CaseStudy2\_MSDS

Simerpreet Reddy

March 11, 2020

Executive Summary: At DDS Anlytics we specialize in Talent Management. Using our cutting edge technology and Frito Lay’s employee data, we have developed a Talent Management solution that can help Frito Lay to identify and then control Attrition, manage employee training, analytically define salary and assess any job role related trends.

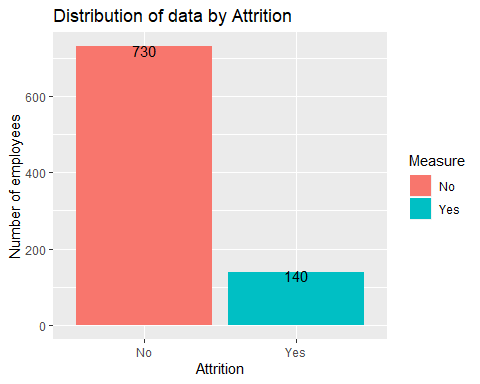
Purpose: The purpose of this document is to provide analytical insights into Frito Lay’s employee attrition data and to also provide insights on job role related trends. This document also provides the models to predict attrition and salary of employees.

This makes a great tool for Talent Management.

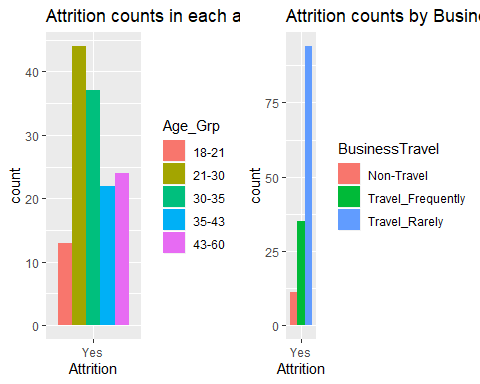
The presenation for this analysis can be watched at this youtube video link: <https://www.youtube.com/watch?v=SBntBfO1AiM&feature=youtu.be>

Job Role trends and Top attrition factors

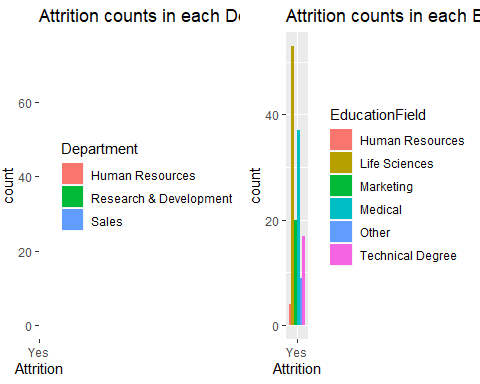
#Include necessary libraries  
library(tidyverse)  
library(ggplot2)  
library(caret)  
library(tidyr)  
library(dplyr)  
library(e1071)  
library(class)  
library(gridExtra)  
  
#Read source file  
CaseStudy2\_raw\_hist <- read.csv(file.choose(), header = TRUE, sep=',')  
  
#nrow(CaseStudy2\_raw\_hist)  
#head(CaseStudy2\_raw\_hist)  
  
#Rename columns for readability  
colnames(CaseStudy2\_raw\_hist) <- c("ID", "Age", "Attrition", "BusinessTravel", "DailyRate", "Department", "DistanceFromHome", "Education", "EducationField", "EmployeeCount", "EmployeeNumber", "EnvironmentSatisfaction", "Gender", "HourlyRate", "JobInvolvement", "JobLevel", "JobRole", "JobSatisfaction", "MaritalStatus", "MonthlyIncome", "MonthlyRate", "NumCompaniesWorked", "Over18", "OverTime", "PercentSalaryHike", "PerformanceRating", "RelationshipSatisfaction", "StandardHours", "StockOptionLevel", "TotalWorkingYears", "TrainingTimesLastYear", "WorkLifeBalance", "YearsAtCompany", "YearsInCurrentRole", "YearsSinceLastPromotion", "YearsWithCurrManager")  
  
#Attrition Analysis  
#Plot the distribution of data by Attrition  
att\_cnt <- data.frame("Measure" = c("Yes","No"),   
 "Counts"= c(nrow(CaseStudy2\_raw\_hist %>% filter(Attrition=='Yes')),nrow(CaseStudy2\_raw\_hist %>% filter(Attrition=='No')) ))  
  
p\_Att<- ggplot(data=att\_cnt,mapping=aes(x=Measure, y= Counts, fill=Measure)) + geom\_bar(stat="identity") +   
 ggtitle("Distribution of data by Attrition") +  
 geom\_text(aes(label = Counts), nudge\_y = -10, color = "black") + xlab("Attrition") + ylab("Number of employees")  
  
p\_Att



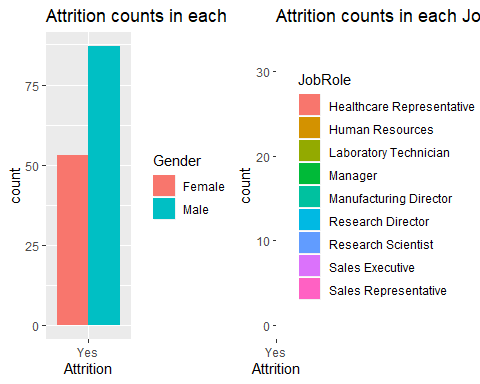
#Filter out the row with null values  
CaseStudy2\_raw\_hist <- CaseStudy2\_raw\_hist %>% filter(ID!=870)  
#summary(CaseStudy2\_raw\_hist)  
#str(CaseStudy2\_raw\_hist)  
  
#Cut the differemt continous variable into groups/categories to be able to build histograms  
#summary(CaseStudy2\_raw\_hist$Age)  
CaseStudy2\_raw\_hist$Age\_Grp <- cut(CaseStudy2\_raw\_hist$Age, breaks = c(17,21,30,35,43,61), labels = c("18-21","21-30" ,"30-35", "35-43", "43-60"))  
  
#CaseStudy2\_raw\_hist$Age\_Grp\_Cat <- cut(CaseStudy2\_raw\_hist$Age, breaks = c(17,21,30,35,43,61), labels = c(1,2 ,3, 4, 5))  
#summarise(CaseStudy2\_raw\_hist$Age\_Grp\_Cat)  
  
#summary(CaseStudy2\_raw\_hist$DistanceFromHome)  
CaseStudy2\_raw\_hist$DistfrmHome <- cut(CaseStudy2\_raw\_hist$DistanceFromHome, breaks = c(0,2,7,10,14,20,30), labels = c("0-2","2-7" ,"7-10", "10-14", "14-20","20-30"))  
#summary(CaseStudy2\_raw\_hist$DistfrmHome)  
  
#summary(CaseStudy2\_raw\_hist$MonthlyIncome)  
CaseStudy2\_raw\_hist$MonthlyIncome\_f <- cut(CaseStudy2\_raw\_hist$MonthlyIncome, breaks = c(1080,2838,4950,8189,20000), labels = c("1080-2838","2838-4950" ,"4950-8189", "8189-20000"))  
#summary(CaseStudy2\_raw\_hist$MonthlyIncome\_f)  
  
#summary(CaseStudy2\_raw\_hist$MonthlyRate)   
CaseStudy2\_raw\_hist$MonthlyRate\_f <- cut(CaseStudy2\_raw\_hist$MonthlyRate, breaks = c(2093,8059,14074,20460,26998), labels = c("2093-8059","8059-14074" ,"14074-20460", "20460-26998"))  
#summary(CaseStudy2\_raw\_hist$MonthlyRate\_f)   
  
#summary(CaseStudy2\_raw\_hist$NumCompaniesWorked)  
CaseStudy2\_raw\_hist$NumCompaniesWorked\_f <- cut(CaseStudy2\_raw\_hist$NumCompaniesWorked, breaks = c(-1,0,2,4,10), labels = c("0","1-2" ,"3-4", "5-10"))  
#summary(CaseStudy2\_raw\_hist$NumCompaniesWorked\_f)  
  
#summary(CaseStudy2\_raw\_hist$TotalWorkingYears)  
CaseStudy2\_raw\_hist$TotalWorkingYears\_f <- cut(CaseStudy2\_raw\_hist$TotalWorkingYears, breaks = c(-1,1,6,10,15,41), labels = c("0","1-6","6-10" ,"10-15", "15-41"))  
#summary(CaseStudy2\_raw\_hist$TotalWorkingYears\_f )  
  
  
#summary(CaseStudy2\_raw\_hist$TrainingTimesLastYear)  
CaseStudy2\_raw\_hist$TrainingTimesLastYear\_f <- cut(CaseStudy2\_raw\_hist$TrainingTimesLastYear, breaks = c(-1,1,2,3,6), labels = c("0-1","2","2-3" ,"3-7"))  
#summary(CaseStudy2\_raw\_hist$TrainingTimesLastYear\_f)  
#select(CaseStudy2\_raw\_hist, TrainingTimesLastYear,TrainingTimesLastYear\_f)  
  
#summary(CaseStudy2\_raw\_hist$YearsAtCompany )  
CaseStudy2\_raw\_hist$YearsAtCompany\_f <- cut(CaseStudy2\_raw\_hist$YearsAtCompany, breaks = c(-1,1,3,5,10,41), labels = c("0","1-3" ,"3-5", "5-10", "10-41"))  
#summary(CaseStudy2\_raw\_hist$YearsAtCompany\_f )  
  
#summary(CaseStudy2\_raw\_hist$YearsInCurrentRole )  
CaseStudy2\_raw\_hist$YearsInCurrentRole\_f <- cut(CaseStudy2\_raw\_hist$YearsInCurrentRole, breaks = c(-1,2,3,7,19), labels = c("0-2" ,"2-3", "3-7", "7-19"))  
#select(CaseStudy2\_raw\_hist, YearsInCurrentRole,YearsInCurrentRole\_f)  
#summary(CaseStudy2\_raw\_hist$YearsInCurrentRole\_f)  
  
#summary(CaseStudy2\_raw\_hist$YearsSinceLastPromotion)  
CaseStudy2\_raw\_hist$YearsSinceLastPromotion\_f <- cut(CaseStudy2\_raw\_hist$YearsSinceLastPromotion, breaks = c(-1,0,1,3,16), labels = c("0","1","2-3" ,"3-15"))  
#summary(CaseStudy2\_raw\_hist$YearsSinceLastPromotion\_f)  
#select(CaseStudy2\_raw\_hist, YearsSinceLastPromotion,YearsSinceLastPromotion\_f)  
  
#summary(CaseStudy2\_raw\_hist$YearsWithCurrManager)  
CaseStudy2\_raw\_hist$YearsWithCurrManager\_f <- cut(CaseStudy2\_raw\_hist$YearsWithCurrManager, breaks = c(-1,1,3,7,18), labels = c("0-1","2-3" ,"3-7", "7-17"))  
#summary(CaseStudy2\_raw\_hist$YearsWithCurrManager\_f)  
  
#Convert numberic columns into factors  
CaseStudy2\_raw\_hist$JobInvolvement <- as.factor(CaseStudy2\_raw\_hist$JobInvolvement)  
CaseStudy2\_raw\_hist$JobLevel <- as.factor(CaseStudy2\_raw\_hist$JobLevel)  
CaseStudy2\_raw\_hist$EnvironmentSatisfaction <- as.factor(CaseStudy2\_raw\_hist$EnvironmentSatisfaction)  
CaseStudy2\_raw\_hist$Education <- as.factor(CaseStudy2\_raw\_hist$Education)  
CaseStudy2\_raw\_hist$EnvironmentSatisfaction<- as.factor(CaseStudy2\_raw\_hist$EnvironmentSatisfaction)  
CaseStudy2\_raw\_hist$JobInvolvement<- as.factor(CaseStudy2\_raw\_hist$JobInvolvement)  
#CaseStudy2\_raw\_hist$JobLevel<- as.factor(CaseStudy2\_raw\_hist$JobLevel)  
CaseStudy2\_raw\_hist$JobSatisfaction<- as.factor(CaseStudy2\_raw\_hist$JobSatisfaction)  
CaseStudy2\_raw\_hist$RelationshipSatisfaction<- as.factor(CaseStudy2\_raw\_hist$RelationshipSatisfaction)  
CaseStudy2\_raw\_hist$StockOptionLevel<- as.factor(CaseStudy2\_raw\_hist$StockOptionLevel)  
CaseStudy2\_raw\_hist$WorkLifeBalance<- as.factor(CaseStudy2\_raw\_hist$WorkLifeBalance)  
  
#Filter out rows where Attrition=Yes  
CaseStudy2\_raw\_hist\_ATT\_Y <- CaseStudy2\_raw\_hist %>% filter(CaseStudy2\_raw\_hist$Attrition=='Yes')  
  
#Age\_group\_by <- group\_by(CaseStudy2\_raw\_hist\_ATT\_Y,Age\_Grp\_Cat)  
#Perc\_Att <- summarise(Age\_group\_by, perc\_att=   
#plot(CaseStudy2\_raw\_hist["Attrition","Age\_Grp"])  
#CaseStudy2\_raw\_hist %>% filter(is.na(Age\_Grp))  
  
#Plot different measures to see how much they contributed to attrition  
p\_Age\_grp <- ggplot(data=CaseStudy2\_raw\_hist\_ATT\_Y) + geom\_bar(mapping=aes(x=Attrition, fill= Age\_Grp), position="dodge") + ggtitle("Attrition counts in each age group")  
p\_BusinessTravel <- ggplot(data=CaseStudy2\_raw\_hist\_ATT\_Y) + geom\_bar(mapping=aes(x=Attrition, fill= BusinessTravel), position="dodge") + ggtitle("Attrition counts by BusinessTravel group")  
p\_Department<- ggplot(data=CaseStudy2\_raw\_hist\_ATT\_Y) + geom\_bar(mapping=aes(x=Attrition, fill= Department), position="dodge") + ggtitle("Attrition counts in each Department")  
p\_EducationField<- ggplot(data=CaseStudy2\_raw\_hist\_ATT\_Y) + geom\_bar(mapping=aes(x=Attrition, fill= EducationField), position="dodge") + ggtitle("Attrition counts in each EducationField")  
p\_Gender<- ggplot(data=CaseStudy2\_raw\_hist\_ATT\_Y) + geom\_bar(mapping=aes(x=Attrition, fill= Gender), position="dodge") + ggtitle("Attrition counts in each Gender")  
p\_JobRole<- ggplot(data=CaseStudy2\_raw\_hist\_ATT\_Y) + geom\_bar(mapping=aes(x=Attrition, fill= JobRole), position="dodge") + ggtitle("Attrition counts in each JobRole")  
p\_MaritalStatus<- ggplot(data=CaseStudy2\_raw\_hist\_ATT\_Y) + geom\_bar(mapping=aes(x=Attrition, fill= MaritalStatus), position="dodge") + ggtitle("Attrition counts in each MaritalStatus")  
p\_OverTime<- ggplot(data=CaseStudy2\_raw\_hist\_ATT\_Y) + geom\_bar(mapping=aes(x=Attrition, fill= OverTime), position="dodge") + ggtitle("Attrition counts in each OverTime")   
p\_JobInvolvement<- ggplot(data=CaseStudy2\_raw\_hist\_ATT\_Y) + geom\_bar(mapping=aes(x=Attrition, fill= JobInvolvement), position="dodge") + ggtitle("Attrition counts in each JobInvolvement")  
p\_DistfrmHome <- ggplot(data=CaseStudy2\_raw\_hist\_ATT\_Y) + geom\_bar(mapping=aes(x=Attrition, fill= DistfrmHome), position="dodge") + ggtitle("Attrition counts in each DistfrmHome category")  
p\_Education <- ggplot(data=CaseStudy2\_raw\_hist\_ATT\_Y) + geom\_bar(mapping=aes(x=Attrition, fill= Education), position="dodge") + ggtitle("Attrition counts in each Education level")  
p\_EnvironmentSatisfaction <- ggplot(data=CaseStudy2\_raw\_hist\_ATT\_Y) + geom\_bar(mapping=aes(x=Attrition, fill= EnvironmentSatisfaction), position="dodge") + ggtitle("Attrition counts in each EnvironmentSatisfaction level")  
p\_JobInvolvement<- ggplot(data=CaseStudy2\_raw\_hist\_ATT\_Y) + geom\_bar(mapping=aes(x=Attrition, fill= JobInvolvement), position="dodge") + ggtitle("Attrition counts in each JobInvolvement")  
p\_JobLevel<- ggplot(data=CaseStudy2\_raw\_hist\_ATT\_Y) + geom\_bar(mapping=aes(x=Attrition, fill= JobLevel), position="dodge") + ggtitle("Attrition counts in each JobLevel")  
p\_JobSatisfaction<- ggplot(data=CaseStudy2\_raw\_hist\_ATT\_Y) + geom\_bar(mapping=aes(x=Attrition, fill= JobSatisfaction), position="dodge") + ggtitle("Attrition counts in each JobSatisfaction")  
p\_RelationshipSatisfaction<- ggplot(data=CaseStudy2\_raw\_hist\_ATT\_Y) + geom\_bar(mapping=aes(x=Attrition, fill= RelationshipSatisfaction), position="dodge") + ggtitle("Attrition counts in each RelationshipSatisfaction level")  
p\_StockOptionLevel<- ggplot(data=CaseStudy2\_raw\_hist\_ATT\_Y) + geom\_bar(mapping=aes(x=Attrition, fill= StockOptionLevel), position="dodge") + ggtitle("Attrition counts in each StockOptionLevel")  
p\_WorkLifeBalance<- ggplot(data=CaseStudy2\_raw\_hist\_ATT\_Y) + geom\_bar(mapping=aes(x=Attrition, fill= WorkLifeBalance), position="dodge") + ggtitle("Attrition counts in each WorkLifeBalance")  
p\_MonthlyIncome\_f<- ggplot(data=CaseStudy2\_raw\_hist\_ATT\_Y) + geom\_bar(mapping=aes(x=Attrition, fill= MonthlyIncome\_f), position="dodge") + ggtitle("Attrition counts in each MonthlyIncome category")  
p\_MonthlyRate\_f<- ggplot(data=CaseStudy2\_raw\_hist\_ATT\_Y) + geom\_bar(mapping=aes(x=Attrition, fill= MonthlyRate\_f), position="dodge") + ggtitle("Attrition counts in each MonthlyRate category")  
p\_NumCompaniesWorked\_f<- ggplot(data=CaseStudy2\_raw\_hist\_ATT\_Y) + geom\_bar(mapping=aes(x=Attrition, fill= NumCompaniesWorked\_f), position="dodge") + ggtitle("Attrition counts in each NumCompaniesWorked category")  
p\_TotalWorkingYears\_f<- ggplot(data=CaseStudy2\_raw\_hist\_ATT\_Y) + geom\_bar(mapping=aes(x=Attrition, fill= TotalWorkingYears\_f), position="dodge") + ggtitle("Attrition counts in each TotalWorkingYears category")  
p\_TrainingTimesLastYear\_f<- ggplot(data=CaseStudy2\_raw\_hist\_ATT\_Y) + geom\_bar(mapping=aes(x=Attrition, fill= TrainingTimesLastYear\_f), position="dodge") + ggtitle("Attrition counts in each TrainingTimesLastYear category")  
p\_YearsAtCompany\_f<- ggplot(data=CaseStudy2\_raw\_hist\_ATT\_Y) + geom\_bar(mapping=aes(x=Attrition, fill= YearsAtCompany\_f), position="dodge") + ggtitle("Attrition counts in each YearsAtCompany category")  
p\_YearsInCurrentRole\_f<- ggplot(data=CaseStudy2\_raw\_hist\_ATT\_Y) + geom\_bar(mapping=aes(x=Attrition, fill= YearsInCurrentRole\_f), position="dodge") + ggtitle("Attrition counts in each YearsInCurrentRole category")  
p\_YearsSinceLastPromotion\_f<- ggplot(data=CaseStudy2\_raw\_hist\_ATT\_Y) + geom\_bar(mapping=aes(x=Attrition, fill= YearsSinceLastPromotion\_f), position="dodge") + ggtitle("Attrition counts in each YearsSinceLastPromotion category")  
p\_YearsWithCurrManager\_f<- ggplot(data=CaseStudy2\_raw\_hist\_ATT\_Y) + geom\_bar(mapping=aes(x=Attrition, fill= YearsWithCurrManager\_f), position="dodge") + ggtitle("Attrition counts in each YearsWithCurrManager category")  
  
