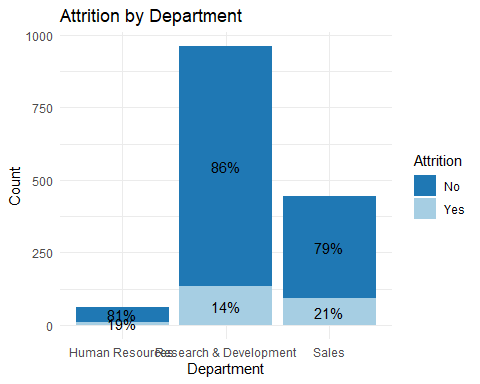


# Attrition by Performance Rating  
percentData <- alldata %>% group\_by(PerformanceRating) %>% count(Attrition) %>%  
 mutate(ratio=scales::percent(n/sum(n)))  
ggplot(alldata,aes(x=PerformanceRating,fill=Attrition)) +  
 ggtitle("Attrition by Performance Rating") + ylab("Count") + # add main & axis titles  
 scale\_fill\_brewer(palette="Paired",direction=-1)+ theme\_minimal() + # change color palette  
 geom\_bar(aes(y = (..count..)),position=position\_dodge()) + # side-by-side bars  
 geom\_bar(aes(y = (..count..))) +   
 geom\_text(data=percentData,aes(y=n,label=ratio),position=position\_stack(vjust=.5)) # % labels

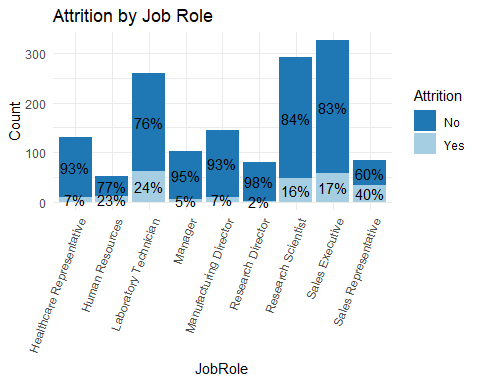


Work Attrition Plots

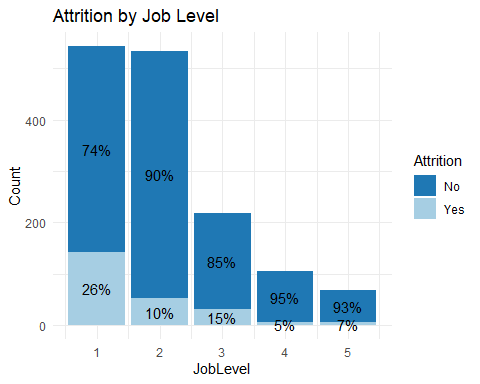
par(mfrow=c(1,3))  
# Attrition by Department  
percentData <- alldata %>% group\_by(Department) %>% count(Attrition) %>%  
 mutate(ratio=scales::percent(n/sum(n)))  
ggplot(alldata,aes(x=Department,fill=Attrition)) +  
 ggtitle("Attrition by Department") + ylab("Count") + # add main & axis titles  
 scale\_fill\_brewer(palette="Paired",direction=-1)+ theme\_minimal() + # change color palette  
 geom\_bar(aes(y = (..count..)),position=position\_dodge()) + # side-by-side bars  
 geom\_bar(aes(y = (..count..))) +   
 geom\_text(data=percentData,aes(y=n,label=ratio),position=position\_stack(vjust=.5)) # % labels



# Attrition by Job Role  
percentData <- alldata %>% group\_by(JobRole) %>% count(Attrition) %>%  
 mutate(ratio=scales::percent(n/sum(n)))  
ggplot(alldata,aes(x=JobRole,fill=Attrition)) +  
 ggtitle("Attrition by Job Role") + ylab("Count") + # add main & axis titles  
 scale\_fill\_brewer(palette="Paired",direction=-1)+ theme\_minimal() + # change color palette  
 geom\_bar(aes(y = (..count..)),position=position\_dodge()) + # side-by-side bars  
 geom\_bar(aes(y = (..count..))) +   
 geom\_text(data=percentData,aes(y=n,label=ratio),position=position\_stack(vjust=.5)) + # % labels  
 theme(axis.text.x=element\_text(angle=70,hjust=1))



