DDS\_Casestudy 2\_New

John Olanipekun

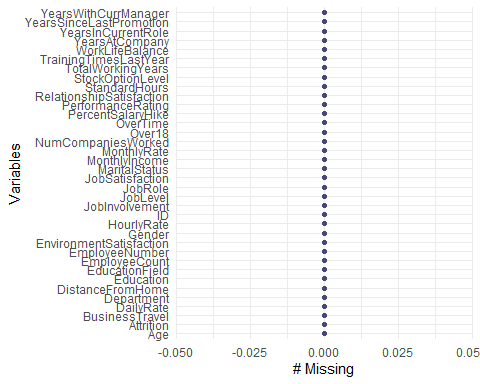
12/2/2020

attrition\_df <- read.csv("C:/Users/olani/OneDrive/Documents/Data Science/SMU-Data Science/Doing Data Science/MSDS\_6306\_DDS/Unit 14 and 15 Case Study 2/CaseStudy2-data.csv")  
  
class(attrition\_df)

## [1] "data.frame"

str(attrition\_df)

gg\_miss\_var(attrition\_df) #no missing data



dim(attrition\_df)

## [1] 870 36

###Recode relevant categorical variables and convert them to factors

attrition\_df\_recode <- attrition\_df %>%  
 mutate(Education = as.factor(if\_else(Education == 1,"Below College", if\_else(Education == 2, "College", if\_else(Education == 3, "Bachelor", if\_else(Education == 4, "Master","Doctor"))))),  
 EnvironmentSatisfaction = as.factor(if\_else(EnvironmentSatisfaction == 1,"Low",if\_else(EnvironmentSatisfaction == 2, "Medium", if\_else(EnvironmentSatisfaction == 3, "High", "Very High")))),  
 JobInvolvement = as.factor(if\_else(JobInvolvement == 1,"Low",if\_else(JobInvolvement == 2, "Medium",if\_else(JobInvolvement == 3, "High", "Very High")))),  
 JobSatisfaction = as.factor(if\_else(JobSatisfaction == 1, "Low",if\_else(JobSatisfaction == 2, "Medium",if\_else(JobSatisfaction == 3, "High","Very High")))),  
 PerformanceRating = as.factor(if\_else(PerformanceRating == 1, "Low",if\_else(PerformanceRating == 2, "Good", if\_else(PerformanceRating == 3, "Excellent", "Outstanding")))),  
 RelationshipSatisfaction = as.factor(if\_else(RelationshipSatisfaction == 1, "Low",if\_else(RelationshipSatisfaction == 2, "Medium", if\_else(RelationshipSatisfaction == 3, "High", "Very High")))),   
 WorkLifeBalance = as.factor(if\_else(WorkLifeBalance == 1, "Bad",if\_else(WorkLifeBalance == 2, "Good", if\_else(WorkLifeBalance == 3, "Better", "Best")))),  
 JobLevel = as.factor(JobLevel),  
 MonthlyIncomeFact = as.factor(cut(MonthlyIncome, breaks = c(0,3000,6000,10000, 20000), labels = c("Low","Average", "High", "Very High"))),  
 JobRole = as.factor(JobRole),  
 Attrition\_recode = as.factor(ifelse(grepl("No",Attrition, ignore.case=TRUE), 1, 0))  
 ) %>%  
 dplyr::select(-ID, -EmployeeCount, -EmployeeNumber, -Over18, -StandardHours, -EmployeeCount, -StockOptionLevel)

###Let us look at the structure and descriptive statistics of each feature

str(attrition\_df\_recode)

## 'data.frame': 870 obs. of 32 variables:  
## $ Age : int 32 40 35 32 24 27 41 37 34 34 ...  
## $ Attrition : chr "No" "No" "No" "No" ...  
## $ BusinessTravel : chr "Travel\_Rarely" "Travel\_Rarely" "Travel\_Frequently" "Travel\_Rarely" ...  
## $ DailyRate : int 117 1308 200 801 567 294 1283 309 1333 653 ...  
## $ Department : chr "Sales" "Research & Development" "Research & Development" "Sales" ...  
## $ DistanceFromHome : int 13 14 18 1 2 10 5 10 10 10 ...  
## $ Education : Factor w/ 5 levels "Bachelor","Below College",..: 5 1 3 5 2 3 4 5 5 5 ...  
## $ EducationField : chr "Life Sciences" "Medical" "Life Sciences" "Marketing" ...  
## $ EnvironmentSatisfaction : Factor w/ 4 levels "High","Low","Medium",..: 3 1 1 1 2 4 3 4 1 4 ...  
## $ Gender : chr "Male" "Male" "Male" "Female" ...  
## $ HourlyRate : int 73 44 60 48 32 32 90 88 87 92 ...  
## $ JobInvolvement : Factor w/ 4 levels "High","Low","Medium",..: 1 3 1 1 1 1 4 3 1 3 ...  
## $ JobLevel : Factor w/ 5 levels "1","2","3","4",..: 2 5 3 3 1 3 1 2 1 2 ...  
## $ JobRole : Factor w/ 9 levels "Healthcare Representative",..: 8 6 5 8 7 5 7 8 9 1 ...  
## $ JobSatisfaction : Factor w/ 4 levels "High","Low","Medium",..: 4 1 4 4 4 2 1 4 1 1 ...  
## $ MaritalStatus : chr "Divorced" "Single" "Single" "Married" ...  
## $ MonthlyIncome : int 4403 19626 9362 10422 3760 8793 2127 6694 2220 5063 ...  
## $ MonthlyRate : int 9250 17544 19944 24032 17218 4809 5561 24223 18410 15332 ...  
## $ NumCompaniesWorked : int 2 1 2 1 1 1 2 2 1 1 ...  
## $ OverTime : chr "No" "No" "No" "No" ...  
## $ PercentSalaryHike : int 11 14 11 19 13 21 12 14 19 14 ...  
## $ PerformanceRating : Factor w/ 2 levels "Excellent","Outstanding": 1 1 1 1 1 2 1 1 1 1 ...  
## $ RelationshipSatisfaction: Factor w/ 4 levels "High","Low","Medium",..: 1 2 1 1 1 1 2 1 4 3 ...  
## $ TotalWorkingYears : int 8 21 10 14 6 9 7 8 1 8 ...  
## $ TrainingTimesLastYear : int 3 2 2 3 2 4 5 5 2 3 ...  
## $ WorkLifeBalance : Factor w/ 4 levels "Bad","Best","Better",..: 4 2 3 3 3 4 4 3 3 4 ...  
## $ YearsAtCompany : int 5 20 2 14 6 9 4 1 1 8 ...  
## $ YearsInCurrentRole : int 2 7 2 10 3 7 2 0 1 2 ...  
## $ YearsSinceLastPromotion : int 0 4 2 5 1 1 0 0 0 7 ...  
## $ YearsWithCurrManager : int 3 9 2 7 3 7 3 0 0 7 ...  
## $ MonthlyIncomeFact : Factor w/ 4 levels "Low","Average",..: 2 4 3 4 2 3 1 3 1 2 ...  
## $ Attrition\_recode : Factor w/ 2 levels "0","1": 2 2 2 2 2 2 2 2 2 2 ...

describe(attrition\_df\_recode)