#grid.arrange(p\_Age\_grp,p\_BusinessTravel,p\_Department,p\_EducationField,p\_Gender,p\_JobRole,p\_MaritalStatus,p\_OverTime,p\_JobInvolvement,ncol = 3, nrow=3)  
#grid.arrange(p\_DistfrmHome,p\_Education,p\_EnvironmentSatisfaction,p\_JobInvolvement, p\_JobLevel,p\_JobSatisfaction,p\_RelationshipSatisfaction,p\_StockOptionLevel,p\_WorkLifeBalance,ncol = 3, nrow=3)  
#grid.arrange(p\_MonthlyIncome\_f,p\_MonthlyRate\_f,p\_NumCompaniesWorked\_f,p\_TrainingTimesLastYear\_f,p\_YearsAtCompany\_f,p\_YearsInCurrentRole\_f,p\_YearsSinceLastPromotion\_f,p\_YearsWithCurrManager\_f,ncol = 3, nrow=3)  
  
grid.arrange(p\_Age\_grp,p\_BusinessTravel,ncol = 2, nrow=1)



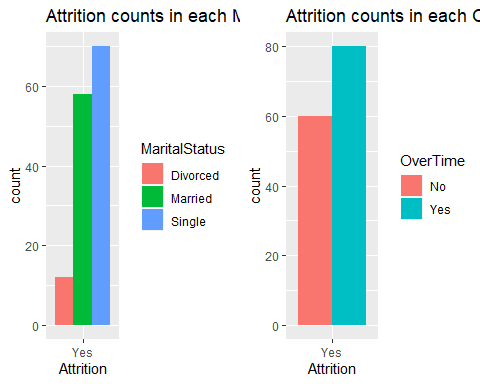
grid.arrange(p\_Department,p\_EducationField, ncol = 2, nrow=1)



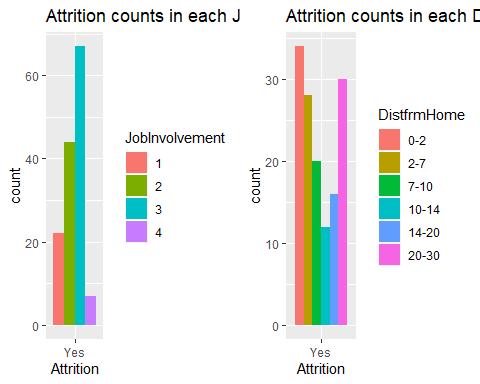
grid.arrange(p\_Gender,p\_JobRole,ncol = 2, nrow=1)



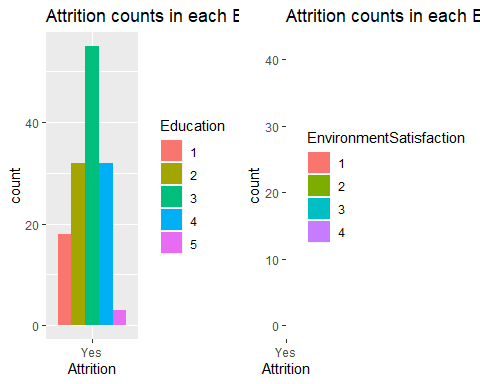
grid.arrange(p\_MaritalStatus,p\_OverTime,ncol = 2, nrow=1)



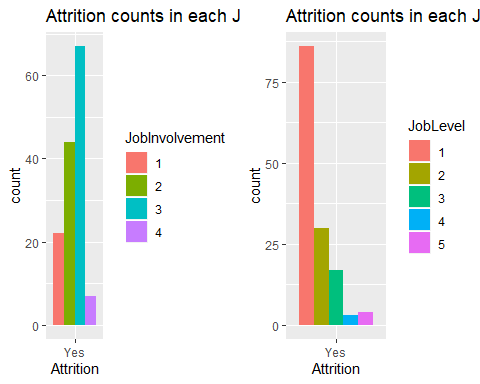
grid.arrange(p\_JobInvolvement,p\_DistfrmHome,ncol = 2, nrow=1)



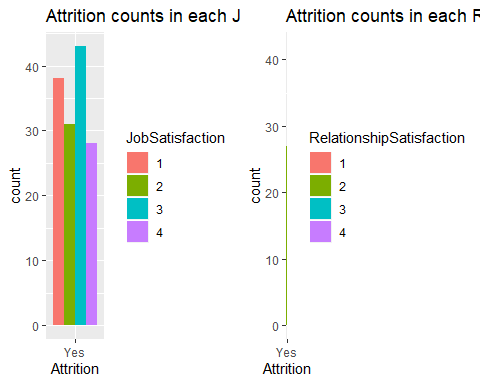
grid.arrange(p\_Education,p\_EnvironmentSatisfaction,ncol = 2, nrow=1)



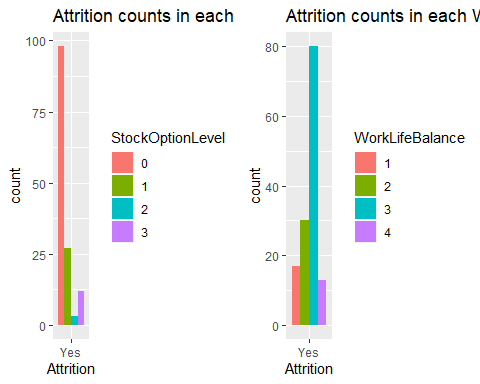
grid.arrange(p\_JobInvolvement, p\_JobLevel,ncol = 2, nrow=1)



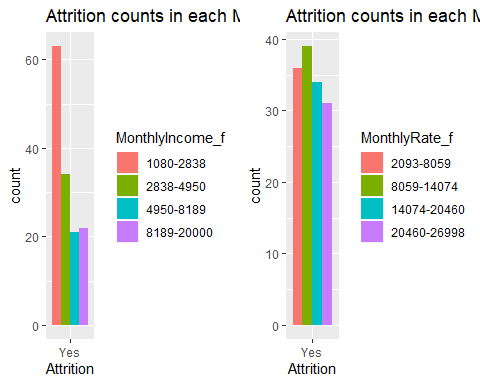
grid.arrange(p\_JobSatisfaction,p\_RelationshipSatisfaction, ncol = 2, nrow=1)



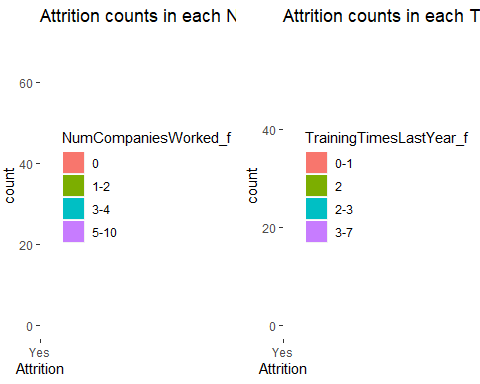
grid.arrange(p\_StockOptionLevel,p\_WorkLifeBalance,ncol = 2, nrow=1)



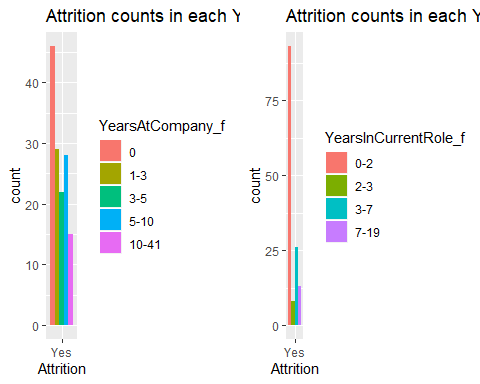
grid.arrange(p\_MonthlyIncome\_f,p\_MonthlyRate\_f,ncol = 2, nrow=1)



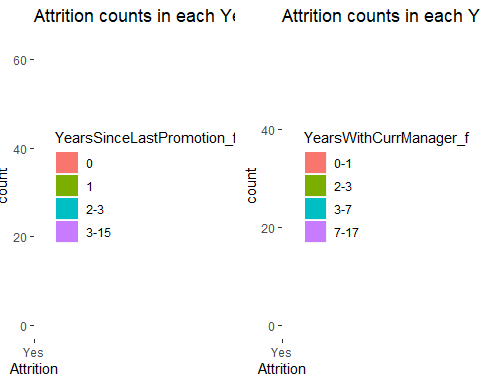
grid.arrange(p\_NumCompaniesWorked\_f,p\_TrainingTimesLastYear\_f,ncol = 2, nrow=1)



grid.arrange(p\_YearsAtCompany\_f,p\_YearsInCurrentRole\_f,ncol = 2, nrow=1)



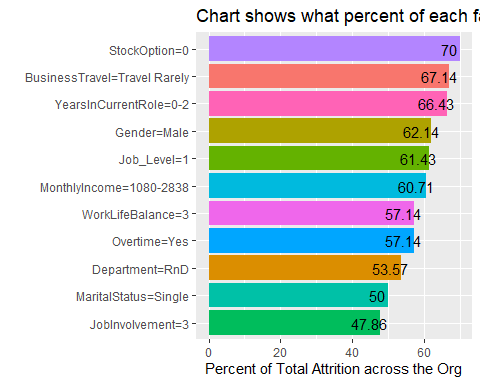
grid.arrange(p\_YearsSinceLastPromotion\_f,p\_YearsWithCurrManager\_f,ncol = 2, nrow=1)



msg <- "We see that for a lot of measures, certain categories standout when seeing only the attirion data. \n Out of all the measures, the following measures made for more than 75 out of total 141 records where attrition=Yes.\n MaritalStatus=Single\n Dept=RnD \n Overtime=Yes\n WorkLifeBalance=3 \n Job\_Level=1 \n Gender=Male \n YrsInCurrentRole=0-2 \n Business Travel=Travel Rarely \n StockOption=0 \n Let's study them all in detail"  
strsplit(msg,"\n")

## [[1]]  
## [1] "We see that for a lot of measures, certain categories standout when seeing only the attirion data. "   
## [2] " Out of all the measures, the following measures made for more than 75 out of total 141 records where attrition=Yes."  
## [3] " MaritalStatus=Single"   
## [4] " Dept=RnD "   
## [5] " Overtime=Yes"   
## [6] " WorkLifeBalance=3 "   
## [7] " Job\_Level=1 "   
## [8] " Gender=Male "   
## [9] " YrsInCurrentRole=0-2 "   
## [10] " Business Travel=Travel Rarely "   
## [11] " StockOption=0 "   
## [12] " Let's study them all in detail"

Total\_Att <-nrow(CaseStudy2\_raw\_hist\_ATT\_Y)  
Att\_Y\_BusinessTravel\_Travel\_Rarely <- round(((nrow(CaseStudy2\_raw\_hist\_ATT\_Y %>% filter(BusinessTravel=='Travel\_Rarely')))/Total\_Att)\*100,2)  
Att\_Y\_Department\_RnD <- round( ((nrow(CaseStudy2\_raw\_hist\_ATT\_Y %>% filter(Department=='Research & Development')))/Total\_Att)\*100,2)  
Att\_Y\_Gender\_M <- round(((nrow(CaseStudy2\_raw\_hist\_ATT\_Y %>% filter(Gender=='Male')))/Total\_Att)\*100,2)  
Att\_Y\_Overtime\_Y <- round( ((nrow(CaseStudy2\_raw\_hist\_ATT\_Y %>% filter(OverTime=='Yes')))/Total\_Att)\*100,2)  
Att\_Y\_Marital\_Status\_S <- round(((nrow(CaseStudy2\_raw\_hist\_ATT\_Y %>% filter(MaritalStatus=='Single')))/Total\_Att)\*100,2)  
Att\_Y\_Job\_Level\_1 <- round(((nrow(CaseStudy2\_raw\_hist\_ATT\_Y %>% filter(JobLevel=='1')))/Total\_Att)\*100,2)  
Att\_Y\_Stock\_Option\_0 <- round(((nrow(CaseStudy2\_raw\_hist\_ATT\_Y %>% filter(StockOptionLevel=='0')))/Total\_Att)\*100,2)  
Att\_Y\_WrrkLB\_3 <- round(((nrow(CaseStudy2\_raw\_hist\_ATT\_Y %>% filter(WorkLifeBalance=='3')))/Total\_Att)\*100,2)  
Att\_Y\_YearsInCurrentRole\_0\_2 <-round(((nrow(CaseStudy2\_raw\_hist\_ATT\_Y %>% filter(YearsInCurrentRole =="0" | YearsInCurrentRole =="1" | YearsInCurrentRole =="2" )))/Total\_Att)\*100,2)  
Att\_Y\_JobInvolvement\_3 <-round(((nrow(CaseStudy2\_raw\_hist\_ATT\_Y %>% filter(JobInvolvement =="3" )))/Total\_Att)\*100,2)  
Att\_Y\_p\_MonthlyIncome\_f\_1080\_2838 <-round(((nrow(CaseStudy2\_raw\_hist\_ATT\_Y %>% filter(MonthlyIncome\_f =="1080-2838" | YearsInCurrentRole =="1" | YearsInCurrentRole =="2" )))/Total\_Att)\*100,2)  
  
  
Top\_Att\_Measures <- data.frame("Measure\_Name" = c("BusinessTravel=Travel Rarely", "Department=RnD","Gender=Male","Overtime=Yes", "MaritalStatus=Single", "Job\_Level=1", "StockOption=0",  
 "WorkLifeBalance=3","YearsInCurrentRole=0-2","JobInvolvement=3","MonthlyIncome=1080-2838" ),   
"Att\_perc"= c(Att\_Y\_BusinessTravel\_Travel\_Rarely,Att\_Y\_Department\_RnD,Att\_Y\_Gender\_M,Att\_Y\_Overtime\_Y,Att\_Y\_Marital\_Status\_S,  
 Att\_Y\_Job\_Level\_1,Att\_Y\_Stock\_Option\_0,Att\_Y\_WrrkLB\_3 ,Att\_Y\_YearsInCurrentRole\_0\_2,Att\_Y\_JobInvolvement\_3,Att\_Y\_p\_MonthlyIncome\_f\_1080\_2838))  
  
  
positions=c("JobInvolvement=3","MaritalStatus=Single","Department=RnD","Overtime=Yes","WorkLifeBalance=3","MonthlyIncome=1080-2838" ,"Job\_Level=1","Gender=Male","YearsInCurrentRole=0-2","BusinessTravel=Travel Rarely","StockOption=0")  
ggplot(data=Top\_Att\_Measures,mapping=aes(y=Att\_perc, x= Measure\_Name, fill=Measure\_Name)) + geom\_bar( stat='identity',show.legend = FALSE) +  
 geom\_text(aes(label = Att\_perc), nudge\_y = -3, color = "black") + scale\_x\_discrete(limits = positions) +  
 ggtitle("Chart shows what percent of each factor contributed to total attrition") +   
 xlab("") +ylab("Percent of Total Attrition across the Org") +  
 coord\_flip()



print("Conclusion: To control Attrition, we should start by focussing on employees who fall in following categories: 1) Stock Option Level=0, 2) Those who go rarely on a business travel 3) 0-2 years in their current role")

## [1] "Conclusion: To control Attrition, we should start by focussing on employees who fall in following categories: 1) Stock Option Level=0, 2) Those who go rarely on a business travel 3) 0-2 years in their current role"

# 2 Attrition Analysis by Department  
#Divide data by department  
data\_dept\_Sales <- CaseStudy2\_raw\_hist %>% filter(Department=='Sales')  
data\_dept\_HR <- CaseStudy2\_raw\_hist %>% filter(Department=='Human Resources')  
data\_dept\_RnD <- CaseStudy2\_raw\_hist %>% filter(Department=='Research & Development')  
#nrow(data\_dept\_Sales)  
#nrow(data\_dept\_HR)  
#nrow(data\_dept\_RnD)  
  
#Calculate total attrition and attrition % in each department  
#Calculate total attrition in the entire data set  
Cnt\_Att\_Yes <- nrow(CaseStudy2\_raw\_hist %>% filter(Attrition=="Yes"))  
Cnt\_Total <-nrow(CaseStudy2\_raw\_hist)  
Attrition\_Perc <- round((Cnt\_Att\_Yes/Cnt\_Total)\*100,0)  
  
Cnt\_Att\_Yes\_Sales <- nrow(data\_dept\_Sales %>% filter(Attrition=="Yes"))  
Cnt\_Total\_Sales <-nrow(data\_dept\_Sales)  
Cnt\_Total\_Sales\_Perc <- round((Cnt\_Total\_Sales/Cnt\_Total)\*100,0)  
Attrition\_Perc\_Sales <- round((Cnt\_Att\_Yes\_Sales/Cnt\_Total\_Sales)\*100,0)  
Attrition\_Perc\_Sales #21.61

## [1] 22

Overall\_Attrition\_Perc\_Sales <- round((Cnt\_Att\_Yes\_Sales/Cnt\_Att\_Yes)\*100,0)  
  
Cnt\_Att\_Yes\_HR <- nrow(data\_dept\_HR %>% filter(Attrition=="Yes"))  
Cnt\_Total\_HR <-nrow(data\_dept\_HR)  
Cnt\_Total\_HR\_Perc <- round((Cnt\_Total\_HR/Cnt\_Total)\*100,0)  
Attrition\_Perc\_HR <- round((Cnt\_Att\_Yes\_HR/Cnt\_Total\_HR)\*100,0)  
Attrition\_Perc\_HR #17.14

## [1] 17

Overall\_Attrition\_Perc\_HR <- round((Cnt\_Att\_Yes\_HR/Cnt\_Att\_Yes)\*100,0)  
  
Cnt\_Att\_Yes\_RnD <- nrow(data\_dept\_RnD %>% filter(Attrition=="Yes"))  
Cnt\_Total\_RnD <-nrow(data\_dept\_RnD)  
Cnt\_Total\_RnD\_Perc <- round((Cnt\_Total\_RnD/Cnt\_Total)\*100,0)  
Attrition\_Perc\_RnD <- round((Cnt\_Att\_Yes\_RnD/Cnt\_Total\_RnD)\*100,0)  
Attrition\_Perc\_RnD #13.37

## [1] 13

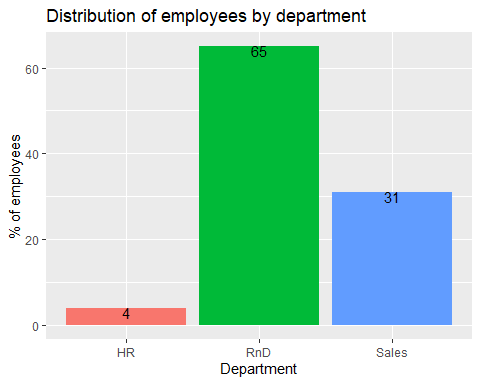
Overall\_Attrition\_Perc\_RnD <- round((Cnt\_Att\_Yes\_RnD/Cnt\_Att\_Yes)\*100,0)  
  
