CaseStudy2Lull

Ellen Lull

4/18/2020

# MSDS 6306 Case Study 2: Analysis of Employee Data for Attrition and Salary Predictions

## Ellen Lull Section 401

### Read Files and Set up Libraries

.  
Executive Summary This is a DDS Analytics project to predict Attrition and Monthly Income based on data 870 Employee Data Records. This data is to be used to create a model for classifying attrition by significant variables. Next it creates a prediction model for Attrition based on existing variables. It will identify the top three turnover predictors. Then we will create a model to predict monthly income for each of the employees. Predictions on test datasets are output to CSV files for both Attrition and Monthly Income. Finally, we will be looking at three Job roles with high level jobs that have low attrition levels and their employee job satisfaction ratings. This project consists of analysis to cover all of these areas and a video presentation  
[Click Here for Video Presentation](https://www.youtube.com/watch?v=aOa0AA3os3k&feature=youtu.be)

library(tidyverse)

## Warning: package 'tidyverse' was built under R version 3.6.2

## -- Attaching packages ------------------------------ tidyverse 1.3.0 --

## v ggplot2 3.2.1 v purrr 0.3.3  
## v tibble 2.1.3 v dplyr 0.8.3  
## v tidyr 1.0.2 v stringr 1.4.0  
## v readr 1.3.1 v forcats 0.4.0

## Warning: package 'tidyr' was built under R version 3.6.2

## Warning: package 'purrr' was built under R version 3.6.2

## Warning: package 'stringr' was built under R version 3.6.2

## -- Conflicts --------------------------------- tidyverse\_conflicts() --  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

library(ggplot2)  
library(stringr)  
library(dplyr)  
library(caret)

## Warning: package 'caret' was built under R version 3.6.2

## Loading required package: lattice

##   
## Attaching package: 'caret'

## The following object is masked from 'package:purrr':  
##   
## lift

library(class)

## Warning: package 'class' was built under R version 3.6.2

library(randomForest)

## Warning: package 'randomForest' was built under R version 3.6.3

## randomForest 4.6-14

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

##   
## Attaching package: 'randomForest'

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

## The following object is masked from 'package:ggplot2':  
##   
## margin

library(glmnet)

## Warning: package 'glmnet' was built under R version 3.6.2

## Loading required package: Matrix

##   
## Attaching package: 'Matrix'

## The following objects are masked from 'package:tidyr':  
##   
## expand, pack, unpack

## Loaded glmnet 3.0-2

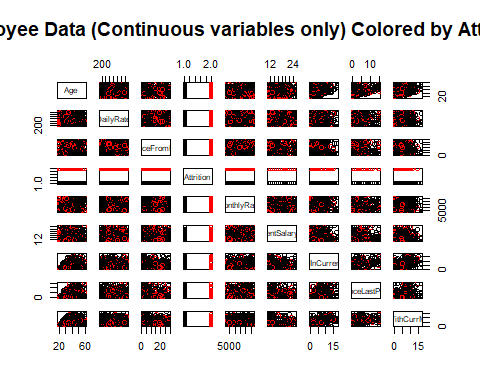
library(e1071)

## Warning: package 'e1071' was built under R version 3.6.2

employee <- read.delim("C:/School Stuff/DS/Doing DS/Project2/CaseStudy2-data.csv",header=TRUE,sep=",")  
  
employeenoinc <- read.delim("C:/School Stuff/DS/Doing DS/Project2/CaseStudy2CompSet No Salary.csv",header=TRUE,sep=",")  
  
employeenoatt<- read.delim("C:/School Stuff/DS/Doing DS/Project2/CaseStudy2CompSet No Attrition.csv",header=TRUE,sep=",")

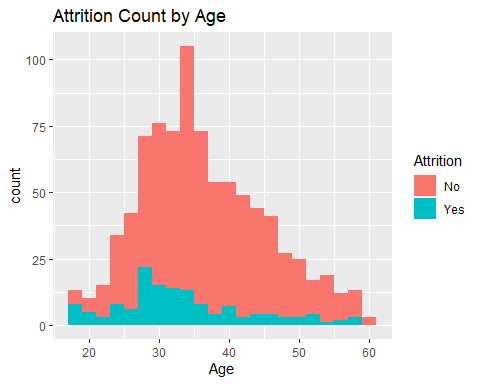
### EDA: Run Pairs Plot for Continuous Variables

employee\_reduced = employee[,c('Age','DailyRate','DistanceFromHome','Attrition','MonthlyRate','PercentSalaryHike','YearsInCurrentRole','YearsSinceLastPromotion','YearsWithCurrManager')]  
  
pairs(employee\_reduced,col=employee$Attrition,main='Employee Data (Continuous variables only) Colored by Attrition')

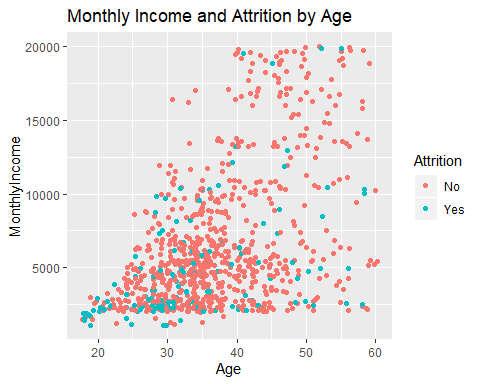


Plot Every Variable as Related to Attrition and Monthly Income ### EDA: Plot Relationships related to Attrition and Monthly Income for all 36 Variables

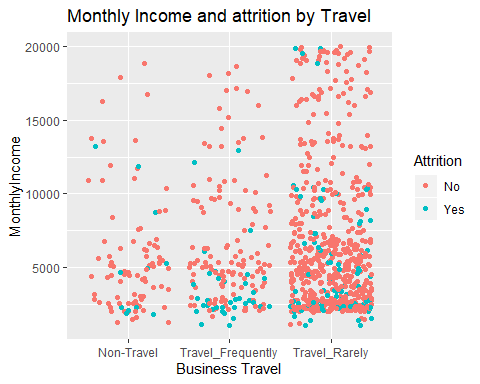
employee %>% ggplot(aes(x = Age,fill=Attrition)) + geom\_histogram(binwidth=2) + ggtitle("Attrition Count by Age") +  
xlab ("Age")



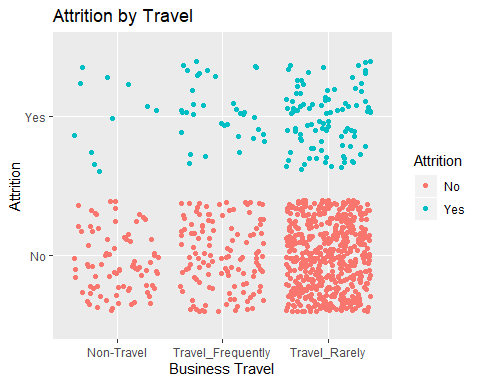
employee %>% ggplot(aes(x = Age,y=MonthlyIncome,col=Attrition)) + geom\_point(pos='Jitter') + ggtitle("Monthly Income and Attrition by Age") +  
 xlab ("Age")



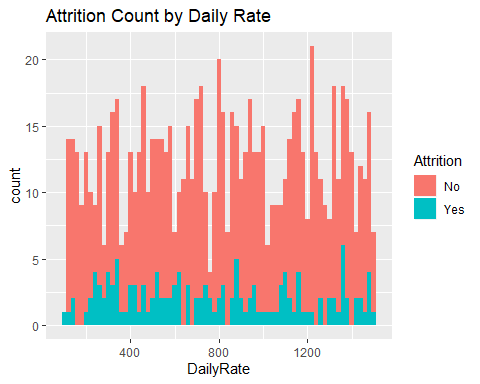
employee %>% ggplot(aes(x = BusinessTravel,y=MonthlyIncome,col=Attrition)) + geom\_point(pos='Jitter') + ggtitle("Monthly Income and attrition by Travel") +  
 xlab ("Business Travel")



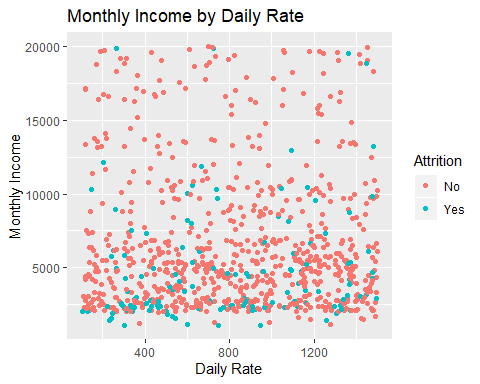
employee %>% ggplot(aes(x = BusinessTravel,y=Attrition,col=Attrition)) + geom\_point(pos='Jitter') + ggtitle("Attrition by Travel") +  
xlab ("Business Travel")



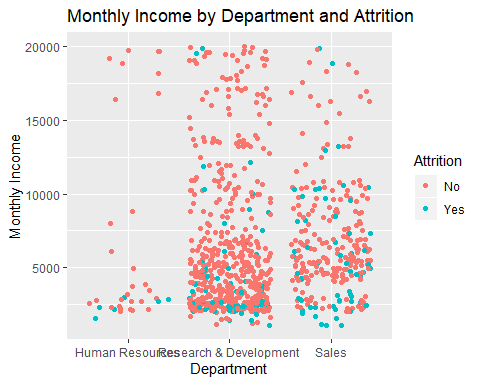
### Continous: DailyRate  
employee %>% ggplot(aes(x = DailyRate,fill=Attrition)) + geom\_histogram(binwidth=20) + ggtitle("Attrition Count by Daily Rate")



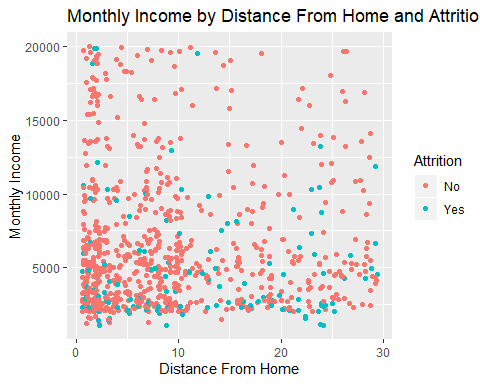
### Continous: DailyRate  
   
employee %>% ggplot(aes(x = DailyRate,y=MonthlyIncome,col=Attrition)) + geom\_point(pos='Jitter') + ggtitle("Monthly Income by Daily Rate") +  
xlab ("Daily Rate") +  
ylab ("Monthly Income")



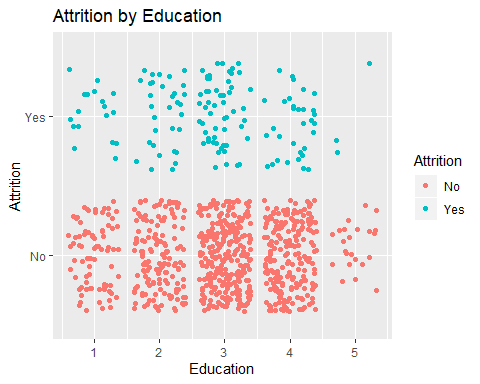
### Categorical: Department  
employee %>% ggplot(aes(x = Department,y=MonthlyIncome,col=Attrition)) + geom\_point(pos='Jitter') + ggtitle("Monthly Income by Department and Attrition") +  
xlab ("Department") +  
ylab ("Monthly Income")



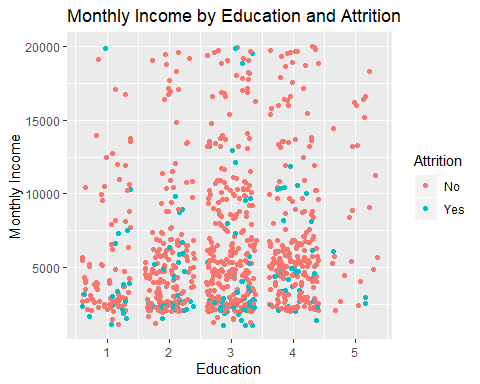
### Continuous: DistanceFromHome  
employee %>% ggplot(aes(x = DistanceFromHome,y=MonthlyIncome,col=Attrition)) + geom\_point(pos='Jitter') + ggtitle("Monthly Income by Distance From Home and Attrition") +  
xlab ("Distance From Home") +  
ylab ("Monthly Income")



