Casestudy2DDS

Dylan Scott

4/6/2021

## Youtube link

Dylan’s Case study 2 Video

## Project Overview

## Data Breakdown

870 observation with 36 variables

library(knitr)  
library(tidyverse)

## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.3 v purrr 0.3.4  
## v tibble 3.1.0 v dplyr 1.0.5  
## v tidyr 1.1.3 v stringr 1.4.0  
## v readr 1.4.0 v forcats 0.5.1

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

library(readxl)  
library(curl)

##   
## Attaching package: 'curl'

## The following object is masked from 'package:readr':  
##   
## parse\_date

library(GGally)

## Registered S3 method overwritten by 'GGally':  
## method from   
## +.gg ggplot2

library(e1071)  
library(caret)

## Loading required package: lattice

##   
## Attaching package: 'caret'

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

library(ggcorrplot)  
library(gridExtra)

##   
## Attaching package: 'gridExtra'

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

library(reticulate)  
library(gt)  
  
  
data <-read.csv(curl("https://raw.githubusercontent.com/Scottdyl/CaseStudy2DDS/main/data/CaseStudy2-data.csv"))  
  
#summary(data)  
#no missing values found  
  
#convert raw data variables to factors  
data$Education = as.factor(data$Education)  
data$EnvironmentSatisfaction = as.factor(data$EnvironmentSatisfaction)  
data$JobInvolvement = as.factor(data$JobInvolvement)  
data$JobLevel = as.factor(data$JobLevel)  
data$JobSatisfaction = as.factor(data$JobSatisfaction)  
data$PerformanceRating = as.factor(data$PerformanceRating)  
data$RelationshipSatisfaction = as.factor(data$RelationshipSatisfaction)  
data$WorkLifeBalance = as.factor(data$WorkLifeBalance)  
  
  
# no attr data  
Comp\_attr <- read.csv(curl("https://raw.githubusercontent.com/Scottdyl/CaseStudy2DDS/main/data/CaseStudy2CompSet%20No%20Attrition.csv"))  
Comp\_attr$Education = as.factor(Comp\_attr$Education)  
Comp\_attr$EnvironmentSatisfaction = as.factor(Comp\_attr$EnvironmentSatisfaction)  
Comp\_attr$JobInvolvement = as.factor(Comp\_attr$JobInvolvement)  
Comp\_attr$JobLevel = as.factor(Comp\_attr$JobLevel)  
Comp\_attr$JobSatisfaction = as.factor(Comp\_attr$JobSatisfaction)  
Comp\_attr$PerformanceRating = as.factor(Comp\_attr$PerformanceRating)  
Comp\_attr$RelationshipSatisfaction = as.factor(Comp\_attr$RelationshipSatisfaction)  
Comp\_attr$WorkLifeBalance = as.factor(Comp\_attr$WorkLifeBalance)  
  
  
# no sal data  
Comp\_sal <-read.csv(curl("https://raw.githubusercontent.com/Scottdyl/CaseStudy2DDS/main/data/CaseStudy2CompSet%20No%20Salary.csv"))  
Comp\_sal$Education = as.factor(Comp\_sal$Education)  
Comp\_sal$EnvironmentSatisfaction = as.factor(Comp\_sal$EnvironmentSatisfaction)  
Comp\_sal$JobInvolvement = as.factor(Comp\_sal$JobInvolvement)  
Comp\_sal$JobLevel = as.factor(Comp\_sal$JobLevel)  
Comp\_sal$JobSatisfaction = as.factor(Comp\_sal$JobSatisfaction)  
Comp\_sal$PerformanceRating = as.factor(Comp\_sal$PerformanceRating)  
Comp\_sal$RelationshipSatisfaction = as.factor(Comp\_sal$RelationshipSatisfaction)  
Comp\_sal$WorkLifeBalance = as.factor(Comp\_sal$WorkLifeBalance)  
  
head(data)

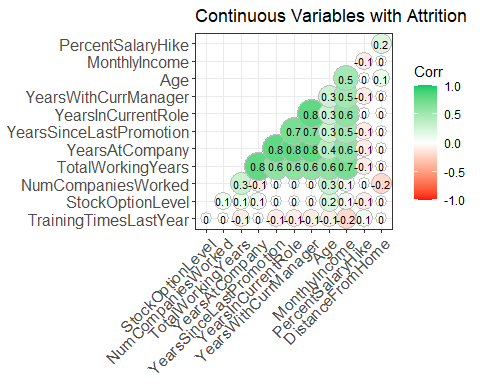
## ID Age Attrition BusinessTravel DailyRate Department  
## 1 1 32 No Travel\_Rarely 117 Sales  
## 2 2 40 No Travel\_Rarely 1308 Research & Development  
## 3 3 35 No Travel\_Frequently 200 Research & Development  
## 4 4 32 No Travel\_Rarely 801 Sales  
## 5 5 24 No Travel\_Frequently 567 Research & Development  
## 6 6 27 No Travel\_Frequently 294 Research & Development  
## DistanceFromHome Education EducationField EmployeeCount EmployeeNumber  
## 1 13 4 Life Sciences 1 859  
## 2 14 3 Medical 1 1128  
## 3 18 2 Life Sciences 1 1412  
## 4 1 4 Marketing 1 2016  
## 5 2 1 Technical Degree 1 1646  
## 6 10 2 Life Sciences 1 733  
## EnvironmentSatisfaction Gender HourlyRate JobInvolvement JobLevel  
## 1 2 Male 73 3 2  
## 2 3 Male 44 2 5  
## 3 3 Male 60 3 3  
## 4 3 Female 48 3 3  
## 5 1 Female 32 3 1  
## 6 4 Male 32 3 3  
## JobRole JobSatisfaction MaritalStatus MonthlyIncome  
## 1 Sales Executive 4 Divorced 4403  
## 2 Research Director 3 Single 19626  
## 3 Manufacturing Director 4 Single 9362  
## 4 Sales Executive 4 Married 10422  
## 5 Research Scientist 4 Single 3760  
## 6 Manufacturing Director 1 Divorced 8793  
## MonthlyRate NumCompaniesWorked Over18 OverTime PercentSalaryHike  
## 1 9250 2 Y No 11  
## 2 17544 1 Y No 14  
## 3 19944 2 Y No 11  
## 4 24032 1 Y No 19  
## 5 17218 1 Y Yes 13  
## 6 4809 1 Y No 21  
## PerformanceRating RelationshipSatisfaction StandardHours StockOptionLevel  
## 1 3 3 80 1  
## 2 3 1 80 0  
## 3 3 3 80 0  
## 4 3 3 80 2  
## 5 3 3 80 0  
## 6 4 3 80 2  
## TotalWorkingYears TrainingTimesLastYear WorkLifeBalance YearsAtCompany  
## 1 8 3 2 5  
## 2 21 2 4 20  
## 3 10 2 3 2  
## 4 14 3 3 14  
## 5 6 2 3 6  
## 6 9 4 2 9  
## YearsInCurrentRole YearsSinceLastPromotion YearsWithCurrManager  
## 1 2 0 3  
## 2 7 4 9  
## 3 2 2 2  
## 4 10 5 7  
## 5 3 1 3  
## 6 7 1 7