  
# Overall Attrition  
OrgWide\_Attrition <- data.frame( "Total\_employees"= Cnt\_Total,"Total\_Attr"=Cnt\_Att\_Yes, "Attr\_rate"= Attrition\_Perc)  
   
#Attrition % by dept  
Attr\_rate\_by\_dept <- data.frame( "Dept" = c("Sales","HR", "RnD", "OrgWide"),  
 "Total\_emp"= c(Cnt\_Total\_Sales,Cnt\_Total\_HR,Cnt\_Total\_RnD,Cnt\_Total),  
 "Total\_emp\_Perc"= c(Cnt\_Total\_Sales\_Perc,Cnt\_Total\_HR\_Perc,Cnt\_Total\_RnD\_Perc,Cnt\_Total),  
 "Attr\_num"=c(Cnt\_Att\_Yes\_Sales, Cnt\_Att\_Yes\_HR,Cnt\_Att\_Yes\_RnD,Cnt\_Att\_Yes),   
 "Attr\_rate\_dept"= c( Attrition\_Perc\_Sales,Attrition\_Perc\_HR,Attrition\_Perc\_RnD,Attrition\_Perc),  
 "Overall\_Att\_Perc" = c(Overall\_Attrition\_Perc\_Sales,Overall\_Attrition\_Perc\_HR,Overall\_Attrition\_Perc\_RnD,Attrition\_Perc))  
  
Attr\_rate\_by\_dept

## Dept Total\_emp Total\_emp\_Perc Attr\_num Attr\_rate\_dept  
## 1 Sales 273 31 59 22  
## 2 HR 35 4 6 17  
## 3 RnD 561 65 75 13  
## 4 OrgWide 869 869 140 16  
## Overall\_Att\_Perc  
## 1 42  
## 2 4  
## 3 54  
## 4 16

summary(Attr\_rate\_by\_dept)

## Dept Total\_emp Total\_emp\_Perc Attr\_num   
## HR :1 Min. : 35.0 Min. : 4.00 Min. : 6.00   
## OrgWide:1 1st Qu.:213.5 1st Qu.: 24.25 1st Qu.: 45.75   
## RnD :1 Median :417.0 Median : 48.00 Median : 67.00   
## Sales :1 Mean :434.5 Mean :242.25 Mean : 70.00   
## 3rd Qu.:638.0 3rd Qu.:266.00 3rd Qu.: 91.25   
## Max. :869.0 Max. :869.00 Max. :140.00   
## Attr\_rate\_dept Overall\_Att\_Perc  
## Min. :13.00 Min. : 4   
## 1st Qu.:15.25 1st Qu.:13   
## Median :16.50 Median :29   
## Mean :17.00 Mean :29   
## 3rd Qu.:18.25 3rd Qu.:45   
## Max. :22.00 Max. :54

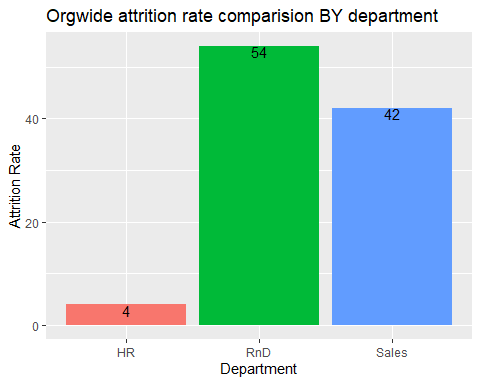
Attr\_rate\_by\_dept\_only <- Attr\_rate\_by\_dept %>% filter(Dept!="OrgWide")  
  
#Distribution of employees by department  
#ggplot(data=Attr\_rate\_by\_dept\_only) + geom\_histogram(mapping=aes(y=Total\_emp\_Perc, x= Dept, fill=Dept), stat="identity") + xlab("Department") + ylab("Attrition Rate") + ggtitle("Distribution of employees by department")  
  
p1 <- ggplot(data=Attr\_rate\_by\_dept\_only, mapping=aes(y=Total\_emp\_Perc, x= Dept, fill=Dept)) + geom\_bar(stat="identity",show.legend = FALSE) + xlab("Department") +  
 ylab("% of employees") + ggtitle("Distribution of employees by department") +  
 geom\_text(aes(label = Total\_emp\_Perc), nudge\_y = -1, color = "black")  
p1



print("RnD department has the most number of employess")

## [1] "RnD department has the most number of employess"

#Orgwide attrition rate comparision BY department  
p2 <- ggplot(data=Attr\_rate\_by\_dept\_only,mapping=aes(y=Overall\_Att\_Perc, x= Dept, fill=Dept)) + geom\_bar(stat="identity",show.legend = FALSE) + xlab("Department") +  
 ylab("Attrition Rate") + ggtitle("Orgwide attrition rate comparision BY department") +  
 geom\_text(aes(label = Overall\_Att\_Perc), nudge\_y = -1, color = "black")  
p2



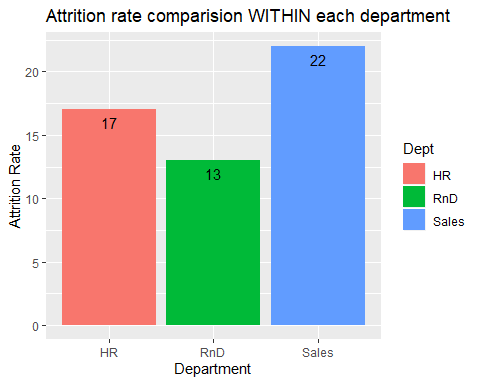
print("Orgwide attrition rate is highest in RnD department")

## [1] "Orgwide attrition rate is highest in RnD department"

Attr\_rate\_by\_dept

## Dept Total\_emp Total\_emp\_Perc Attr\_num Attr\_rate\_dept  
## 1 Sales 273 31 59 22  
## 2 HR 35 4 6 17  
## 3 RnD 561 65 75 13  
## 4 OrgWide 869 869 140 16  
## Overall\_Att\_Perc  
## 1 42  
## 2 4  
## 3 54  
## 4 16

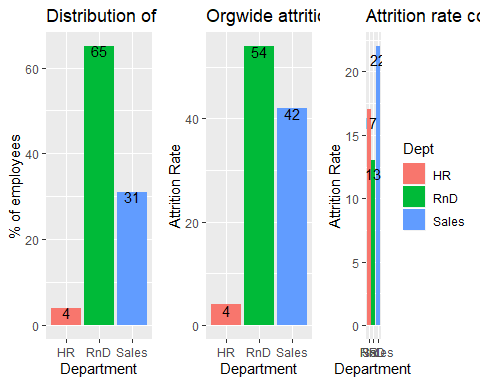
#Attrition rate comparision WITHIN each department  
p3<- ggplot(data=Attr\_rate\_by\_dept\_only,mapping=aes(y=Attr\_rate\_dept, x= Dept, fill=Dept)) + geom\_histogram(stat="identity") + xlab("Department") +  
 ylab("Attrition Rate") + ggtitle("Attrition rate comparision WITHIN each department") +  
 geom\_text(aes(label = Attr\_rate\_dept), nudge\_y = -1, color = "black")  
p3



print("Attrition is highest in Sales department, let's dive deep into this")

## [1] "Attrition is highest in Sales department, let's dive deep into this"

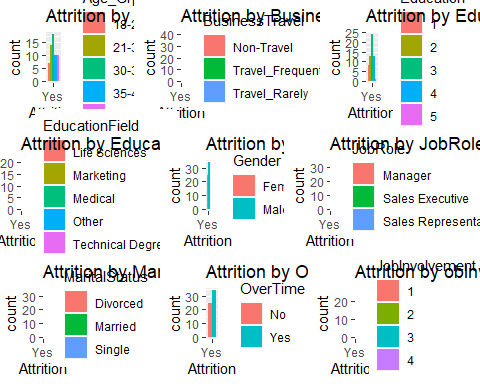
grid.arrange(p1,p2, p3, ncol = 3)



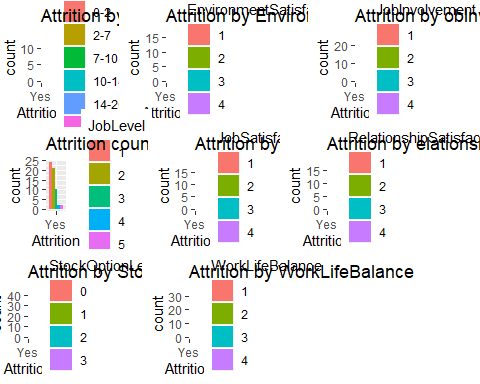
print("RnD department has the most number of employess, and has the most the number of attritions as well.But comparing the attrition within departments, we see that the Sales department had the highest attrition rate")

## [1] "RnD department has the most number of employess, and has the most the number of attritions as well.But comparing the attrition within departments, we see that the Sales department had the highest attrition rate"

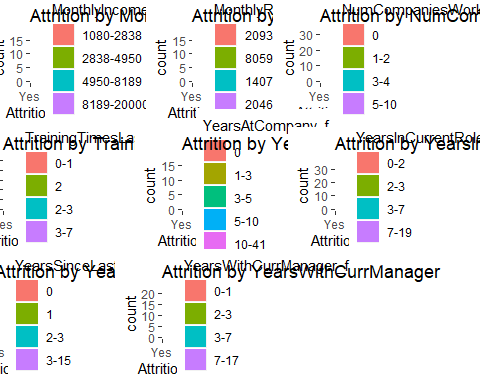
#Attrition analysis in Sales department  
Sales\_att <- CaseStudy2\_raw\_hist %>% filter(Department=='Sales' & Attrition=='Yes')  
#nrow(Sales\_att)  
#Plot different measures to see how much they contributed to attrition  
Sales\_p\_Age\_grp <- ggplot(data=Sales\_att) + geom\_bar(mapping=aes(x=Attrition, fill= Age\_Grp), position="dodge") + ggtitle("Attrition by age")  
Sales\_p\_BusinessTravel <- ggplot(data=Sales\_att) + geom\_bar(mapping=aes(x=Attrition, fill= BusinessTravel), position="dodge") + ggtitle("Attrition by BusinessTravel")  
Sales\_p\_EducationField<- ggplot(data=Sales\_att) + geom\_bar(mapping=aes(x=Attrition, fill= EducationField), position="dodge") + ggtitle("Attrition by EducationField")  
Sales\_p\_Gender<- ggplot(data=Sales\_att) + geom\_bar(mapping=aes(x=Attrition, fill= Gender), position="dodge") + ggtitle("Attrition by Gender")  
Sales\_p\_JobRole<- ggplot(data=Sales\_att) + geom\_bar(mapping=aes(x=Attrition, fill= JobRole), position="dodge") + ggtitle("Attrition by JobRole")  
Sales\_p\_MaritalStatus<- ggplot(data=Sales\_att) + geom\_bar(mapping=aes(x=Attrition, fill= MaritalStatus), position="dodge") + ggtitle("Attrition by MaritalStatus")  
Sales\_p\_OverTime<- ggplot(data=Sales\_att) + geom\_bar(mapping=aes(x=Attrition, fill= OverTime), position="dodge") + ggtitle("Attrition by OverTime")   
Sales\_p\_JobInvolvement<- ggplot(data=Sales\_att) + geom\_bar(mapping=aes(x=Attrition, fill= JobInvolvement), position="dodge") + ggtitle("Attrition by JobInvolvement")  
Sales\_p\_DistfrmHome <- ggplot(data=Sales\_att) + geom\_bar(mapping=aes(x=Attrition, fill= DistfrmHome), position="dodge") + ggtitle("Attrition by age DistancefrmHome")  
Sales\_p\_Education <- ggplot(data=Sales\_att) + geom\_bar(mapping=aes(x=Attrition, fill= Education), position="dodge") + ggtitle("Attrition by Education level")  
Sales\_p\_EnvironmentSatisfaction <- ggplot(data=Sales\_att) + geom\_bar(mapping=aes(x=Attrition, fill= EnvironmentSatisfaction), position="dodge") + ggtitle("Attrition by EnvironmentSatisfaction")  
Sales\_p\_JobInvolvement<- ggplot(data=Sales\_att) + geom\_bar(mapping=aes(x=Attrition, fill= JobInvolvement), position="dodge") + ggtitle("Attrition by obInvolvement")  
Sales\_p\_JobLevel<- ggplot(data=Sales\_att) + geom\_bar(mapping=aes(x=Attrition, fill= JobLevel), position="dodge") + ggtitle("Attrition counts by obLevel")  
Sales\_p\_JobSatisfaction<- ggplot(data=Sales\_att) + geom\_bar(mapping=aes(x=Attrition, fill= JobSatisfaction), position="dodge") + ggtitle("Attrition by JobSatisfaction")  
Sales\_p\_RelationshipSatisfaction<- ggplot(data=Sales\_att) + geom\_bar(mapping=aes(x=Attrition, fill= RelationshipSatisfaction), position="dodge") + ggtitle("Attrition by elationshipSatisfaction")  
Sales\_p\_StockOptionLevel<- ggplot(data=Sales\_att) + geom\_bar(mapping=aes(x=Attrition, fill= StockOptionLevel), position="dodge") + ggtitle("Attrition by StockOptionLevel")  
Sales\_p\_WorkLifeBalance<- ggplot(data=Sales\_att) + geom\_bar(mapping=aes(x=Attrition, fill= WorkLifeBalance), position="dodge") + ggtitle("Attrition by WorkLifeBalance")  
Sales\_p\_MonthlyIncome\_f<- ggplot(data=Sales\_att) + geom\_bar(mapping=aes(x=Attrition, fill= MonthlyIncome\_f), position="dodge") + ggtitle("Attrition by MonthlyIncome")  
Sales\_p\_MonthlyRate\_f<- ggplot(data=Sales\_att) + geom\_bar(mapping=aes(x=Attrition, fill= MonthlyRate\_f), position="dodge") + ggtitle("Attrition by MonthlyRate")  
Sales\_p\_NumCompaniesWorked\_f<- ggplot(data=Sales\_att) + geom\_bar(mapping=aes(x=Attrition, fill= NumCompaniesWorked\_f), position="dodge") + ggtitle("Attrition by NumCompaniesWorked")  
Sales\_p\_TotalWorkingYears\_f<- ggplot(data=Sales\_att) + geom\_bar(mapping=aes(x=Attrition, fill= TotalWorkingYears\_f), position="dodge") + ggtitle("Attrition by TotalWorkingYears")  
Sales\_p\_TrainingTimesLastYear\_f<- ggplot(data=Sales\_att) + geom\_bar(mapping=aes(x=Attrition, fill= TrainingTimesLastYear\_f), position="dodge") + ggtitle("Attrition by TrainingTimesLastYear")  
Sales\_p\_YearsAtCompany\_f<- ggplot(data=Sales\_att) + geom\_bar(mapping=aes(x=Attrition, fill= YearsAtCompany\_f), position="dodge") + ggtitle("Attrition by YearsAtCompany")  
Sales\_p\_YearsInCurrentRole\_f<- ggplot(data=Sales\_att) + geom\_bar(mapping=aes(x=Attrition, fill= YearsInCurrentRole\_f), position="dodge") + ggtitle("Attrition by YearsInCurrentRole")  
Sales\_p\_YearsSinceLastPromotion\_f<- ggplot(data=Sales\_att) + geom\_bar(mapping=aes(x=Attrition, fill= YearsSinceLastPromotion\_f), position="dodge") + ggtitle("Attrition by YearsSinceLastPromotion")  
Sales\_p\_YearsWithCurrManager\_f<- ggplot(data=Sales\_att) + geom\_bar(mapping=aes(x=Attrition, fill= YearsWithCurrManager\_f), position="dodge") + ggtitle("Attrition by YearsWithCurrManager")  
  
grid.arrange(Sales\_p\_Age\_grp,Sales\_p\_BusinessTravel,Sales\_p\_Education,Sales\_p\_EducationField,Sales\_p\_Gender,Sales\_p\_JobRole,Sales\_p\_MaritalStatus,Sales\_p\_OverTime,Sales\_p\_JobInvolvement,ncol = 3, nrow=3)



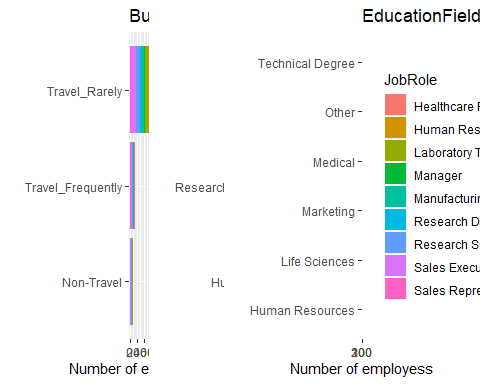
grid.arrange(Sales\_p\_DistfrmHome,Sales\_p\_EnvironmentSatisfaction,Sales\_p\_JobInvolvement, Sales\_p\_JobLevel,Sales\_p\_JobSatisfaction,Sales\_p\_RelationshipSatisfaction,Sales\_p\_StockOptionLevel,Sales\_p\_WorkLifeBalance,ncol = 3, nrow=3)



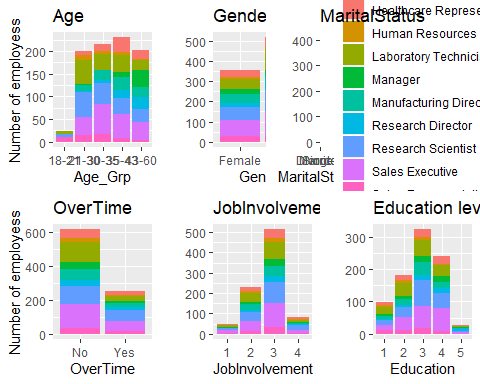
grid.arrange(Sales\_p\_MonthlyIncome\_f,Sales\_p\_MonthlyRate\_f,Sales\_p\_NumCompaniesWorked\_f,Sales\_p\_TrainingTimesLastYear\_f,Sales\_p\_YearsAtCompany\_f,Sales\_p\_YearsInCurrentRole\_f,Sales\_p\_YearsSinceLastPromotion\_f,Sales\_p\_YearsWithCurrManager\_f,ncol = 3, nrow=3)



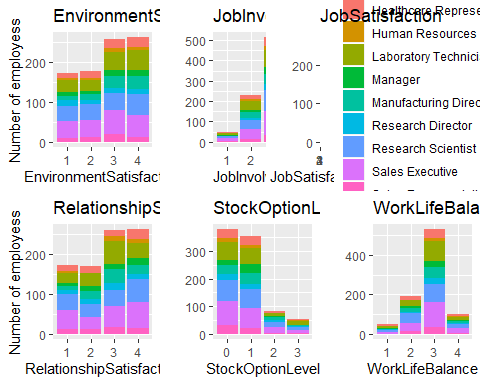
#3 Job Role specific trends  
Job\_Lvl\_p\_Age\_grp <- ggplot(data= CaseStudy2\_raw\_hist) + geom\_bar(mapping=aes(fill=JobRole, x= Age\_Grp),show.legend = FALSE) + ggtitle("Age") + ylab("Number of employess")  
Job\_Lvl\_p\_Gender<- ggplot(data= CaseStudy2\_raw\_hist) + geom\_bar(mapping=aes(fill=JobRole, x= Gender),show.legend = FALSE) + ggtitle("Gender") + ylab("")  
Job\_Lvl\_p\_MaritalStatus<- ggplot(data= CaseStudy2\_raw\_hist) + geom\_bar(mapping=aes(fill=JobRole, x= MaritalStatus)) + ggtitle("MaritalStatus") + ylab("")  
Job\_Lvl\_p\_OverTime<- ggplot(data= CaseStudy2\_raw\_hist) + geom\_bar(mapping=aes(fill=JobRole, x= OverTime),show.legend = FALSE) + ggtitle("OverTime") + ylab("Number of employess")  
Job\_Lvl\_p\_JobInvolvement<- ggplot(data= CaseStudy2\_raw\_hist) + geom\_bar(mapping=aes(fill=JobRole, x= JobInvolvement),show.legend = FALSE) + ggtitle("JobInvolvement") + ylab("")  
Job\_Lvl\_p\_Education <- ggplot(data= CaseStudy2\_raw\_hist) + geom\_bar(mapping=aes(fill=JobRole, x= Education),show.legend = FALSE) + ggtitle("Education level") + ylab("")  
  