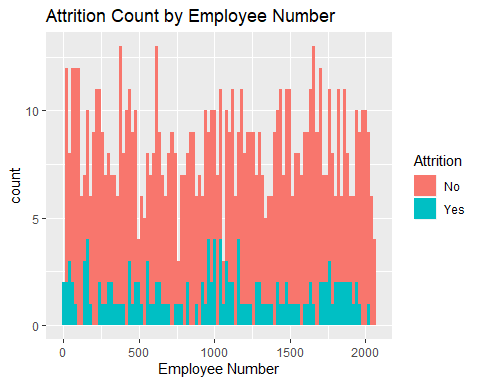
### Categorical: Education  
### Look at further to see < 4 years versus 4/5 years  
employee %>% ggplot(aes(x = Education,y=Attrition,col=Attrition)) + geom\_point(pos='Jitter') + ggtitle("Attrition by Education")



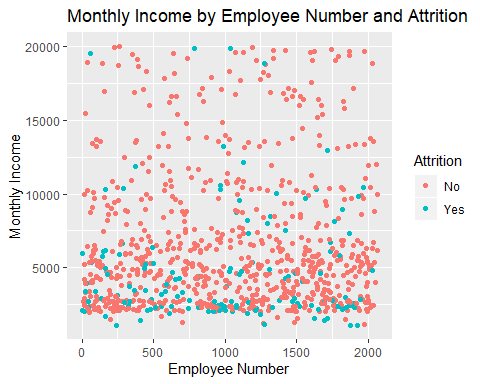
employee %>% ggplot(aes(x = Education,y=MonthlyIncome,col=Attrition)) + geom\_point(pos='Jitter') + ggtitle("Monthly Income by Education and Attrition") +  
xlab ("Education") +  
ylab ("Monthly Income")



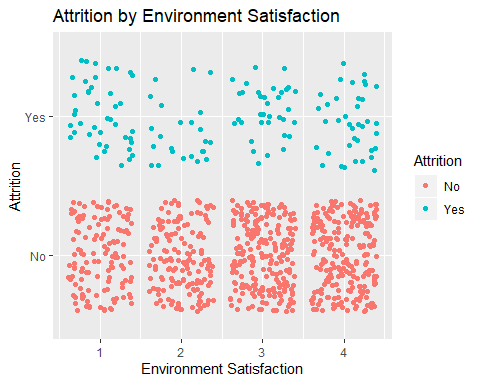
### Employee Count is always 1  
  
### Continuous: EmployeeNumber   
  
employee %>% ggplot(aes(x = EmployeeNumber,fill=Attrition)) + geom\_histogram(binwidth=20) + ggtitle("Attrition Count by Employee Number") +  
xlab ("Employee Number")



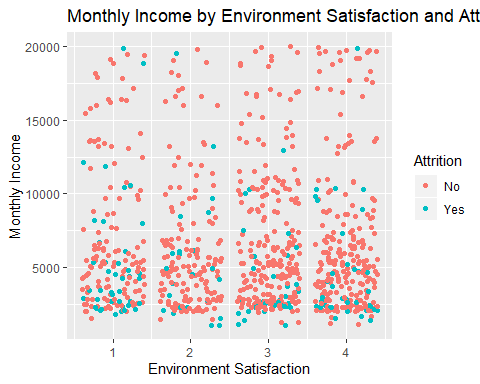
employee %>% ggplot(aes(x = EmployeeNumber,y=MonthlyIncome,col=Attrition)) + geom\_point(pos='Jitter') + ggtitle("Monthly Income by Employee Number and Attrition") +  
xlab ("Employee Number") +  
ylab ("Monthly Income")



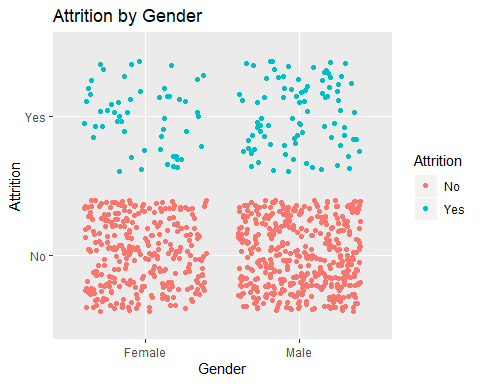
### Categorical: EnvironmentSatisfaction  
employee %>% ggplot(aes(x = EnvironmentSatisfaction,y=Attrition,col=Attrition)) + geom\_point(pos='Jitter') + ggtitle("Attrition by Environment Satisfaction") +  
xlab ("Environment Satisfaction")



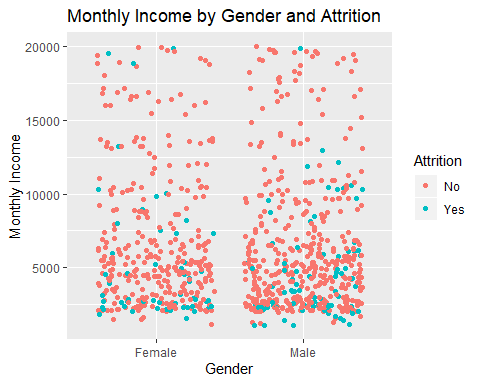
employee %>% ggplot(aes(x = EnvironmentSatisfaction,y=MonthlyIncome,col=Attrition)) + geom\_point(pos='Jitter') + ggtitle("Monthly Income by Environment Satisfaction and Attrition") +  
xlab ("Environment Satisfaction") +  
ylab ("Monthly Income")



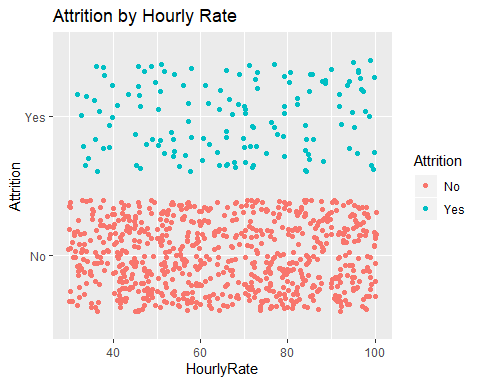
### Categorical: Gender   
### Males have a higher rate  
  
employee %>% ggplot(aes(x = Gender,y=Attrition,col=Attrition)) + geom\_point(pos='Jitter') + ggtitle("Attrition by Gender")



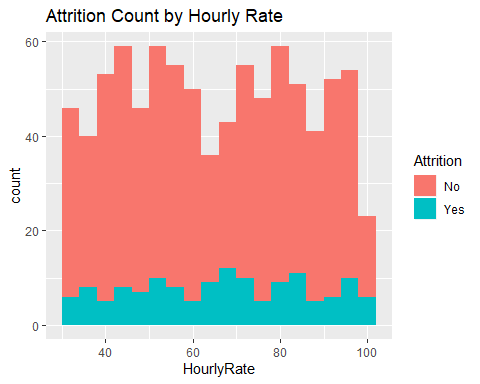
employee %>% ggplot(aes(x = Gender,y=MonthlyIncome,col=Attrition)) + geom\_point(pos='Jitter') + ggtitle("Monthly Income by Gender and Attrition") +  
xlab ("Gender") +  
ylab ("Monthly Income")



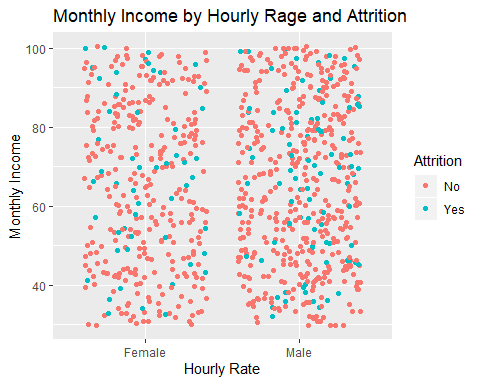
### Continuous: HourlyRate  
  
employee %>% ggplot(aes(x = HourlyRate,y=Attrition,col=Attrition)) + geom\_point(pos='Jitter') + ggtitle("Attrition by Hourly Rate")



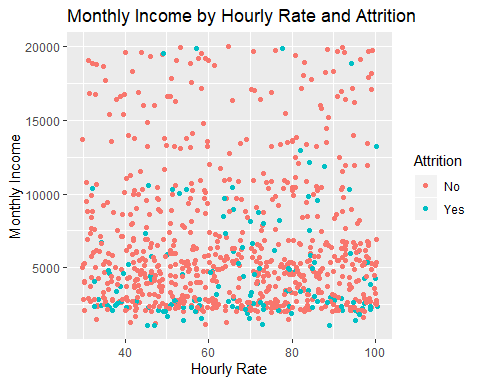
employee %>% ggplot(aes(x = HourlyRate,fill=Attrition)) + geom\_histogram(binwidth=4) + ggtitle("Attrition Count by Hourly Rate")



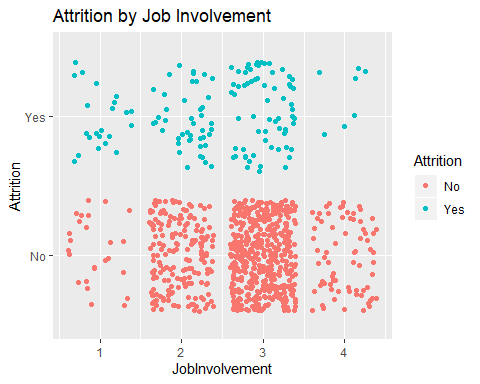
employee %>% ggplot(aes(x = Gender,HourlyRate,col=Attrition)) + geom\_point(pos='Jitter') + ggtitle("Monthly Income by Hourly Rage and Attrition") +  
xlab ("Hourly Rate") +  
ylab ("Monthly Income")



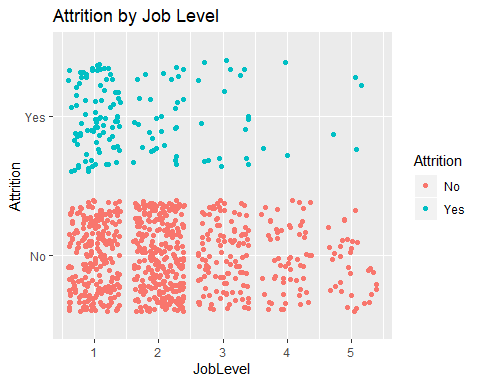
### Likely a billing rate   
  
employee %>% ggplot(aes(x = HourlyRate,y=MonthlyIncome,col=Attrition)) + geom\_point(pos='Jitter') + ggtitle("Monthly Income by Hourly Rate and Attrition") +  
 xlab ("Hourly Rate") +  
 ylab ("Monthly Income")



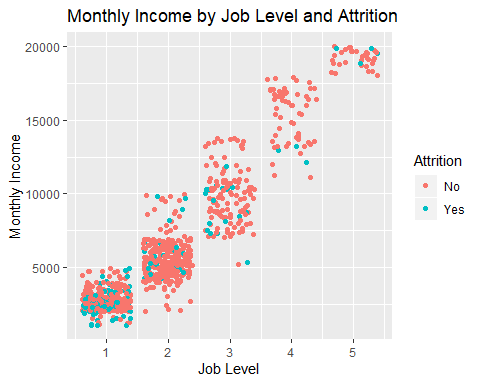
### Categorical JobInvolvement  
### Look at high job involvement  
  
employee %>% ggplot(aes(x = JobInvolvement,y=Attrition,col=Attrition)) + geom\_point(pos='Jitter') + ggtitle("Attrition by Job Involvement")



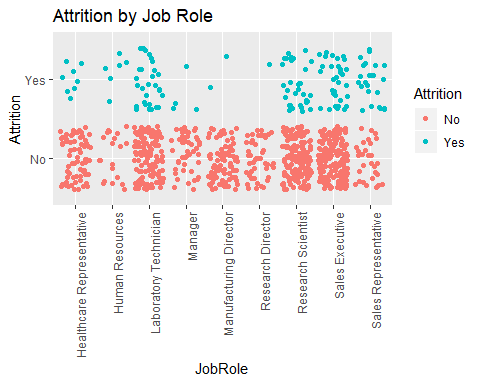
### Categorical JobLevel  
### Look at High Job Levels  
  
employee %>% ggplot(aes(x = JobLevel,y=Attrition,col=Attrition)) + geom\_point(pos='Jitter') + ggtitle("Attrition by Job Level")