## vars n mean sd median trimmed mad  
## Age 1 870 36.83 8.93 35.0 36.39 8.90  
## Attrition\* 2 870 1.16 0.37 1.0 1.08 0.00  
## BusinessTravel\* 3 870 2.60 0.68 3.0 2.75 0.00  
## DailyRate 4 870 815.23 401.12 817.5 817.85 514.46  
## Department\* 5 870 2.27 0.53 2.0 2.27 0.00  
## DistanceFromHome 6 870 9.34 8.14 7.0 8.28 7.41  
## Education\* 7 870 2.72 1.63 3.0 2.66 2.97  
## EducationField\* 8 870 3.24 1.32 3.0 3.09 1.48  
## EnvironmentSatisfaction\* 9 870 2.51 1.20 3.0 2.51 1.48  
## Gender\* 10 870 1.59 0.49 2.0 1.62 0.00  
## HourlyRate 11 870 65.61 20.13 66.0 65.63 25.20  
## JobInvolvement\* 12 870 1.86 1.10 1.0 1.71 0.00  
## JobLevel\* 13 870 2.04 1.09 2.0 1.87 1.48  
## JobRole\* 14 870 5.53 2.46 6.0 5.70 2.97  
## JobSatisfaction\* 15 870 2.52 1.21 3.0 2.53 1.48  
## MaritalStatus\* 16 870 2.09 0.72 2.0 2.11 1.48  
## MonthlyIncome 17 870 6390.26 4597.70 4945.5 5582.97 3304.72  
## MonthlyRate 18 870 14325.62 7108.38 14074.5 14285.47 9273.66  
## NumCompaniesWorked 19 870 2.73 2.52 2.0 2.40 1.48  
## OverTime\* 20 870 1.29 0.45 1.0 1.24 0.00  
## PercentSalaryHike 21 870 15.20 3.68 14.0 14.78 2.97  
## PerformanceRating\* 22 870 1.15 0.36 1.0 1.06 0.00  
## RelationshipSatisfaction\* 23 870 2.50 1.21 2.5 2.50 2.22  
## TotalWorkingYears 24 870 11.05 7.51 10.0 10.19 5.93  
## TrainingTimesLastYear 25 870 2.83 1.27 3.0 2.76 1.48  
## WorkLifeBalance\* 26 870 3.00 0.74 3.0 3.07 0.00  
## YearsAtCompany 27 870 6.96 6.02 5.0 6.00 4.45  
## YearsInCurrentRole 28 870 4.20 3.64 3.0 3.83 4.45  
## YearsSinceLastPromotion 29 870 2.17 3.19 1.0 1.47 1.48  
## YearsWithCurrManager 30 870 4.14 3.57 3.0 3.81 4.45  
## MonthlyIncomeFact\* 31 870 2.28 1.06 2.0 2.23 1.48  
## Attrition\_recode\* 32 870 1.84 0.37 2.0 1.92 0.00  
## min max range skew kurtosis se  
## Age 18 60 42 0.42 -0.31 0.30  
## Attrition\* 1 2 1 1.84 1.40 0.01  
## BusinessTravel\* 1 3 2 -1.43 0.62 0.02  
## DailyRate 103 1499 1396 -0.03 -1.18 13.60  
## Department\* 1 3 2 0.16 -0.48 0.02  
## DistanceFromHome 1 29 28 0.91 -0.35 0.28  
## Education\* 1 5 4 0.31 -1.48 0.06  
## EducationField\* 1 6 5 0.55 -0.69 0.04  
## EnvironmentSatisfaction\* 1 4 3 -0.02 -1.54 0.04  
## Gender\* 1 2 1 -0.38 -1.86 0.02  
## HourlyRate 30 100 70 0.01 -1.21 0.68  
## JobInvolvement\* 1 4 3 0.71 -1.13 0.04  
## JobLevel\* 1 5 4 1.03 0.44 0.04  
## JobRole\* 1 9 8 -0.40 -1.16 0.08  
## JobSatisfaction\* 1 4 3 -0.02 -1.55 0.04  
## MaritalStatus\* 1 3 2 -0.14 -1.08 0.02  
## MonthlyIncome 1081 19999 18918 1.39 1.14 155.88  
## MonthlyRate 2094 26997 24903 0.04 -1.21 241.00  
## NumCompaniesWorked 0 9 9 1.00 -0.07 0.09  
## OverTime\* 1 2 1 0.93 -1.14 0.02  
## PercentSalaryHike 11 25 14 0.83 -0.28 0.12  
## PerformanceRating\* 1 2 1 1.94 1.76 0.01  
## RelationshipSatisfaction\* 1 4 3 0.00 -1.55 0.04  
## TotalWorkingYears 0 40 40 1.13 1.09 0.25  
## TrainingTimesLastYear 0 6 6 0.52 0.48 0.04  
## WorkLifeBalance\* 1 4 3 -0.80 0.94 0.03  
## YearsAtCompany 0 40 40 1.62 3.36 0.20  
## YearsInCurrentRole 0 18 18 0.87 0.31 0.12  
## YearsSinceLastPromotion 0 15 15 1.99 3.71 0.11  
## YearsWithCurrManager 0 17 17 0.73 -0.15 0.12  
## MonthlyIncomeFact\* 1 4 3 0.35 -1.11 0.04  
## Attrition\_recode\* 1 2 1 -1.84 1.40 0.01

summary(attrition\_df\_recode)

###Let us look at contingency table for attrition. This will enable us to know the majority class and watch out for class imbalance which can compromise the significance of our model accuracy

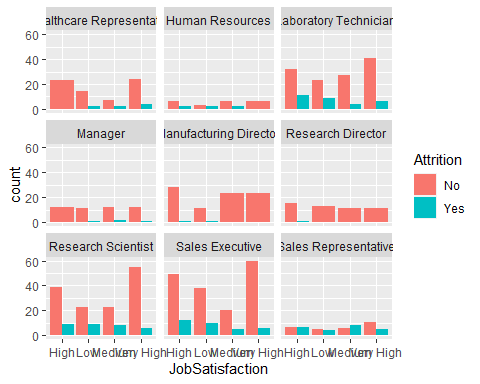
table(attrition\_df\_recode$Attrition)

##   
## No Yes   
## 730 140

There is class imbalance No = 730 Yes = 140 but we will focus on other metrics like sensitivity and specificity

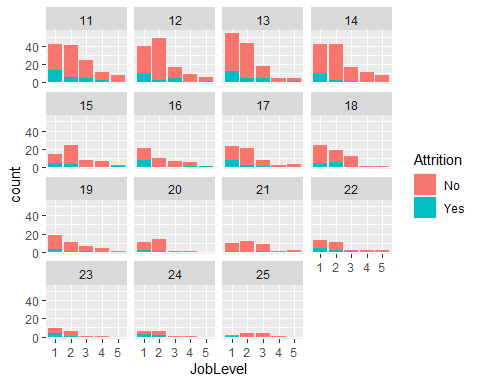
###plot to see general trends in the data set and how these are related to attrition

#JobRole, JobSatistfaction, Attrition  
attrition\_df\_recode %>%   
ggplot(aes(x = JobSatisfaction, fill = Attrition)) +   
geom\_bar(position = position\_dodge()) +   
facet\_wrap(vars(JobRole))

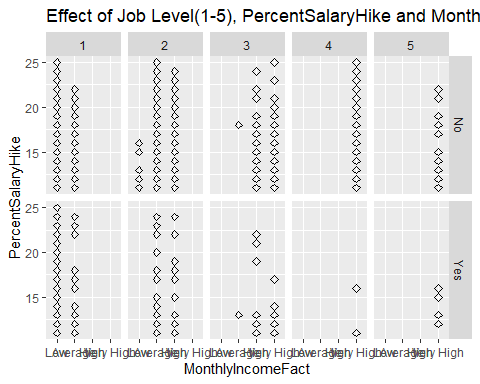
 Visual examination of the above plot suggests that attrition is most common among Laboratory technicians and Sales representative and level of job satisfaction plays little to no role in the attrition. Relatively speaking, attrition is uncommon among upper management staff (Manager, Sales Executive and research director)

#effect of job level, PercentSalaryHike and MonthlyIncome on Attrition (Y/N)  
  