Job\_Lvl\_p\_EnvironmentSatisfaction <- ggplot(data= CaseStudy2\_raw\_hist) + geom\_bar(mapping=aes(fill=JobRole, x= EnvironmentSatisfaction),show.legend = FALSE) + ggtitle("EnvironmentSatisfaction") + ylab("Number of employess")  
#Job\_Lvl\_p\_JobInvolvement<- ggplot(data= CaseStudy2\_raw\_hist) + geom\_bar(mapping=aes(fill=JobRole, x= JobInvolvement),show.legend = FALSE) + ggtitle("JobInvolvement") + ylab("")  
Job\_Lvl\_p\_JobSatisfaction<- ggplot(data= CaseStudy2\_raw\_hist) + geom\_bar(mapping=aes(fill=JobRole, x= JobSatisfaction)) + ggtitle("JobSatisfaction") + ylab("")  
Job\_Lvl\_p\_RelationshipSatisfaction<- ggplot(data= CaseStudy2\_raw\_hist) + geom\_bar(mapping=aes(fill=JobRole, x= RelationshipSatisfaction),show.legend = FALSE) + ggtitle("RelationshipSatisfaction") + ylab("Number of employess")  
Job\_Lvl\_p\_StockOptionLevel<- ggplot(data= CaseStudy2\_raw\_hist) + geom\_bar(mapping=aes(fill=JobRole, x= StockOptionLevel),show.legend = FALSE) + ggtitle("StockOptionLevel") + ylab("")  
Job\_Lvl\_p\_WorkLifeBalance<- ggplot(data= CaseStudy2\_raw\_hist) + geom\_bar(mapping=aes(fill=JobRole, x = WorkLifeBalance),show.legend = FALSE) + ggtitle("WorkLifeBalance") + ylab("")  
  
Job\_Lvl\_p\_MonthlyIncome\_f<- ggplot(data= CaseStudy2\_raw\_hist) + geom\_bar(mapping=aes(fill=JobRole, x= MonthlyIncome\_f),show.legend = FALSE) + ggtitle("MonthlyIncome") + ylab("Number of employess")  
Job\_Lvl\_p\_MonthlyRate\_f<- ggplot(data= CaseStudy2\_raw\_hist) + geom\_bar(mapping=aes(fill=JobRole, x= MonthlyRate\_f),show.legend = FALSE) + ggtitle("MonthlyRate") + ylab("")  
Job\_Lvl\_p\_NumCompaniesWorked\_f<- ggplot(data= CaseStudy2\_raw\_hist) + geom\_bar(mapping=aes(fill=JobRole, x= NumCompaniesWorked\_f)) + ggtitle("NumCompaniesWorked") + ylab("")  
Job\_Lvl\_p\_TotalWorkingYears\_f<- ggplot(data= CaseStudy2\_raw\_hist) + geom\_bar(mapping=aes(fill=JobRole, x= TotalWorkingYears\_f),show.legend = FALSE) + ggtitle("TotalWorkingYears") + ylab("Number of employess")  
Job\_Lvl\_p\_TrainingTimesLastYear\_f<- ggplot(data= CaseStudy2\_raw\_hist) + geom\_bar(mapping=aes(fill=JobRole, x= TrainingTimesLastYear\_f),show.legend = FALSE) + ggtitle("TrainingTimesLastYear") + ylab("")  
Job\_Lvl\_p\_YearsAtCompany\_f<- ggplot(data= CaseStudy2\_raw\_hist) + geom\_bar(mapping=aes(fill=JobRole, x= YearsAtCompany\_f),show.legend = FALSE) + ggtitle("YearsAtCompany") + ylab("")  
  
Job\_Lvl\_p\_YearsInCurrentRole\_f<- ggplot(data= CaseStudy2\_raw\_hist) + geom\_bar(mapping=aes(fill=JobRole, x= YearsInCurrentRole\_f),show.legend = FALSE) + ggtitle("YearsInCurrentRole") + ylab("Number of employess")  
Job\_Lvl\_p\_YearsSinceLastPromotion\_f<- ggplot(data= CaseStudy2\_raw\_hist) + geom\_bar(mapping=aes(fill=JobRole, x= YearsSinceLastPromotion\_f)) + ggtitle("YearsSinceLastPromotion") + ylab("")  
Job\_Lvl\_p\_YearsWithCurrManager\_f<- ggplot(data= CaseStudy2\_raw\_hist) + geom\_bar(mapping=aes(fill=JobRole, x= YearsWithCurrManager\_f),show.legend = FALSE) + ggtitle("YearsWithCurrManager") + ylab("")  
Job\_Lvl\_p\_DistfrmHome <- ggplot(data= CaseStudy2\_raw\_hist) + geom\_bar(mapping=aes(fill=JobRole, x= DistfrmHome),show.legend = FALSE) + ggtitle("DistanceFromHome") + ylab("Number of employess") +xlab("DistanceFromHome")  
  
Job\_Lvl\_p\_BusinessTravel <- ggplot(data= CaseStudy2\_raw\_hist) + geom\_bar(mapping=aes(fill=JobRole, x= BusinessTravel),show.legend = FALSE) + ggtitle("BusinessTravel") + coord\_flip() + xlab("") + ylab("Number of employess")  
Job\_Lvl\_p\_Department<- ggplot(data= CaseStudy2\_raw\_hist) + geom\_bar(mapping=aes(fill=JobRole, x= Department),show.legend = FALSE) + ggtitle("Department")+ coord\_flip() + xlab("") + ylab("Number of employess")  
Job\_Lvl\_p\_EducationField<- ggplot(data= CaseStudy2\_raw\_hist) + geom\_bar(mapping=aes(fill=JobRole, x= EducationField)) + ggtitle("EducationField") + coord\_flip() + xlab("") + ylab("Number of employess")  
  
  
grid.arrange(Job\_Lvl\_p\_BusinessTravel,Job\_Lvl\_p\_Department,Job\_Lvl\_p\_EducationField,ncol = 3, nrow=1)



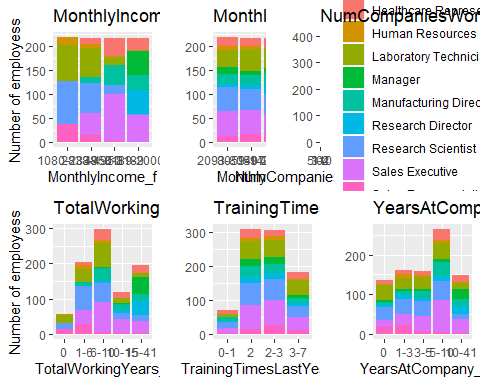
grid.arrange(Job\_Lvl\_p\_Age\_grp,Job\_Lvl\_p\_Gender,Job\_Lvl\_p\_MaritalStatus,Job\_Lvl\_p\_OverTime,Job\_Lvl\_p\_JobInvolvement,Job\_Lvl\_p\_Education,ncol = 3, nrow=2)



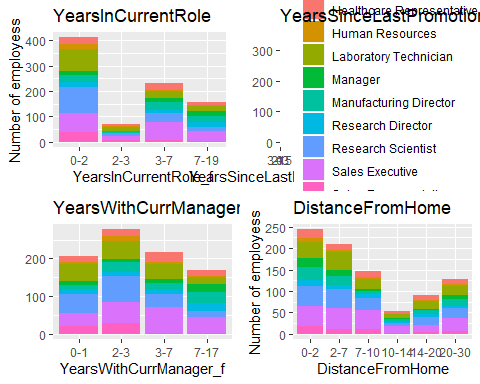
grid.arrange(Job\_Lvl\_p\_EnvironmentSatisfaction,Job\_Lvl\_p\_JobInvolvement,Job\_Lvl\_p\_JobSatisfaction,Job\_Lvl\_p\_RelationshipSatisfaction,Job\_Lvl\_p\_StockOptionLevel,Job\_Lvl\_p\_WorkLifeBalance,ncol = 3, nrow=2)



grid.arrange(Job\_Lvl\_p\_MonthlyIncome\_f,Job\_Lvl\_p\_MonthlyRate\_f,Job\_Lvl\_p\_NumCompaniesWorked\_f,Job\_Lvl\_p\_TotalWorkingYears\_f,Job\_Lvl\_p\_TrainingTimesLastYear\_f,Job\_Lvl\_p\_YearsAtCompany\_f,ncol = 3, nrow=2)



grid.arrange(Job\_Lvl\_p\_YearsInCurrentRole\_f,Job\_Lvl\_p\_YearsSinceLastPromotion\_f,Job\_Lvl\_p\_YearsWithCurrManager\_f,Job\_Lvl\_p\_DistfrmHome,ncol = 2, nrow=2)



#Classification and Regression Model

#Include encessary libraries  
library(tidyverse)  
library(ggplot2)  
library(caret)  
library(tidyr)  
library(dplyr)  
library(e1071)  
library(class)  
  
#Read Source file  
CaseStudy2\_raw <- read.csv(file.choose(), header = TRUE, sep=',')  
  
  
#Read No Attritions File  
CaseStudy2\_Comp\_set<- read.csv(file.choose(), header = TRUE, sep=',')  
  
  
#Read No Salary file  
CaseStudy2\_Comp\_Sal\_set<- read.csv(file.choose(), header = TRUE, sep=',')  
  
#colnames(CaseStudy2\_raw)  
#nrow(CaseStudy2\_raw)  
#head(CaseStudy2\_raw)  
colnames(CaseStudy2\_raw) <- c("ID", "Age", "Attrition", "BusinessTravel", "DailyRate", "Department", "DistanceFromHome", "Education", "EducationField", "EmployeeCount", "EmployeeNumber", "EnvironmentSatisfaction", "Gender", "HourlyRate", "JobInvolvement", "JobLevel", "JobRole", "JobSatisfaction", "MaritalStatus", "MonthlyIncome", "MonthlyRate", "NumCompaniesWorked", "Over18", "OverTime", "PercentSalaryHike", "PerformanceRating", "RelationshipSatisfaction", "StandardHours", "StockOptionLevel", "TotalWorkingYears", "TrainingTimesLastYear", "WorkLifeBalance", "YearsAtCompany", "YearsInCurrentRole", "YearsSinceLastPromotion", "YearsWithCurrManager")  
#Rename for test file  
#colnames(CaseStudy2\_raw) <- c("ID", "Age","BusinessTravel", "DailyRate", "Department", "DistanceFromHome", "Education", "EducationField", "EmployeeCount", "EmployeeNumber", "EnvironmentSatisfaction", "Gender", "HourlyRate", "JobInvolvement", "JobLevel", "JobRole", "JobSatisfaction", "MaritalStatus", "MonthlyIncome", "MonthlyRate", "NumCompaniesWorked", "Over18", "OverTime", "PercentSalaryHike", "PerformanceRating", "RelationshipSatisfaction", "StandardHours", "StockOptionLevel", "TotalWorkingYears", "TrainingTimesLastYear", "WorkLifeBalance", "YearsAtCompany", "YearsInCurrentRole", "YearsSinceLastPromotion", "YearsWithCurrManager")  
  
  
CaseStudy2\_raw <- CaseStudy2\_raw %>% filter(ID!=870)  
#summary(CaseStudy2\_raw)  
  
#Final\_Model\_Results <- data.frame( "Test\_Type" = character(), "Accuracy\_knn" = double(),"Accuracy\_NB" = double(),   
# "Specificity\_knn" = double(), "Specificity\_NB" = double(),   
# "Sensitivity\_knn" = double(), "Sensitivity\_NB" = double(), "DAta\_Set\_Name"=character())  
  
#Replace char values with numerical values - Business Travel(1-3), Gender(1,2), Marital Status -(1-3), OverTime (1,2), JobRole,EducationField,Department  
CaseStudy2\_raw$BusinessTravel\_N <- ifelse(str\_detect(CaseStudy2\_raw$BusinessTravel,'Non-Travel'),1, ifelse(str\_detect(CaseStudy2\_raw$BusinessTravel,'Travel\_Rarely'),2,ifelse(str\_detect(CaseStudy2\_raw$BusinessTravel,'Travel\_Frequently'),3,4)))  
CaseStudy2\_raw$Gender\_N <-ifelse(str\_detect(CaseStudy2\_raw$Gender,"^Male"),1, ifelse(str\_detect(CaseStudy2\_raw$Gender,"^Female"),2,3))  
CaseStudy2\_raw$Over18\_N <-ifelse(str\_detect(CaseStudy2\_raw$Over18,"Y"),1, 2)  
CaseStudy2\_raw$MaritalStatus\_N <- ifelse(str\_detect(CaseStudy2\_raw$MaritalStatus,'Single'),1, ifelse(str\_detect(CaseStudy2\_raw$MaritalStatus,'Married'),2,ifelse(str\_detect(CaseStudy2\_raw$MaritalStatus,'Divorced'),3,4)))  
CaseStudy2\_raw$OverTime\_N <-ifelse(str\_detect(CaseStudy2\_raw$OverTime,"No"),1, ifelse(str\_detect(CaseStudy2\_raw$OverTime,"Yes"),2,3))  
CaseStudy2\_raw$Department\_N <- ifelse(str\_detect(CaseStudy2\_raw$Department,"Human Resources"),1, ifelse(str\_detect(CaseStudy2\_raw$Department,"Research & Development"),2,3))  
CaseStudy2\_raw$EducationField\_N <- ifelse(str\_detect(CaseStudy2\_raw$Department,"Human Resources"),1,ifelse(str\_detect(CaseStudy2\_raw$Department,"Life Sciences"),2,  
 ifelse(str\_detect(CaseStudy2\_raw$Department,"Marketing"),3,   
 ifelse(str\_detect(CaseStudy2\_raw$Department, "Medical"),4,   
 ifelse(str\_detect(CaseStudy2\_raw$Department,"Other"),5, 6)))))   
  
CaseStudy2\_raw$JobRole\_N <- ifelse(str\_detect(CaseStudy2\_raw$JobRole,"Sales Executive"),1,ifelse(str\_detect(CaseStudy2\_raw$Department,"Research Scientist"),2,  
 ifelse(str\_detect(CaseStudy2\_raw$Department,"Laboratory Technician"),3,   
 ifelse(str\_detect(CaseStudy2\_raw$Department, "Manufacturing Director"),4,   
 ifelse(str\_detect(CaseStudy2\_raw$Department,"Healthcare Representative"),5,   
 ifelse(str\_detect(CaseStudy2\_raw$Department,"Sales Representative"),6,   
 7))))) )   
  
#CaseStudy2\_raw %>% select(JobRole\_N,JobRole)  
#summary(CaseStudy2\_raw)  
  
#Exclude employee number and employee cnt from the data  
#CaseStudy2\_N <- CaseStudy2\_raw %>% select("ID", "Attrition","Age","BusinessTravel\_N", "DailyRate", "Department\_N", "DistanceFromHome", "Education", "EducationField\_N", "EmployeeCount","EnvironmentSatisfaction", "Gender\_N", "HourlyRate", "JobInvolvement", "JobLevel", "JobRole\_N", "JobSatisfaction", "MaritalStatus\_N", "MonthlyIncome", "MonthlyRate", "NumCompaniesWorked", "Over18\_N", "OverTime\_N", "PercentSalaryHike", "PerformanceRating", "RelationshipSatisfaction", "StandardHours", "StockOptionLevel", "TotalWorkingYears", "TrainingTimesLastYear", "WorkLifeBalance", "YearsAtCompany", "YearsInCurrentRole", "YearsSinceLastPromotion", "YearsWithCurrManager")  
  
CaseStudy2\_N <- CaseStudy2\_raw %>% select("ID", "Attrition","Age","BusinessTravel\_N", "DailyRate", "Department\_N", "DistanceFromHome", "Education", "EducationField\_N","EnvironmentSatisfaction", "Gender\_N", "HourlyRate", "JobInvolvement", "JobLevel", "JobRole\_N", "JobSatisfaction", "MaritalStatus\_N", "MonthlyIncome", "MonthlyRate", "NumCompaniesWorked", "Over18\_N", "OverTime\_N", "PercentSalaryHike", "PerformanceRating", "RelationshipSatisfaction", "StandardHours", "StockOptionLevel", "TotalWorkingYears", "TrainingTimesLastYear", "WorkLifeBalance", "YearsAtCompany", "YearsInCurrentRole", "YearsSinceLastPromotion", "YearsWithCurrManager")  
  
  
#Plotting the data  
#plot(CaseStudy2\_N)  
  
  
##Base\_Data <- CaseStudy2\_N #data\_Grp\_1 # OR data\_Grp\_1 or data\_Grp\_2\_Nmbr  
#Base\_Data\_Name <- "CaseStudy2\_N" # or Data\_Group\_2  
#Number\_of\_Cols <- ncol(Base\_Data)  
#Start Modelling the data   
#1 Normalize the data  
#normalize <- function(x) {return ((x - min(x)) / (max(x) - min(x))) }  
#Base\_Data\_n <- as.data.frame(lapply(Asnumber\_Base\_Data[,3:Number\_of\_Cols], normalize))  
#head(Base\_Data\_n)  
#Base\_Data\_knn <- cbind(Base\_Data[1],Base\_Data[2],Base\_Data\_n)  
#head(Base\_Data\_knn)  
  
#Normalize the data using zscore  
library(jmotif)  
#Asnumber <- function(x) {return (as.numeric(x)) }  
#Asnumber\_Base\_Data <- as.data.frame(lapply(Base\_Data[,3:Number\_of\_Cols], Asnumber))  
#Number\_of\_Cols <- ncol(Asnumber\_Base\_Data)  
#Base\_Data\_Z <- znorm(Asnumber\_Base\_Data[,3]):Number\_of\_Cols]))  
#normalize(Base\_Data, method="Z-score")  
#summary(Asnumber\_Base\_Data)  
  
#CaseStudy2\_N   
Number\_of\_Cols\_1 <- ncol(CaseStudy2\_N)  
Asnumber <- function(x) {return (as.numeric(x)) } # function convert measures to numberic  
CaseStudy2\_N\_nmbr <- as.data.frame(lapply(CaseStudy2\_N[,3:Number\_of\_Cols\_1], Asnumber))  
znorm\_f <-function(x) {return (znorm(x)) } # function to Z-norm the data  
Number\_of\_Cols\_2 <- ncol(CaseStudy2\_N\_nmbr)  
znorm\_CaseStudy2\_N\_nmbr <- as.data.frame(lapply(CaseStudy2\_N\_nmbr[,1:Number\_of\_Cols\_2], znorm\_f))  
CaseStudy2\_N\_Z <- cbind(CaseStudy2\_N[1],CaseStudy2\_N[2],znorm\_CaseStudy2\_N\_nmbr) #combine z-norm data with ID and Attrition values  
#head(CaseStudy2\_N\_Z)  
Base\_Data <- CaseStudy2\_N\_Z  
Number\_of\_Cols <- ncol(CaseStudy2\_N\_Z)  
  
#2 Finding K value with highest accuracy  
Acc= matrix(nrow=100, ncol=100)  
for (seed in 1:100)  
{  
   