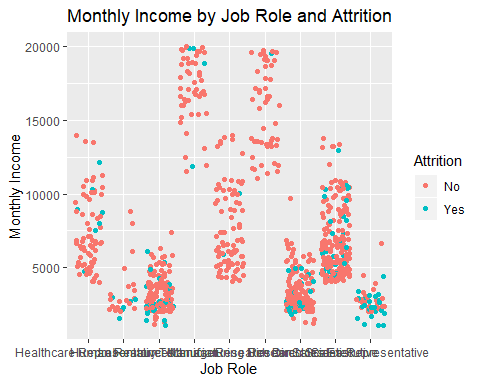
employee %>% ggplot(aes(x = JobLevel,y=MonthlyIncome,col=Attrition)) + geom\_point(pos='Jitter') + ggtitle("Monthly Income by Job Level and Attrition") +  
 xlab ("Job Level") +  
 ylab ("Monthly Income")



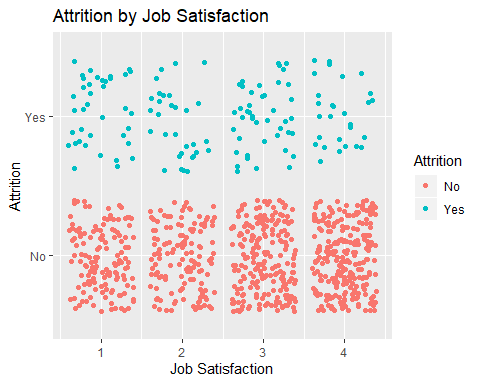
### Categorical JobRole  
### Look at Manager Manufacturing Director and Research Director  
  
employee %>% ggplot(aes(x = JobRole,y=Attrition,col=Attrition)) + geom\_point(pos='Jitter') + ggtitle("Attrition by Job Role") +  
 theme(axis.text.x=element\_text(angle=90, hjust=1))



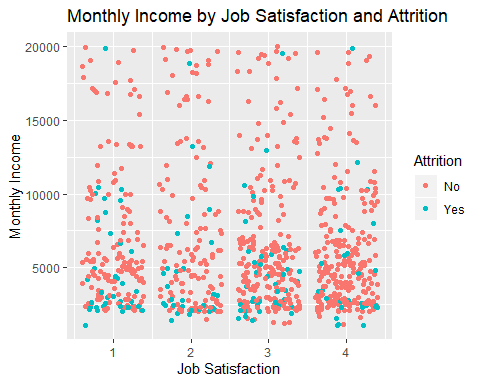
employee %>% ggplot(aes(x = JobRole,y=MonthlyIncome,col=Attrition)) + geom\_point(pos='Jitter') + ggtitle("Monthly Income by Job Role and Attrition") +  
 xlab ("Job Role") +  
 ylab ("Monthly Income")



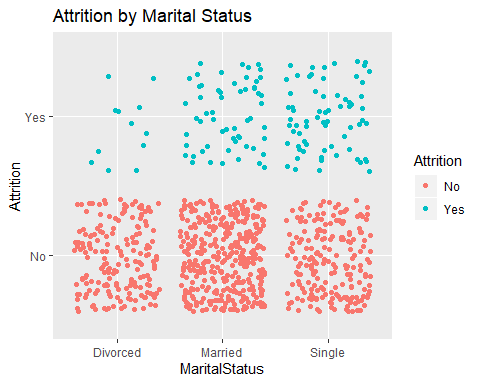
### Categorical JobSatisfaction  
  
employee %>% ggplot(aes(x = JobSatisfaction,y=Attrition,col=Attrition)) + geom\_point(pos='Jitter') + ggtitle("Attrition by Job Satisfaction") +  
xlab ("Job Satisfaction")



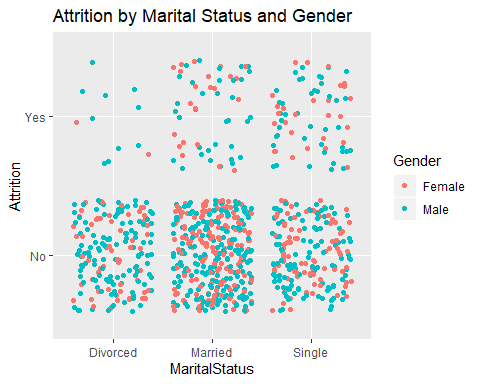
employee %>% ggplot(aes(x = JobSatisfaction,y=MonthlyIncome,col=Attrition)) + geom\_point(pos='Jitter') + ggtitle("Monthly Income by Job Satisfaction and Attrition") +  
 xlab ("Job Satisfaction") +  
 ylab ("Monthly Income")



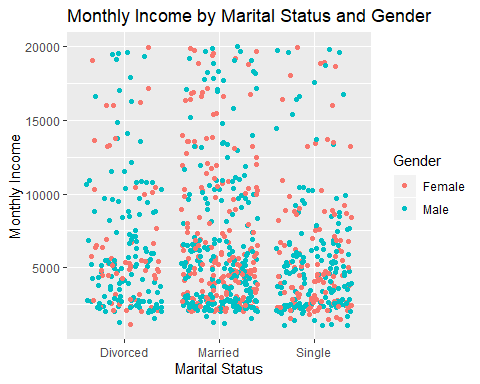
### Categorical: MaritalStatus  
### Look at Divorced Women. Divorced Men lower too  
  
employee %>% ggplot(aes(x = MaritalStatus,y=Attrition,col=Attrition)) + geom\_point(pos='Jitter') + ggtitle("Attrition by Marital Status")



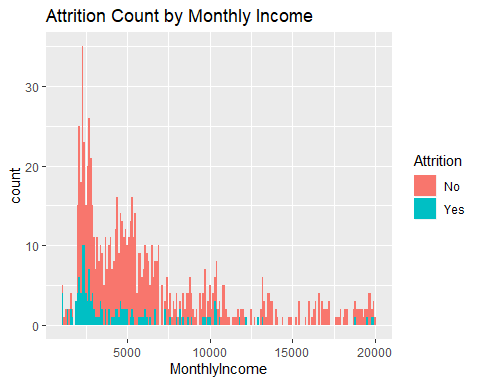
employee %>% ggplot(aes(x = MaritalStatus,y=Attrition,col=Gender)) + geom\_point(pos='Jitter') + ggtitle("Attrition by Marital Status and Gender")



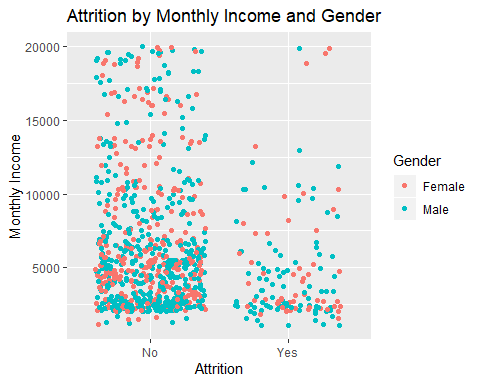
employee %>% ggplot(aes(x = MaritalStatus,y=MonthlyIncome,col=Gender)) + geom\_point(pos='Jitter') + ggtitle("Monthly Income by Marital Status and Gender") +  
 xlab ("Marital Status") +  
 ylab ("Monthly Income")



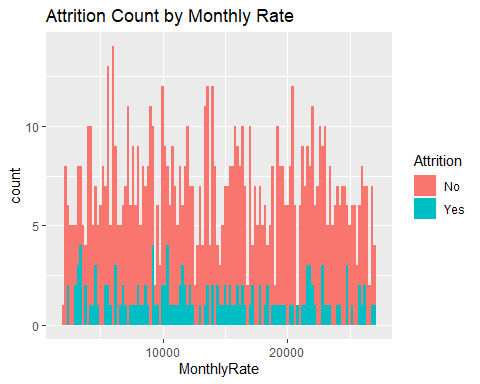
### Continous: MonthlyIncome  
employee %>% ggplot(aes(x = MonthlyIncome,fill=Attrition)) + geom\_histogram(binwidth=100) + ggtitle("Attrition Count by Monthly Income")



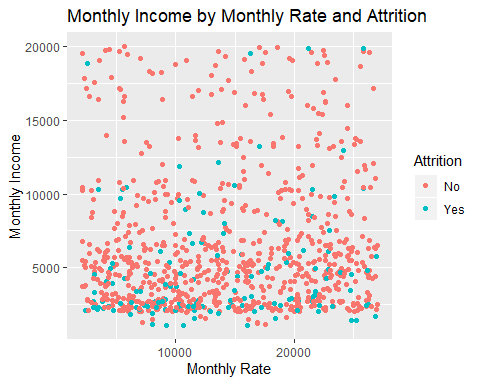
employee %>% ggplot(aes(x = Attrition,y=MonthlyIncome,col=Gender)) + geom\_point(pos='Jitter') + ggtitle("Attrition by Monthly Income and Gender") +  
 xlab ("Attrition") +  
 ylab ("Monthly Income")



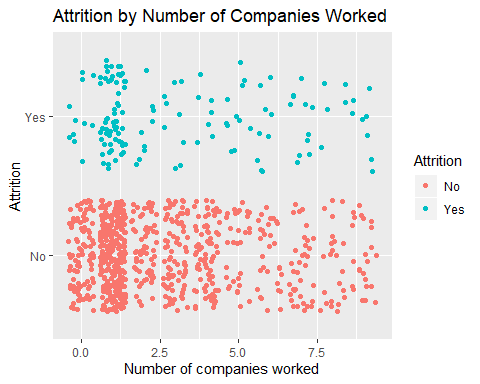
### Continuous: MonthlyRate   
employee %>% ggplot(aes(x = MonthlyRate,fill=Attrition)) + geom\_histogram(binwidth=200) + ggtitle("Attrition Count by Monthly Rate")



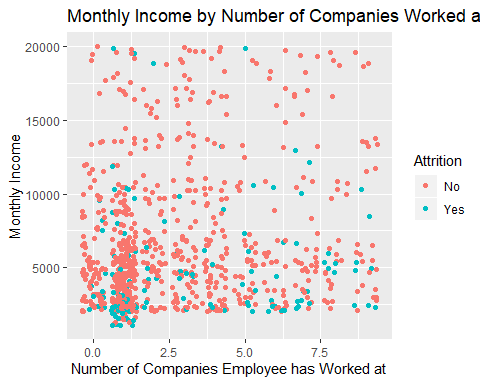
employee %>% ggplot(aes(x = MonthlyRate,y=MonthlyIncome,col=Attrition)) + geom\_point(pos='Jitter') + ggtitle("Monthly Income by Monthly Rate and Attrition") +  
 xlab ("Monthly Rate") +  
 ylab ("Monthly Income")



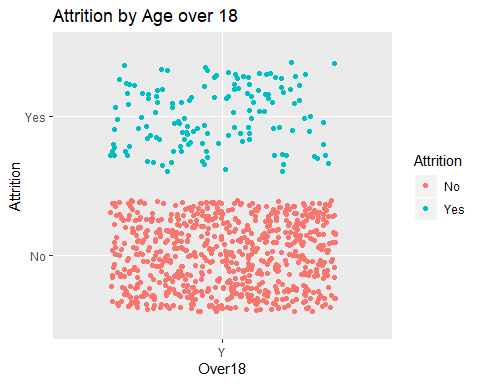
### Catigorical (numeric but only 9): NumCompaniesWorked   
employee %>% ggplot(aes(x = NumCompaniesWorked,y=Attrition,col=Attrition)) + geom\_point(pos='Jitter') + ggtitle("Attrition by Number of Companies Worked") +  
 xlab ("Number of companies worked")



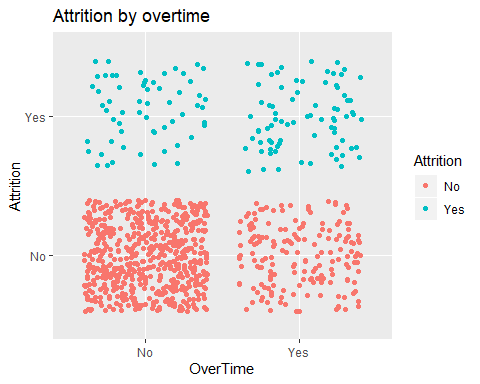
employee %>% ggplot(aes(x = NumCompaniesWorked,y=MonthlyIncome,col=Attrition)) + geom\_point(pos='Jitter') + ggtitle("Monthly Income by Number of Companies Worked and Attrition") +  
 xlab ("Number of Companies Employee has Worked at") +  
 ylab ("Monthly Income")