#JobRole, JobLevel, Attrition  
attrition\_df\_recode %>%   
mutate(MonthlyIncomeFact = cut(MonthlyIncome, breaks = c(0,3000,6000,10000, 20000), labels = c("Low","Average", "High", "Very High"))) %>%   
ggplot(aes(x = JobLevel, fill = Attrition)) +   
geom\_bar(binwidth = 1) +   
facet\_wrap(vars(PercentSalaryHike))

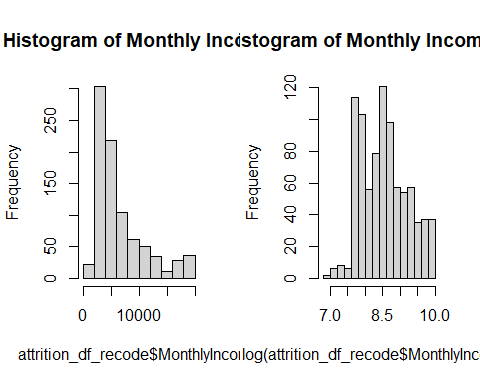
## Warning: Ignoring unknown parameters: binwidth



attrition\_df\_recode %>%  
 mutate(MonthlyIncomeFact = cut(MonthlyIncome, breaks = c(0,3000,6000,10000, 20000), labels = c("Low","Average", "High", "Very High"))) %>%   
ggplot(aes(x=MonthlyIncomeFact, y=PercentSalaryHike)) + geom\_point(shape=5)+ ggtitle("Effect of Job Level(1-5), PercentSalaryHike and MonthlyIncome on Attrition(Y/N)") +   
facet\_grid(Attrition ~ JobLevel)



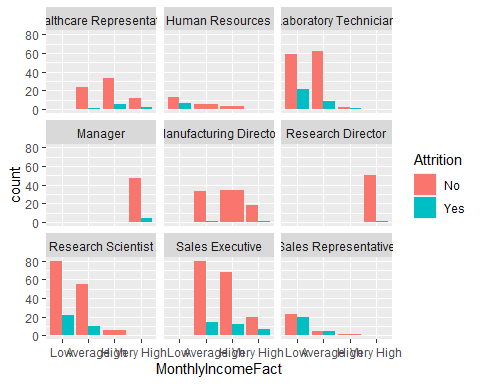
#Job levels 1 and 2 who are least paid also have the highest proportion of Attrition = Yes  
  
#checking the data distribution of MonthlyIncome to see if it satisfies normality assumption needed for linear models.  
  
par(mfrow = c(1, 2))  
p\_s <- hist(attrition\_df\_recode$MonthlyIncome, main = "Histogram of Monthly Income")  
p\_s.l <- hist(log(attrition\_df\_recode$MonthlyIncome), main = "Histogram of Monthly Income(log)")



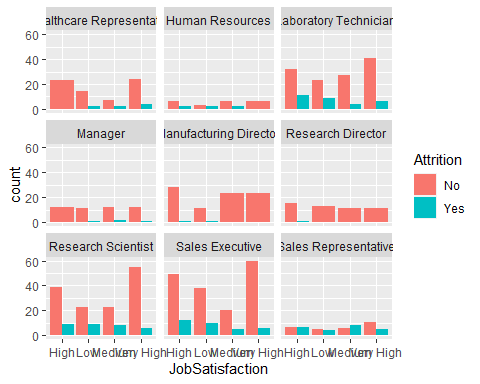
par(mfrow = c(1, 1))  
  
#The distribution approaches normality when transformed

###Other multivariate plots

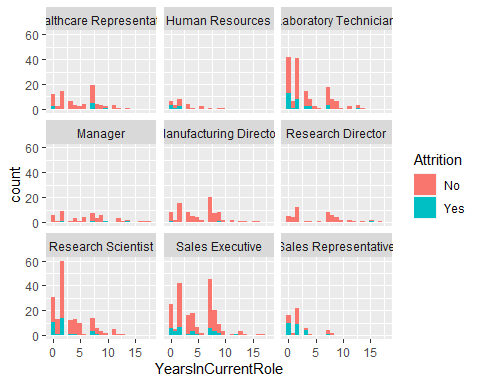
#Monthly income, Jobrole, Attrition  
  
attrition\_df\_recode %>%   
ggplot(aes(x = MonthlyIncomeFact, fill = Attrition)) +   
geom\_bar(position = position\_dodge()) +   
facet\_wrap(vars(JobRole))



#JobRole, JobSatistfaction, Attrition  
attrition\_df\_recode %>%   
ggplot(aes(x = JobSatisfaction, fill = Attrition)) +   
geom\_bar(position = position\_dodge()) +   
facet\_wrap(vars(JobRole))



#For the above. For sales reps, distinctively, whether they are high or low Job satisfaction, they still quit. Research Scientists are less likely to quit based on JobSat. So also Healthcare reps. JobSat is not conclusively responsible for Attrition among Lab techs. Upper management almost never quit.  
  
  
  
#JobRole, YearsinCurrentRole, Attrition  
attrition\_df\_recode %>%   
ggplot(aes(x = YearsInCurrentRole, fill = Attrition)) +   
geom\_histogram(binwidth = 0.8) +   
facet\_wrap(vars(JobRole))

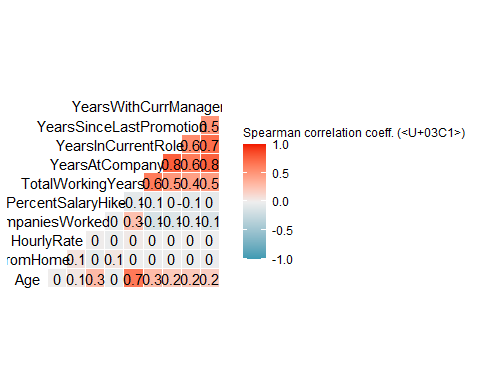


#Except for upper management, people that spend less than 5 years in a role tend to have high attrition=yes rate.

###more EDA ##Correlation matrix To view multi-colinearity among the variables and see which of the variables have strong relationship with Attrition

#plotting correlation matrix. Note that r documentation stated that Pearson method is not good for ordinal variables, So I used Spearman method instead.  
  
attrition\_df\_recode %>%   
 dplyr::select(Age,Attrition\_recode,Department,DistanceFromHome,Education,Gender,HourlyRate,JobInvolvement, MaritalStatus, NumCompaniesWorked,PercentSalaryHike, PerformanceRating,RelationshipSatisfaction, TotalWorkingYears,YearsAtCompany, YearsInCurrentRole, YearsSinceLastPromotion,YearsWithCurrManager, JobRole) %>%  
 ggcorr(palette = "RdBu", label = TRUE, hjust= 0.9, layout.exp = 1.2, name = "Spearman correlation coeff. (ρ)")

## Warning in ggcorr(., palette = "RdBu", label = TRUE, hjust = 0.9, layout.exp  
## = 1.2, : data in column(s) 'Attrition\_recode', 'Department', 'Education',  
## 'Gender', 'JobInvolvement', 'MaritalStatus', 'PerformanceRating',  
## 'RelationshipSatisfaction', 'JobRole' are not numeric and were ignored



attrition\_df\_recode %>% distinct(Gender)

## Gender  
## 1 Male  
## 2 Female

#the cor matrix suggests attrition correlates with Monthly income, JobRole, Years wih current mgr, Yearsincurrent role, Years at company,Totalworkingyears, Job Involvement. But Years at company and Totalworking years are likely confounded by Age.