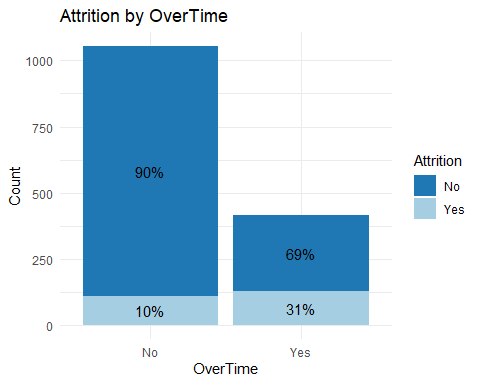
# Attrition by Job Level  
percentData <- alldata %>% group\_by(JobLevel) %>% count(Attrition) %>%  
 mutate(ratio=scales::percent(n/sum(n)))  
ggplot(alldata,aes(x=JobLevel,fill=Attrition)) +  
 ggtitle("Attrition by Job Level") + ylab("Count") + # add main & axis titles  
 scale\_fill\_brewer(palette="Paired",direction=-1)+ theme\_minimal() + # change color palette  
 geom\_bar(aes(y = (..count..)),position=position\_dodge()) + # side-by-side bars  
 geom\_bar(aes(y = (..count..))) +   
 geom\_text(data=percentData,aes(y=n,label=ratio),position=position\_stack(vjust=.5)) # % labels



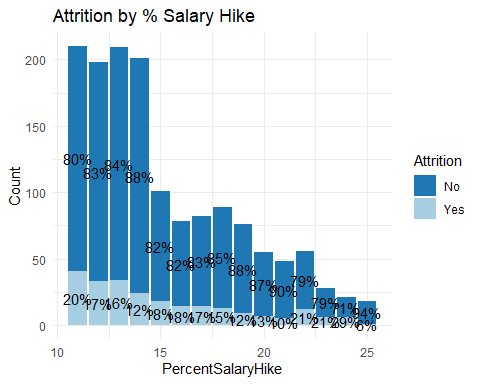
# Attrition by Travel  
percentData <- alldata %>% group\_by(BusinessTravel) %>% count(Attrition) %>%  
 mutate(ratio=scales::percent(n/sum(n)))  
aTravel<-ggplot(alldata,aes(x=BusinessTravel,fill=Attrition)) +  
 ggtitle("Attrition by Business Travel") + ylab("Count") + # add main & axis titles  
 scale\_fill\_brewer(palette="Paired",direction=-1)+ theme\_minimal() + # change color palette  
 geom\_bar(aes(y = (..count..))) +   
 geom\_text(data=percentData,aes(y=n,label=ratio),position=position\_stack(vjust=.5)) # % labels

Financial Attrition Plots

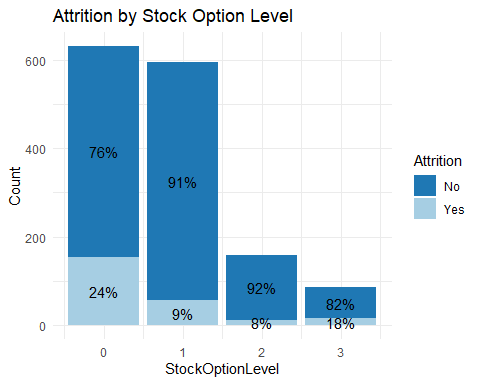
# Attrition by Overtime  
percentData <- alldata %>% group\_by(OverTime) %>% count(Attrition) %>%  
 mutate(ratio=scales::percent(n/sum(n)))  
ggplot(alldata,aes(x=OverTime,fill=Attrition)) +  
 ggtitle("Attrition by OverTime") + ylab("Count") + # add main & axis titles  
 scale\_fill\_brewer(palette="Paired",direction=-1)+ theme\_minimal() + # change color palette  
 geom\_bar(aes(y = (..count..)),position=position\_dodge()) + # side-by-side bars  
 geom\_bar(aes(y = (..count..))) +   
 geom\_text(data=percentData,aes(y=n,label=ratio),position=position\_stack(vjust=.5)) # % labels



# Attrition by Percent Salary Hike  
percentData <- alldata %>% group\_by(PercentSalaryHike) %>% count(Attrition) %>%  
 mutate(ratio=scales::percent(n/sum(n)))  
ggplot(alldata,aes(x=PercentSalaryHike,fill=Attrition)) +  
 ggtitle("Attrition by % Salary Hike") + ylab("Count") + # add main & axis titles  
 scale\_fill\_brewer(palette="Paired",direction=-1)+ theme\_minimal() + # change color palette  
 geom\_bar(aes(y = (..count..)),position=position\_dodge()) + # side-by-side bars  
 geom\_bar(aes(y = (..count..))) +   
 geom\_text(data=percentData,aes(y=n,label=ratio),position=position\_stack(vjust=.5)) # % labels