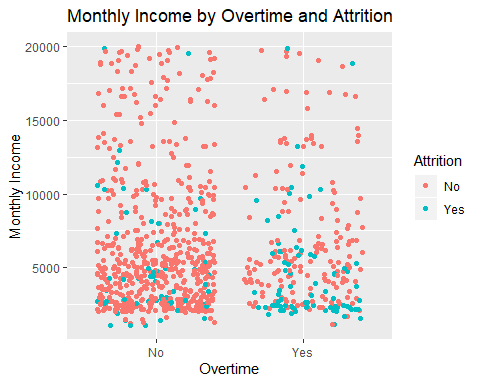
### Catigorical Over18   
## All over 18  
  
employee %>% ggplot(aes(x = Over18,y=Attrition,col=Attrition)) + geom\_point(pos='Jitter') + ggtitle("Attrition by Age over 18")



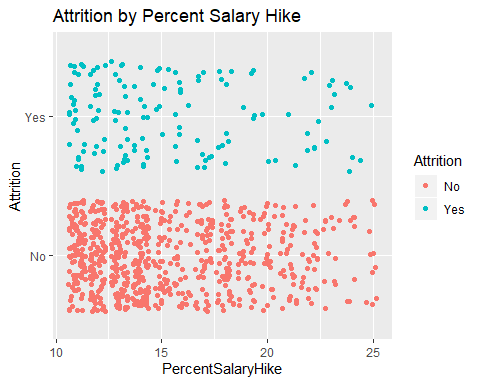
### Catigorical OverTime   
   
employee %>% ggplot(aes(x = OverTime,y=Attrition,col=Attrition)) + geom\_point(pos='Jitter') + ggtitle("Attrition by overtime")



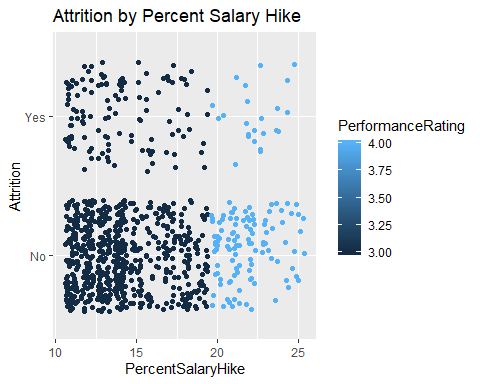
employee %>% ggplot(aes(x = OverTime,y=MonthlyIncome,col=Attrition)) + geom\_point(pos='Jitter') + ggtitle("Monthly Income by Overtime and Attrition") +  
 xlab ("Overtime") +  
 ylab ("Monthly Income")



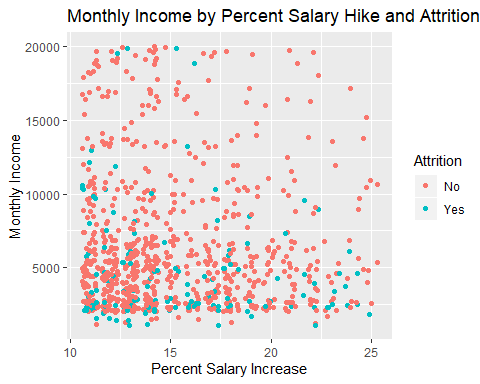
### Continuous: PercentSalaryHike  
### Look at salary hike with interaction of Permformance Rating  
   
employee %>% ggplot(aes(x = PercentSalaryHike,y=Attrition,col=Attrition)) + geom\_point(pos='Jitter') + ggtitle("Attrition by Percent Salary Hike")



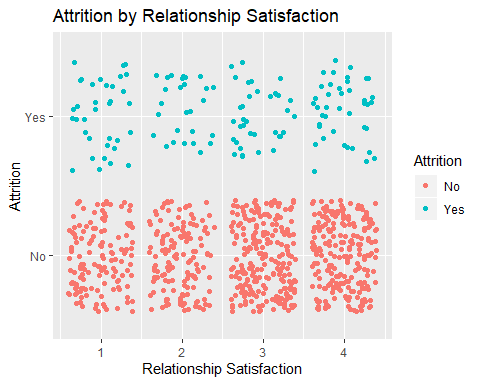
employee %>% ggplot(aes(x = PercentSalaryHike,y=Attrition,col=PerformanceRating)) + geom\_point(pos='Jitter') + ggtitle("Attrition by Percent Salary Hike")



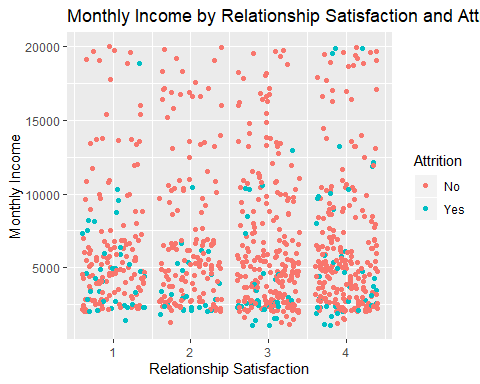
employee %>% ggplot(aes(x = PercentSalaryHike,y=MonthlyIncome,col=Attrition)) + geom\_point(pos='Jitter') + ggtitle("Monthly Income by Percent Salary Hike and Attrition") +  
 xlab ("Percent Salary Increase") +  
 ylab ("Monthly Income")



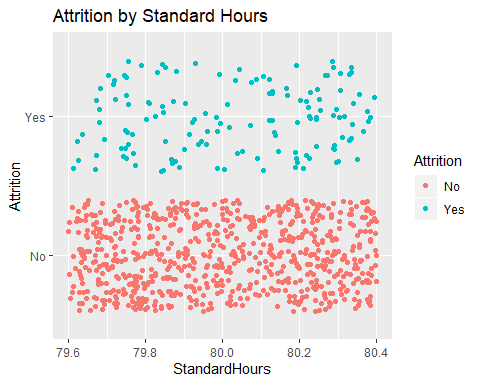
### Catigorical RelationshipSatisfaction   
   
employee %>% ggplot(aes(x = RelationshipSatisfaction,y=Attrition,col=Attrition)) + geom\_point(pos='Jitter') + ggtitle("Attrition by Relationship Satisfaction") +  
 xlab("Relationship Satisfaction")



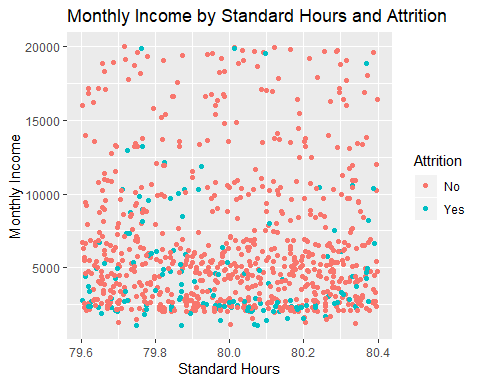
employee %>% ggplot(aes(x = RelationshipSatisfaction,y=MonthlyIncome,col=Attrition)) + geom\_point(pos='Jitter') + ggtitle("Monthly Income by Relationship Satisfaction and Attrition") +  
 xlab ("Relationship Satisfaction") +  
 ylab ("Monthly Income")



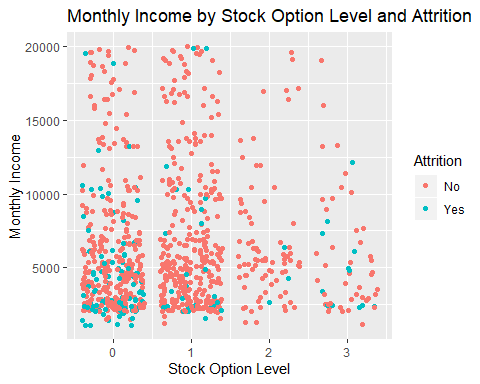
### Catigorical StandardHours   
 employee %>% ggplot(aes(x = StandardHours,y=Attrition,col=Attrition)) + geom\_point(pos='Jitter') + ggtitle("Attrition by Standard Hours")



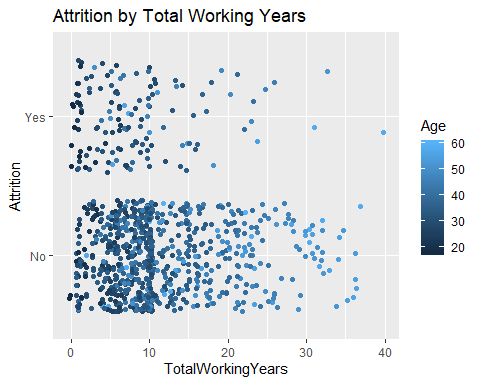
employee %>% ggplot(aes(x = StandardHours,y=MonthlyIncome,col=Attrition)) + geom\_point(pos='Jitter') + ggtitle("Monthly Income by Standard Hours and Attrition") +  
 xlab ("Standard Hours") +  
 ylab ("Monthly Income")



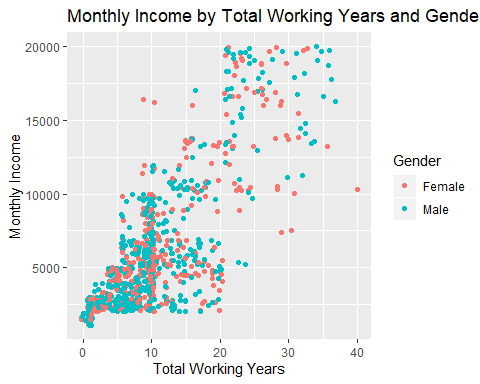
### Catigorical StandardHours   
### Stock Option Level 2  
   
employee %>% ggplot(aes(x = StockOptionLevel,y=MonthlyIncome,col=Attrition)) + geom\_point(pos='Jitter') + ggtitle("Monthly Income by Stock Option Level and Attrition") +  
 xlab ("Stock Option Level") +  
 ylab ("Monthly Income")



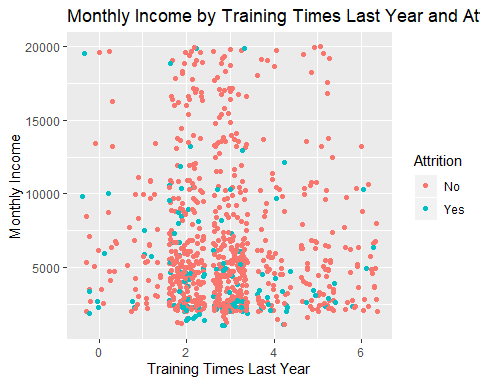
### Continuous: TotalWorkingYears  
### Obviously strong correlation with Age  
   
employee %>% ggplot(aes(x = TotalWorkingYears,y=Attrition,col=Age)) + geom\_point(pos='Jitter') + ggtitle("Attrition by Total Working Years")



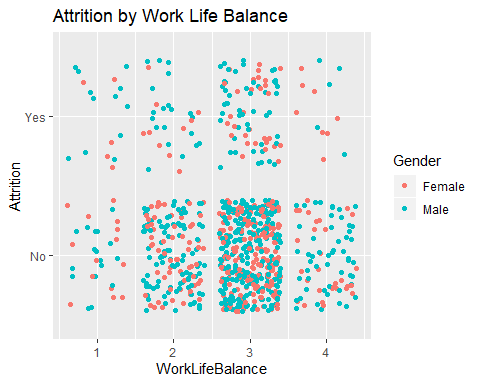
employee %>% ggplot(aes(x = TotalWorkingYears,y=MonthlyIncome,col=Gender)) + geom\_point(pos='Jitter') + ggtitle("Monthly Income by Total Working Years and Gender") +  
 xlab ("Total Working Years") +  
 ylab ("Monthly Income")



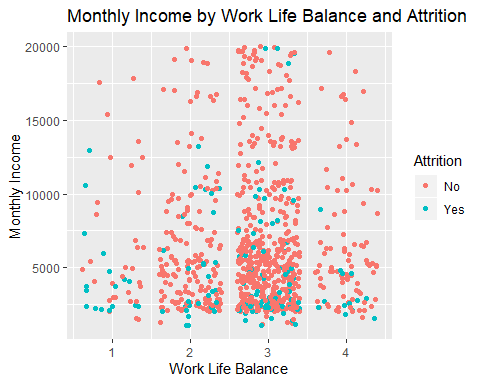
### TrainingTimesLastYear; values 1-6  
employee %>% ggplot(aes(x = TrainingTimesLastYear,y=MonthlyIncome,col=Attrition)) + geom\_point(pos='Jitter') + ggtitle("Monthly Income by Training Times Last Year and Attrition") +  
 xlab ("Training Times Last Year") +  
 ylab ("Monthly Income")