###Encoding categorical variables The categorical variables are in levels and each level has a different proportion of contribution it offers to the response.To treat each level as a predictor entity we will re-code the levels. I will be using One hot encoding for encoding categorical variables wherein, each category of a categorical variable is converted into a new binary column (1/0).

dmy <- dummyVars(~., data = attrition\_df\_recode[,c(-2,-32)])  
trsf <- data.frame(predict(dmy, newdata = attrition\_df\_recode[, c(-2,-32)]))

###Removing Skewness Skewness in variables is undesirable for predictive modeling. Many of our varibales are skewed as seeing in the Monthly income histogram. There are many techniques for transforming a dataset but I choose log transform combined with centering and scaling as I do not know how the Box Cox works yet.

trsf <- trsf %>%  
 mutate(Age = log(Age + 1)  
 ,DailyRate = log(DailyRate + 1)  
 ,DistanceFromHome = log(DistanceFromHome + 1)  
 ,HourlyRate = log(HourlyRate + 1)  
 ,MonthlyIncome = log(MonthlyIncome + 1)  
 ,MonthlyRate = log(MonthlyRate + 1)  
 ,NumCompaniesWorked = log(NumCompaniesWorked + 1)  
 ,PercentSalaryHike = log(PercentSalaryHike + 1)  
 ,TotalWorkingYears = log(TotalWorkingYears + 1)  
 ,TrainingTimesLastYear = log(TrainingTimesLastYear + 1)  
 ,YearsAtCompany = log(YearsAtCompany +1)  
 ,YearsInCurrentRole = log(YearsInCurrentRole + 1)  
 ,YearsSinceLastPromotion = log(YearsSinceLastPromotion + 1)  
 ,YearsWithCurrManager = log(YearsWithCurrManager + 1)  
 )

prep\_num = preProcess(trsf, method=c("center", "scale"))  
final\_dataset = predict(prep\_num, trsf)

###Removing co related independent variables It is not desirable to have correlated features if we are using linear regressions. We will first find out variables which have a corelation of 0.85 or higher

set.seed(200)  
cor\_mat<- cor(final\_dataset)  
high\_corr <- findCorrelation(cor\_mat, cutoff = 0.85)  
names(trsf)[high\_corr]

## [1] "DepartmentSales" "DepartmentHuman.Resources"   
## [3] "GenderMale" "PerformanceRating.Outstanding"  
## [5] "OverTimeYes"

Removing the highly correlated variables

final\_dataset <- cbind(trsf, attrition\_df\_recode[2])  
final\_dataset <- cbind(final\_dataset, attrition\_df[1])  
final\_dataset <- final\_dataset %>%  
 mutate(Attrition = as.factor(if\_else(Attrition == "Yes",1,0))) %>%  
 dplyr::select(-DepartmentSales,-JobRole.Human.Resources,-PerformanceRating.Outstanding,-GenderMale,-OverTimeYes)  
str(final\_dataset$Attrition)#Attrition is now in factor with 2 levels

## Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...

###Predictive modelling

#Splitting the dataset  
set.seed(200)  
  
Train <- createDataPartition(final\_dataset$Attrition, p=0.7, list=FALSE)  
training <- final\_dataset[ Train, ]  
testing <- final\_dataset[ -Train, ]

##Let’s check if the proportion of response data is maintained in the sets we created.

prop.table(table(final\_dataset$Attrition))

##   
## 0 1   
## 0.8390805 0.1609195

prop.table(table(training$Attrition))

##   
## 0 1   
## 0.8390805 0.1609195

Yes so the proportion seems to be maintained, but some class imbalanced still persists.

###Model building Two models will be designed to predict the attrition of an employee. These models are : Linear regression, k-NN

##Logistic regression

trsf1 <- attrition\_df\_recode%>%  
 mutate(Age = log(Age + 1)  
 ,DailyRate = log(DailyRate + 1)  
 ,DistanceFromHome = log(DistanceFromHome + 1)  
 ,HourlyRate = log(HourlyRate + 1)  
 ,MonthlyIncome = log(MonthlyIncome + 1)  
 ,MonthlyRate = log(MonthlyRate + 1)  
 ,NumCompaniesWorked = log(NumCompaniesWorked + 1)  
 ,PercentSalaryHike = log(PercentSalaryHike + 1)  
 ,TotalWorkingYears = log(TotalWorkingYears + 1)  
 ,TrainingTimesLastYear = log(TrainingTimesLastYear + 1)  
 ,YearsAtCompany = log(YearsAtCompany +1)  
 ,YearsInCurrentRole = log(YearsInCurrentRole + 1)  
 ,YearsSinceLastPromotion = log(YearsSinceLastPromotion + 1)  
 ,YearsWithCurrManager = log(YearsWithCurrManager + 1)  
 ,Attrition = as.factor(Attrition)) %>%  
 dplyr::select(-Attrition\_recode)

###Logistic Regression ##Model Training

describe(trsf1)

## vars n mean sd median trimmed mad min max  
## Age 1 870 3.61 0.24 3.58 3.61 0.25 2.94 4.11  
## Attrition\* 2 870 1.16 0.37 1.00 1.08 0.00 1.00 2.00  
## BusinessTravel\* 3 870 2.60 0.68 3.00 2.75 0.00 1.00 3.00  
## DailyRate 4 870 6.54 0.64 6.71 6.62 0.60 4.64 7.31  
## Department\* 5 870 2.27 0.53 2.00 2.27 0.00 1.00 3.00  
## DistanceFromHome 6 870 2.00 0.85 2.08 2.00 1.12 0.69 3.40  
## Education\* 7 870 2.72 1.63 3.00 2.66 2.97 1.00 5.00  
## EducationField\* 8 870 3.24 1.32 3.00 3.09 1.48 1.00 6.00  
## EnvironmentSatisfaction\* 9 870 2.51 1.20 3.00 2.51 1.48 1.00 4.00  
## Gender\* 10 870 1.59 0.49 2.00 1.62 0.00 1.00 2.00  
## HourlyRate 11 870 4.15 0.32 4.20 4.17 0.38 3.43 4.62  
## JobInvolvement\* 12 870 1.86 1.10 1.00 1.71 0.00 1.00 4.00  
## JobLevel\* 13 870 2.04 1.09 2.00 1.87 1.48 1.00 5.00  
## JobRole\* 14 870 5.53 2.46 6.00 5.70 2.97 1.00 9.00  
## JobSatisfaction\* 15 870 2.52 1.21 3.00 2.53 1.48 1.00 4.00  
## MaritalStatus\* 16 870 2.09 0.72 2.00 2.11 1.48 1.00 3.00  
## MonthlyIncome 17 870 8.54 0.66 8.51 8.51 0.80 6.99 9.90  
## MonthlyRate 18 870 9.41 0.63 9.55 9.47 0.64 7.65 10.20  
## NumCompaniesWorked 19 870 1.09 0.67 1.10 1.09 0.60 0.00 2.30  
## OverTime\* 20 870 1.29 0.45 1.00 1.24 0.00 1.00 2.00  
## PercentSalaryHike 21 870 2.76 0.22 2.71 2.75 0.21 2.48 3.26  
## PerformanceRating\* 22 870 1.15 0.36 1.00 1.06 0.00 1.00 2.00  
## RelationshipSatisfaction\* 23 870 2.50 1.21 2.50 2.50 2.22 1.00 4.00  
## TotalWorkingYears 24 870 2.28 0.69 2.40 2.33 0.65 0.00 3.71  
## TrainingTimesLastYear 25 870 1.28 0.37 1.39 1.30 0.43 0.00 1.95  
## WorkLifeBalance\* 26 870 3.00 0.74 3.00 3.07 0.00 1.00 4.00  
## YearsAtCompany 27 870 1.80 0.77 1.79 1.81 0.90 0.00 3.71  
## YearsInCurrentRole 28 870 1.37 0.81 1.39 1.40 1.03 0.00 2.94  
## YearsSinceLastPromotion 29 870 0.79 0.81 0.69 0.69 1.03 0.00 2.77  
## YearsWithCurrManager 30 870 1.35 0.82 1.39 1.37 1.03 0.00 2.89  
## MonthlyIncomeFact\* 31 870 2.28 1.06 2.00 2.23 1.48 1.00 4.00  
## range skew kurtosis se  
## Age 1.17 -0.16 -0.23 0.01  
## Attrition\* 1.00 1.84 1.40 0.01  
## BusinessTravel\* 2.00 -1.43 0.62 0.02  
## DailyRate 2.67 -0.94 0.08 0.02  
## Department\* 2.00 0.16 -0.48 0.02  
## DistanceFromHome 2.71 -0.04 -1.23 0.03  
## Education\* 4.00 0.31 -1.48 0.06  
## EducationField\* 5.00 0.55 -0.69 0.04  
## EnvironmentSatisfaction\* 3.00 -0.02 -1.54 0.04  
## Gender\* 1.00 -0.38 -1.86 0.02  
## HourlyRate 1.18 -0.40 -0.95 0.01  
## JobInvolvement\* 3.00 0.71 -1.13 0.04  
## JobLevel\* 4.00 1.03 0.44 0.04  
## JobRole\* 8.00 -0.40 -1.16 0.08  
## JobSatisfaction\* 3.00 -0.02 -1.55 0.04  
## MaritalStatus\* 2.00 -0.14 -1.08 0.02  
## MonthlyIncome 2.92 0.27 -0.71 0.02  
## MonthlyRate 2.56 -0.81 -0.20 0.02  
## NumCompaniesWorked 2.30 0.10 -0.95 0.02  
## OverTime\* 1.00 0.93 -1.14 0.02  
## PercentSalaryHike 0.77 0.52 -0.80 0.01  
## PerformanceRating\* 1.00 1.94 1.76 0.01  
## RelationshipSatisfaction\* 3.00 0.00 -1.55 0.04  
## TotalWorkingYears 3.71 -0.67 0.60 0.02  
## TrainingTimesLastYear 1.95 -1.12 2.99 0.01  
## WorkLifeBalance\* 3.00 -0.80 0.94 0.03  
## YearsAtCompany 3.71 -0.23 -0.42 0.03  
## YearsInCurrentRole 2.94 -0.36 -0.88 0.03  
## YearsSinceLastPromotion 2.77 0.71 -0.61 0.03  
## YearsWithCurrManager 2.89 -0.37 -1.00 0.03  
## MonthlyIncomeFact\* 3.00 0.35 -1.11 0.04