## exploring all continious varables.

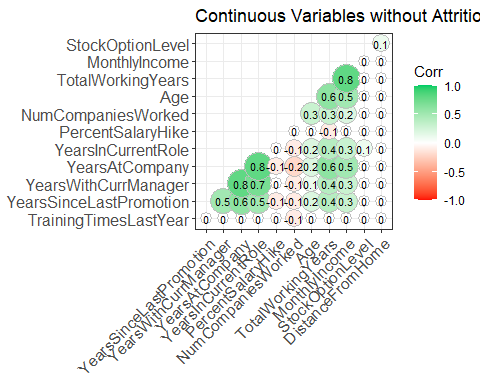
## Plotting and looking at all continious varbles. The most significant will be shown in the next chunk.

#commented out least impactful looking varables  
#data %>% ggplot(aes(Age))+geom\_density(aes(fill=Attrition))  
#data %>% ggplot(aes(YearsSinceLastPromotion))+geom\_density(aes(fill=Attrition))  
#data %>% ggplot(aes(TrainingTimesLastYear))+geom\_density(aes(fill=Attrition))  
#data %>% ggplot(aes(StockOptionLevel))+geom\_density(aes(fill=Attrition))  
#data %>% ggplot(aes(PercentSalaryHike))+geom\_density(aes(fill=Attrition))  
#data %>% ggplot(aes(NumCompaniesWorked))+geom\_density(aes(fill=Attrition))  
#data %>% ggplot(aes(MonthlyRate))+geom\_density(aes(fill=Attrition))  
#data %>% ggplot(aes(HourlyRate))+geom\_density(aes(fill=Attrition))  
#data %>% ggplot(aes(DailyRate))+geom\_density(aes(fill=Attrition))  
#data %>% ggplot(aes(DistanceFromHome))+geom\_density(aes(fill=Attrition))

#summary(as.matrix(data))  
corr\_data\_y <- data %>% filter(Attrition =="Yes") %>% select(Age, TotalWorkingYears, YearsAtCompany, YearsSinceLastPromotion, YearsInCurrentRole, YearsWithCurrManager, MonthlyIncome,TrainingTimesLastYear,StockOptionLevel,PercentSalaryHike,NumCompaniesWorked,DistanceFromHome)  
  
corr <- round(cor(corr\_data\_y),1)  
  
ggcorrplot(corr, hc.order = TRUE,   
 type = "lower",   
 lab = TRUE,   
 lab\_size = 3,   
 method="circle",   
 colors = c("red", "white", "springgreen3"),   
 title="Continuous Variables with Attrition",   
 ggtheme=theme\_bw)



corr\_data\_n <- data %>% filter(Attrition =="No") %>% select(Age, TotalWorkingYears, YearsAtCompany, YearsSinceLastPromotion, YearsInCurrentRole, YearsWithCurrManager, MonthlyIncome,TrainingTimesLastYear,StockOptionLevel,PercentSalaryHike,NumCompaniesWorked,DistanceFromHome)  
  
corr <- round(cor(corr\_data\_n),1)  
  
ggcorrplot(corr, hc.order = TRUE,   
 type = "lower",   
 lab = TRUE,   
 lab\_size = 3,   
 method="circle",   
 colors = c("red", "white", "springgreen3"),   
 title="Continuous Variables without Attrition",   
 ggtheme=theme\_bw)

 ## TTest - testing the 5 most significant varables

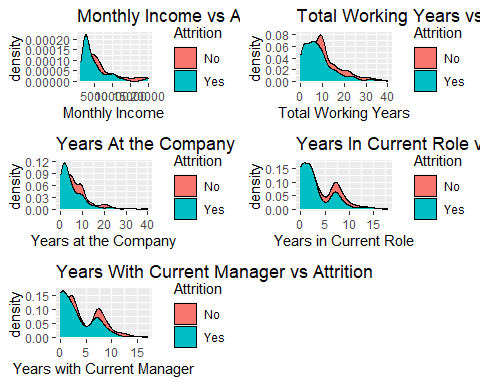
# we will be using a T test here because we would like to know: does this variable lead to a difference in attrition?  
# if we have a samll P value that means this varable does show a difference in means betwwen attrition and not.  
  