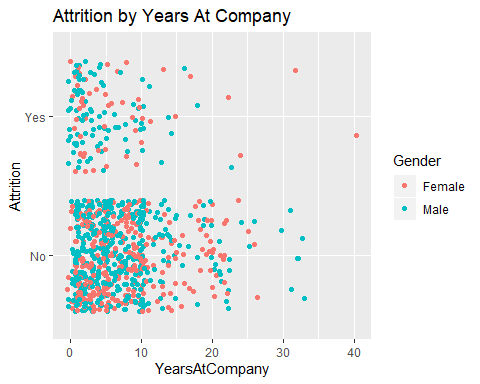
### Catigorical work life balance  
### High is good for both genders  
  
employee %>% ggplot(aes(x = WorkLifeBalance,y=Attrition,col=Gender)) + geom\_point(pos='Jitter') + ggtitle("Attrition by Work Life Balance")



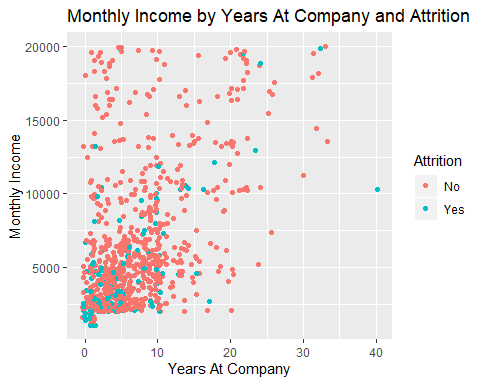
employee %>% ggplot(aes(x = WorkLifeBalance,y=MonthlyIncome,col=Attrition)) + geom\_point(pos='Jitter') + ggtitle("Monthly Income by Work Life Balance and Attrition") +  
 xlab ("Work Life Balance") +  
 ylab ("Monthly Income")



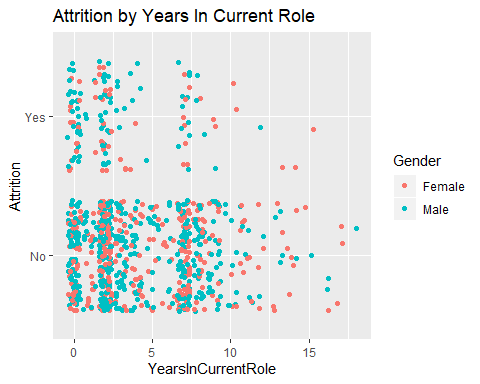
### Continuous: YearsAtCompany   
### Light attrition after 20 years  
  
employee %>% ggplot(aes(x = YearsAtCompany,y=Attrition,col=Gender)) + geom\_point(pos='Jitter') + ggtitle("Attrition by Years At Company")



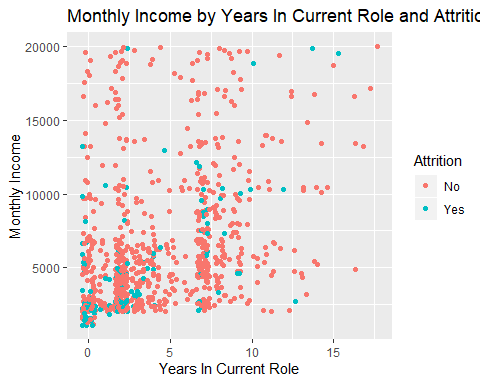
employee %>% ggplot(aes(x = YearsAtCompany,y=MonthlyIncome,col=Attrition)) + geom\_point(pos='Jitter') + ggtitle("Monthly Income by Years At Company and Attrition") +  
 xlab ("Years At Company") +  
 ylab ("Monthly Income")



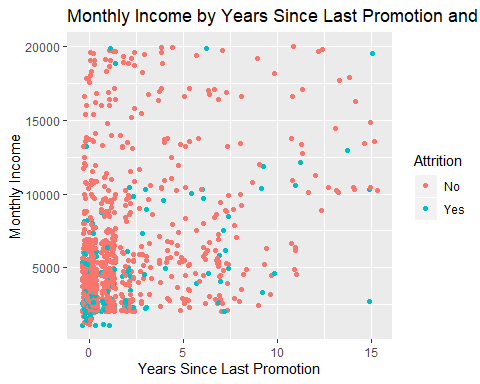
### Years in current role  
### Not much movement in larger years  
employee %>% ggplot(aes(x = YearsInCurrentRole,y=Attrition,col=Gender)) + geom\_point(pos='Jitter') + ggtitle("Attrition by Years In Current Role")



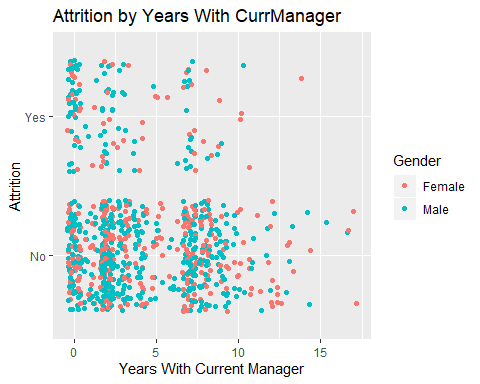
employee %>% ggplot(aes(x = YearsInCurrentRole,y=MonthlyIncome,col=Attrition)) + geom\_point(pos='Jitter') + ggtitle("Monthly Income by Years In Current Role and Attrition") +  
 xlab ("Years In Current Role") +  
 ylab ("Monthly Income")



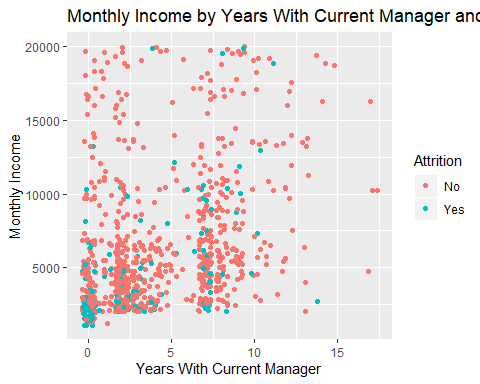
### YearsSinceLastPromotion  
employee %>% ggplot(aes(x = YearsSinceLastPromotion,y=MonthlyIncome,col=Attrition)) + geom\_point(pos='Jitter') + ggtitle("Monthly Income by Years Since Last Promotion and Attrition") +  
 xlab ("Years Since Last Promotion") +  
 ylab ("Monthly Income")



### YearsWithCurrManager  
employee %>% ggplot(aes(x = YearsWithCurrManager,y=Attrition,col=Gender)) + geom\_point(pos='Jitter') + ggtitle("Attrition by Years With CurrManager") +  
xlab ("Years With Current Manager")



employee %>% ggplot(aes(x = YearsWithCurrManager,y=MonthlyIncome,col=Attrition)) + geom\_point(pos='Jitter') + ggtitle("Monthly Income by Years With Current Manager and Attrition") +  
 xlab ("Years With Current Manager") +  
 ylab ("Monthly Income")

 ### Create Test and Training Data Sets Data is Skewed for Attrition Reponse, so first split the data into two sets for Yes and No responses. Then divide 80 percent to training and 20 percent test on the individual datasets. Then put the two back together

This will insure the YES responses are adequately represented. If we do not do this, we risk having no YES responses in one of the created datasets

employee\_reduced2 = employee[,c('Age','WorkLifeBalance','Education','StockOptionLevel','TotalWorkingYears','YearsAtCompany','YearsWithCurrManager','Attrition')]  
  
#split datasets yes/no  
AttritionYes = employee\_reduced2 %>% filter(Attrition == "Yes")  
AttritionNo = employee\_reduced2 %>% filter(Attrition == "No")  
  
# Balance the Dataset with taking the test/train split of 80%/ 20% on both Yes and No  
# Attrition so both values are in the test and training datasets  
  
set.seed(9)  
trainInd = sample(seq(1,dim(AttritionYes)[1],1),round(.7\*dim(AttritionYes)[1]))  
trainYES = AttritionYes[trainInd,]  
testYES = AttritionYes[-trainInd,]  
  
trainInd = sample(seq(1,dim(AttritionNo)[1],1),round(.7\*dim(AttritionNo)[1]))  
train = AttritionNo[trainInd,]  
test = AttritionNo[-trainInd,]  
  
train = rbind(train,trainYES)  
test = rbind(test,testYES)  
table(test$Attrition)

##   
## No Yes   
## 219 42

cltest=knn(train[,c(1:7)],test[,c(1:7)],train$Attrition, prob = TRUE, k = 25)  
table(cltest,test$Attrition)

##   
## cltest No Yes  
## No 217 40  
## Yes 2 2

CM = confusionMatrix(table(cltest,test$Attrition))  
CM

## Confusion Matrix and Statistics  
##   
##   
## cltest No Yes  
## No 217 40  
## Yes 2 2  
##   
## Accuracy : 0.8391   
## 95% CI : (0.7888, 0.8815)  
## No Information Rate : 0.8391   
## P-Value [Acc > NIR] : 0.5411   
##   
## Kappa : 0.0607   
##   
## Mcnemar's Test P-Value : 1.135e-08   
##   
## Sensitivity : 0.99087   
## Specificity : 0.04762   
## Pos Pred Value : 0.84436   
## Neg Pred Value : 0.50000   
## Prevalence : 0.83908   
## Detection Rate : 0.83142   
## Detection Prevalence : 0.98467   
## Balanced Accuracy : 0.51924   
##   
## 'Positive' Class : No   
##

employee\_reducedz = data.frame(ZAge = scale(employee$Age), ZWorkLifeBalance = scale(employee$WorkLifeBalance), ZEducation=scale(employee$Education) ,ZStockOptionLevel=scale(employee$StockOptionLevel), ZTotalWorkingYears = scale(employee$TotalWorkingYears) , ZYearsAtCompany = scale(employee$YearsAtCompany) , ZYearsWithCurrManager=scale(employee$YearsWithCurrManager), Attrition = employee$Attrition)  
  
#split datasets yes/no  
AttritionYes = employee\_reducedz %>% filter(Attrition == "Yes")  
AttritionNo = employee\_reducedz %>% filter(Attrition == "No")  
  
# Balance the Dataset with taking the test/train split of 80%/ 20% on both Yes and No  
# Attrition so both values are in the test and training datasets  
  
set.seed(9)  
trainInd = sample(seq(1,dim(AttritionYes)[1],1),round(.7\*dim(AttritionYes)[1]))  
trainYES = AttritionYes[trainInd,]  
testYES = AttritionYes[-trainInd,]  
  
trainInd = sample(seq(1,dim(AttritionNo)[1],1),round(.7\*dim(AttritionNo)[1]))  
train = AttritionNo[trainInd,]  
test = AttritionNo[-trainInd,]  
  
train = rbind(train,trainYES)  
test = rbind(test,testYES)  
table(test$Attrition)

##   
## No Yes   
## 219 42

cltest=knn(train[,c(1:7)],test[,c(1:7)],train$Attrition, prob = TRUE, k = 25)  
table(cltest,test$Attrition)

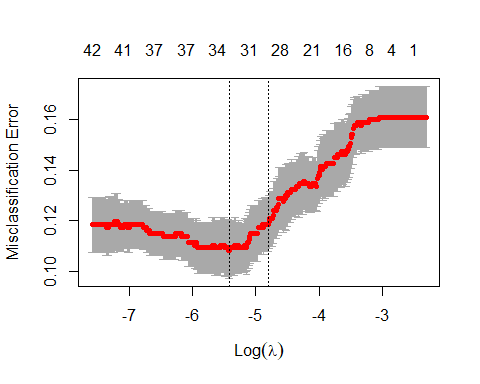
##   
## cltest No Yes  
## No 219 40  
## Yes 0 2

CM = confusionMatrix(table(cltest,test$Attrition))  
CM

## Confusion Matrix and Statistics  
##   
##   
## cltest No Yes  
## No 219 40  
## Yes 0 2  
##   
## Accuracy : 0.8467   
## 95% CI : (0.7972, 0.8882)  
## No Information Rate : 0.8391   
## P-Value [Acc > NIR] : 0.4072   
##   
## Kappa : 0.0774   
##   
## Mcnemar's Test P-Value : 6.984e-10   
##   
## Sensitivity : 1.00000   
## Specificity : 0.04762   
## Pos Pred Value : 0.84556   
## Neg Pred Value : 1.00000   
## Prevalence : 0.83908   
## Detection Rate : 0.83908   
## Detection Prevalence : 0.99234   
## Balanced Accuracy : 0.52381   
##   
## 'Positive' Class : No   
##