set.seed(200)  
  
# Train <- createDataPartition(trsf1$Attrition, p=0.7, list=FALSE)  
# training1 <- trsf1[ Train, ]  
# testing1 <- trsf1[ -Train, ]  
#   
# training1 <- trsf1  
#BIC and AIC for this model  
  
attrition\_mod1 = glm(Attrition~., data = training, family = "binomial")

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

model\_summary <- summary(attrition\_mod1)  
  
step<- stepAIC(attrition\_mod1,direction = "backward",trace=FALSE)

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

summary(step)

##   
## Call:  
## glm(formula = Attrition ~ Age + BusinessTravelTravel\_Frequently +   
## DepartmentHuman.Resources + DepartmentResearch...Development +   
## DistanceFromHome + EnvironmentSatisfaction.Low + HourlyRate +   
## JobInvolvement.Low + JobInvolvement.Medium + JobLevel.1 +   
## JobLevel.2 + JobLevel.3 + JobLevel.4 + JobRole.Healthcare.Representative +   
## JobRole.Laboratory.Technician + JobRole.Manager + JobRole.Manufacturing.Director +   
## JobRole.Research.Director + JobRole.Research.Scientist +   
## JobRole.Sales.Executive + JobRole.Sales.Representative +   
## JobSatisfaction.High + JobSatisfaction.Low + MaritalStatusDivorced +   
## MaritalStatusMarried + MonthlyIncome + NumCompaniesWorked +   
## OverTimeNo + RelationshipSatisfaction.Low + RelationshipSatisfaction.Medium +   
## TrainingTimesLastYear + WorkLifeBalance.Bad + YearsSinceLastPromotion +   
## YearsWithCurrManager + ID, family = "binomial", data = training)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.72275 -0.37967 -0.10854 -0.00001 3.06902   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 6.044e+01 6.545e+03 0.009 0.992633   
## Age -2.171e+00 7.802e-01 -2.783 0.005381 \*\*   
## BusinessTravelTravel\_Frequently 9.263e-01 4.328e-01 2.140 0.032355 \*   
## DepartmentHuman.Resources -2.286e+01 6.343e+03 -0.004 0.997125   
## DepartmentResearch...Development -3.656e+00 2.671e+00 -1.369 0.171128   
## DistanceFromHome 4.688e-01 2.024e-01 2.316 0.020576 \*   
## EnvironmentSatisfaction.Low 1.491e+00 4.114e-01 3.624 0.000290 \*\*\*  
## HourlyRate 1.026e+00 5.341e-01 1.921 0.054715 .   
## JobInvolvement.Low 2.842e+00 6.530e-01 4.352 1.35e-05 \*\*\*  
## JobInvolvement.Medium 7.297e-01 3.860e-01 1.890 0.058721 .   
## JobLevel.1 -2.410e+01 1.616e+03 -0.015 0.988101   
## JobLevel.2 -2.560e+01 1.616e+03 -0.016 0.987358   
## JobLevel.3 -2.323e+01 1.616e+03 -0.014 0.988527   
## JobLevel.4 -2.314e+01 1.616e+03 -0.014 0.988571   
## JobRole.Healthcare.Representative -2.022e+01 6.343e+03 -0.003 0.997456   
## JobRole.Laboratory.Technician -2.084e+01 6.343e+03 -0.003 0.997378   
## JobRole.Manager -4.119e+01 6.545e+03 -0.006 0.994979   
## JobRole.Manufacturing.Director -3.843e+01 6.566e+03 -0.006 0.995330   
## JobRole.Research.Director -5.980e+01 6.799e+03 -0.009 0.992982   
## JobRole.Research.Scientist -2.123e+01 6.343e+03 -0.003 0.997329   
## JobRole.Sales.Executive -2.302e+01 6.343e+03 -0.004 0.997104   
## JobRole.Sales.Representative -2.283e+01 6.343e+03 -0.004 0.997128   
## JobSatisfaction.High 8.188e-01 3.890e-01 2.105 0.035308 \*   
## JobSatisfaction.Low 1.764e+00 4.500e-01 3.920 8.84e-05 \*\*\*  
## MaritalStatusDivorced -2.366e+00 5.674e-01 -4.169 3.05e-05 \*\*\*  
## MaritalStatusMarried -1.129e+00 3.742e-01 -3.018 0.002542 \*\*   
## MonthlyIncome -1.330e+00 7.071e-01 -1.881 0.059942 .   
## NumCompaniesWorked 9.505e-01 2.655e-01 3.580 0.000344 \*\*\*  
## OverTimeNo -2.432e+00 3.847e-01 -6.320 2.61e-10 \*\*\*  
## RelationshipSatisfaction.Low 1.074e+00 4.240e-01 2.533 0.011319 \*   
## RelationshipSatisfaction.Medium 8.036e-01 4.785e-01 1.679 0.093055 .   
## TrainingTimesLastYear -8.047e-01 4.350e-01 -1.850 0.064359 .   
## WorkLifeBalance.Bad 1.975e+00 5.626e-01 3.511 0.000447 \*\*\*  
## YearsSinceLastPromotion 9.863e-01 2.983e-01 3.307 0.000943 \*\*\*  
## YearsWithCurrManager -1.229e+00 2.766e-01 -4.442 8.93e-06 \*\*\*  
## ID 1.590e-03 6.957e-04 2.286 0.022281 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 537.37 on 608 degrees of freedom  
## Residual deviance: 265.10 on 573 degrees of freedom  
## AIC: 337.1  
##   
## Number of Fisher Scoring iterations: 19