att\_mi <- data %>% filter(Attrition=="Yes") %>% select(MonthlyIncome)  
stay\_mi <- data %>% filter(Attrition=="No") %>% select(MonthlyIncome)  
mi\_t <-t.test(att\_mi, stay\_mi, alternative="two.sided")  
  
att\_twy <- data %>% filter(Attrition=="Yes") %>% select(TotalWorkingYears)  
stay\_twy <- data %>% filter(Attrition=="No") %>% select(TotalWorkingYears)  
twy\_t <-t.test(att\_twy, stay\_twy, alternative="two.sided")  
  
att\_yicr <- data %>% filter(Attrition=="Yes") %>% select(YearsInCurrentRole)  
stay\_yicr <- data %>% filter(Attrition=="No") %>% select(YearsInCurrentRole)  
yicr <- t.test(att\_yicr, stay\_yicr, alternative="two.sided")  
  
att\_ywcm <- data %>% filter(Attrition=="Yes") %>% select(YearsWithCurrManager)  
stay\_ywcm <- data %>% filter(Attrition=="No") %>% select(YearsWithCurrManager)  
ywcm <- t.test(att\_ywcm, stay\_ywcm, alternative="two.sided")  
  
att\_yac <- data %>% filter(Attrition=="Yes") %>% select(YearsAtCompany)  
stay\_yac <- data %>% filter(Attrition=="No") %>% select(YearsAtCompany)  
yac <- t.test(att\_yac, stay\_yac, alternative="two.sided")  
  
cont\_var = c("MonthlyIncome", "TotalWorkingYears","YearsInCurrentRole","YearsWithCurrManager","YearsAtCompany")  
ttest\_p = c(mi\_t$p.value, twy\_t$p.value, yicr$p.value, ywcm$p.value, yac$p.value)  
  
df1\_ttest = data.frame(Variable=cont\_var, "T-Test pvalue"=ttest\_p)  
gt(df1\_ttest)

df1\_ttest

## Variable T.Test.pvalue  
## 1 MonthlyIncome 2.412488e-07  
## 2 TotalWorkingYears 6.595682e-07  
## 3 YearsInCurrentRole 1.522152e-06  
## 4 YearsWithCurrManager 5.084229e-06  
## 5 YearsAtCompany 2.563021e-04

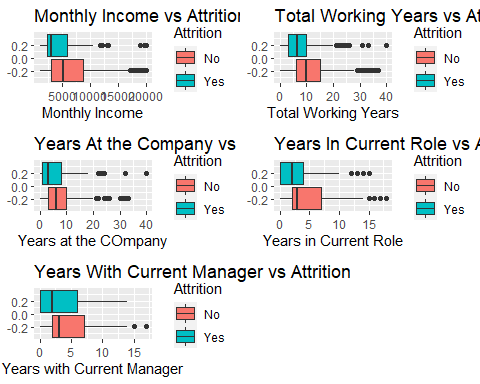
## showing off the most significant continious varables.

#these varables appear to be significant  
mi\_plot <-data %>% ggplot(aes(MonthlyIncome))+  
 geom\_density(aes(fill=Attrition))+  
 labs(title="Monthly Income vs Attrition")+  
 xlab("Monthly Income")  
  
twy\_plot <-data %>% ggplot(aes(TotalWorkingYears))+  
 geom\_density(aes(fill=Attrition))+  
 labs(title="Total Working Years vs Attrition")+  
 xlab("Total Working Years")  
  
yac\_plot <- data %>% ggplot(aes(YearsAtCompany))+  
 geom\_density(aes(fill=Attrition))+  
 labs(title="Years At the Company vs Attrition")+  
 xlab("Years at the Company")  
  
yicr\_plot <- data %>% ggplot(aes(YearsInCurrentRole))+  
 geom\_density(aes(fill=Attrition))+  
 labs(title="Years In Current Role vs Attrition")+  
 xlab("Years in Current Role")  
  
ywcm\_plot <- data %>% ggplot(aes(YearsWithCurrManager))+  
 geom\_density(aes(fill=Attrition))+  
 labs(title="Years With Current Manager vs Attrition")+  
 xlab("Years with Current Manager")  
  
grid.arrange(mi\_plot,twy\_plot,yac\_plot,yicr\_plot,ywcm\_plot)



# show both density and box charts

# show the same data as above but in boxplot  
mi\_plot\_box <-data %>% ggplot(aes(MonthlyIncome))+  
 geom\_boxplot(aes(fill=Attrition))+  
 labs(title="Monthly Income vs Attrition")+  
 xlab("Monthly Income")  
  
twy\_plot\_box <-data %>% ggplot(aes(TotalWorkingYears))+  
 geom\_boxplot(aes(fill=Attrition))+  
 labs(title="Total Working Years vs Attrition")+  
 xlab("Total Working Years")  
  
yac\_plot\_box <- data %>% ggplot(aes(YearsAtCompany))+  
 geom\_boxplot(aes(fill=Attrition))+  
 labs(title="Years At the Company vs Attrition")+  
 xlab("Years at the COmpany")  
  
yicr\_plot\_box <- data %>% ggplot(aes(YearsInCurrentRole))+  
 geom\_boxplot(aes(fill=Attrition))+  
 labs(title="Years In Current Role vs Attrition")+  
 xlab("Years in Current Role")  
  
ywcm\_plot\_box <- data %>% ggplot(aes(YearsWithCurrManager))+  
 geom\_boxplot(aes(fill=Attrition))+  
 labs(title="Years With Current Manager vs Attrition")+  
 xlab("Years with Current Manager")  
  
grid.arrange(mi\_plot\_box,twy\_plot\_box,yac\_plot\_box,yicr\_plot\_box,ywcm\_plot\_box)