# Attrition by Percent Salary Hike  
percentData <- alldata %>% group\_by(StockOptionLevel) %>% count(Attrition) %>%  
 mutate(ratio=scales::percent(n/sum(n)))  
ggplot(alldata,aes(x=StockOptionLevel,fill=Attrition)) +  
 ggtitle("Attrition by Stock Option Level") + ylab("Count") + # add main & axis titles  
 scale\_fill\_brewer(palette="Paired",direction=-1)+ theme\_minimal() + # change color palette  
 geom\_bar(aes(y = (..count..)),position=position\_dodge()) + # side-by-side bars  
 geom\_bar(aes(y = (..count..))) +   
 geom\_text(data=percentData,aes(y=n,label=ratio),position=position\_stack(vjust=.5)) # % labels

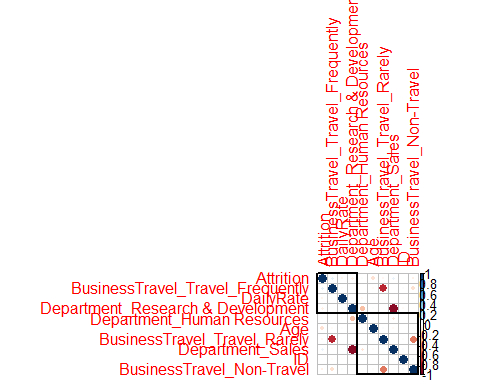


Feature Analysis

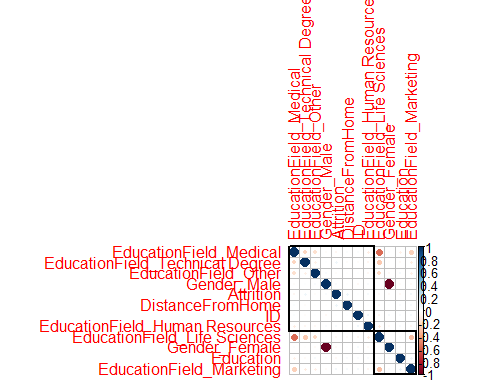
Correlation plots provide a quick way to identify correlation between variables. You want correlation with Attrition but you want to be careful about including variables that are collinear with each other since they can negatively affect the model. Selecting the optimal features requires an iteration between initial EDA, review of collinearity plots, variable inflation factors, variable p-values, model output.

Here we show collinearity plots. We recoded Attrition and used dummy variables for the categorical data and applied the corrplot function (since it uses only numerical data). In order to see the plots clearer, we split the data into 5 different different plots but keeping Attrition in each plot.

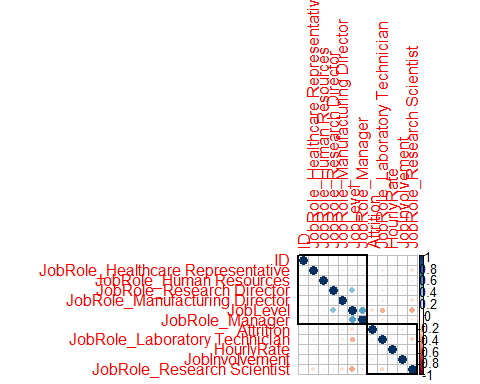
# Split data to create 2 correlation matrices, more readable  
data<-alldata[,-c(10,11,23,28,37)] # remove EmpCount, Emp #, Over18, StdHrs,Rand  
data$Attrition <- ifelse(data [,3] == "Yes", 1, ifelse(data[,3] == "No", 0, 99)) # Recode Attrition with 0s and 1s  
data <- one\_hot(as.data.table(data)) # Create dummy vars for categories  
  
# Looking for linear relationships so will not normaalize  
#data\_n <- lapply(data,normalize) # normalize numeric data  
#data <- data.frame(data\_n) # create dataframe of normalized data  
  
# create sets of data for corr plot viewing  
data1<-data[,c(1:10)]  
data2<-data[,c(1,3,11:20)]  
data3<-data[,c(1,3,21:30)]  
data4<-data[,c(1,3,31:41)]  
data5<-data[,c(1,3,42:50)]  
  
M\_df1 <- select\_if(data1, is.numeric) # get all numeric data  
M1 <- cor(M\_df1) # convert to matrix  
corrplot(M1, order="hclust",addrect=2) # Display the correlation coefficient