step$coefficients

## (Intercept) Age   
## 60.436829600 -2.171448033   
## BusinessTravelTravel\_Frequently DepartmentHuman.Resources   
## 0.926284643 -22.857837310   
## DepartmentResearch...Development DistanceFromHome   
## -3.655741237 0.468804070   
## EnvironmentSatisfaction.Low HourlyRate   
## 1.491046016 1.026113323   
## JobInvolvement.Low JobInvolvement.Medium   
## 2.842146367 0.729659195   
## JobLevel.1 JobLevel.2   
## -24.095212440 -25.599968240   
## JobLevel.3 JobLevel.4   
## -23.232885260 -23.142879814   
## JobRole.Healthcare.Representative JobRole.Laboratory.Technician   
## -20.224038719 -20.839500922   
## JobRole.Manager JobRole.Manufacturing.Director   
## -41.189532704 -38.430179117   
## JobRole.Research.Director JobRole.Research.Scientist   
## -59.798375618 -21.233724177   
## JobRole.Sales.Executive JobRole.Sales.Representative   
## -23.017607815 -22.828405391   
## JobSatisfaction.High JobSatisfaction.Low   
## 0.818773340 1.764055916   
## MaritalStatusDivorced MaritalStatusMarried   
## -2.365732273 -1.129372267   
## MonthlyIncome NumCompaniesWorked   
## -1.330207939 0.950534444   
## OverTimeNo RelationshipSatisfaction.Low   
## -2.431734257 1.073840533   
## RelationshipSatisfaction.Medium TrainingTimesLastYear   
## 0.803586290 -0.804673973   
## WorkLifeBalance.Bad YearsSinceLastPromotion   
## 1.975105348 0.986345215   
## YearsWithCurrManager ID   
## -1.228699181 0.001590051

step$anova

## Stepwise Model Path   
## Analysis of Deviance Table  
##   
## Initial Model:  
## Attrition ~ Age + BusinessTravelNon.Travel + BusinessTravelTravel\_Frequently +   
## BusinessTravelTravel\_Rarely + DailyRate + DepartmentHuman.Resources +   
## DepartmentResearch...Development + DistanceFromHome + Education.Bachelor +   
## Education.Below.College + Education.College + Education.Doctor +   
## Education.Master + EducationFieldHuman.Resources + EducationFieldLife.Sciences +   
## EducationFieldMarketing + EducationFieldMedical + EducationFieldOther +   
## EducationFieldTechnical.Degree + EnvironmentSatisfaction.High +   
## EnvironmentSatisfaction.Low + EnvironmentSatisfaction.Medium +   
## EnvironmentSatisfaction.Very.High + GenderFemale + HourlyRate +   
## JobInvolvement.High + JobInvolvement.Low + JobInvolvement.Medium +   
## JobInvolvement.Very.High + JobLevel.1 + JobLevel.2 + JobLevel.3 +   
## JobLevel.4 + JobLevel.5 + JobRole.Healthcare.Representative +   
## JobRole.Laboratory.Technician + JobRole.Manager + JobRole.Manufacturing.Director +   
## JobRole.Research.Director + JobRole.Research.Scientist +   
## JobRole.Sales.Executive + JobRole.Sales.Representative +   
## JobSatisfaction.High + JobSatisfaction.Low + JobSatisfaction.Medium +   
## JobSatisfaction.Very.High + MaritalStatusDivorced + MaritalStatusMarried +   
## MaritalStatusSingle + MonthlyIncome + MonthlyRate + NumCompaniesWorked +   
## OverTimeNo + PercentSalaryHike + PerformanceRating.Excellent +   
## RelationshipSatisfaction.High + RelationshipSatisfaction.Low +   
## RelationshipSatisfaction.Medium + RelationshipSatisfaction.Very.High +   
## TotalWorkingYears + TrainingTimesLastYear + WorkLifeBalance.Bad +   
## WorkLifeBalance.Best + WorkLifeBalance.Better + WorkLifeBalance.Good +   
## YearsAtCompany + YearsInCurrentRole + YearsSinceLastPromotion +   
## YearsWithCurrManager + MonthlyIncomeFact.Low + MonthlyIncomeFact.Average +   
## MonthlyIncomeFact.High + MonthlyIncomeFact.Very.High + ID  
##   
## Final Model:  
## Attrition ~ Age + BusinessTravelTravel\_Frequently + DepartmentHuman.Resources +   
## DepartmentResearch...Development + DistanceFromHome + EnvironmentSatisfaction.Low +   
## HourlyRate + JobInvolvement.Low + JobInvolvement.Medium +   
## JobLevel.1 + JobLevel.2 + JobLevel.3 + JobLevel.4 + JobRole.Healthcare.Representative +   
## JobRole.Laboratory.Technician + JobRole.Manager + JobRole.Manufacturing.Director +   
## JobRole.Research.Director + JobRole.Research.Scientist +   
## JobRole.Sales.Executive + JobRole.Sales.Representative +   
## JobSatisfaction.High + JobSatisfaction.Low + MaritalStatusDivorced +   
## MaritalStatusMarried + MonthlyIncome + NumCompaniesWorked +   
## OverTimeNo + RelationshipSatisfaction.Low + RelationshipSatisfaction.Medium +   
## TrainingTimesLastYear + WorkLifeBalance.Bad + YearsSinceLastPromotion +   
## YearsWithCurrManager + ID  
##   
##   
## Step Df Deviance Resid. Df Resid. Dev  
## 1 545 248.7297  
## 2 - MonthlyIncomeFact.Very.High 0 0.000000000 545 248.7297  
## 3 - WorkLifeBalance.Good 0 0.000000000 545 248.7297  
## 4 - RelationshipSatisfaction.Very.High 0 0.000000000 545 248.7297  
## 5 - MaritalStatusSingle 0 0.000000000 545 248.7297  
## 6 - JobSatisfaction.Very.High 0 0.000000000 545 248.7297  
## 7 - JobLevel.5 0 0.000000000 545 248.7297  
## 8 - JobInvolvement.Very.High 0 0.000000000 545 248.7297  
## 9 - EnvironmentSatisfaction.Very.High 0 0.000000000 545 248.7297  
## 10 - EducationFieldTechnical.Degree 0 0.000000000 545 248.7297  
## 11 - Education.Master 0 0.000000000 545 248.7297  
## 12 - BusinessTravelTravel\_Rarely 0 0.000000000 545 248.7297  
## 13 - EducationFieldHuman.Resources 1 0.001039077 546 248.7307  
## 14 - PercentSalaryHike 1 0.007566596 547 248.7383  
## 15 - Education.Bachelor 1 0.010247345 548 248.7485  
## 16 - JobSatisfaction.Medium 1 0.041040980 549 248.7896  
## 17 - EducationFieldOther 1 0.039382723 550 248.8290  
## 18 - Education.College 1 0.093695424 551 248.9227  
## 19 - RelationshipSatisfaction.High 1 0.094619199 552 249.0173  
## 20 - MonthlyIncomeFact.Low 1 0.112756040 553 249.1300  
## 21 - MonthlyIncomeFact.High 1 0.052702998 554 249.1827  
## 22 - GenderFemale 1 0.380982369 555 249.5637  
## 23 - EnvironmentSatisfaction.Medium 1 0.383848735 556 249.9476  
## 24 - EnvironmentSatisfaction.High 1 0.199635590 557 250.1472  
## 25 - JobInvolvement.High 1 0.365098743 558 250.5123  
## 26 - YearsInCurrentRole 1 0.396622659 559 250.9089  
## 27 - YearsAtCompany 1 0.460250728 560 251.3692  
## 28 - TotalWorkingYears 1 0.425849089 561 251.7950  
## 29 - MonthlyIncomeFact.Average 1 0.628314337 562 252.4233  
## 30 - Education.Doctor 1 0.781225098 563 253.2046  
## 31 - MonthlyRate 1 0.725577750 564 253.9302  
## 32 - DailyRate 1 0.934604028 565 254.8648  
## 33 - WorkLifeBalance.Better 1 1.265332121 566 256.1301  
## 34 - WorkLifeBalance.Best 1 0.647080450 567 256.7772  
## 35 - Education.Below.College 1 1.248072142 568 258.0252  
## 36 - EducationFieldMarketing 1 1.308371158 569 259.3336  
## 37 - EducationFieldMedical 1 1.067738825 570 260.4013  
## 38 - EducationFieldLife.Sciences 1 1.601466620 571 262.0028  
## 39 - PerformanceRating.Excellent 1 1.322370761 572 263.3252  
## 40 - BusinessTravelNon.Travel 1 1.773431362 573 265.0986  
## AIC  
## 1 376.7297  
## 2 376.7297  
## 3 376.7297  
## 4 376.7297  
## 5 376.7297  
## 6 376.7297  
## 7 376.7297  
## 8 376.7297  
## 9 376.7297  
## 10 376.7297  
## 11 376.7297  
## 12 376.7297  
## 13 374.7307  
## 14 372.7383  
## 15 370.7485  
## 16 368.7896  
## 17 366.8290  
## 18 364.9227  
## 19 363.0173  
## 20 361.1300  
## 21 359.1827  
## 22 357.5637  
## 23 355.9476  
## 24 354.1472  
## 25 352.5123  
## 26 350.9089  
## 27 349.3692  
## 28 347.7950  
## 29 346.4233  
## 30 345.2046  
## 31 343.9302  
## 32 342.8648  
## 33 342.1301  
## 34 340.7772  
## 35 340.0252  
## 36 339.3336  
## 37 338.4013  
## 38 338.0028  
## 39 337.3252  
## 40 337.0986