LASSO Feature Selection to use in Logistical Regression Attrition Model and basis for Linear Regression Income Model

lasso.y <-employee[,c("Attrition")]  
   
lasso.x <- model.matrix(Attrition~ Age + BusinessTravel + DailyRate + Department + DistanceFromHome + Education + EducationField +EmployeeCount + EmployeeNumber + EnvironmentSatisfaction + Gender + HourlyRate + JobInvolvement + JobLevel + JobRole + JobSatisfaction + MaritalStatus + MonthlyIncome + MonthlyRate + NumCompaniesWorked + OverTime + PerformanceRating + RelationshipSatisfaction + StandardHours + StockOptionLevel + TotalWorkingYears + TrainingTimesLastYear +WorkLifeBalance + YearsAtCompany + YearsInCurrentRole + YearsSinceLastPromotion + YearsWithCurrManager  
,employee)  
  
  
### LASSO Model for Feature Selection  
lasso.mdl.cvfit <- cv.glmnet(lasso.x, lasso.y, family = "binomial", type.measure = "class", nlambda = 1000)  
plot(lasso.mdl.cvfit)



coef(lasso.mdl.cvfit)

## 48 x 1 sparse Matrix of class "dgCMatrix"  
## 1  
## (Intercept) 2.431166e+00  
## (Intercept) .   
## Age -2.140829e-02  
## BusinessTravelTravel\_Frequently 4.287978e-01  
## BusinessTravelTravel\_Rarely .   
## DailyRate -5.934412e-05  
## DepartmentResearch & Development -1.754303e-01  
## DepartmentSales .   
## DistanceFromHome 2.711976e-02  
## Education .   
## EducationFieldLife Sciences .   
## EducationFieldMarketing .   
## EducationFieldMedical .   
## EducationFieldOther .   
## EducationFieldTechnical Degree 2.160689e-01  
## EmployeeCount .   
## EmployeeNumber .   
## EnvironmentSatisfaction -1.773350e-01  
## GenderMale 1.715702e-02  
## HourlyRate 3.889946e-03  
## JobInvolvement -6.150345e-01  
## JobLevel .   
## JobRoleHuman Resources 2.857016e-01  
## JobRoleLaboratory Technician 2.682875e-01  
## JobRoleManager .   
## JobRoleManufacturing Director -1.052099e+00  
## JobRoleResearch Director -6.450810e-01  
## JobRoleResearch Scientist .   
## JobRoleSales Executive .   
## JobRoleSales Representative 1.019709e+00  
## JobSatisfaction -3.112936e-01  
## MaritalStatusMarried 1.749374e-01  
## MaritalStatusSingle 7.594999e-01  
## MonthlyIncome .   
## MonthlyRate -3.337732e-06  
## NumCompaniesWorked 1.266096e-01  
## OverTimeYes 1.502123e+00  
## PerformanceRating .   
## RelationshipSatisfaction -1.175047e-01  
## StandardHours .   
## StockOptionLevel -1.376822e-01  
## TotalWorkingYears -4.208574e-02  
## TrainingTimesLastYear -1.453440e-01  
## WorkLifeBalance -3.195260e-01  
## YearsAtCompany .   
## YearsInCurrentRole -4.711800e-02  
## YearsSinceLastPromotion 1.241853e-01  
## YearsWithCurrManager -5.379575e-02

### Create Test and Training Data Sets ### Logistical Regression based on LASSO feature Selection

Data is Skewed for Attrition Reponse, so first split the data into two sets for Yes and No responses. Then divide 80 percent to training and 20 percent test on the individual datasets. Then put the two back together

This will insure the YES responses are adequately represented. If we do not do this, we risk having no YES responses in one of the created datasets

#Train and Test Split 80%/20%, with a seed of 10 so all members of the group can use to compare results on the same basis  
#The split wa done using the Yes and No Attrition Values Seperately to keep it balanced  
  
AttritionYes = employee %>% filter(Attrition == "Yes")  
AttritionNo = employee %>% filter(Attrition == "No")  
set.seed(9)  
trainInd = sample(dim(AttritionYes)[1],round(.8\*dim(AttritionYes)[1]))  
trainYES = AttritionYes[trainInd,]  
testYES = AttritionYes[-trainInd,]  
  
trainInd = sample(dim(AttritionNo)[1],round(.8\*dim(AttritionNo)[1]))  
train = AttritionNo[trainInd,]  
test = AttritionNo[-trainInd,]  
  
train = rbind(train,trainYES)  
test = rbind(test,testYES)  
table(test$Attritiontion)

## < table of extent 0 >

lr.employee <-glm(Attrition ~ Age + BusinessTravel + DistanceFromHome + EducationField + EnvironmentSatisfaction + HourlyRate + JobInvolvement + JobRole + JobSatisfaction + MaritalStatus\*Gender + MonthlyRate + NumCompaniesWorked + OverTime + RelationshipSatisfaction + StockOptionLevel + TotalWorkingYears + TrainingTimesLastYear + WorkLifeBalance + YearsInCurrentRole + YearsSinceLastPromotion + YearsWithCurrManager,data=train,family=binomial(link="logit"))  
  
lr.employee.pred2 <- data.frame(predict(lr.employee, newdata = test, type = "response"))  
lr.employee.pred2 = lr.employee.pred2 %>% mutate(pred = ifelse(lr.employee.pred2 <0.25, "No", "Yes"))  
  
table(lr.employee.pred2$pred)

##   
## No Yes   
## 130 44

predtble = as.factor(lr.employee.pred2$pred)  
predtble <-relevel(predtble, ref = "No")   
Truth<-test$Attrition  
confmtx = as.matrix(table(predtble,Truth))  
CM = confusionMatrix(confmtx)  
CM

## Confusion Matrix and Statistics  
##   
## Truth  
## predtble No Yes  
## No 123 7  
## Yes 23 21  
##   
## Accuracy : 0.8276   
## 95% CI : (0.7631, 0.8805)  
## No Information Rate : 0.8391   
## P-Value [Acc > NIR] : 0.70289   
##   
## Kappa : 0.4813   
##   
## Mcnemar's Test P-Value : 0.00617   
##   
## Sensitivity : 0.8425   
## Specificity : 0.7500   
## Pos Pred Value : 0.9462   
## Neg Pred Value : 0.4773   
## Prevalence : 0.8391   
## Detection Rate : 0.7069   
## Detection Prevalence : 0.7471   
## Balanced Accuracy : 0.7962   
##   
## 'Positive' Class : No   
##

### Create Predcition Spreadsheet from Logistical Regression Model I

lr.employee.predA <- data.frame(predict(lr.employee, newdata = employeenoatt, type = "response"))  
lr.employee.predA = lr.employee.predA %>% mutate(pred = ifelse(lr.employee.predA <0.25, "No", "Yes"))  
  
preddfA = NewAttr=data.frame(lr.employee.predA[,c(2)])  
names(preddfA) <- c("PredictAttrition")  
AttrPred <- cbind(employeenoatt[1],preddfA)  
  
write.csv(AttrPred,'c:/School Stuff/DS/Doing DS/Project2/Case2PredictionsLull Attrition.csv')

Running a second LR Model on the same test/train datasets as above This model was created by hand using only Data Analysis graphs

lr.employee <-glm(Attrition~Age\*TotalWorkingYears+ WorkLifeBalance + NumCompaniesWorked + StockOptionLevel +JobLevel +JobInvolvement + MaritalStatus\*Gender + YearsAtCompany + YearsWithCurrManager,data=train,family=binomial(link="logit"))  
  
lr.employee.pred2 <- data.frame(predict(lr.employee, newdata = test, type = "response"))  
lr.employee.pred2 = lr.employee.pred2 %>% mutate(pred = ifelse(lr.employee.pred2 <0.25, "No", "Yes"))  
  
table(lr.employee.pred2$pred)

##   
## No Yes   
## 122 52

predtble = as.factor(lr.employee.pred2$pred)  
predtble <-relevel(predtble, ref = "No")   
Truth<-test$Attrition  
confmtx = as.matrix(table(predtble,Truth))  
CM = confusionMatrix(confmtx)  
CM

## Confusion Matrix and Statistics  
##   
## Truth  
## predtble No Yes  
## No 110 12  
## Yes 36 16  
##   
## Accuracy : 0.7241   
## 95% CI : (0.6514, 0.7891)  
## No Information Rate : 0.8391   
## P-Value [Acc > NIR] : 0.9999580   
##   
## Kappa : 0.2413   
##   
## Mcnemar's Test P-Value : 0.0009009   
##   
## Sensitivity : 0.7534   
## Specificity : 0.5714   
## Pos Pred Value : 0.9016   
## Neg Pred Value : 0.3077   
## Prevalence : 0.8391   
## Detection Rate : 0.6322   
## Detection Prevalence : 0.7011   
## Balanced Accuracy : 0.6624   
##   
## 'Positive' Class : No   
##

### Test for Attrition Classification on Random Forest based on reduced dataset by EDA Plots

#### Try Random Forest to see what model looks like  
employee\_reducedrf = employee[,c('WorkLifeBalance','Education','StockOptionLevel','JobLevel','Department','OverTime','TotalWorkingYears','YearsAtCompany','YearsWithCurrManager','Attrition')]  
  
  
employee.rf <-randomForest(Attrition~.,data=employee\_reducedrf,mtry=10,ntree=500,importance=T)

## Warning in randomForest.default(m, y, ...): invalid mtry: reset to within valid  
## range

summary(employee.rf)

## Length Class Mode   
## call 6 -none- call   
## type 1 -none- character  
## predicted 870 factor numeric   
## err.rate 1500 -none- numeric   
## confusion 6 -none- numeric   
## votes 1740 matrix numeric   
## oob.times 870 -none- numeric   
## classes 2 -none- character  
## importance 36 -none- numeric   
## importanceSD 27 -none- numeric   
## localImportance 0 -none- NULL   
## proximity 0 -none- NULL   
## ntree 1 -none- numeric   
## mtry 1 -none- numeric   
## forest 14 -none- list   
## y 870 factor numeric   
## test 0 -none- NULL   
## inbag 0 -none- NULL   
## terms 3 terms call

table(employee.rf$predicted,employee\_reducedrf$Attrition)

##   
## No Yes  
## No 680 92  
## Yes 50 48

CM = confusionMatrix(table(employee.rf$predicted,employee\_reducedrf$Attrition))  
CM

## Confusion Matrix and Statistics  
##   
##   
## No Yes  
## No 680 92  
## Yes 50 48  
##   
## Accuracy : 0.8368   
## 95% CI : (0.8105, 0.8607)  
## No Information Rate : 0.8391   
## P-Value [Acc > NIR] : 0.5950295   
##   
## Kappa : 0.3122   
##   
## Mcnemar's Test P-Value : 0.0005803   
##   
## Sensitivity : 0.9315   
## Specificity : 0.3429   
## Pos Pred Value : 0.8808   
## Neg Pred Value : 0.4898   
## Prevalence : 0.8391   
## Detection Rate : 0.7816   
## Detection Prevalence : 0.8874   
## Balanced Accuracy : 0.6372   
##   
## 'Positive' Class : No   
##

### Linear Regression Model for Income Prediction

Run Linear Regression Model for Monthly Income Predictions This model was based on the same LASSO Feature selection criteria used for Attrition. However several interactions were added based on EDA to decrease the RMSE. The Histogram of residuals showed skewed data, which is a violation of the Normality Assumption.  
To correct for that normality violation, a log transformation was done on Monthly Income. This smoothed out the residuals and gave a substatially lower RMSE.