 #seed=4  
 set.seed(seed)  
 trainIndices = sample(seq(1:nrow(Base\_Data)),round(.7\*nrow(Base\_Data)))  
 train\_data = Base\_Data[trainIndices,]  
 test\_data = Base\_Data[-trainIndices,]   
 #nrow(train\_CaseStudy2\_raw) #609  
 #nrow(test\_CaseStudy2\_raw) #261  
 ncol(train\_data)  
   
 #For loop to run KNN model for k=1-100   
 for( i in 1:100) {  
 # knn1 <- knn(train\_CaseStudy2\_raw[,c(1,18)],test\_CaseStudy2\_raw[,c(1,18)],train\_CaseStudy2\_raw$Attrition, k=i )  
 knn1 <- knn(train\_data[,3:Number\_of\_Cols],test\_data[,3:Number\_of\_Cols],train\_data$Attrition, k=i )#ERROR if data not normalized  
 t\_knn <- table(knn1, test\_data$Attrition)  
 CM\_knn <- confusionMatrix(t\_knn)  
 #Capture Accuracy  
 #ACC <- CM\_knn$overall[1]  
 Acc[seed,i] = CM\_knn$overall[1]  
   
 }   
}  
#Caluclate mean accuracy for each of the seed/knn iteritions  
mean\_Acc = colMeans(Acc)  
i\_highest\_acc <- which.max(mean\_Acc)  
  
#3. Run Knn and NB on the data set with different seed values  
Conclusion\_job\_sat <- data.frame("Seed" = double() , "Accuracy\_knn" = double(), "Accuracy\_NB" = double(),"Specificity\_knn" = double() ,  
 "Specificity\_NB" = double(), "Sensitivity\_knn" = double() , "Sensitivity\_NB" = double() )  
  
for (seed in 1:100)  
{  
 set.seed(seed)  
 #Split the data set into train and test  
 trainIndices = sample(seq(1:nrow(Base\_Data)),round(.7\*nrow(Base\_Data)))  
 train\_data = Base\_Data[trainIndices,]  
 test\_data = Base\_Data[-trainIndices,]   
   
 #KNN  
 knn1 <- knn(train\_data[,3:Number\_of\_Cols],test\_data[,3:Number\_of\_Cols],train\_data$Attrition, k=i\_highest\_acc )#ERROR - TRY TO NORMALISE THE DATA  
 t\_knn <- table(knn1, test\_data$Attrition)  
 CM\_knn <- confusionMatrix(t\_knn)  
   
 #NB  
 model = naiveBayes(train\_data[,3:Number\_of\_Cols], train\_data$Attrition)  
 NB\_Predict <- predict(model,test\_data[,3:Number\_of\_Cols])  
 t\_Predict <- table(NB\_Predict,test\_data$Attrition)  
 CM\_NB <- confusionMatrix(t\_Predict)  
   
 #Capture the comparision in a dataframe  
 Conclusion <- data.frame( "Seed" = seed, "Accuracy\_knn" = CM\_knn$overall[1],"Accuracy\_NB" = CM\_NB$overall[1],   
 "Specificity\_knn" = CM\_knn$byClass[2], "Specificity\_NB" = CM\_NB$byClass[2],   
 "Sensitivity\_knn" = CM\_knn$byClass[1], "Sensitivity\_NB" = CM\_NB$byClass[1])  
 #Merge the 2 dataframes  
 Conclusion\_job\_sat <- rbind(Conclusion\_job\_sat,Conclusion)   
}  
  
#Calculate Mean Accuracy, Specificity and Senitivity across all seeds for each model  
mean\_Accuracy\_knn <- mean(Conclusion\_job\_sat$Accuracy\_knn)  
mean\_Accuracy\_NB <- mean(Conclusion\_job\_sat$Accuracy\_NB)  
mean\_Specificity\_knn <- mean(Conclusion\_job\_sat$Specificity\_knn)  
mean\_Specificity\_NB <- mean(Conclusion\_job\_sat$Specificity\_NB)  
mean\_Sensitivity\_knn <- mean(Conclusion\_job\_sat$Sensitivity\_knn)  
mean\_Sensitivity\_NB <- mean(Conclusion\_job\_sat$Sensitivity\_NB)  
  
#Putting all Means in one dataframe  
Mean\_Knn\_NB <- data.frame( Test\_Type = c("KNN", "NB"), Mean\_Accuracy = c(mean\_Accuracy\_knn, mean\_Accuracy\_NB),   
 Mean\_Specificity=c(mean\_Specificity\_knn,mean\_Specificity\_NB ),  
 Mean\_Sensitivity=c(mean\_Sensitivity\_knn, mean\_Sensitivity\_NB))  
  
#print(paste("Modelling results for data set",Base\_Data\_Name) )   
#Mean\_Knn\_NB$DAta\_Set\_Name <- Base\_Data\_Name  
Mean\_Knn\_NB

## Test\_Type Mean\_Accuracy Mean\_Specificity Mean\_Sensitivity  
## 1 KNN 0.8490038 0.08641722 0.9975938  
## 2 NB 0.8224521 0.56830219 0.8722358

#Capturing results for different data groups and different models  
#Final\_Model\_Results <- rbind(Final\_Model\_Results,Mean\_Knn\_NB)  
#Final\_Model\_Results  
  
#NB Prediction:   
#Testing the provided file  
  
colnames(CaseStudy2\_Comp\_set)

## [1] "ï..ID" "Age"   
## [3] "BusinessTravel" "DailyRate"   
## [5] "Department" "DistanceFromHome"   
## [7] "Education" "EducationField"   
## [9] "EmployeeCount" "EmployeeNumber"   
## [11] "EnvironmentSatisfaction" "Gender"   
## [13] "HourlyRate" "JobInvolvement"   
## [15] "JobLevel" "JobRole"   
## [17] "JobSatisfaction" "MaritalStatus"   
## [19] "MonthlyIncome" "MonthlyRate"   
## [21] "NumCompaniesWorked" "Over18"   
## [23] "OverTime" "PercentSalaryHike"   
## [25] "PerformanceRating" "RelationshipSatisfaction"  
## [27] "StandardHours" "StockOptionLevel"   
## [29] "TotalWorkingYears" "TrainingTimesLastYear"   
## [31] "WorkLifeBalance" "YearsAtCompany"   
## [33] "YearsInCurrentRole" "YearsSinceLastPromotion"   
## [35] "YearsWithCurrManager"

colnames(CaseStudy2\_Comp\_set) <- c("ID", "Age","BusinessTravel", "DailyRate", "Department", "DistanceFromHome", "Education", "EducationField", "EmployeeCount", "EmployeeNumber", "EnvironmentSatisfaction", "Gender", "HourlyRate", "JobInvolvement", "JobLevel", "JobRole", "JobSatisfaction", "MaritalStatus", "MonthlyIncome", "MonthlyRate", "NumCompaniesWorked", "Over18", "OverTime", "PercentSalaryHike", "PerformanceRating", "RelationshipSatisfaction", "StandardHours", "StockOptionLevel", "TotalWorkingYears", "TrainingTimesLastYear", "WorkLifeBalance", "YearsAtCompany", "YearsInCurrentRole", "YearsSinceLastPromotion", "YearsWithCurrManager")  
CaseStudy2\_Comp\_set$BusinessTravel\_N <- ifelse(str\_detect(CaseStudy2\_Comp\_set$BusinessTravel,'Non-Travel'),1, ifelse(str\_detect(CaseStudy2\_Comp\_set$BusinessTravel,'Travel\_Rarely'),2,ifelse(str\_detect(CaseStudy2\_Comp\_set$BusinessTravel,'Travel\_Frequently'),3,4)))  
CaseStudy2\_Comp\_set$Gender\_N <-ifelse(str\_detect(CaseStudy2\_Comp\_set$Gender,"^Male"),1, ifelse(str\_detect(CaseStudy2\_Comp\_set$Gender,"^Female"),2,3))  
CaseStudy2\_Comp\_set$Over18\_N <-ifelse(str\_detect(CaseStudy2\_Comp\_set$Over18,"Y"),1, 2)  
CaseStudy2\_Comp\_set$MaritalStatus\_N <- ifelse(str\_detect(CaseStudy2\_Comp\_set$MaritalStatus,'Single'),1, ifelse(str\_detect(CaseStudy2\_Comp\_set$MaritalStatus,'Married'),2,ifelse(str\_detect(CaseStudy2\_Comp\_set$MaritalStatus,'Divorced'),3,4)))  
CaseStudy2\_Comp\_set$OverTime\_N <-ifelse(str\_detect(CaseStudy2\_Comp\_set$OverTime,"No"),1, ifelse(str\_detect(CaseStudy2\_Comp\_set$OverTime,"Yes"),2,3))  
CaseStudy2\_Comp\_set$Department\_N <- ifelse(str\_detect(CaseStudy2\_Comp\_set$Department,"Human Resources"),1, ifelse(str\_detect(CaseStudy2\_Comp\_set$Department,"Research & Development"),2,3))  
CaseStudy2\_Comp\_set$EducationField\_N <- ifelse(str\_detect(CaseStudy2\_Comp\_set$Department,"Human Resources"),1,ifelse(str\_detect(CaseStudy2\_Comp\_set$Department,"Life Sciences"),2,  
 ifelse(str\_detect(CaseStudy2\_Comp\_set$Department,"Marketing"),3,   
 ifelse(str\_detect(CaseStudy2\_Comp\_set$Department, "Medical"),4,   
 ifelse(str\_detect(CaseStudy2\_Comp\_set$Department,"Other"),5, 6)))))   
  
CaseStudy2\_Comp\_set$JobRole\_N <- ifelse(str\_detect(CaseStudy2\_Comp\_set$JobRole,"Sales Executive"),1,ifelse(str\_detect(CaseStudy2\_Comp\_set$Department,"Research Scientist"),2,  
 ifelse(str\_detect(CaseStudy2\_Comp\_set$Department,"Laboratory Technician"),3,   
 ifelse(str\_detect(CaseStudy2\_Comp\_set$Department, "Manufacturing Director"),4,   
 ifelse(str\_detect(CaseStudy2\_Comp\_set$Department,"Healthcare Representative"),5,   
 ifelse(str\_detect(CaseStudy2\_Comp\_set$Department,"Sales Representative"),6,   
 7))))) )   
  
  
#Exclude employee number and employee cnt from the data  
CaseStudy2\_comp\_N <- CaseStudy2\_Comp\_set %>% select("ID", "Age","BusinessTravel\_N", "DailyRate", "Department\_N", "DistanceFromHome", "Education", "EducationField\_N","EnvironmentSatisfaction", "Gender\_N", "HourlyRate", "JobInvolvement", "JobLevel", "JobRole\_N", "JobSatisfaction", "MaritalStatus\_N", "MonthlyIncome", "MonthlyRate", "NumCompaniesWorked", "Over18\_N", "OverTime\_N", "PercentSalaryHike", "PerformanceRating", "RelationshipSatisfaction", "StandardHours", "StockOptionLevel", "TotalWorkingYears", "TrainingTimesLastYear", "WorkLifeBalance", "YearsAtCompany", "YearsInCurrentRole", "YearsSinceLastPromotion", "YearsWithCurrManager")  
  
Number\_of\_Cols\_1 <- ncol(CaseStudy2\_comp\_N)  
Asnumber <- function(x) {return (as.numeric(x)) } # function convert measures to numberic  
CaseStudy2\_comp\_N\_nmbr <- as.data.frame(lapply(CaseStudy2\_comp\_N[,2:Number\_of\_Cols\_1], Asnumber))  
znorm\_f <-function(x) {return (znorm(x)) } # function to Z-norm the data  
Number\_of\_Cols\_2 <- ncol(CaseStudy2\_comp\_N\_nmbr)  
znorm\_CaseStudy2\_comp\_N\_nmbr <- as.data.frame(lapply(CaseStudy2\_comp\_N\_nmbr[,1:Number\_of\_Cols\_2], znorm\_f))  
CaseStudy2\_comp\_N\_Z <- cbind(CaseStudy2\_comp\_N[1],znorm\_CaseStudy2\_comp\_N\_nmbr) #combine z-norm data with ID and Attrition values  
#head(CaseStudy2\_N\_Z)  
Base\_Data <- CaseStudy2\_comp\_N\_Z  
Number\_of\_Cols <- ncol(CaseStudy2\_comp\_N\_Z)  
  
NB\_Predict\_Att <- predict(model,CaseStudy2\_comp\_N\_Z[,2:Number\_of\_Cols])  
nrow(CaseStudy2\_comp\_N\_Z)

## [1] 300

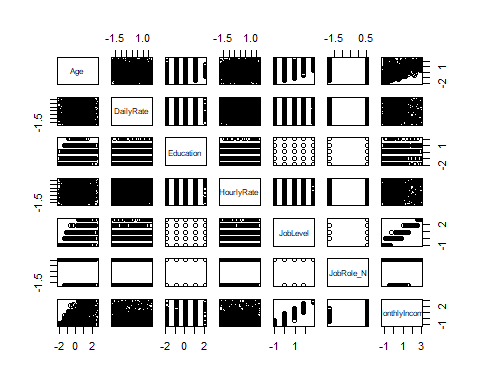
answers <- cbind(NB\_Predict\_Att,CaseStudy2\_comp\_N\_Z[1])  
head(answers)

## NB\_Predict\_Att ID  
## 1 No 1171  
## 2 No 1172  
## 3 No 1173  
## 4 No 1174  
## 5 No 1175  
## 6 No 1176

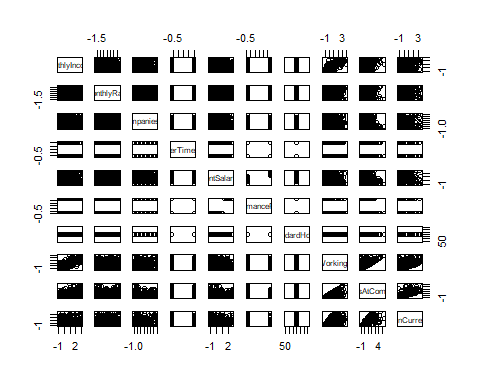
colnames(answers) <- c('Predcited Attrition','ID')  
joined\_data <- left\_join(answers, CaseStudy2\_Comp\_set, by = "ID")  
head(joined\_data)

## Predcited Attrition ID Age BusinessTravel DailyRate  
## 1 No 1171 35 Travel\_Rarely 750  
## 2 No 1172 33 Travel\_Rarely 147  
## 3 No 1173 26 Travel\_Rarely 1330  
## 4 No 1174 55 Travel\_Rarely 1311  
## 5 No 1175 29 Travel\_Rarely 1246  
## 6 No 1176 51 Travel\_Frequently 1456  
## Department DistanceFromHome Education EducationField  
## 1 Research & Development 28 3 Life Sciences  
## 2 Human Resources 2 3 Human Resources  
## 3 Research & Development 21 3 Medical  
## 4 Research & Development 2 3 Life Sciences  
## 5 Sales 19 3 Life Sciences  
## 6 Research & Development 1 4 Medical  
## EmployeeCount EmployeeNumber EnvironmentSatisfaction Gender HourlyRate  
## 1 1 1596 2 Male 46  
## 2 1 1207 2 Male 99  
## 3 1 1107 1 Male 37  
## 4 1 505 3 Female 97  
## 5 1 1497 3 Male 77  
## 6 1 145 1 Female 30  
## JobInvolvement JobLevel JobRole JobSatisfaction  
## 1 4 2 Laboratory Technician 3  
## 2 3 1 Human Resources 3  
## 3 3 1 Laboratory Technician 3  
## 4 3 4 Manager 4  
## 5 2 2 Sales Executive 3  
## 6 2 3 Healthcare Representative 1  
## MaritalStatus MonthlyIncome MonthlyRate NumCompaniesWorked Over18  
## 1 Married 3407 25348 1 Y  
## 2 Married 3600 8429 1 Y  
## 3 Divorced 2377 19373 1 Y  
## 4 Single 16659 23258 2 Y  
## 5 Divorced 8620 23757 1 Y  
## 6 Single 7484 25796 3 Y  
## OverTime PercentSalaryHike PerformanceRating RelationshipSatisfaction  
## 1 No 17 3 4  
## 2 No 13 3 4  
## 3 No 20 4 3  
## 4 Yes 13 3 3  
## 5 No 14 3 3  
## 6 No 20 4 3  
## StandardHours StockOptionLevel TotalWorkingYears TrainingTimesLastYear  
## 1 80 2 10 3  
## 2 80 1 5 2  
## 3 80 1 1 0  
## 4 80 0 30 2  
## 5 80 2 10 3  
## 6 80 0 23 1  
## WorkLifeBalance YearsAtCompany YearsInCurrentRole  
## 1 2 10 9  
## 2 3 5 4  
## 3 2 1 1  
## 4 3 5 4  
## 5 3 10 7  
## 6 2 13 12  
## YearsSinceLastPromotion YearsWithCurrManager BusinessTravel\_N Gender\_N  
## 1 6 8 2 1  
## 2 1 4 2 1  
## 3 0 0 2 1  
## 4 1 2 2 2  
## 5 0 4 2 1  
## 6 12 8 3 2  
## Over18\_N MaritalStatus\_N OverTime\_N Department\_N EducationField\_N  
## 1 1 2 1 2 6  
## 2 1 2 1 1 1  
## 3 1 3 1 2 6  
## 4 1 1 2 2 6  
## 5 1 3 1 3 6  
## 6 1 1 1 2 6  
## JobRole\_N  
## 1 7  
## 2 7  
## 3 7  
## 4 7  
## 5 1  
## 6 7

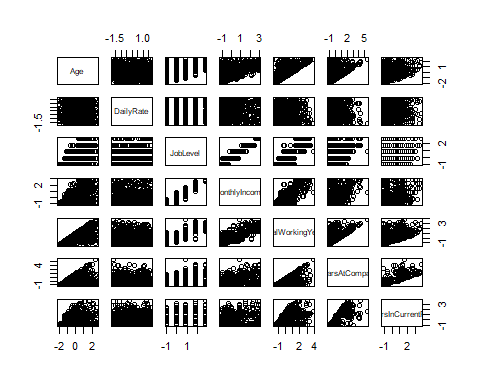
output <- select (joined\_data,'Predcited Attrition',"ID", "Age","BusinessTravel", "DailyRate", "Department", "DistanceFromHome", "Education", "EducationField", "EmployeeCount", "EmployeeNumber", "EnvironmentSatisfaction", "Gender", "HourlyRate", "JobInvolvement", "JobLevel", "JobRole", "JobSatisfaction", "MaritalStatus", "MonthlyIncome", "MonthlyRate", "NumCompaniesWorked", "Over18", "OverTime", "PercentSalaryHike", "PerformanceRating", "RelationshipSatisfaction", "StandardHours", "StockOptionLevel", "TotalWorkingYears", "TrainingTimesLastYear", "WorkLifeBalance", "YearsAtCompany", "YearsInCurrentRole", "YearsSinceLastPromotion", "YearsWithCurrManager")   
#PUT THE ANSWERS IN CSV  
  
#Write output  
write.csv(output, file = "C:/Users/simer/Documents/SMU/Doing\_DataScience/Case\_Study\_2/Case2PredictionsClassifyREDDY.csv")  
  
#EDA for Linear Regression  
  
  
#EDA  
M\_p1 <- plot(CaseStudy2\_N\_Z[,c(3,5,8,12,14,15,18)])