#this resulted in the below model  
  
step\_select <- glm(Attrition ~ Age + BusinessTravel + DistanceFromHome + EnvironmentSatisfaction +   
 HourlyRate + JobInvolvement + JobLevel + JobRole + JobSatisfaction +   
 MaritalStatus + NumCompaniesWorked + OverTime + PerformanceRating +   
 RelationshipSatisfaction + TotalWorkingYears + TrainingTimesLastYear +   
 WorkLifeBalance + YearsInCurrentRole + YearsSinceLastPromotion +   
 YearsWithCurrManager, data = training, family = "binomial")

## Error in eval(predvars, data, env): object 'BusinessTravel' not found

step\_select\_summary <- summary(step\_select)

## Error in summary(step\_select): object 'step\_select' not found

step\_select\_summary

## Error in eval(expr, envir, enclos): object 'step\_select\_summary' not found

#to calculate R correlation:  
#1 - (Residual Deviance/Null Deviance)  
  
step\_select.r <- with(summary(step\_select), 1 - deviance/null.deviance) #r =0.4469, r^2 = 0.16

## Error in summary(step\_select): object 'step\_select' not found

#Residual plots and Cook's D plots to check for assumptions.  
  
par(mfrow = c(2, 1))  
p\_r <- plot(step\_select$fitted.values,step\_select$residuals, main = "Residual Plot for Attrition classification")

## Error in plot(step\_select$fitted.values, step\_select$residuals, main = "Residual Plot for Attrition classification"): object 'step\_select' not found

p\_c <- plot(cooks.distance(step\_select), main = "Cooks' D for Attrition classification")

## Error in cooks.distance(step\_select): object 'step\_select' not found

par(mfrow = c(1, 1))  
  
#There are points that are far from the regression line and there are no random clouds of residuals around -2 to +2.  
#This shows that the linear model is not a good model to classify the response.  
#on the positive side the Cooks'D plot did not show any strongly high leverage point,so there are no extreme outlier in the plotted data.  
  
  
library(qpcR)

## Warning: package 'qpcR' was built under R version 4.0.3

## Loading required package: minpack.lm

## Warning: package 'minpack.lm' was built under R version 4.0.3

## Loading required package: rgl

## Warning: package 'rgl' was built under R version 4.0.3

## Loading required package: robustbase

## Warning: package 'robustbase' was built under R version 4.0.3

## Loading required package: Matrix

##   
## Attaching package: 'Matrix'

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

##   
## Attaching package: 'qpcR'

## The following object is masked from 'package:caret':  
##   
## RMSE

qpcR::RMSE(attrition\_mod1) #calculated from the residuals

## [1] 0.6390799

#Effect: The greater the parameter estimate(step\_select output) the greater its effect on attrition=Yes when other predictors are kept constant.

These top variables are:JobInvolvementLow, BusinessTravelTravel,OverTimeYes,MaritalStatusSingle

###Below is KNN classification model

set.seed(400)  
control <- trainControl(method="repeatedcv",repeats = 3) #,classProbs=TRUE,summaryFunction = twoClassSummary)  
knnFit <- train(Attrition~., data = training, method = "knn", trControl = control, preProcess = c("center","scale"), tuneLength = 20)  
#The final value used for the model was k = 7.  
pred\_rf <- predict(knnFit, newdata=testing)  
confusionMatrix(table(pred\_rf,testing$Attrition))

## Confusion Matrix and Statistics  
##   
##   
## pred\_rf 0 1  
## 0 216 38  
## 1 3 4  
##   
## Accuracy : 0.8429   
## 95% CI : (0.793, 0.8849)  
## No Information Rate : 0.8391   
## P-Value [Acc > NIR] : 0.474   
##   
## Kappa : 0.1229   
##   
## Mcnemar's Test P-Value : 1.097e-07   
##   
## Sensitivity : 0.98630   
## Specificity : 0.09524   
## Pos Pred Value : 0.85039   
## Neg Pred Value : 0.57143   
## Prevalence : 0.83908   
## Detection Rate : 0.82759   
## Detection Prevalence : 0.97318   
## Balanced Accuracy : 0.54077   
##   
## 'Positive' Class : 0   
##

#The above implementation gave me an accuracy of 0.85 and a Sensitivity of 0.995

###Classifying the new unknown data set that doesnt contain attrition label.

no.attrition\_df <- read.csv("C:/Users/olani/OneDrive/Documents/Data Science/SMU-Data Science/Doing Data Science/MSDS\_6306\_DDS/Unit 14 and 15 Case Study 2/CaseStudy2CompSet No Attrition.csv")  
  
class(no.attrition\_df)

## [1] "data.frame"

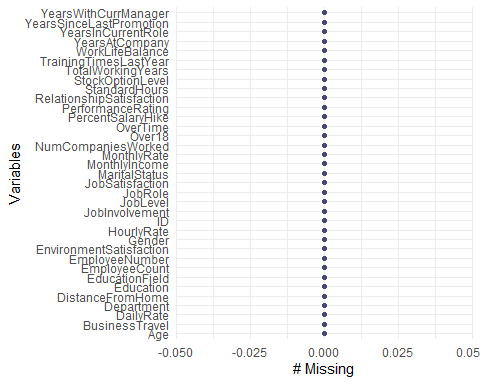
str(attrition\_df)