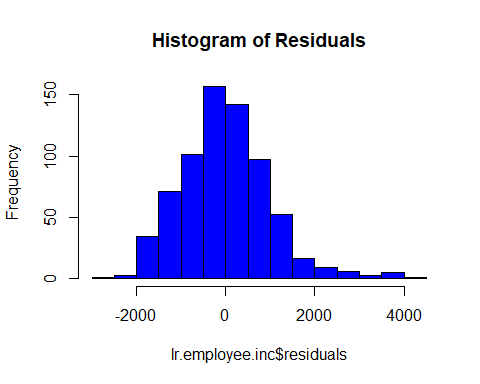
#split datasets yes/no  
AttritionYes = employee %>% filter(Attrition == "Yes")  
AttritionNo = employee %>% filter(Attrition == "No")  
  
#Redo Test/Train split - same as done before  
set.seed(9)  
trainInd = sample(dim(AttritionYes)[1],round(.8\*dim(AttritionYes)[1]))  
trainYES = employee[trainInd,]  
testYES = employee[-trainInd,]  
  
trainInd = sample(dim(AttritionNo)[1],round(.8\*dim(AttritionNo)[1]))  
train = AttritionNo[trainInd,]  
test = AttritionNo[-trainInd,]  
  
train = rbind(train,trainYES)  
test = rbind(test,testYES)  
table(test$Attribution)

## < table of extent 0 >

### Added Interactions for Joblevel and Job Role, Business Travel and Gender, Total Working Years and Age, Years in current Role and Current Manager  
  
lr.employee.inc <-lm(MonthlyIncome ~ JobLevel\*JobRole + Department + BusinessTravel\*Gender + DistanceFromHome + EducationField\*Education + EnvironmentSatisfaction + HourlyRate\*MonthlyRate + JobInvolvement + JobSatisfaction + MaritalStatus + NumCompaniesWorked + OverTime + RelationshipSatisfaction + StockOptionLevel + TotalWorkingYears\*Age + TrainingTimesLastYear + WorkLifeBalance + YearsSinceLastPromotion + YearsInCurrentRole\*YearsWithCurrManager,data=train)  
  
RMSE = sqrt(mean(lr.employee.inc$residuals^2))  
RMSE

## [1] 988.333

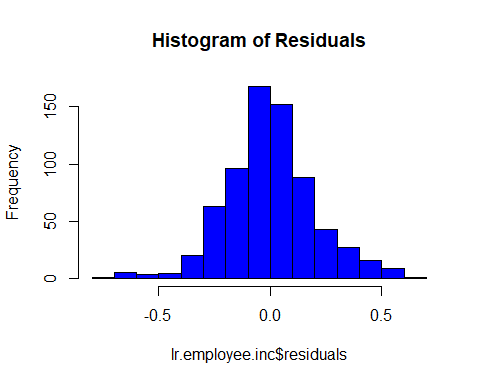
#### Histogram of Residuals  
  
 hist(lr.employee.inc$residuals, col = "blue", main = "Histogram of Residuals")



#### Shows voiation of Normality so Log Monthly Income  
  
lr.employee.inc <-lm(log(MonthlyIncome) ~ JobLevel\*JobRole + Department + BusinessTravel\*Gender + DistanceFromHome + EducationField\*Education + EnvironmentSatisfaction + HourlyRate\*MonthlyRate + JobInvolvement + JobSatisfaction + MaritalStatus + NumCompaniesWorked + OverTime + RelationshipSatisfaction + StockOptionLevel + TotalWorkingYears\*Age + TrainingTimesLastYear + WorkLifeBalance + YearsSinceLastPromotion + YearsInCurrentRole\*YearsWithCurrManager,data=train)  
  
  
RMSE = sqrt(mean(lr.employee.inc$residuals^2))  
RMSE

## [1] 0.1990818

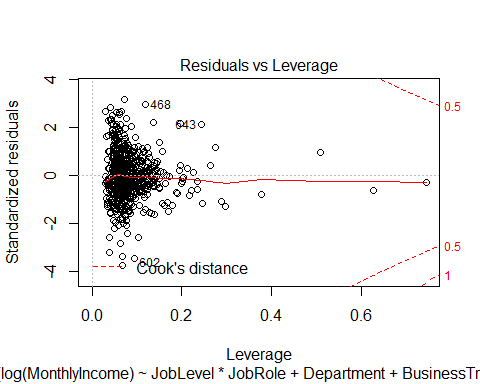
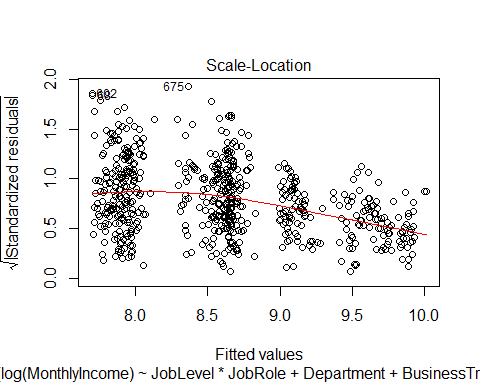
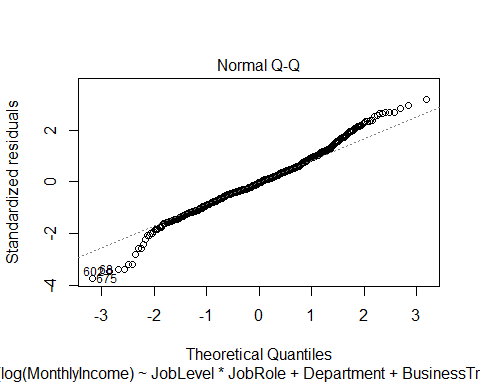
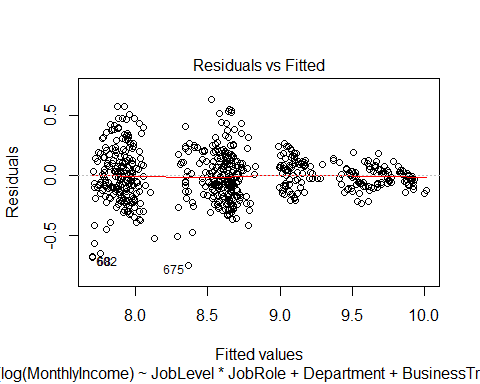
#### Histogram of Residuals  
  
 hist(lr.employee.inc$residuals, col = "blue", main = "Histogram of Residuals")



summary(lr.employee.inc)

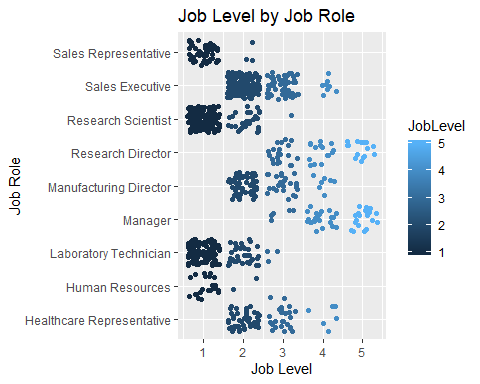
##   
## Call:  
## lm(formula = log(MonthlyIncome) ~ JobLevel \* JobRole + Department +   
## BusinessTravel \* Gender + DistanceFromHome + EducationField \*   
## Education + EnvironmentSatisfaction + HourlyRate \* MonthlyRate +   
## JobInvolvement + JobSatisfaction + MaritalStatus + NumCompaniesWorked +   
## OverTime + RelationshipSatisfaction + StockOptionLevel +   
## TotalWorkingYears \* Age + TrainingTimesLastYear + WorkLifeBalance +   
## YearsSinceLastPromotion + YearsInCurrentRole \* YearsWithCurrManager,   
## data = train)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.74798 -0.11415 -0.00636 0.10787 0.63484   
##   
## Coefficients:  
## Estimate Std. Error t value  
## (Intercept) 7.259e+00 3.178e-01 22.840  
## JobLevel 3.995e-01 4.472e-02 8.932  
## JobRoleHuman Resources -2.913e-01 1.946e-01 -1.497  
## JobRoleLaboratory Technician -2.197e-01 1.226e-01 -1.792  
## JobRoleManager 1.112e+00 2.647e-01 4.199  
## JobRoleManufacturing Director -1.878e-02 1.399e-01 -0.134  
## JobRoleResearch Director 9.425e-01 1.864e-01 5.055  
## JobRoleResearch Scientist -4.559e-01 1.242e-01 -3.671  
## JobRoleSales Executive 6.219e-02 1.475e-01 0.422  
## JobRoleSales Representative -4.671e-01 1.860e-01 -2.511  
## DepartmentResearch & Development 4.288e-02 1.003e-01 0.427  
## DepartmentSales -1.523e-02 1.016e-01 -0.150  
## BusinessTravelTravel\_Frequently -2.483e-02 5.285e-02 -0.470  
## BusinessTravelTravel\_Rarely -3.176e-02 4.570e-02 -0.695  
## GenderMale -5.114e-02 5.344e-02 -0.957  
## DistanceFromHome -2.281e-04 1.064e-03 -0.214  
## EducationFieldLife Sciences 2.528e-01 2.411e-01 1.049  
## EducationFieldMarketing 3.732e-01 2.562e-01 1.457  
## EducationFieldMedical 1.654e-01 2.425e-01 0.682  
## EducationFieldOther 1.163e-01 2.700e-01 0.431  
## EducationFieldTechnical Degree 2.672e-01 2.569e-01 1.040  
## Education 7.224e-02 8.048e-02 0.898  
## EnvironmentSatisfaction -1.759e-02 7.747e-03 -2.270  
## HourlyRate 2.162e-04 9.496e-04 0.228  
## MonthlyRate 2.938e-06 4.042e-06 0.727  
## JobInvolvement 1.148e-02 1.233e-02 0.931  
## JobSatisfaction -7.512e-03 7.497e-03 -1.002  
## MaritalStatusMarried 6.230e-03 2.174e-02 0.287  
## MaritalStatusSingle 2.311e-02 3.007e-02 0.768  
## NumCompaniesWorked 1.326e-03 3.838e-03 0.346  
## OverTimeYes 4.002e-02 1.939e-02 2.064  
## RelationshipSatisfaction -4.694e-03 7.544e-03 -0.622  
## StockOptionLevel -3.258e-03 1.289e-02 -0.253  
## TotalWorkingYears 3.172e-02 6.924e-03 4.581  
## Age 4.339e-03 1.830e-03 2.370  
## TrainingTimesLastYear -3.516e-03 6.751e-03 -0.521  
## WorkLifeBalance 1.872e-03 1.195e-02 0.157  
## YearsSinceLastPromotion -2.711e-03 3.290e-03 -0.824  
## YearsInCurrentRole 1.247e-02 4.825e-03 2.585  
## YearsWithCurrManager 1.770e-03 4.859e-03 0.364  
## JobLevel:JobRoleHuman Resources 1.321e-01 1.051e-01 1.257  
## JobLevel:JobRoleLaboratory Technician -3.605e-02 5.929e-02 -0.608  
## JobLevel:JobRoleManager -2.259e-01 6.979e-02 -3.237  
## JobLevel:JobRoleManufacturing Director 2.161e-02 5.470e-02 0.395  
## JobLevel:JobRoleResearch Director -2.019e-01 5.737e-02 -3.519  
## JobLevel:JobRoleResearch Scientist 1.568e-01 6.273e-02 2.499  
## JobLevel:JobRoleSales Executive -2.611e-03 5.266e-02 -0.050  
## JobLevel:JobRoleSales Representative 2.028e-01 1.237e-01 1.640  
## BusinessTravelTravel\_Frequently:GenderMale 3.531e-02 6.744e-02 0.524  
## BusinessTravelTravel\_Rarely:GenderMale 9.679e-02 5.711e-02 1.695  
## EducationFieldLife Sciences:Education -6.753e-02 8.105e-02 -0.833  
## EducationFieldMarketing:Education -1.075e-01 8.463e-02 -1.271  
## EducationFieldMedical:Education -4.456e-02 8.172e-02 -0.545  
## EducationFieldOther:Education -1.745e-02 9.020e-02 -0.193  
## EducationFieldTechnical Degree:Education -7.508e-02 8.621e-02 -0.871  
## HourlyRate:MonthlyRate -3.352e-08 5.843e-08 -0.574  
## TotalWorkingYears:Age -5.213e-04 1.380e-04 -3.778  
## YearsInCurrentRole:YearsWithCurrManager -1.220e-03 6.671e-04 -1.830  
## Pr(>|t|)   
## (Intercept) < 2e-16 \*\*\*  
## JobLevel < 2e-16 \*\*\*  
## JobRoleHuman Resources 0.134810   
## JobRoleLaboratory Technician 0.073546 .   
## JobRoleManager 3.06e-05 \*\*\*  
## JobRoleManufacturing Director 0.893219   
## JobRoleResearch Director 5.62e-07 \*\*\*  
## JobRoleResearch Scientist 0.000262 \*\*\*  
## JobRoleSales Executive 0.673498   
## JobRoleSales Representative 0.012269 \*   
## DepartmentResearch & Development 0.669164   
## DepartmentSales 0.880867   
## BusinessTravelTravel\_Frequently 0.638643   
## BusinessTravelTravel\_Rarely 0.487421   
## GenderMale 0.338927   
## DistanceFromHome 0.830355   
## EducationFieldLife Sciences 0.294672   
## EducationFieldMarketing 0.145696   
## EducationFieldMedical 0.495600   
## EducationFieldOther 0.666780   
## EducationFieldTechnical Degree 0.298840   
## Education 0.369715   
## EnvironmentSatisfaction 0.023522 \*   
## HourlyRate 0.820003   
## MonthlyRate 0.467669   
## JobInvolvement 0.352011   
## JobSatisfaction 0.316744   
## MaritalStatusMarried 0.774559   
## MaritalStatusSingle 0.442493   
## NumCompaniesWorked 0.729798   
## OverTimeYes 0.039449 \*   
## RelationshipSatisfaction 0.534033   
## StockOptionLevel 0.800567   
## TotalWorkingYears 5.58e-06 \*\*\*  
## Age 0.018066 \*   
## TrainingTimesLastYear 0.602716   
## WorkLifeBalance 0.875553   
## YearsSinceLastPromotion 0.410237   
## YearsInCurrentRole 0.009948 \*\*   
## YearsWithCurrManager 0.715778   
## JobLevel:JobRoleHuman Resources 0.209142   
## JobLevel:JobRoleLaboratory Technician 0.543441   
## JobLevel:JobRoleManager 0.001271 \*\*   
## JobLevel:JobRoleManufacturing Director 0.692975   
## JobLevel:JobRoleResearch Director 0.000464 \*\*\*  
## JobLevel:JobRoleResearch Scientist 0.012702 \*   
## JobLevel:JobRoleSales Executive 0.960480   
## JobLevel:JobRoleSales Representative 0.101477   
## BusinessTravelTravel\_Frequently:GenderMale 0.600745   
## BusinessTravelTravel\_Rarely:GenderMale 0.090563 .   
## EducationFieldLife Sciences:Education 0.405017   
## EducationFieldMarketing:Education 0.204253   
## EducationFieldMedical:Education 0.585787   
## EducationFieldOther:Education 0.846660   
## EducationFieldTechnical Degree:Education 0.384109   
## HourlyRate:MonthlyRate 0.566363   
## TotalWorkingYears:Age 0.000173 \*\*\*  
## YearsInCurrentRole:YearsWithCurrManager 0.067782 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.2079 on 638 degrees of freedom  
## Multiple R-squared: 0.9092, Adjusted R-squared: 0.9011   
## F-statistic: 112.1 on 57 and 638 DF, p-value: < 2.2e-16