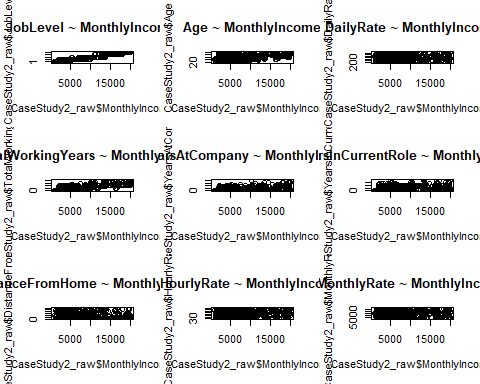
M\_p2 <- plot(CaseStudy2\_N\_Z[,c(18,19,20,22,23,24,26,28,31,32)])



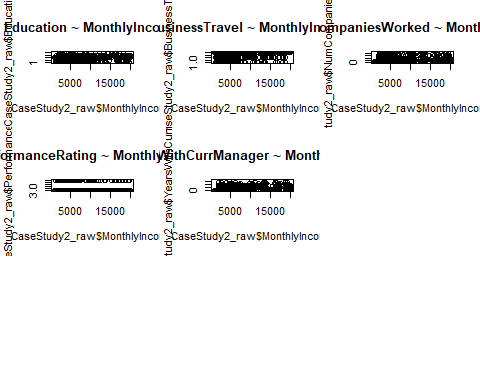
plot(CaseStudy2\_N\_Z[,c(3, 5, 14, 18,28,31,32)])



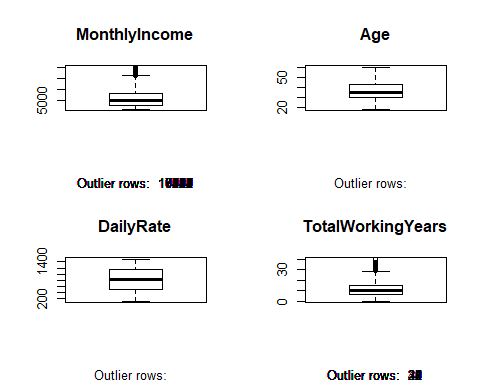
#summary(CaseStudy2\_raw)  
#head(CaseStudy2\_N\_Z)  
#Columns of interest : MonthlyIncome,JobLevel,Age,DailyRate,TotalWorkingYears,YearsAtCompany,YearsInCurrentRole  
par(mfrow=c(3, 3))   
  
#Scatter Plot of Shortlisted measures  
scatter.smooth(x=CaseStudy2\_raw$MonthlyIncome, y=CaseStudy2\_raw$JobLevel, main="JobLevel ~ MonthlyIncome")  
scatter.smooth(x=CaseStudy2\_raw$MonthlyIncome, y=CaseStudy2\_raw$Age, main="Age ~ MonthlyIncome")  
scatter.smooth(x=CaseStudy2\_raw$MonthlyIncome, y=CaseStudy2\_raw$DailyRate, main="DailyRate ~ MonthlyIncome") #pretty flat  
scatter.smooth(x=CaseStudy2\_raw$MonthlyIncome, y=CaseStudy2\_raw$TotalWorkingYears, main="TotalWorkingYears ~ MonthlyIncome")  
scatter.smooth(x=CaseStudy2\_raw$MonthlyIncome, y=CaseStudy2\_raw$YearsAtCompany, main="YearsAtCompany ~ MonthlyIncome")  
scatter.smooth(x=CaseStudy2\_raw$MonthlyIncome, y=CaseStudy2\_raw$YearsInCurrentRole, main="YearsInCurrentRole ~ MonthlyIncome")  
  
#Extra measures  
scatter.smooth(x=CaseStudy2\_raw$MonthlyIncome, y=CaseStudy2\_raw$DistanceFromHome, main="DistanceFromHome ~ MonthlyIncome")  
scatter.smooth(x=CaseStudy2\_raw$MonthlyIncome, y=CaseStudy2\_raw$HourlyRate, main="HourlyRate ~ MonthlyIncome")  
scatter.smooth(x=CaseStudy2\_raw$MonthlyIncome, y=CaseStudy2\_raw$MonthlyRate, main="MonthlyRate ~ MonthlyIncome")



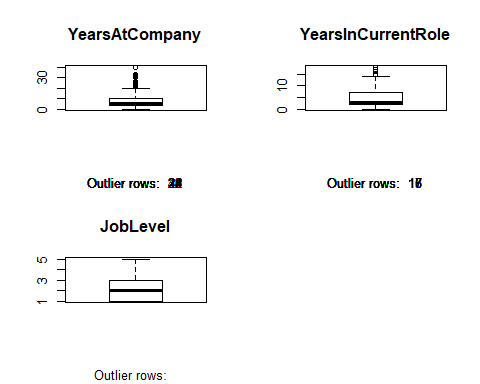
scatter.smooth(x=CaseStudy2\_raw$MonthlyIncome, y=CaseStudy2\_raw$Education, main="Education ~ MonthlyIncome")  
scatter.smooth(x=CaseStudy2\_raw$MonthlyIncome, y=CaseStudy2\_raw$BusinessTravel, main="BusinessTravel ~ MonthlyIncome")  
scatter.smooth(x=CaseStudy2\_raw$MonthlyIncome, y=CaseStudy2\_raw$NumCompaniesWorked, main="NumCompaniesWorked ~ MonthlyIncome")  
scatter.smooth(x=CaseStudy2\_raw$MonthlyIncome, y=CaseStudy2\_raw$PerformanceRating, main="PerformanceRating ~ MonthlyIncome")  
scatter.smooth(x=CaseStudy2\_raw$MonthlyIncome, y=CaseStudy2\_raw$YearsWithCurrManager, main="YearsWithCurrManager ~ MonthlyIncome")  
  
# BOX PLOT to see outliers : Let's look at Job Level, Age, Total working years  
par(mfrow=c(2, 2)) # divide graph area in 2 columns



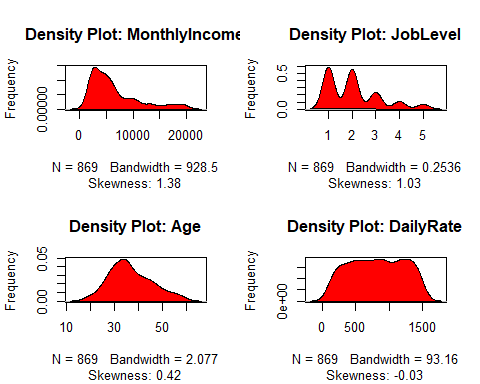
boxplot(CaseStudy2\_raw$MonthlyIncome, main="MonthlyIncome", sub=paste("Outlier rows: ", boxplot.stats(CaseStudy2\_raw$MonthlyIncome)$out)) # box plot for 'speed'  
boxplot(CaseStudy2\_raw$Age, main="Age", sub=paste("Outlier rows: ", boxplot.stats(CaseStudy2\_raw$Age)$out))   
boxplot(CaseStudy2\_raw$DailyRate, main="DailyRate", sub=paste("Outlier rows: ", boxplot.stats(CaseStudy2\_raw$DailyRate)$out))   
boxplot(CaseStudy2\_raw$TotalWorkingYears, main="TotalWorkingYears", sub=paste("Outlier rows: ", boxplot.stats(CaseStudy2\_raw$TotalWorkingYears)$out))



boxplot(CaseStudy2\_raw$YearsAtCompany, main="YearsAtCompany", sub=paste("Outlier rows: ", boxplot.stats(CaseStudy2\_raw$YearsAtCompany)$out))   
boxplot(CaseStudy2\_raw$YearsInCurrentRole, main="YearsInCurrentRole", sub=paste("Outlier rows: ", boxplot.stats(CaseStudy2\_raw$YearsInCurrentRole)$out))   
boxplot(CaseStudy2\_raw$JobLevel, main="JobLevel", sub=paste("Outlier rows: ", boxplot.stats(CaseStudy2\_raw$JobLevel)$out))   
  
# Density Plot  
par(mfrow=c(2, 2)) # divide graph area in 2 columns



plot(density(CaseStudy2\_raw$MonthlyIncome), main="Density Plot: MonthlyIncome", ylab="Frequency", sub=paste("Skewness:", round(e1071::skewness(CaseStudy2\_raw$MonthlyIncome), 2))) # density plot for 'MonthlyIncome'  
polygon(density(CaseStudy2\_raw$MonthlyIncome), col="red")  
  
plot(density(CaseStudy2\_raw$JobLevel), main="Density Plot: JobLevel", ylab="Frequency", sub=paste("Skewness:", round(e1071::skewness(CaseStudy2\_raw$JobLevel), 2))) # density plot for 'MonthlyIncome'  
polygon(density(CaseStudy2\_raw$JobLevel), col="red")  
  
plot(density(CaseStudy2\_raw$Age), main="Density Plot: Age", ylab="Frequency", sub=paste("Skewness:", round(e1071::skewness(CaseStudy2\_raw$Age), 2))) # density plot for 'Age'  
polygon(density(CaseStudy2\_raw$Age), col="red")  
  
plot(density(CaseStudy2\_raw$DailyRate), main="Density Plot: DailyRate", ylab="Frequency", sub=paste("Skewness:", round(e1071::skewness(CaseStudy2\_raw$DailyRate), 2))) # density plot for 'DailyRate'  
polygon(density(CaseStudy2\_raw$DailyRate), col="red")



plot(density(CaseStudy2\_raw$TotalWorkingYears), main="Density Plot: TotalWorkingYears", ylab="Frequency", sub=paste("Skewness:", round(e1071::skewness(CaseStudy2\_raw$TotalWorkingYears), 2))) # density plot for 'TotalWorkingYears'  
polygon(density(CaseStudy2\_raw$TotalWorkingYears), col="red")  
  
plot(density(CaseStudy2\_raw$YearsAtCompany), main="Density Plot: YearsAtCompany", ylab="Frequency", sub=paste("Skewness:", round(e1071::skewness(CaseStudy2\_raw$YearsAtCompany), 2))) # density plot for 'YearsAtCompany'  
polygon(density(CaseStudy2\_raw$YearsAtCompany), col="red")  
  
plot(density(CaseStudy2\_raw$YearsInCurrentRole), main="Density Plot: YearsInCurrentRole", ylab="Frequency", sub=paste("Skewness:", round(e1071::skewness(CaseStudy2\_raw$YearsInCurrentRole), 2))) # density plot for 'CaseStudy2\_raw$YearsInCurrentRole'  
polygon(density(CaseStudy2\_raw$YearsInCurrentRole), col="red")  
  
#Correlation  
cor(CaseStudy2\_raw$MonthlyIncome, CaseStudy2\_raw$JobLevel) #0.9517233

## [1] 0.9517233

cor(CaseStudy2\_raw$MonthlyIncome, CaseStudy2\_raw$TotalWorkingYears) #0.7785329

## [1] 0.7785329

cor(CaseStudy2\_raw$MonthlyIncome, CaseStudy2\_raw$Age) #0.4842501

## [1] 0.4842501

#Let's start building lm with variable with highest correlation factor  
#JOB LEVEL  
linearMod\_JL <- lm(MonthlyIncome ~ JobLevel, data=CaseStudy2\_raw) # build linear regression model on full data  
print(linearMod\_JL)

##   
## Call:  
## lm(formula = MonthlyIncome ~ JobLevel, data = CaseStudy2\_raw)  
##   
## Coefficients:  
## (Intercept) JobLevel   
## -1792 4014

#MonthlyIncome = Intercept + (β ∗ JobLevel)  
#MonthlyIncome = -1792 + 4014 \* JobLevel  
modelSummary\_JL <- summary(linearMod\_JL)  
modelCoeffs\_JL <- modelSummary\_JL$coefficients  
AIC\_JL <- AIC(linearMod\_JL)  
BIC\_JL <- BIC(linearMod\_JL)  
pred\_MonthlyIncome <- predict(linearMod\_JL, newdata = CaseStudy2\_raw)  
t\_JL\_Model\_1 <- train(MonthlyIncome ~ JobLevel, method = "lm", data = CaseStudy2\_raw, trControl = trainControl(method = "LOOCV"))  
summary(t\_JL\_Model\_1)

##   
## Call:  
## lm(formula = .outcome ~ ., data = dat)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -5039.1 -926.5 80.1 695.5 3721.5   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -1791.71 101.65 -17.63 <2e-16 \*\*\*  
## JobLevel 4013.60 43.96 91.29 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1413 on 867 degrees of freedom  
## Multiple R-squared: 0.9058, Adjusted R-squared: 0.9057   
## F-statistic: 8335 on 1 and 867 DF, p-value: < 2.2e-16

t\_JL\_Model\_1$results #RMSE 1414.445/ r-sq 0.9053364

## intercept RMSE Rsquared MAE  
## 1 TRUE 1414.445 0.9053364 1075.208

MonthlyIncome\_preds = predict(linearMod\_JL)  
CaseStudy2\_raw %>% ggplot(aes(x = MonthlyIncome, y = JobLevel)) + geom\_point() + geom\_line(data = CaseStudy2\_raw, aes( x = MonthlyIncome\_preds, y = JobLevel, col = "red")) + ggtitle("LR Model: JobLevel vs MonthlyIncome") + scale\_color\_discrete(name = "Predicted")  
  
linearMod\_JL\_sq <- lm(MonthlyIncome ~ JobLevel + I(JobLevel^2), data=CaseStudy2\_raw) #  
linearMod\_JL\_sq\_summary <- summary(linearMod\_JL\_sq)  
linearMod\_JL\_sq\_summary$coefficients

## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 296.6038 190.66100 1.555661 1.201542e-01  
## JobLevel 1948.8735 169.17125 11.520122 1.136307e-28  
## I(JobLevel^2) 396.8903 31.57535 12.569627 2.100722e-33

CaseStudy2\_raw$JobLevelSQ <- CaseStudy2\_raw$JobLevel^2  
t\_JL\_Model\_2 <- train(MonthlyIncome ~ JobLevel + JobLevelSQ, method = "lm", data = CaseStudy2\_raw, trControl = trainControl(method = "LOOCV"))  
summary(t\_JL\_Model\_2)

##   
## Call:  
## lm(formula = .outcome ~ ., data = dat)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -4505.2 -739.9 -130.2 655.6 4175.1   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 296.60 190.66 1.556 0.12   
## JobLevel 1948.87 169.17 11.520 <2e-16 \*\*\*  
## JobLevelSQ 396.89 31.58 12.570 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1300 on 866 degrees of freedom  
## Multiple R-squared: 0.9203, Adjusted R-squared: 0.9201   
## F-statistic: 5001 on 2 and 866 DF, p-value: < 2.2e-16

CaseStudy2\_raw <- CaseStudy2\_raw %>% mutate(JobLevelSQ = JobLevel^2)  
MonthlyIncome\_preds2 = predict(t\_JL\_Model\_2)  
t\_JL\_Model\_2$results # #RMSE 1436.79/ r-sq 0.9262231

## intercept RMSE Rsquared MAE  
## 1 TRUE 1302.034 0.9197854 963.1192

#MODEL 1 IS BETTER  
#t\_JL\_Model\_3 <- train(MonthlyIncome ~ JobLevel + I(JobLevel^2) + I(JobLevel^3), method = "lm", data = CaseStudy2\_raw, trControl = trainControl(method = "LOOCV"))  
#t\_JL\_Model\_3$results # #RMSE 1396.326  
#summary(t\_JL\_Model\_3)  
  
#t\_JL\_Model\_4 <- train(MonthlyIncome ~ JobLevel + I(JobLevel^2) + I(JobLevel^3) + I(JobLevel^4), method = "lm", data = CaseStudy2\_raw, trControl = trainControl(method = "LOOCV"))  
#t\_JL\_Model\_4$results # #RMSE 1378.202  
#summary(t\_JL\_Model\_4)  
  
#t\_JL\_Model\_5 <- train(MonthlyIncome ~ JobLevel + I(JobLevel^2) + I(JobLevel^3) + I(JobLevel^4) + I(JobLevel^5), method = "lm", data = CaseStudy2\_raw, trControl = trainControl(method = "LOOCV"))  
#t\_JL\_Model\_5$results #RMSE 1378.202  
#summary(t\_JL\_Model\_5)  
  
#TotalWorkingYears  
linearMod\_T\_Wrkng\_Yrs\_t <- train(MonthlyIncome ~ TotalWorkingYears , method = "lm", data = CaseStudy2\_raw, trControl = trainControl(method = "LOOCV"))  
summary(linearMod\_T\_Wrkng\_Yrs\_t)

##   
## Call:  
## lm(formula = .outcome ~ ., data = dat)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -9868.9 -1706.1 -44.7 1454.3 11007.9   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1126.93 174.31 6.465 1.69e-10 \*\*\*  
## TotalWorkingYears 476.35 13.04 36.526 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2889 on 867 degrees of freedom  
## Multiple R-squared: 0.6061, Adjusted R-squared: 0.6057   
## F-statistic: 1334 on 1 and 867 DF, p-value: < 2.2e-16

linearMod\_T\_Wrkng\_Yrs\_t$results #RmSE 2893.807 increased

## intercept RMSE Rsquared MAE  
## 1 TRUE 2893.807 0.6037701 2164.422

CaseStudy2\_raw <- CaseStudy2\_raw %>% mutate(TotalWorkingYearsSQ = TotalWorkingYears^2)  
linearMod\_T\_Wrkng\_Yrs\_t2 <- train(MonthlyIncome ~ TotalWorkingYears + TotalWorkingYearsSQ, method = "lm", data = CaseStudy2\_raw, trControl = trainControl(method = "LOOCV"))  
summary(linearMod\_T\_Wrkng\_Yrs\_t2)

##   
## Call:  
## lm(formula = .outcome ~ ., data = dat)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -10925.4 -1700.6 -102.1 1421.0 11076.6   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1443.096 270.105 5.343 1.17e-07 \*\*\*  
## TotalWorkingYears 415.809 41.623 9.990 < 2e-16 \*\*\*  
## TotalWorkingYearsSQ 1.976 1.290 1.531 0.126   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2886 on 866 degrees of freedom  
## Multiple R-squared: 0.6072, Adjusted R-squared: 0.6063   
## F-statistic: 669.3 on 2 and 866 DF, p-value: < 2.2e-16

linearMod\_T\_Wrkng\_Yrs\_t2$results # rmsse 2896, never mind

## intercept RMSE Rsquared MAE  
## 1 TRUE 2896.12 0.60315 2148.472

t\_JL\_TWrkngHrs <- train(MonthlyIncome ~ JobLevel + TotalWorkingYears, method = "lm", data = CaseStudy2\_raw, trControl = trainControl(method = "LOOCV"))  
t\_JL\_TWrkngHrs$results # RMSE 1392 BETTER THAN MODEL 1 JOB LEVEL

## intercept RMSE Rsquared MAE  
## 1 TRUE 1392.143 0.908298 1057.2

#T\_Wrkng\_Yrs\_preds1 = predict(linearMod\_T\_Wrkng\_Yrs\_t)  
#T\_Wrkng\_Yrs\_preds2 = predict(linearMod\_T\_Wrkng\_Yrs\_t2)  
#CaseStudy2\_raw %>% ggplot(aes(x = MonthlyIncome, y = TotalWorkingYears)) + geom\_point() + geom\_line(data = CaseStudy2\_raw, aes( x = MonthlyIncome\_preds, y = T\_Wrkng\_Yrs\_preds1, col = "red")) + ggtitle("LR Model: JobLevel + Horsepower^2 + Horsepower^3 vs MPG") + scale\_color\_discrete(name = "Predicted")  
#CaseStudy2\_raw %>% ggplot(aes(x = MonthlyIncome, y = TotalWorkingYears)) + geom\_point() + geom\_line(data = CaseStudy2\_raw, aes( x = MonthlyIncome\_preds, y = T\_Wrkng\_Yrs\_preds2, col = "red")) + ggtitle("LR Model: JobLevel + Horsepower^2 + Horsepower^3 vs MPG") + scale\_color\_discrete(name = "Predicted")  
  
  
linearMod\_T\_Wrkng\_Yrs <- lm(MonthlyIncome ~ TotalWorkingYears, data=CaseStudy2\_raw) # build linear regression model on full data  
#print(linearMod\_T\_Wrkng\_Yrs)  
#MonthlyIncome = Intercept + (β ∗ JobLevel)  
#MonthlyIncome = -1902 + 4093 \* JobLevel  
modelSummary\_T\_Wrkng\_Yrs <- summary(linearMod\_T\_Wrkng\_Yrs)  
modelCoeffs\_T\_Wrkng\_Yrs <- modelSummary\_T\_Wrkng\_Yrs$coefficients  
AIC\_T\_Wrkng\_Yrs <- AIC(linearMod\_T\_Wrkng\_Yrs)  
BIC\_T\_Wrkng\_Yrs <- BIC(linearMod\_T\_Wrkng\_Yrs)  
  