## wrangling categorical varables

## calculating the proportion

## varables with little relationship

#data %>% group\_by(PerformanceRating) %>% count(Attrition) %>% mutate(sum=sum(n)) %>% mutate(proportion=n/sum\*100) %>% filter(Attrition=="Yes")  
#data %>% group\_by(RelationshipSatisfaction) %>% count(Attrition) %>% mutate(sum=sum(n)) %>% mutate(proportion=n/sum\*100) %>% filter(Attrition=="Yes")  
#data %>% group\_by(Education) %>% count(Attrition) %>% mutate(sum=sum(n)) %>% mutate(proportion=n/sum\*100) %>% filter(Attrition=="Yes")  
#data %>% group\_by(EducationField) %>% count(Attrition) %>% mutate(sum=sum(n)) %>% mutate(proportion=n/sum\*100) %>% filter(Attrition=="Yes")  
#data %>% group\_by(Gender) %>% count(Attrition) %>% mutate(sum=sum(n)) %>% mutate(proportion=n/sum\*100) %>% filter(Attrition=="Yes")  
#data %>% group\_by(BusinessTravel) %>% count(Attrition) %>% mutate(sum=sum(n)) %>% mutate(proportion=n/sum\*100) %>% filter(Attrition=="Yes")  
#data %>% group\_by(Department) %>% count(Attrition) %>% mutate(sum=sum(n)) %>% mutate(proportion=n/sum\*100) %>% filter(Attrition=="Yes")  
#data %>% group\_by(WorkLifeBalance) %>% count(Attrition) %>% mutate(sum=sum(n)) %>% mutate(proportion=n/sum\*100) %>% filter(Attrition=="Yes")  
#data %>% group\_by(JobSatisfaction) %>% count(Attrition) %>% mutate(sum=sum(n)) %>% mutate(proportion=n/sum\*100) %>% filter(Attrition=="Yes")

## Categorical varables that show a relatinship

data %>% group\_by(JobLevel) %>%   
 count(Attrition) %>%   
 mutate(sum=sum(n)) %>%   
 mutate(proportion=n/sum\*100) %>%   
 filter(Attrition=="Yes")

## # A tibble: 5 x 5  
## # Groups: JobLevel [5]  
## JobLevel Attrition n sum proportion  
## <fct> <chr> <int> <int> <dbl>  
## 1 1 Yes 86 329 26.1   
## 2 2 Yes 30 312 9.62  
## 3 3 Yes 17 132 12.9   
## 4 4 Yes 3 60 5   
## 5 5 Yes 4 37 10.8

data %>% group\_by(OverTime) %>%   
 count(Attrition) %>%   
 mutate(sum=sum(n)) %>%   
 mutate(proportion=n/sum\*100) %>%   
 filter(Attrition=="Yes")

## # A tibble: 2 x 5  
## # Groups: OverTime [2]  
## OverTime Attrition n sum proportion  
## <chr> <chr> <int> <int> <dbl>  
## 1 No Yes 60 618 9.71  
## 2 Yes Yes 80 252 31.7

data %>% group\_by(JobInvolvement) %>%   
 count(Attrition) %>%   
 mutate(sum=sum(n)) %>%   
 mutate(proportion=n/sum\*100) %>%   
 filter(Attrition=="Yes")

## # A tibble: 4 x 5  
## # Groups: JobInvolvement [4]  
## JobInvolvement Attrition n sum proportion  
## <fct> <chr> <int> <int> <dbl>  
## 1 1 Yes 22 47 46.8   
## 2 2 Yes 44 228 19.3   
## 3 3 Yes 67 514 13.0   
## 4 4 Yes 7 81 8.64

data %>% group\_by(JobRole) %>%   
 count(Attrition) %>%   
 mutate(sum=sum(n)) %>%   
 mutate(proportion=n/sum\*100) %>%   
 filter(Attrition=="Yes")

## # A tibble: 9 x 5  
## # Groups: JobRole [9]  
## JobRole Attrition n sum proportion  
## <chr> <chr> <int> <int> <dbl>  
## 1 Healthcare Representative Yes 8 76 10.5   
## 2 Human Resources Yes 6 27 22.2   
## 3 Laboratory Technician Yes 30 153 19.6   
## 4 Manager Yes 4 51 7.84  
## 5 Manufacturing Director Yes 2 87 2.30  
## 6 Research Director Yes 1 51 1.96  
## 7 Research Scientist Yes 32 172 18.6   
## 8 Sales Executive Yes 33 200 16.5   
## 9 Sales Representative Yes 24 53 45.3

data %>% group\_by(MaritalStatus) %>%   
 count(Attrition) %>%   
 mutate(sum=sum(n)) %>%   
 mutate(proportion=n/sum\*100) %>%   
 filter(Attrition=="Yes")

## # A tibble: 3 x 5  
## # Groups: MaritalStatus [3]  
## MaritalStatus Attrition n sum proportion  
## <chr> <chr> <int> <int> <dbl>  
## 1 Divorced Yes 12 191 6.28  
## 2 Married Yes 58 410 14.1   
## 3 Single Yes 70 269 26.0

## Plotting the Categorical varables

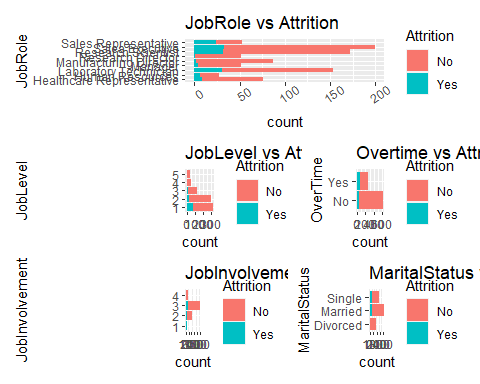
library(patchwork)  
  