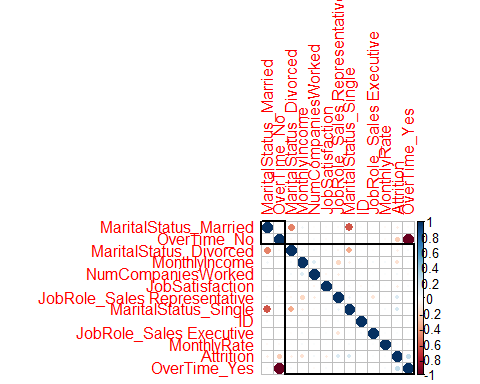
M\_df2 <- select\_if(data2, is.numeric) # get all numeric data  
M2 <- cor(M\_df2) # convert to matrix  
corrplot(M2, order="hclust",addrect=2) # Display the correlation coefficient



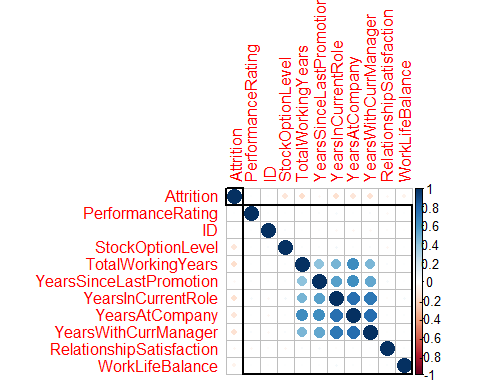
M\_df3 <- select\_if(data3, is.numeric) # get all numeric data  
M3 <- cor(M\_df3) # convert to matrix  
corrplot(M3, order="hclust",addrect=2) # Display the correlation coefficient



M\_df4 <- select\_if(data4, is.numeric) # get all numeric data  
M4 <- cor(M\_df4) # convert to matrix  
corrplot(M4, order="hclust",addrect=2) # Display the correlation coefficient



M\_df5 <- select\_if(data5, is.numeric) # get all numeric data  
M5 <- cor(M\_df5) # convert to matrix  
corrplot(M5, order="hclust",addrect=2) # Display the correlation coefficient



Data Preparation

We chose the KNN model due to it’s non-parametric technique and it provides a classification output which matches our overall goal with predicting attrition. The algorithm assumes that similar things exist in close proximity. Because of this assumption, the data must be normalized for proximity determination.

# function to normalize data for KNN  
normalize <- function(x) {  
 y <- (x-min(x))/(max(x)-min(x))  
 y  
}  
  
# Create dummy vars for categories  
train.dummies <- one\_hot(as.data.table(train))  
val.dummies <- one\_hot(as.data.table(validation))  
  
# Normalize the data  
train\_n <- lapply(train.dummies[,c(2,9,10,11,12,22,23,24,25,26,27,28,29,30,31,32,33,34,35,36,37,38,39,40,41,42,43,45,46,47,48,49,51,52,53,54,55,56,57,58)],normalize)   
val\_n <- lapply(val.dummies[,c(2,9,10,11,12,22,23,24,25,26,27,28,29,30,31,32,33,34,35,36,37,38,39,40,41,42,43,45,46,47,48,49,51,52,53,54,55,56,57,58)],normalize)  
  
dfTrain <- data.frame(train\_n) # create dataframe of normalized data  
dfVal <- data.frame(val\_n)  
  
rownames(dfTrain) <- train$ID # add IDs  
rownames(dfVal) <- validation$ID  
  
train.attrition <- train$Attrition # isolate attrition  
val.attrition <- validation$Attrition  
  
names(train.attrition) <- train$ID # add IDs  
names(val.attrition) <- validation$ID  
  
#dfTrain[1:4] # check normalized data  
#dfVal[1:4]  
  
valIDs <- rownames(dfVal) # IDs used below  
  
print("Check Training & Validation Attrition Data")

## [1] "Check Training & Validation Attrition Data"

train.attrition[1:10] # check data

## 1 2 3 4 5 6 7 8 9 10   
## Yes No No No No Yes No No Yes No   
## Levels: No Yes

val.attrition[1:10]

## 1171 1172 1173 1174 1175 1176 1177 1178 1179 1180   
## No No No No No No No No No Yes   
## Levels: No Yes

Execution of Model and Results

The KNN algorithm allows you to select the number of nearest neighbors for classification. Our goal was not only to obtain a high accuracy rate but achieve the best specificity as well. Our goal was to achieve greater than 80% accuracy and 50% for specificity. Using a k=3 and the selected features, were able to meet those goals. Our accuracy is 84.7% having a 95% confidence interval of (80%, 88%) with specificity of 57%.