T\_Wrkng\_Yrs\_preds = predict(linearMod\_T\_Wrkng\_Yrs)  
#CaseStudy2\_raw %>% ggplot(aes(x = MonthlyIncome, y = TotalWorkingYears)) + geom\_smooth() + geom\_line(data = CaseStudy2\_raw, aes( x = MonthlyIncome, y = T\_Wrkng\_Yrs\_preds, col = "red")) + ggtitle("LR Model: Horsepower + Horsepower^2 + Horsepower^3 vs MPG") + scale\_color\_discrete(name = "Predicted")  
#CaseStudy2\_raw %>% ggplot(aes(x = MonthlyIncome, y = JobLevel)) + geom\_smooth() + geom\_line(data = CaseStudy2\_raw, aes( x = MonthlyIncome, y = T\_Wrkng\_Yrs\_preds, col = "red")) + ggtitle("LR Model: Horsepower + Horsepower^2 + Horsepower^3 vs MPG") + scale\_color\_discrete(name = "Predicted")  
  
#TotalWorkingYears  
linearMod\_Age <- lm(MonthlyIncome ~ Age, data=CaseStudy2\_raw) # build linear regression model on full data  
print(linearMod\_Age)

##   
## Call:  
## lm(formula = MonthlyIncome ~ Age, data = CaseStudy2\_raw)  
##   
## Coefficients:  
## (Intercept) Age   
## -2793.6 249.4

linearMod\_Age\_JL <- lm(MonthlyIncome ~ Age + JobLevel, data=CaseStudy2\_raw) # build linear regression model on full data  
print(linearMod\_Age\_JL)

##   
## Call:  
## lm(formula = MonthlyIncome ~ Age + JobLevel, data = CaseStudy2\_raw)  
##   
## Coefficients:  
## (Intercept) Age JobLevel   
## -2330.9 18.7 3940.2

AIC(linearMod\_Age\_JL)

## [1] 15068.95

BIC(linearMod\_Age\_JL)

## [1] 15088.02

linearMod\_Age\_JL\_TWH <- lm(MonthlyIncome ~ Age + JobLevel + TotalWorkingYears, data=CaseStudy2\_raw) # build linear regression model on full data  
print(linearMod\_Age\_JL\_TWH)

##   
## Call:  
## lm(formula = MonthlyIncome ~ Age + JobLevel + TotalWorkingYears,   
## data = CaseStudy2\_raw)  
##   
## Coefficients:  
## (Intercept) Age JobLevel   
## -1865.786 2.419 3716.000   
## TotalWorkingYears   
## 53.538

AIC(linearMod\_Age\_JL\_TWH)

## [1] 15050.01

BIC(linearMod\_Age\_JL\_TWH)

## [1] 15073.84

linearMod\_Age\_t <- train(MonthlyIncome ~ Age , method = "lm", data = CaseStudy2\_raw, trControl = trainControl(method = "LOOCV"))  
summary(linearMod\_Age\_t)

##   
## Call:  
## lm(formula = .outcome ~ ., data = dat)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -9744.8 -2624.8 -623.5 1990.8 12650.0   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -2793.6 580.0 -4.817 1.72e-06 \*\*\*  
## Age 249.4 15.3 16.297 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 4027 on 867 degrees of freedom  
## Multiple R-squared: 0.2345, Adjusted R-squared: 0.2336   
## F-statistic: 265.6 on 1 and 867 DF, p-value: < 2.2e-16

linearMod\_Age\_t$results #RmSE 4033.305

## intercept RMSE Rsquared MAE  
## 1 TRUE 4033.305 0.2302999 3056.027

#MonthlyIncome = Intercept + (β ∗ JobLevel)  
#MonthlyIncome = -1902 + 4093 \* JobLevel  
modelSummary\_Age <- summary(linearMod\_Age)  
modelCoeffs\_Age <- modelSummary\_Age$coefficients  
AIC\_Age <- AIC(linearMod\_Age)  
BIC\_Age <- BIC(linearMod\_Age)  
  
AIC\_BIC\_LM\_JL <- data.frame("Measure" = c("AIC","BIC"), "lm\_on" = c("JobLevel","JobLevel"), "Value" = c(AIC\_JL,BIC\_JL))  
AIC\_BIC\_LM\_T\_Wrkng\_Yrs <- data.frame("Measure" = c("AIC","BIC"), "lm\_on" = c("TotalWorkingYears","TotalWorkingYears"), "Value" = c(AIC\_T\_Wrkng\_Yrs,BIC\_T\_Wrkng\_Yrs))  
AIC\_BIC\_LM\_Age <- data.frame("Measure" = c("AIC","BIC"), "lm\_on" = c("Age","Age"), "Value" = c(AIC\_Age,BIC\_Age))  
AIC\_BIC <- rbind(AIC\_BIC\_LM\_JL,AIC\_BIC\_LM\_T\_Wrkng\_Yrs,AIC\_BIC\_LM\_Age)  
AIC(linearMod\_Age\_JL)

## [1] 15068.95

BIC(linearMod\_Age\_JL)

## [1] 15088.02

AIC(linearMod\_Age\_JL\_TWH)

## [1] 15050.01

BIC(linearMod\_Age\_JL\_TWH)

## [1] 15073.84

AIC\_BIC

## Measure lm\_on Value  
## 1 AIC JobLevel 15076.37  
## 2 BIC JobLevel 15090.67  
## 3 AIC TotalWorkingYears 16319.39  
## 4 BIC TotalWorkingYears 16333.69  
## 5 AIC Age 16896.81  
## 6 BIC Age 16911.11

linearMod\_Age\_T\_Wrkng\_Yrs\_jl <- train(MonthlyIncome ~ Age + TotalWorkingYears + JobLevel , method = "lm", data = CaseStudy2\_raw, trControl = trainControl(method = "LOOCV"))  
summary(linearMod\_Age\_T\_Wrkng\_Yrs\_jl)

##   
## Call:  
## lm(formula = .outcome ~ ., data = dat)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -5466.0 -875.9 63.5 725.3 3943.4   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -1865.786 224.414 -8.314 3.57e-16 \*\*\*  
## Age 2.419 6.986 0.346 0.729   
## TotalWorkingYears 53.538 11.656 4.593 5.01e-06 \*\*\*  
## JobLevel 3716.000 69.362 53.574 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1390 on 865 degrees of freedom  
## Multiple R-squared: 0.909, Adjusted R-squared: 0.9087   
## F-statistic: 2881 on 3 and 865 DF, p-value: < 2.2e-16

linearMod\_Age\_T\_Wrkng\_Yrs\_jl$results #RMSE 1393

## intercept RMSE Rsquared MAE  
## 1 TRUE 1393.27 0.9081496 1058.506

t\_JL\_TWrkngHrs <- train(MonthlyIncome ~ JobLevel + TotalWorkingYears, method = "lm", data = CaseStudy2\_raw, trControl = trainControl(method = "LOOCV"))  
t\_JL\_TWrkngHrs$results # RMSE 1392.143

## intercept RMSE Rsquared MAE  
## 1 TRUE 1392.143 0.908298 1057.2

#COMBINING ALL 3 MEASURES  
linearMod\_Age\_JL\_TWH <- lm(MonthlyIncome ~ Age + JobLevel + TotalWorkingYears, data=CaseStudy2\_raw) # build linear regression model on full data  
print(linearMod\_Age\_JL\_TWH)

##   
## Call:  
## lm(formula = MonthlyIncome ~ Age + JobLevel + TotalWorkingYears,   
## data = CaseStudy2\_raw)  
##   
## Coefficients:  
## (Intercept) Age JobLevel   
## -1865.786 2.419 3716.000   
## TotalWorkingYears   
## 53.538

AIC(linearMod\_Age\_JL\_TWH)

## [1] 15050.01

BIC(linearMod\_Age\_JL\_TWH)

## [1] 15073.84

CaseStudy2\_raw$JobLevelCub <- CaseStudy2\_raw$JobLevel^3  
CaseStudy2\_raw$JobLevelCub <- CaseStudy2\_raw$JobLevel^3  
  
linearMod\_Age\_T\_Wrkng\_Yrs\_jl2 <- train(MonthlyIncome ~ Age + TotalWorkingYears + JobLevel + JobLevelCub , method = "lm", data = CaseStudy2\_raw, trControl = trainControl(method = "LOOCV"))  
summary(linearMod\_Age\_T\_Wrkng\_Yrs\_jl)

##   
## Call:  
## lm(formula = .outcome ~ ., data = dat)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -5466.0 -875.9 63.5 725.3 3943.4   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -1865.786 224.414 -8.314 3.57e-16 \*\*\*  
## Age 2.419 6.986 0.346 0.729   
## TotalWorkingYears 53.538 11.656 4.593 5.01e-06 \*\*\*  
## JobLevel 3716.000 69.362 53.574 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1390 on 865 degrees of freedom  
## Multiple R-squared: 0.909, Adjusted R-squared: 0.9087   
## F-statistic: 2881 on 3 and 865 DF, p-value: < 2.2e-16

linearMod\_Age\_T\_Wrkng\_Yrs\_jl$results #RMSE 1303

## intercept RMSE Rsquared MAE  
## 1 TRUE 1393.27 0.9081496 1058.506

#Predict\_JL\_Age\_Model\_1 <- predict(JL\_Age\_Model\_1)  
Sum\_ge\_T\_Wrkng\_Yrs\_jl <- summary(linearMod\_Age\_T\_Wrkng\_Yrs\_jl)  
Sum\_ge\_T\_Wrkng\_Yrs\_jl$r.squared #r-sq 0.92

## [1] 0.9090122

MonthlyIncome\_preds = predict(linearMod\_JL)  
CaseStudy2\_raw %>% ggplot(aes(x = MonthlyIncome, y = JobLevel)) + geom\_point() + geom\_line(data = CaseStudy2\_raw, aes( x = MonthlyIncome\_preds, y = JobLevel, col = "red")) + ggtitle("LR Model: JobLevel vs MonthlyIncome") + scale\_color\_discrete(name = "Predicted")  
  
MonthlyIncome\_preds2 = predict(t\_JL\_Model\_2)  
CaseStudy2\_raw %>% ggplot(aes(x = MonthlyIncome, y = JobLevel)) + geom\_point() + geom\_line(data = CaseStudy2\_raw, aes( x = MonthlyIncome\_preds2, y = JobLevel, col = "red")) + ggtitle("LR Model: JobLevel + JobLevel^2 vs MonthlyIncome") + scale\_color\_discrete(name = "Predicted")  
  
  
CaseStudy2\_raw$TotalWorkingYearsSQ <- CaseStudy2\_raw$TotalWorkingYears^2  
TWH\_Model\_2 <- train(MonthlyIncome ~ TotalWorkingYears + TotalWorkingYearsSQ, method = "lm", data = CaseStudy2\_raw, trControl = trainControl(method = "LOOCV"))  
Predict\_TWH\_Model\_2 <- predict(TWH\_Model\_2)  
CaseStudy2\_raw %>% ggplot(aes(x = TotalWorkingYears, y = Predict\_TWH\_Model\_2)) + geom\_smooth() + geom\_point()  
  
CaseStudy2\_raw$TotalWorkingYearsCub <- CaseStudy2\_raw$TotalWorkingYears^3  
TWH\_Model\_3 <- train(MonthlyIncome ~ TotalWorkingYears + TotalWorkingYearsSQ + TotalWorkingYearsCub, method = "lm", data = CaseStudy2\_raw, trControl = trainControl(method = "LOOCV"))  
Predict\_JL\_Model\_3 <- predict(TWH\_Model\_3)  
CaseStudy2\_raw %>% ggplot(aes(x = TotalWorkingYears, y = Predict\_JL\_Model\_3)) + geom\_smooth() + geom\_point()  
  
  
CaseStudy2\_raw %>% ggplot(aes(x = MonthlyIncome, y = Age)) + geom\_smooth()  
  
Age\_Model\_1 <- train(MonthlyIncome ~ Age, method = "lm", data = CaseStudy2\_raw, trControl = trainControl(method = "LOOCV"))  
Predict\_Age\_Model\_1 <- predict(Age\_Model\_1)  
#CaseStudy2\_raw %>% ggplot(aes(x = Predict\_TWH\_Model\_1, y = TotalWorkingYears)) + geom\_smooth() + geom\_point()  
CaseStudy2\_raw %>% ggplot(aes(x = Age, y = Predict\_Age\_Model\_1)) + geom\_smooth() + geom\_point()  
  
  
CaseStudy2\_raw$TotalWorkingYearsSQ <- CaseStudy2\_raw$TotalWorkingYears^2  
TWH\_Model\_2 <- train(MonthlyIncome ~ TotalWorkingYears + TotalWorkingYearsSQ, method = "lm", data = CaseStudy2\_raw, trControl = trainControl(method = "LOOCV"))  
Predict\_TWH\_Model\_2 <- predict(TWH\_Model\_2)  
CaseStudy2\_raw %>% ggplot(aes(x = TotalWorkingYears, y = Predict\_TWH\_Model\_2)) + geom\_smooth() + geom\_point()  
  
CaseStudy2\_raw$TotalWorkingYearsCub <- CaseStudy2\_raw$TotalWorkingYears^3  
TWH\_Model\_3 <- train(MonthlyIncome ~ TotalWorkingYears + TotalWorkingYearsSQ + TotalWorkingYearsCub, method = "lm", data = CaseStudy2\_raw, trControl = trainControl(method = "LOOCV"))  
Predict\_JL\_Model\_3 <- predict(TWH\_Model\_3)  
CaseStudy2\_raw %>% ggplot(aes(x = TotalWorkingYears, y = Predict\_JL\_Model\_3)) + geom\_smooth() + geom\_point()  
  
  
  
summary(t\_JL\_Model\_2)

##   
## Call:  
## lm(formula = .outcome ~ ., data = dat)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -4505.2 -739.9 -130.2 655.6 4175.1   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 296.60 190.66 1.556 0.12   
## JobLevel 1948.87 169.17 11.520 <2e-16 \*\*\*  
## JobLevelSQ 396.89 31.58 12.570 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1300 on 866 degrees of freedom  
## Multiple R-squared: 0.9203, Adjusted R-squared: 0.9201   
## F-statistic: 5001 on 2 and 866 DF, p-value: < 2.2e-16

CaseStudy2\_raw <- CaseStudy2\_raw %>% mutate(JobLevelSQ = JobLevel^2)  
MonthlyIncome\_preds2 = predict(t\_JL\_Model\_2)  
  
Predict\_JL\_1 <- predict(t\_JL\_Model\_1)  
CaseStudy2\_raw %>% ggplot(aes(x = MonthlyIncome, y = JobLevel)) + geom\_smooth()  
#CaseStudy2\_raw %>% ggplot(aes(x = Final\_Predict, y = JobLevel)) + geom\_smooth()  
CaseStudy2\_raw %>% ggplot(aes(x = MonthlyIncome\_preds2, y = JobLevel)) + geom\_smooth()  
  
#More EDA  
CaseStudy2\_raw %>% ggplot(aes(x = MonthlyIncome, y = JobLevel)) + geom\_smooth()  
  
JL\_Model\_1 <- train(MonthlyIncome ~ JobLevel, method = "lm", data = CaseStudy2\_raw, trControl = trainControl(method = "LOOCV"))  
Predict\_JL\_Model\_1 <- predict(JL\_Model\_1)  
Sum\_JL\_Model\_1 <-summary(JL\_Model\_1)  
CaseStudy2\_raw %>% ggplot(aes(x = Predict\_JL\_Model\_1, y = JobLevel)) + geom\_smooth() + geom\_point()  
CaseStudy2\_raw %>% ggplot(aes(x = JobLevel, y = Predict\_JL\_Model\_1)) + geom\_smooth() + geom\_point()  
  
  
CaseStudy2\_raw$JobLevelSQ <- CaseStudy2\_raw$JobLevel^2  
JL\_Model\_2 <- train(MonthlyIncome ~ JobLevel + JobLevelSQ, method = "lm", data = CaseStudy2\_raw, trControl = trainControl(method = "LOOCV"))  
Predict\_JL\_Model\_2 <- predict(JL\_Model\_2)  
Sum\_JL\_Model\_2 <-summary(JL\_Model\_2)  
CaseStudy2\_raw %>% ggplot(aes(x = JobLevel, y = Predict\_JL\_Model\_2)) + geom\_smooth() + geom\_point()  
  
CaseStudy2\_raw$JobLevelCub <- CaseStudy2\_raw$JobLevel^3  
JL\_Model\_3 <- train(MonthlyIncome ~ JobLevel + JobLevelSQ + JobLevelCub, method = "lm", data = CaseStudy2\_raw, trControl = trainControl(method = "LOOCV"))  
Predict\_JL\_Model\_3 <- predict(JL\_Model\_3)  
Sum\_JL\_Model\_3 <- summary(JL\_Model\_3)  
CaseStudy2\_raw %>% ggplot(aes(x = JobLevel, y = Predict\_JL\_Model\_3)) + geom\_smooth()  
  
  
CaseStudy2\_raw %>% ggplot(aes(x = MonthlyIncome, y = TotalWorkingYears)) + geom\_smooth()  
  
TWY\_Model\_1 <- train(MonthlyIncome ~ TotalWorkingYears, method = "lm", data = CaseStudy2\_raw, trControl = trainControl(method = "LOOCV"))  
Predict\_TWY\_Model\_1 <- predict(TWY\_Model\_1)  
Sum\_TWY\_Model\_1 <- summary(TWY\_Model\_1)  
  
CaseStudy2\_raw$TotalWorkingYearsSQ <- CaseStudy2\_raw$TotalWorkingYears^2  
TWY\_Model\_2 <- train(MonthlyIncome ~ TotalWorkingYears + TotalWorkingYearsSQ , method = "lm", data = CaseStudy2\_raw, trControl = trainControl(method = "LOOCV"))  
Predict\_TWY\_Model\_2 <- predict(TWY\_Model\_2)  
Sum\_TWY\_Model\_2 <- summary(TWY\_Model\_2)  
#CaseStudy2\_raw %>% ggplot(aes(x = Predict\_TWH\_Model\_1, y = TotalWorkingYears)) + geom\_smooth() + geom\_point()  
#CaseStudy2\_raw %>% ggplot(aes(x = TotalWorkingYears, y = Predict\_TWH\_Model\_1)) + geom\_smooth() + geom\_point()  
  
JL\_TWH\_Model\_1 <- train(MonthlyIncome ~ JobLevel + TotalWorkingYears, method = "lm", data = CaseStudy2\_raw, trControl = trainControl(method = "LOOCV"))  
Predict\_JL\_TWH\_Model\_1 <- predict(JL\_TWH\_Model\_1)  
Sum\_JL\_TWH\_Model\_1 <- summary(JL\_TWH\_Model\_1)  
  