JL\_chi <- chisq.test(data$JobLevel, data$Attrition)  
O\_chi <- chisq.test(data$OverTime, data$Attrition)  
JI\_chi <- chisq.test(data$JobInvolvement, data$Attrition)  
JR\_chi <- chisq.test(data$JobRole, data$Attrition)

## Warning in chisq.test(data$JobRole, data$Attrition): Chi-squared approximation  
## may be incorrect

MS\_chi <- chisq.test(data$MaritalStatus, data$Attrition)  
  
cat\_var = c("JobLevel", "OverTime", "JobInvolvement", "JobRole", "MaritalStatus")  
chi\_p = c(JL\_chi$p.value, O\_chi$p.value, JI\_chi$p.value, JR\_chi$p.value, MS\_chi$p.value)  
df\_chitest = data.frame(Variable=cat\_var, Chisq.pvalue=chi\_p)  
  
JL\_plot <- data %>% ggplot(aes(JobLevel))+geom\_bar(aes(fill=Attrition)) + labs(title="JobLevel vs Attrition")+coord\_flip()  
O\_plot <- data %>% ggplot(aes(OverTime))+geom\_bar(aes(fill=Attrition)) + labs(title="Overtime vs Attrition")+coord\_flip()  
JI\_plot <- data %>% ggplot(aes(JobInvolvement))+geom\_bar(aes(fill=Attrition)) + labs(title="JobInvolvement vs Attrition")+coord\_flip()  
JS\_plot <- data %>% ggplot(aes(JobSatisfaction))+geom\_bar(aes(fill=Attrition)) + labs(title="JobSatisfaction vs Attrition")+coord\_flip()  
JR\_plot <-data %>% ggplot(aes(JobRole))+geom\_bar(aes(fill=Attrition)) + labs(title="JobRole vs Attrition")+ coord\_flip()+ theme(axis.text.x = element\_text(angle = 35))+ theme(axis.text.y = element\_text(vjust = .5,hjust = 1))  
MS\_plot <- data %>% ggplot(aes(MaritalStatus))+geom\_bar(aes(fill=Attrition)) + labs(title="MaritalStatus vs Attrition")+coord\_flip()  
  
#Chi square test to determine correlation  
JL\_chi <- chisq.test(data$JobLevel, data$Attrition)  
O\_chi <- chisq.test(data$OverTime, data$Attrition)  
JI\_chi <- chisq.test(data$JobInvolvement, data$Attrition)  
JR\_chi <- chisq.test(data$JobRole, data$Attrition)

## Warning in chisq.test(data$JobRole, data$Attrition): Chi-squared approximation  
## may be incorrect

MS\_chi <- chisq.test(data$MaritalStatus, data$Attrition)  
  
cat\_var = c("JobLevel", "OverTime", "JobInvolvement", "JobRole", "MaritalStatus")  
chi\_p = c(JL\_chi$p.value, O\_chi$p.value, JI\_chi$p.value, JR\_chi$p.value, MS\_chi$p.value)  
df\_chitest = data.frame(Variable=cat\_var, Chisq.pvalue=chi\_p)  
  
JR\_plot+(JL\_plot+ O\_plot)+ (JI\_plot+MS\_plot)+ plot\_layout(ncol = 1)



#grid.arrange(JL\_plot, O\_plot, JI\_plot, JR\_plot, MS\_plot,  
# widths = c(1,1,1),c(1,1,1),layout\_matrix = rbind(c(1, 2, 3))),c(4,5,5)))  
df\_chitest

## Variable Chisq.pvalue  
## 1 JobLevel 2.084703e-08  
## 2 OverTime 2.332981e-15  
## 3 JobInvolvement 5.211041e-09  
## 4 JobRole 3.646836e-10  
## 5 MaritalStatus 3.378946e-08

##NaiveBayes variables were:  
\* OverTime  
\* JobRole  
\* JobInvolvement  
\* MonthlyIncome  
\* TotalWorkingYears  
\* YearsInCurrentRole  
\* JobLevel

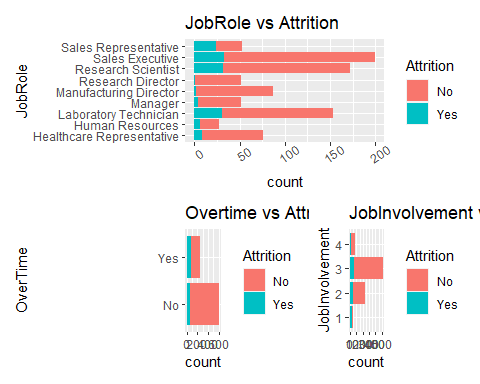
# Goal: reach a sensitivity and specificty greater than 60%  
# add varables one by one until a desiered result is acheived  
set.seed(12)  
splitPerc = .75  
  
trainindex = sample(seq(1,dim(data)[1],1), round(splitPerc\*dim(data)[1]))  
  
trainIndices = sample(seq(1,dim(data)[1],1),round(splitPerc \* dim(data)[1]))  
train = data[trainindex,]  
test = data[-trainindex,]  
#note it is illegal to know about relationship status when interviewing  
  
m = naiveBayes(Attrition~OverTime+JobRole+JobInvolvement+MonthlyIncome+TotalWorkingYears+JobLevel,data=train)  
table(predict(m, newdata=test),test$Attrition)