plot(lr.employee.inc)

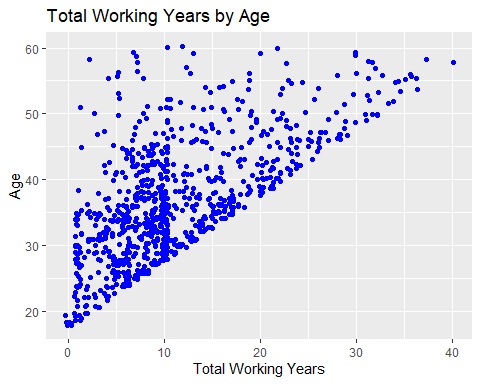


Show detailed plots of Interactions

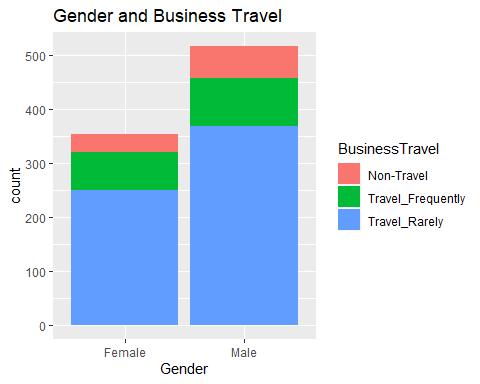
##### Plot the Interactions #####  
  
  
  
employee %>% ggplot(aes(x = JobLevel,y=JobRole,col=JobLevel)) + geom\_point(pos='Jitter') + ggtitle("Job Level by Job Role ") +  
 xlab ("Job Level") +  
 ylab ("Job Role")



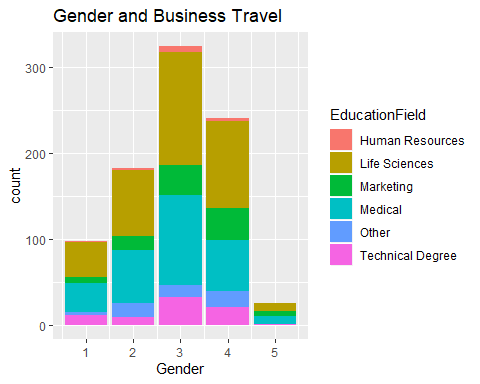
employee %>% ggplot(aes(x = TotalWorkingYears,y=Age)) + geom\_point(col='Blue',pos='Jitter') + ggtitle("Total Working Years by Age ") +  
 xlab ("Total Working Years") +  
 ylab ("Age")



employee %>% ggplot(aes(x = Gender, fill=BusinessTravel)) + geom\_bar(stat='count') +  
 xlab ("Gender") + ggtitle ("Gender and Business Travel")



employee %>% ggplot(aes(x = Education, fill=EducationField)) + geom\_bar(stat='count') +  
 xlab ("Gender") + ggtitle ("Gender and Business Travel")

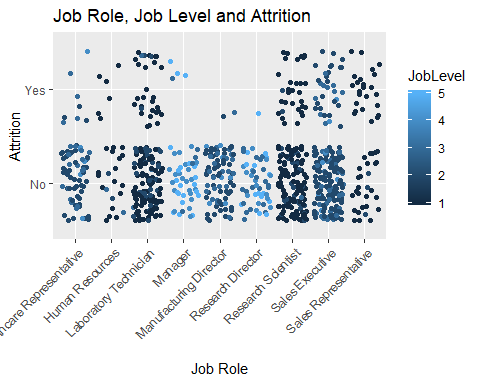


Run the Linear Regression Employee Salary Predictions

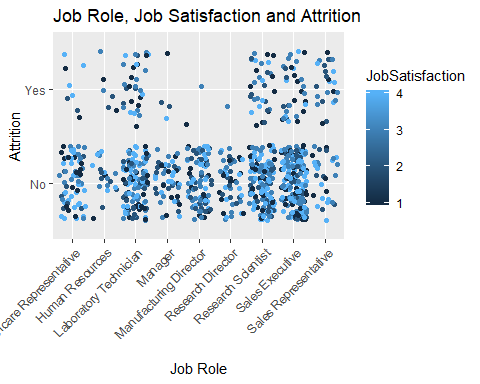
###### Linear Model Predictions  
lr.employee.predI <- data.frame(predict(lr.employee.inc, newdata = test, type = "response"))  
newdf = NewIncome=data.frame(exp(lr.employee.predI))  
  
   
#### Run on competition dataset  
  
lr.employee.predI <- data.frame(predict(lr.employee.inc, newdata = employeenoinc, type = "response"))  
preddf = NewIncome=data.frame(exp(lr.employee.predI))  
names(preddf) <- c("PredictIncome")  
IncomePred <- cbind(employeenoinc[1],preddf)  
  
write.csv(IncomePred,'c:/School Stuff/DS/Doing DS/Project2/Case2PredictionsLull Salary.csv')

Managers and Directors Rarely Leave but Job Satisfaction Rating is not Significant

employee %>% ggplot(aes(x = JobRole,y=Attrition,col=JobLevel)) + geom\_point(pos='Jitter') + ggtitle("Job Role, Job Level and Attrition") +  
 xlab ("Job Role") +  
 ylab ("Attrition") + theme(axis.text.x=element\_text(angle=45, hjust=1))



employee %>% ggplot(aes(x = JobRole,y=Attrition,col=JobSatisfaction)) + geom\_point(pos='Jitter') + ggtitle("Job Role, Job Satisfaction and Attrition") +  
 xlab ("Job Role") +  
 ylab ("Attrition") + theme(axis.text.x=element\_text(angle=45, hjust=1))

 Run a T-Test to show that Job Satisfaction is < 3 for Manager, Manufacturing Director and Research Directors with Job Satisfaction > 3

newemp <- employee %>% filter(JobLevel>3 & JobRole %in% c("Manager", "Manufacturing Director","Research Director"))  
  
t.test(newemp$JobSatisfaction, var.equal=F, mu=2,alternative='less')

##   
## One Sample t-test  
##   
## data: newemp$JobSatisfaction  
## t = 4.918, df = 82, p-value = 1  
## alternative hypothesis: true mean is less than 2  
## 95 percent confidence interval:  
## -Inf 2.790066  
## sample estimates:  
## mean of x   
## 2.590361

### Top 3 Categories Of the varaibles picked in the best model, these changed the sensitivity and specificity the least

AttritionYes = employee %>% filter(Attrition == "Yes")  
AttritionNo = employee %>% filter(Attrition == "No")  
  
### Same test train split as before. Picked top most influential Variables  
set.seed(9)  
trainInd = sample(seq(1,dim(AttritionYes)[1],1),round(.7\*dim(AttritionYes)[1]))  
trainYES = AttritionYes[trainInd,]  
testYES = AttritionYes[-trainInd,]  
  
trainInd = sample(seq(1,dim(AttritionNo)[1],1),round(.7\*dim(AttritionNo)[1]))  
train = AttritionNo[trainInd,]  
test = AttritionNo[-trainInd,]  
  
train = rbind(train,trainYES)  
test = rbind(test,testYES)  
   
  
   
lr.employee <-glm(Attrition~TotalWorkingYears + JobRole + WorkLifeBalance,data=train,family=binomial(link="logit"))  
  
lr.employee.pred2 <- data.frame(predict(lr.employee, newdata = test, type = "response"))  
lr.employee.pred2 = lr.employee.pred2 %>% mutate(pred = ifelse(lr.employee.pred2 <0.25, "No", "Yes"))  
  
table(lr.employee.pred2$pred)

##   
## No Yes   
## 220 41

predtble = as.factor(lr.employee.pred2$pred)  
predtble <-relevel(predtble, ref = "No")   
Truth<-test$Attrition  
confmtx = as.matrix(table(predtble,Truth))  
CM = confusionMatrix(confmtx)  
CM

## Confusion Matrix and Statistics  
##   
## Truth  
## predtble No Yes  
## No 196 24  
## Yes 23 18  
##   
## Accuracy : 0.8199   
## 95% CI : (0.7678, 0.8646)  
## No Information Rate : 0.8391   
## P-Value [Acc > NIR] : 0.8237   
##   
## Kappa : 0.3267   
##   
## Mcnemar's Test P-Value : 1.0000   
##   
## Sensitivity : 0.8950   
## Specificity : 0.4286   
## Pos Pred Value : 0.8909   
## Neg Pred Value : 0.4390   
## Prevalence : 0.8391   
## Detection Rate : 0.7510   
## Detection Prevalence : 0.8429   
## Balanced Accuracy : 0.6618   
##   
## 'Positive' Class : No   
##

Run a Naive Bayes on top 3 factors

employee$WorkYrFactor = cut(employee$TotalWorkingYears, breaks = c(0,10,20,30,50), labels = c("< 10","10-20", "20-30","Above 30"))  
  
nbemploy <- data.frame(WorkLifeBalance = factor(employee$WorkLifeBalance),JobLevel = factor(employee$JobLevel),employee$WorkYrFactor,Attrition=employee$Attrition)  
  
model = naiveBayes(Attrition~.,data = nbemploy)  
tp=predict(model,nbemploy[,c('Attrition')])  
Truth=employee$Attrition  
confmtx = as.matrix(table(tp,Truth))  
CM = confusionMatrix(confmtx)  
CM

## Confusion Matrix and Statistics  
##   
## Truth  
## tp No Yes  
## No 730 140  
## Yes 0 0  
##   
## Accuracy : 0.8391   
## 95% CI : (0.8129, 0.8629)  
## No Information Rate : 0.8391   
## P-Value [Acc > NIR] : 0.5225   
##   
## Kappa : 0   
##   
## Mcnemar's Test P-Value : <2e-16   
##   
## Sensitivity : 1.0000   
## Specificity : 0.0000   
## Pos Pred Value : 0.8391   
## Neg Pred Value : NaN   
## Prevalence : 0.8391   
## Detection Rate : 0.8391   
## Detection Prevalence : 1.0000   
## Balanced Accuracy : 0.5000   
##   
## 'Positive' Class : No   
##