TWH\_JL\_Model\_1 <- train(MonthlyIncome ~ TotalWorkingYears +JobLevel , method = "lm", data = CaseStudy2\_raw, trControl = trainControl(method = "LOOCV"))  
Predict\_TWH\_JL\_Model\_1 <- predict(TWH\_JL\_Model\_1)  
Sum\_TWH\_JL\_Model\_1 <- summary(TWH\_JL\_Model\_1)  
  
Age\_Model\_1 <- train(MonthlyIncome ~ Age, method = "lm", data = CaseStudy2\_raw, trControl = trainControl(method = "LOOCV"))  
Predict\_Age\_Model\_1<- predict(Age\_Model\_1)  
Sum\_Age\_Model\_1 <- summary(Age\_Model\_1)  
  
CaseStudy2\_raw$AgeSQ <- CaseStudy2\_raw$Age^2  
Age\_Model\_2 <- train(MonthlyIncome ~ Age + AgeSQ , method = "lm", data = CaseStudy2\_raw, trControl = trainControl(method = "LOOCV"))  
Predict\_Age\_Model\_2 <- predict(Age\_Model\_2)  
Sum\_Age\_Model\_2 <- summary(Age\_Model\_2)  
  
TWH\_Age\_Model\_1 <- train(MonthlyIncome ~ Age + TotalWorkingYears , method = "lm", data = CaseStudy2\_raw, trControl = trainControl(method = "LOOCV"))  
Predict\_TWH\_Age\_Model\_1 <- predict(TWH\_Age\_Model\_1)  
Sum\_TWH\_Age\_Model\_1 <- summary(TWH\_Age\_Model\_1)  
  
JL\_Age\_Model\_1 <- train(MonthlyIncome ~ Age + JobLevel , method = "lm", data = CaseStudy2\_raw, trControl = trainControl(method = "LOOCV"))  
Predict\_JL\_Age\_Model\_1 <- predict(JL\_Age\_Model\_1)  
Sum\_JL\_Age\_Model\_1 <- summary(JL\_Age\_Model\_1)  
  
Sum\_JL\_Age\_Model\_1$r.squared #0.9141776

## [1] 0.9067929

Sum\_TWH\_Age\_Model\_1$r.squared #0.616751

## [1] 0.6071021

Sum\_Age\_Model\_2$r.squared #0.2755832

## [1] 0.2349385

Sum\_Age\_Model\_1$r.squared #0.2755832

## [1] 0.2344981

Sum\_JL\_TWH\_Model\_1$r.squared #0.915356

## [1] 0.9089995

Sum\_JL\_TWH\_Model\_1$r.squared #0.915356

## [1] 0.9089995

Sum\_JL\_Model\_1$r.squared #0.9141695

## [1] 0.9057772

Sum\_JL\_Model\_2$r.squared #0.9274724

## [1] 0.9203151

Sum\_JL\_Model\_3$r.squared # 0.9320185

## [1] 0.9244172

Sum\_TWY\_Model\_1$r.squared # 0.6122029

## [1] 0.6061136

Sum\_TWY\_Model\_2$r.squared #0.6156517

## [1] 0.6071774

#Predict\_Age\_Model\_1  
Age\_Model\_1 <- train(MonthlyIncome ~ Age, method = "lm", data = CaseStudy2\_raw, trControl = trainControl(method = "LOOCV"))  
Predict\_Age\_Model\_1<- predict(Age\_Model\_1)  
Sum\_Age\_Model\_1 <- summary(Age\_Model\_1)  
CaseStudy2\_raw %>% ggplot(aes(x = Age, y = MonthlyIncome)) + geom\_smooth() + geom\_point()  
CaseStudy2\_raw %>% ggplot(aes(x = Age, y = Predict\_Age\_Model\_1)) + geom\_smooth() + geom\_point()  
  
CaseStudy2\_raw\_Under\_50 <- CaseStudy2\_raw %>% filter(Age <=53)  
CaseStudy2\_raw\_Over\_50 <- CaseStudy2\_raw %>% filter(Age >53)  
nrow(CaseStudy2\_raw\_Under\_50)

## [1] 822

nrow(CaseStudy2\_raw\_Over\_50)

## [1] 47

linearMod\_Age\_T\_Wrkng\_Yrs\_jl2 <- train(MonthlyIncome ~ Age + TotalWorkingYears + JobLevel + JobLevelCub , method = "lm", data = CaseStudy2\_raw, trControl = trainControl(method = "LOOCV"))  
summary(linearMod\_Age\_T\_Wrkng\_Yrs\_jl2)

##   
## Call:  
## lm(formula = .outcome ~ ., data = dat)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -4855.6 -732.5 -137.2 623.4 4252.8   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -587.276 239.125 -2.456 0.0142 \*   
## Age 3.098 6.535 0.474 0.6355   
## TotalWorkingYears 43.129 10.942 3.942 8.75e-05 \*\*\*  
## JobLevel 2779.379 106.061 26.206 < 2e-16 \*\*\*  
## JobLevelCub 42.228 3.783 11.163 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1300 on 864 degrees of freedom  
## Multiple R-squared: 0.9205, Adjusted R-squared: 0.9201   
## F-statistic: 2500 on 4 and 864 DF, p-value: < 2.2e-16

linearMod\_Age\_T\_Wrkng\_Yrs\_jl2$results #RMSE 1303

## intercept RMSE Rsquared MAE  
## 1 TRUE 1303.511 0.9196035 957.084

Predict\_Age\_T\_Wrkng\_Yrs\_jl2 <- predict(linearMod\_Age\_T\_Wrkng\_Yrs\_jl2)  
Sum\_Age\_T\_Wrkng\_Yrs\_jl2 <- summary(linearMod\_Age\_T\_Wrkng\_Yrs\_jl2)  
Sum\_Age\_T\_Wrkng\_Yrs\_jl2$r.squared #r-sq 0.92

## [1] 0.9204816

CaseStudy2\_raw %>% ggplot(aes(x = TotalWorkingYears, y = Predict\_Age\_T\_Wrkng\_Yrs\_jl2)) + geom\_smooth() + geom\_point()  
  
linearMod\_Age\_T\_Wrkng\_Yrs\_jl2 <- train(MonthlyIncome ~ Age + TotalWorkingYears + JobLevel + JobLevelCub , method = "lm", data = CaseStudy2\_raw, trControl = trainControl(method = "LOOCV"))  
summary(linearMod\_Age\_T\_Wrkng\_Yrs\_jl2)

##   
## Call:  
## lm(formula = .outcome ~ ., data = dat)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -4855.6 -732.5 -137.2 623.4 4252.8   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -587.276 239.125 -2.456 0.0142 \*   
## Age 3.098 6.535 0.474 0.6355   
## TotalWorkingYears 43.129 10.942 3.942 8.75e-05 \*\*\*  
## JobLevel 2779.379 106.061 26.206 < 2e-16 \*\*\*  
## JobLevelCub 42.228 3.783 11.163 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1300 on 864 degrees of freedom  
## Multiple R-squared: 0.9205, Adjusted R-squared: 0.9201   
## F-statistic: 2500 on 4 and 864 DF, p-value: < 2.2e-16

linearMod\_Age\_T\_Wrkng\_Yrs\_jl2$results #RMSE 1303

## intercept RMSE Rsquared MAE  
## 1 TRUE 1303.511 0.9196035 957.084

Predict\_Age\_T\_Wrkng\_Yrs\_jl2 <- predict(linearMod\_Age\_T\_Wrkng\_Yrs\_jl2)  
Sum\_Age\_T\_Wrkng\_Yrs\_jl2 <- summary(linearMod\_Age\_T\_Wrkng\_Yrs\_jl2)  
Sum\_Age\_T\_Wrkng\_Yrs\_jl2$r.squared #r-sq 0.92

## [1] 0.9204816

CaseStudy2\_raw %>% ggplot(aes(x = Age, y = Predict\_Age\_T\_Wrkng\_Yrs\_jl2)) + geom\_smooth() + geom\_point()  
  
LM\_Under\_50 <- train(MonthlyIncome ~ Age + TotalWorkingYears + JobLevel + JobLevelCub , method = "lm", data = CaseStudy2\_raw\_Under\_50, trControl = trainControl(method = "LOOCV"))  
summary(LM\_Under\_50)

##   
## Call:  
## lm(formula = .outcome ~ ., data = dat)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -4890.2 -742.4 -127.0 623.8 4283.2   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -603.554 260.543 -2.317 0.020776 \*   
## Age 4.998 7.569 0.660 0.509213   
## TotalWorkingYears 45.821 11.932 3.840 0.000132 \*\*\*  
## JobLevel 2724.920 110.281 24.709 < 2e-16 \*\*\*  
## JobLevelCub 44.606 4.072 10.954 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1298 on 817 degrees of freedom  
## Multiple R-squared: 0.9123, Adjusted R-squared: 0.9119   
## F-statistic: 2125 on 4 and 817 DF, p-value: < 2.2e-16

LM\_Under\_50$results #RMSE 1300

## intercept RMSE Rsquared MAE  
## 1 TRUE 1302.167 0.911252 956.7622

Predict\_LM\_Under\_50 <- predict(LM\_Under\_50)  
Sum\_LM\_Under\_50 <- summary(LM\_Under\_50)  
Sum\_LM\_Under\_50$r.squared #r-sq 0.908

## [1] 0.9123277

CaseStudy2\_raw\_Under\_50 %>% ggplot(aes(x = Age, y = Predict\_LM\_Under\_50)) + geom\_smooth() + geom\_point()  
  
LM\_Over\_50 <- train(MonthlyIncome ~ Age + TotalWorkingYears + JobLevel , method = "lm", data = CaseStudy2\_raw\_Over\_50, trControl = trainControl(method = "LOOCV"))  
summary(LM\_Over\_50)

##   
## Call:  
## lm(formula = .outcome ~ ., data = dat)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3783.6 -565.1 -33.5 644.8 2604.6   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 8085.890 6152.052 1.314 0.196   
## Age -176.837 107.545 -1.644 0.107   
## TotalWorkingYears 1.021 32.895 0.031 0.975   
## JobLevel 4168.235 250.877 16.615 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1358 on 43 degrees of freedom  
## Multiple R-squared: 0.956, Adjusted R-squared: 0.953   
## F-statistic: 311.6 on 3 and 43 DF, p-value: < 2.2e-16

LM\_Over\_50$results #RMSE 1300

## intercept RMSE Rsquared MAE  
## 1 TRUE 1407.908 0.9483537 1028.622

Predict\_LM\_Over\_50 <- predict(LM\_Over\_50)  
Sum\_LM\_Over\_50 <- summary(LM\_Over\_50)  
Sum\_LM\_Over\_50$r.squared #r-sq 0.908

## [1] 0.9560193

CaseStudy2\_raw\_Over\_50 %>% ggplot(aes(x = Age, y = Predict\_LM\_Over\_50)) + geom\_smooth() + geom\_point()  
#Final SALARY MODEL  
CaseStudy2\_raw$JobLevelCub <- CaseStudy2\_raw$JobLevel^3  
linearMod\_Age\_T\_Wrkng\_Yrs\_jl2 <- train(MonthlyIncome ~ Age + TotalWorkingYears + JobLevel + JobLevelCub , method = "lm", data = CaseStudy2\_raw, trControl = trainControl(method = "LOOCV"))  
summary(linearMod\_Age\_T\_Wrkng\_Yrs\_jl2)

##   
## Call:  
## lm(formula = .outcome ~ ., data = dat)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -4855.6 -732.5 -137.2 623.4 4252.8   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -587.276 239.125 -2.456 0.0142 \*   
## Age 3.098 6.535 0.474 0.6355   
## TotalWorkingYears 43.129 10.942 3.942 8.75e-05 \*\*\*  
## JobLevel 2779.379 106.061 26.206 < 2e-16 \*\*\*  
## JobLevelCub 42.228 3.783 11.163 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1300 on 864 degrees of freedom  
## Multiple R-squared: 0.9205, Adjusted R-squared: 0.9201   
## F-statistic: 2500 on 4 and 864 DF, p-value: < 2.2e-16

linearMod\_Age\_T\_Wrkng\_Yrs\_jl2$results #RMSE 1303

## intercept RMSE Rsquared MAE  
## 1 TRUE 1303.511 0.9196035 957.084

Predict\_Age\_T\_Wrkng\_Yrs\_jl2 <- predict(linearMod\_Age\_T\_Wrkng\_Yrs\_jl2)  
Sum\_Age\_T\_Wrkng\_Yrs\_jl2 <- summary(linearMod\_Age\_T\_Wrkng\_Yrs\_jl2)  
Sum\_Age\_T\_Wrkng\_Yrs\_jl2$r.squared #r-sq 0.92

## [1] 0.9204816

#CaseStudy2\_raw %>% ggplot(aes(x = Age, y = Predict\_Age\_T\_Wrkng\_Yrs\_jl2)) + geom\_smooth() + geom\_point() + ylab("Predicted Salary") +ggtitle("Plot showing the predicted salary vs.Age")  
  
#CaseStudy2\_Comp\_Sal\_set<- read.csv(file.choose(), header = TRUE, sep=',')  
colnames(CaseStudy2\_Comp\_Sal\_set)

## [1] "ï..ID" "Age"   
## [3] "Attrition" "BusinessTravel"   
## [5] "DailyRate" "Department"   
## [7] "DistanceFromHome" "Education"   
## [9] "EducationField" "EmployeeCount"   
## [11] "EmployeeNumber" "EnvironmentSatisfaction"   
## [13] "Gender" "HourlyRate"   
## [15] "JobInvolvement" "JobLevel"   
## [17] "JobRole" "JobSatisfaction"   
## [19] "MaritalStatus" "MonthlyRate"   
## [21] "NumCompaniesWorked" "Over18"   
## [23] "OverTime" "PercentSalaryHike"   
## [25] "PerformanceRating" "RelationshipSatisfaction"  
## [27] "StandardHours" "StockOptionLevel"   
## [29] "TotalWorkingYears" "TrainingTimesLastYear"   
## [31] "WorkLifeBalance" "YearsAtCompany"   
## [33] "YearsInCurrentRole" "YearsSinceLastPromotion"   
## [35] "YearsWithCurrManager"

colnames(CaseStudy2\_Comp\_Sal\_set) <-c("ID","Age","Attrition","BusinessTravel","DailyRate","Department","DistanceFromHome","Education","EducationField","EmployeeCount", "EmployeeNumber", "EnvironmentSatisfaction","Gender","HourlyRate","JobInvolvement","JobLevel","JobRole","JobSatisfaction", "MaritalStatus","MonthlyRate", "NumCompaniesWorked","Over18","OverTime","PercentSalaryHike","PerformanceRating","RelationshipSatisfaction","StandardHours","StockOptionLevel","TotalWorkingYears","TrainingTimesLastYear", "WorkLifeBalance","YearsAtCompany","YearsInCurrentRole","YearsSinceLastPromotion", "YearsWithCurrManager")  
  
CaseStudy2\_Comp\_Sal\_set$JobLevelCub <- CaseStudy2\_Comp\_Sal\_set$JobLevel^3  
Predict\_Comp\_Sal\_set <- round(predict(linearMod\_Age\_T\_Wrkng\_Yrs\_jl2, newdata = CaseStudy2\_Comp\_Sal\_set),0)  
  
answers\_sal <- cbind(Predict\_Comp\_Sal\_set,CaseStudy2\_Comp\_Sal\_set[1])  
head(answers\_sal)

## Predict\_Comp\_Sal\_set ID  
## 1 5744 871  
## 2 2552 872  
## 3 14438 873  
## 4 2562 874  
## 5 2620 875  
## 6 6250 876

colnames(answers\_sal) <- c('Predicted Salary','ID')  
joined\_data\_sal <- left\_join(answers\_sal, CaseStudy2\_Comp\_Sal\_set, by = "ID")  
head(joined\_data\_sal)

## Predicted Salary ID Age Attrition BusinessTravel DailyRate  
## 1 5744 871 43 No Travel\_Frequently 1422  
## 2 2552 872 33 No Travel\_Rarely 461  
## 3 14438 873 55 Yes Travel\_Rarely 267  
## 4 2562 874 36 No Non-Travel 1351  
## 5 2620 875 27 No Travel\_Rarely 1302  
## 6 6250 876 39 Yes Travel\_Rarely 895  
## Department DistanceFromHome Education EducationField  
## 1 Sales 2 4 Life Sciences  
## 2 Research & Development 13 1 Life Sciences  
## 3 Sales 13 4 Marketing  
## 4 Research & Development 9 4 Life Sciences  
## 5 Research & Development 19 3 Other  
## 6 Sales 5 3 Technical Degree  
## EmployeeCount EmployeeNumber EnvironmentSatisfaction Gender HourlyRate  
## 1 1 1849 1 Male 92  
## 2 1 995 2 Female 53  
## 3 1 1372 1 Male 85  
## 4 1 1949 1 Male 66  
## 5 1 1619 4 Male 67  
## 6 1 42 4 Male 56  
## JobInvolvement JobLevel JobRole JobSatisfaction  
## 1 3 2 Sales Executive 4  
## 2 3 1 Research Scientist 4  
## 3 4 4 Sales Executive 3  
## 4 4 1 Laboratory Technician 2  
## 5 2 1 Laboratory Technician 1  
## 6 3 2 Sales Representative 4  
## MaritalStatus MonthlyRate NumCompaniesWorked Over18 OverTime  
## 1 Married 19246 1 Y No  
## 2 Single 17241 3 Y No  
## 3 Single 9277 6 Y Yes  
## 4 Married 9238 1 Y No  
## 5 Divorced 16290 1 Y No  
## 6 Married 3335 3 Y No  
## PercentSalaryHike PerformanceRating RelationshipSatisfaction  
## 1 20 4 3  
## 2 18 3 1  
## 3 17 3 3  
## 4 22 4 2  
## 5 11 3 1  
## 6 14 3 3  
## StandardHours StockOptionLevel TotalWorkingYears TrainingTimesLastYear  
## 1 80 1 7 5  
## 2 80 0 5 4  
## 3 80 0 24 2  
## 4 80 0 5 3  
## 5 80 2 7 3  
## 6 80 1 19 6  
## WorkLifeBalance YearsAtCompany YearsInCurrentRole  
## 1 3 7 7  
## 2 3 3 2  
## 3 2 19 7  
## 4 3 5 4  
## 5 3 7 7  
## 6 4 1 0  
## YearsSinceLastPromotion YearsWithCurrManager JobLevelCub  
## 1 7 7 8  
## 2 0 2 1  
## 3 3 8 64  
## 4 0 2 1  
## 5 0 7 1  
## 6 0 0 8

output\_sal <- select (joined\_data\_sal,"Predicted Salary", "ID","Age","Attrition","BusinessTravel","DailyRate","Department","DistanceFromHome","Education","EducationField","EmployeeCount", "EmployeeNumber","EnvironmentSatisfaction", "Gender","HourlyRate","JobInvolvement","JobLevel","JobRole","JobSatisfaction","MaritalStatus","MonthlyRate","NumCompaniesWorked","Over18","OverTime","PercentSalaryHike","PerformanceRating","RelationshipSatisfaction","StandardHours","StockOptionLevel","TotalWorkingYears","TrainingTimesLastYear","WorkLifeBalance","YearsAtCompany","YearsInCurrentRole","YearsSinceLastPromotion","YearsWithCurrManager")  
  
write.csv(output\_sal, file = "C:/Users/simer/Documents/SMU/Doing\_DataScience/Case\_Study\_2/Case2PredictionsRegressREDDY.csv")