##   
## No Yes  
## No 176 11  
## Yes 13 18

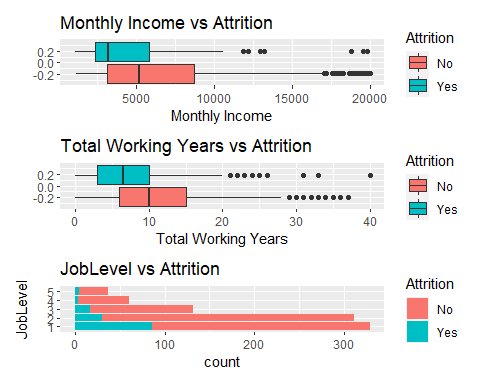
CM = confusionMatrix(table(predict(m, newdata=test),test$Attrition))  
CM

## Confusion Matrix and Statistics  
##   
##   
## No Yes  
## No 176 11  
## Yes 13 18  
##   
## Accuracy : 0.8899   
## 95% CI : (0.8406, 0.9282)  
## No Information Rate : 0.867   
## P-Value [Acc > NIR] : 0.1859   
##   
## Kappa : 0.5363   
##   
## Mcnemar's Test P-Value : 0.8383   
##   
## Sensitivity : 0.9312   
## Specificity : 0.6207   
## Pos Pred Value : 0.9412   
## Neg Pred Value : 0.5806   
## Prevalence : 0.8670   
## Detection Rate : 0.8073   
## Detection Prevalence : 0.8578   
## Balanced Accuracy : 0.7760   
##   
## 'Positive' Class : No   
##

#grid.arrange(O\_plot, JR\_plot,JI\_plot, mi\_plot, twy\_plot,yicr\_plot,JL\_plot)  
  
comp\_NB = naiveBayes(Attrition~OverTime+MonthlyIncome+JobRole+TotalWorkingYears+JobInvolvement+JobLevel,data=data)  
  
pred\_att = data.frame(Attrition =predict(comp\_NB, newdata=Comp\_attr))  
att\_comp <-bind\_cols(Comp\_attr,pred\_att)  
  
JR\_plot+ (O\_plot +JI\_plot)+ plot\_layout(ncol = 1)



(mi\_plot\_box+ twy\_plot\_box)+(JL\_plot)+ plot\_layout(ncol = 1)



#this meets out goal but we could do better. It is dependent on the seed  
#summary(att\_comp)  
  
  
#write.csv(select(att\_comp,ID,Attrition),file="Case2Predictions\_Scott\_Attrition.csv")

## trying a “better model” could be overfitting however.

#Variables such as TotalWorkingYears, YearsAtCompany, YearsInCurrentRole, YearsSinceLastPromotion and YearsWithCurrentManager are highly corelated to each other.  
# removed from the model due to colinearity. Left total working years  
set.seed(12)  
splitPerc = .75  
head(data)

## ID Age Attrition BusinessTravel DailyRate Department  
## 1 1 32 No Travel\_Rarely 117 Sales  
## 2 2 40 No Travel\_Rarely 1308 Research & Development  
## 3 3 35 No Travel\_Frequently 200 Research & Development  
## 4 4 32 No Travel\_Rarely 801 Sales  
## 5 5 24 No Travel\_Frequently 567 Research & Development  
## 6 6 27 No Travel\_Frequently 294 Research & Development  
## DistanceFromHome Education EducationField EmployeeCount EmployeeNumber  
## 1 13 4 Life Sciences 1 859  
## 2 14 3 Medical 1 1128  
## 3 18 2 Life Sciences 1 1412  
## 4 1 4 Marketing 1 2016  
## 5 2 1 Technical Degree 1 1646  
## 6 10 2 Life Sciences 1 733  
## EnvironmentSatisfaction Gender HourlyRate JobInvolvement JobLevel  
## 1 2 Male 73 3 2  
## 2 3 Male 44 2 5  
## 3 3 Male 60 3 3  
## 4 3 Female 48 3 3  
## 5 1 Female 32 3 1  
## 6 4 Male 32 3 3  
## JobRole JobSatisfaction MaritalStatus MonthlyIncome  
## 1 Sales Executive 4 Divorced 4403  
## 2 Research Director 3 Single 19626  
## 3 Manufacturing Director 4 Single 9362  
## 4 Sales Executive 4 Married 10422  
## 5 Research Scientist 4 Single 3760  
## 6 Manufacturing Director 1 Divorced 8793  
## MonthlyRate NumCompaniesWorked Over18 OverTime PercentSalaryHike  
## 1 9250 2 Y No 11  
## 2 17544 1 Y No 14  
## 3 19944 2 Y No 11  
## 4 24032 1 Y No 19  
## 5 17218 1 Y Yes 13  
## 6 4809 1 Y No 21  
## PerformanceRating RelationshipSatisfaction StandardHours StockOptionLevel  
## 1 3 3 80 1  
## 2 3 1 80 0  
## 3 3 3 80 0  
## 4 3 3 80 2  
## 5 3 3 80 0  
## 6 4 3 80 2  
## TotalWorkingYears TrainingTimesLastYear WorkLifeBalance YearsAtCompany  
## 1 8 3 2 5  
## 2 21 2 4 20  
## 3 10 2 3 2  
## 4 14 3 3 14  
## 5 6 2 3 6  
## 6 9 4 2 9  
## YearsInCurrentRole YearsSinceLastPromotion YearsWithCurrManager  
## 1 2 0 3  
## 2 7 4 9  
## 3 2 2 2  
## 4 10 5 7  
## 5 3 1 3  
## 6 7 1 7

trainindex = sample(seq(1,dim(data)[1],1), round(splitPerc\*dim(data)[1]))  
  
trainIndices = sample(seq(1,dim(data)[1],1),round(splitPerc \* dim(data)[1]))  
train = data[trainindex,]  
test = data[-trainindex,]  
#note it is illegal to know about relationship status when interviewing  
m = naiveBayes(Attrition~BusinessTravel+DistanceFromHome+EnvironmentSatisfaction+JobInvolvement+JobSatisfaction+MaritalStatus+NumCompaniesWorked+OverTime+TotalWorkingYears+RelationshipSatisfaction+MonthlyIncome+JobRole+YearsInCurrentRole+JobLevel,data=train)  
table(predict(m, newdata=test),test$Attrition)