k<-3  
predKNN <- knn(dfTrain,dfVal,cl=train.attrition,k) # run knn model  
dfPred <- data.frame(predKNN) # store predictions  
confusionMatrix(table(val.attrition,predKNN)) # confusion matrix

## Confusion Matrix and Statistics  
##   
## predKNN  
## val.attrition No Yes  
## No 242 9  
## Yes 37 12  
##   
## Accuracy : 0.8467   
## 95% CI : (0.8008, 0.8855)  
## No Information Rate : 0.93   
## P-Value [Acc > NIR] : 1   
##   
## Kappa : 0.2715   
##   
## Mcnemar's Test P-Value : 6.865e-05   
##   
## Sensitivity : 0.8674   
## Specificity : 0.5714   
## Pos Pred Value : 0.9641   
## Neg Pred Value : 0.2449   
## Prevalence : 0.9300   
## Detection Rate : 0.8067   
## Detection Prevalence : 0.8367   
## Balanced Accuracy : 0.7194   
##   
## 'Positive' Class : No   
##

Logistic Regression Model is another classification model that works well with both continusous and discrete data like KNN. The package selected provide a nice table making it easy to idenfity significant factors using p-values. The top variables include Commute, Environmental Satisfaction, Overtime, Job Involvement, Work Life Balance, Business Travel and Years Since Last Promotion.

logreg<-glm(Attrition~Age+BusinessTravel+Department+DistanceFromHome+Education+EducationField+EnvironmentSatisfaction+Gender+JobInvolvement+JobLevel+JobRole+JobSatisfaction+MaritalStatus+MonthlyIncome+NumCompaniesWorked+OverTime+PercentSalaryHike+PerformanceRating+RelationshipSatisfaction+StockOptionLevel+TotalWorkingYears+TrainingTimesLastYear+WorkLifeBalance+YearsAtCompany+YearsInCurrentRole+YearsSinceLastPromotion+YearsWithCurrManager, data=train,family=binomial(link='logit'))

Apply training data to Logistic Regression model. The logistic regression model improves upon the KNN model with an accuracy of 87.7% having a 95% confidence interval of (83%, 91%) with specificity of 80%.

# Store the probabilites for every observation in the dataset   
logreg.prob <- predict(logreg, validation, type = "response")  
  
# Tranform the "Yes" and "No" to binary variables  
test.logreg <- validation %>% mutate(model\_pred = 1\*(logreg.prob > .50) + 0,visit\_binary = 1\*(Attrition == "Yes") + 0)  
  
# Determine the accuracy of our model  
test.logreg <- test.logreg %>% mutate(accurate = 1\*(model\_pred == visit\_binary))  
sum(test.logreg$accurate)/nrow(test.logreg) # Accuracy Score

## [1] 0.8766667

val.attrition <- ifelse(val.attrition == "Yes",1, ifelse(val.attrition == "No",0, 99))   
  
# Recode Attrition with Yes's and No's  
confusionMatrix(table(val.attrition,test.logreg$model\_pred)) # confusion matrix

## Confusion Matrix and Statistics  
##   
##   
## val.attrition 0 1  
## 0 247 4  
## 1 33 16  
##   
## Accuracy : 0.8767   
## 95% CI : (0.834, 0.9117)  
## No Information Rate : 0.9333   
## P-Value [Acc > NIR] : 0.9999   
##   
## Kappa : 0.4077   
##   
## Mcnemar's Test P-Value : 4.161e-06   
##   
## Sensitivity : 0.8821   
## Specificity : 0.8000   
## Pos Pred Value : 0.9841   
## Neg Pred Value : 0.3265   
## Prevalence : 0.9333   
## Detection Rate : 0.8233   
## Detection Prevalence : 0.8367   
## Balanced Accuracy : 0.8411   
##   
## 'Positive' Class : 0   
##

dfPred <- data.frame(test.logreg$ID, test.logreg$model\_pred) # store predictions  
names(dfPred) <- c("ID","Predictions") # rename table columns

# Output Prediction File  
dfPred <- data.frame(dfPred) # create dataframe of normalized data  
dfPred$ID <- valIDs # add IDs  
dfPred <- dfPred[,c(1,2)]  
names(dfPred) <- c("ID","PredictedAttrition")