##   
## No Yes  
## No 168 10  
## Yes 21 19

CM = confusionMatrix(table(predict(m, newdata=test),test$Attrition))  
CM

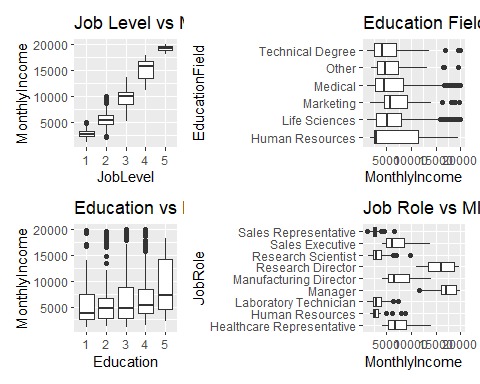
## Confusion Matrix and Statistics  
##   
##   
## No Yes  
## No 168 10  
## Yes 21 19  
##   
## Accuracy : 0.8578   
## 95% CI : (0.8043, 0.9013)  
## No Information Rate : 0.867   
## P-Value [Acc > NIR] : 0.69734   
##   
## Kappa : 0.4688   
##   
## Mcnemar's Test P-Value : 0.07249   
##   
## Sensitivity : 0.8889   
## Specificity : 0.6552   
## Pos Pred Value : 0.9438   
## Neg Pred Value : 0.4750   
## Prevalence : 0.8670   
## Detection Rate : 0.7706   
## Detection Prevalence : 0.8165   
## Balanced Accuracy : 0.7720   
##   
## 'Positive' Class : No   
##

#grid.arrange(O\_plot, JR\_plot,JI\_plot, mi\_plot, twy\_plot,yicr\_plot,JL\_plot)  
  
comp\_NB = naiveBayes(Attrition~BusinessTravel+DistanceFromHome+EnvironmentSatisfaction+JobInvolvement+JobSatisfaction+MaritalStatus+NumCompaniesWorked+OverTime+TotalWorkingYears+RelationshipSatisfaction+MonthlyIncome+JobRole+YearsInCurrentRole+JobLevel,data=data)  
  
pred\_att = data.frame(Attrition =predict(comp\_NB, newdata=Comp\_attr))  
att\_comp <-bind\_cols(Comp\_attr,pred\_att)

## Salary EDA

Categorical Variables Vs MonthlyIncome \* JobLevel  
\* Education  
\* JobRole  
\* EducationField

#Ordinal variable EDA  
#data %>% select(Education, EnvironmentSatisfaction, JobInvolvement, JobSatisfaction, PerformanceRating, WorkLifeBalance, JobLevel, RelationshipSatisfaction, MonthlyIncome) %>% ggpairs(upper = list(continuous="smooth", combo="box", discrete = "facetbar"), lower=list(continuous="smooth", combo="box", discrete = "facetbar"))  
# Correlation to JobLevel, maybe Education  
  
JLS\_plot <- data %>% ggplot(aes(JobLevel, MonthlyIncome))+geom\_boxplot()+labs(title="Job Level vs MI")  
  
ES\_plot <- data %>% ggplot(aes(Education, MonthlyIncome))+geom\_boxplot()+labs(title="Education vs MI")  
  
#Nominal variable EDA  
#data %>% select(Attrition, BusinessTravel, Department, EducationField, Gender, JobRole, MaritalStatus, OverTime, MonthlyIncome) %>% ggpairs(upper = list(continuous="smooth", combo="box", discrete = "facetbar"), lower=list(continuous="smooth", combo="box", discrete = "facetbar"))  
# JobRole has interesting differences. Maybe EducationField  
  
JRS\_plot <- data %>% ggplot(aes(JobRole,MonthlyIncome))+geom\_boxplot()+labs(title="Job Role vs MI")+coord\_flip()  
  
EFS\_plot <- data %>% ggplot(aes(EducationField,MonthlyIncome))+geom\_boxplot()+labs(title="Education Field vs MI")+coord\_flip()  
  
JLS\_plot+EFS\_plot+ ES\_plot + JRS\_plot+ plot\_layout(ncol = 2)



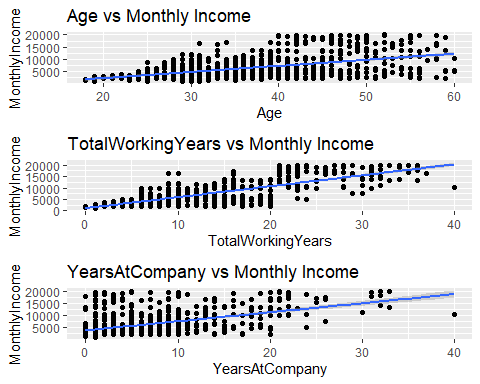
#grid.arrange(JLS\_plot, ES\_plot, JRS\_plot, EFS\_plot)

## Continious Varables

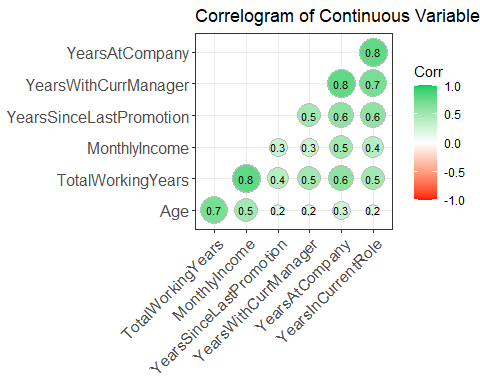
* Age
* TotalWorkingYears
* Years at Company

# there is a possibility that total years worked and years at company could be colinear and we do not want that  
  
#Continous variable EDA  
#data %>% select(Age,DailyRate, DistanceFromHome, HourlyRate, MonthlyRate, NumCompaniesWorked, PercentSalaryHike, MonthlyIncome) %>% ggpairs(upper = list(continuous="smooth", combo="box", discrete = "facetbar"), lower=list(continuous="smooth", combo="box", discrete = "facetbar"))  
# Age has a positive correlation  
  
AgS\_plot <- data %>% ggplot(aes(Age,MonthlyIncome))+geom\_point()+geom\_smooth(method="lm")+labs(title = "Age vs Monthly Income")  
  
#data %>% select(StockOptionLevel, TotalWorkingYears, TrainingTimesLastYear, YearsSinceLastPromotion, YearsInCurrentRole, YearsWithCurrManager, YearsAtCompany, MonthlyIncome) %>% ggpairs(upper = list(continuous="smooth", combo="box", discrete = "facetbar"), lower=list(continuous="smooth", combo="box", discrete = "facetbar"))  
#TotalWorking Years and YearsatCompany have a strong correlation  
#Years SinceLastPromotion, InCurrentRole and WithCurrManager have weak positive correlation  
  
TWYS\_plot <- data %>% ggplot(aes(TotalWorkingYears,MonthlyIncome))+geom\_point()+geom\_smooth(method="lm")+labs(title = "TotalWorkingYears vs Monthly Income")  
  
YaCS\_plot <- data %>% ggplot(aes(YearsAtCompany,MonthlyIncome))+geom\_point()+geom\_smooth(method="lm")+labs(title = "YearsAtCompany vs Monthly Income")  
  
grid.arrange(AgS\_plot,TWYS\_plot, YaCS\_plot)

## `geom\_smooth()` using formula 'y ~ x'  
## `geom\_smooth()` using formula 'y ~ x'  
## `geom\_smooth()` using formula 'y ~ x'



corr\_data <- data %>% select(Age, TotalWorkingYears, YearsAtCompany, YearsSinceLastPromotion, YearsInCurrentRole, YearsWithCurrManager, MonthlyIncome)  
  
corr <- round(cor(corr\_data),1)  
  
ggcorrplot(corr, hc.order = TRUE,   
 type = "lower",   
 lab = TRUE,   
 lab\_size = 3,   
 method="circle",   
 colors = c("red", "white", "springgreen3"),   
 title="Correlogram of Continuous Variables",   
 ggtheme=theme\_bw)



## LM Salary

Salary\_train <- data %>% select(MonthlyIncome, TotalWorkingYears, JobLevel)  
  
fit <- lm(MonthlyIncome~TotalWorkingYears+JobLevel, data=Salary\_train)  
  
summary(fit)

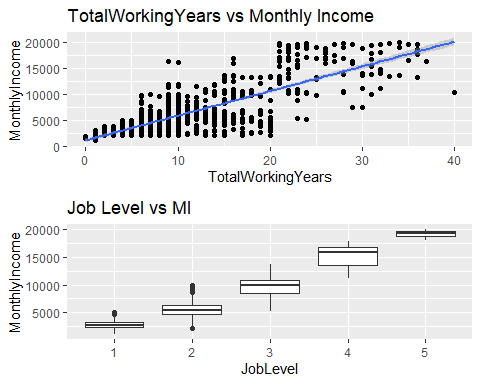
##   
## Call:  
## lm(formula = MonthlyIncome ~ TotalWorkingYears + JobLevel, data = Salary\_train)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -4957.9 -657.8 -134.6 618.2 4525.8   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2544.901 89.085 28.567 < 2e-16 \*\*\*  
## TotalWorkingYears 33.442 9.426 3.548 0.000409 \*\*\*  
## JobLevel2 2652.205 107.666 24.634 < 2e-16 \*\*\*  
## JobLevel3 6820.371 152.732 44.656 < 2e-16 \*\*\*  
## JobLevel4 11858.212 254.564 46.582 < 2e-16 \*\*\*  
## JobLevel5 15800.546 289.997 54.485 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1256 on 864 degrees of freedom  
## Multiple R-squared: 0.9258, Adjusted R-squared: 0.9254   
## F-statistic: 2157 on 5 and 864 DF, p-value: < 2.2e-16

#confint(fit)  
  
train(MonthlyIncome~TotalWorkingYears+JobLevel, method="lm",data=Salary\_train, trControl = trainControl(method = "LOOCV"))

## Linear Regression   
##   
## 870 samples  
## 2 predictor  
##   
## No pre-processing  
## Resampling: Leave-One-Out Cross-Validation   
## Summary of sample sizes: 869, 869, 869, 869, 869, 869, ...   
## Resampling results:  
##   
## RMSE Rsquared MAE   
## 1261.666 0.9246115 915.7436  
##   
## Tuning parameter 'intercept' was held constant at a value of TRUE

grid.arrange(TWYS\_plot, JLS\_plot)

## `geom\_smooth()` using formula 'y ~ x'



pred\_sal = data.frame(MonthlyIncome = predict(fit, newdata = Comp\_sal))  
sal\_comp <-bind\_cols(Comp\_sal, pred\_sal)  
  
#summary(sal\_comp)  
  
#write.csv(select(sal\_comp,ID,MonthlyIncome), file="Case2Predictions\_Scott\_Salary.csv")