## 'data.frame': 870 obs. of 36 variables:  
## $ ID : int 1 2 3 4 5 6 7 8 9 10 ...  
## $ Age : int 32 40 35 32 24 27 41 37 34 34 ...  
## $ Attrition : chr "No" "No" "No" "No" ...  
## $ BusinessTravel : chr "Travel\_Rarely" "Travel\_Rarely" "Travel\_Frequently" "Travel\_Rarely" ...  
## $ DailyRate : int 117 1308 200 801 567 294 1283 309 1333 653 ...  
## $ Department : chr "Sales" "Research & Development" "Research & Development" "Sales" ...  
## $ DistanceFromHome : int 13 14 18 1 2 10 5 10 10 10 ...  
## $ Education : int 4 3 2 4 1 2 5 4 4 4 ...  
## $ EducationField : chr "Life Sciences" "Medical" "Life Sciences" "Marketing" ...  
## $ EmployeeCount : int 1 1 1 1 1 1 1 1 1 1 ...  
## $ EmployeeNumber : int 859 1128 1412 2016 1646 733 1448 1105 1055 1597 ...  
## $ EnvironmentSatisfaction : int 2 3 3 3 1 4 2 4 3 4 ...  
## $ Gender : chr "Male" "Male" "Male" "Female" ...  
## $ HourlyRate : int 73 44 60 48 32 32 90 88 87 92 ...  
## $ JobInvolvement : int 3 2 3 3 3 3 4 2 3 2 ...  
## $ JobLevel : int 2 5 3 3 1 3 1 2 1 2 ...  
## $ JobRole : chr "Sales Executive" "Research Director" "Manufacturing Director" "Sales Executive" ...  
## $ JobSatisfaction : int 4 3 4 4 4 1 3 4 3 3 ...  
## $ MaritalStatus : chr "Divorced" "Single" "Single" "Married" ...  
## $ MonthlyIncome : int 4403 19626 9362 10422 3760 8793 2127 6694 2220 5063 ...  
## $ MonthlyRate : int 9250 17544 19944 24032 17218 4809 5561 24223 18410 15332 ...  
## $ NumCompaniesWorked : int 2 1 2 1 1 1 2 2 1 1 ...  
## $ Over18 : chr "Y" "Y" "Y" "Y" ...  
## $ OverTime : chr "No" "No" "No" "No" ...  
## $ PercentSalaryHike : int 11 14 11 19 13 21 12 14 19 14 ...  
## $ PerformanceRating : int 3 3 3 3 3 4 3 3 3 3 ...  
## $ RelationshipSatisfaction: int 3 1 3 3 3 3 1 3 4 2 ...  
## $ StandardHours : int 80 80 80 80 80 80 80 80 80 80 ...  
## $ StockOptionLevel : int 1 0 0 2 0 2 0 3 1 1 ...  
## $ TotalWorkingYears : int 8 21 10 14 6 9 7 8 1 8 ...  
## $ TrainingTimesLastYear : int 3 2 2 3 2 4 5 5 2 3 ...  
## $ WorkLifeBalance : int 2 4 3 3 3 2 2 3 3 2 ...  
## $ YearsAtCompany : int 5 20 2 14 6 9 4 1 1 8 ...  
## $ YearsInCurrentRole : int 2 7 2 10 3 7 2 0 1 2 ...  
## $ YearsSinceLastPromotion : int 0 4 2 5 1 1 0 0 0 7 ...  
## $ YearsWithCurrManager : int 3 9 2 7 3 7 3 0 0 7 ...

names(no.attrition\_df)

## [1] "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"

gg\_miss\_var(no.attrition\_df) #no missing data



dim(no.attrition\_df)

## [1] 300 35

###I ll do every processing I did to the training data to this unknown data set as well

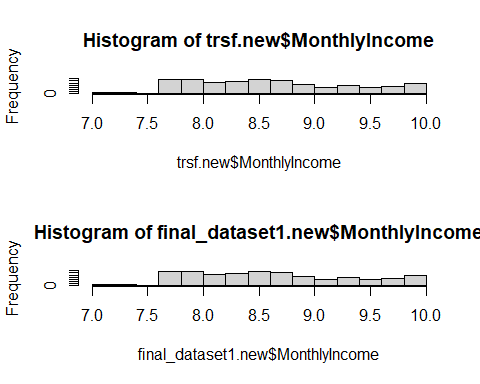
no.attrition\_df\_recode <- no.attrition\_df %>%  
 mutate(Education = as.factor(if\_else(Education == 1,"Below College", if\_else(Education == 2, "College", if\_else(Education == 3, "Bachelor", if\_else(Education == 4, "Master","Doctor"))))),  
 EnvironmentSatisfaction = as.factor(if\_else(EnvironmentSatisfaction == 1,"Low",if\_else(EnvironmentSatisfaction == 2, "Medium", if\_else(EnvironmentSatisfaction == 3, "High", "Very High")))),  
 JobInvolvement = as.factor(if\_else(JobInvolvement == 1,"Low",if\_else(JobInvolvement == 2, "Medium",if\_else(JobInvolvement == 3, "High", "Very High")))),  
 JobSatisfaction = as.factor(if\_else(JobSatisfaction == 1, "Low",if\_else(JobSatisfaction == 2, "Medium",if\_else(JobSatisfaction == 3, "High","Very High")))),  
 PerformanceRating = as.factor(if\_else(PerformanceRating == 1, "Low",if\_else(PerformanceRating == 2, "Good", if\_else(PerformanceRating == 3, "Excellent", "Outstanding")))),  
 RelationshipSatisfaction = as.factor(if\_else(RelationshipSatisfaction == 1, "Low",if\_else(RelationshipSatisfaction == 2, "Medium", if\_else(RelationshipSatisfaction == 3, "High", "Very High")))),   
 WorkLifeBalance = as.factor(if\_else(WorkLifeBalance == 1, "Bad",if\_else(WorkLifeBalance == 2, "Good", if\_else(WorkLifeBalance == 3, "Better", "Best")))),  
 JobLevel = as.factor(JobLevel),  
 MonthlyIncomeFact = as.factor(cut(MonthlyIncome, breaks = c(0,3000,6000,10000, 20000), labels = c("Low","Average", "High", "Very High"))),  
 JobRole = as.factor(JobRole)) %>%  
 dplyr::select(-EmployeeCount, -EmployeeNumber, -Over18, -StandardHours, -EmployeeCount, -StockOptionLevel)  
  
#check if those features are now in levels  
str(no.attrition\_df\_recode)

## 'data.frame': 300 obs. of 31 variables:  
## $ ID : int 1171 1172 1173 1174 1175 1176 1177 1178 1179 1180 ...  
## $ Age : int 35 33 26 55 29 51 52 39 31 31 ...  
## $ BusinessTravel : chr "Travel\_Rarely" "Travel\_Rarely" "Travel\_Rarely" "Travel\_Rarely" ...  
## $ DailyRate : int 750 147 1330 1311 1246 1456 585 1387 1062 534 ...  
## $ Department : chr "Research & Development" "Human Resources" "Research & Development" "Research & Development" ...  
## $ DistanceFromHome : int 28 2 21 2 19 1 29 10 24 20 ...  
## $ Education : Factor w/ 5 levels "Bachelor","Below College",..: 1 1 1 1 1 5 5 4 1 1 ...  
## $ EducationField : chr "Life Sciences" "Human Resources" "Medical" "Life Sciences" ...  
## $ EnvironmentSatisfaction : Factor w/ 4 levels "High","Low","Medium",..: 3 3 2 1 1 2 2 3 1 2 ...  
## $ Gender : chr "Male" "Male" "Male" "Female" ...  
## $ HourlyRate : int 46 99 37 97 77 30 40 76 96 66 ...  
## $ JobInvolvement : Factor w/ 4 levels "High","Low","Medium",..: 4 1 1 1 3 3 1 1 3 1 ...  
## $ JobLevel : Factor w/ 5 levels "1","2","3","4",..: 2 1 1 4 2 3 1 2 2 3 ...  
## $ JobRole : Factor w/ 9 levels "Healthcare Representative",..: 3 2 3 4 8 1 9 5 1 1 ...  
## $ JobSatisfaction : Factor w/ 4 levels "High","Low","Medium",..: 1 1 1 4 1 2 4 2 2 1 ...  
## $ MaritalStatus : chr "Married" "Married" "Divorced" "Single" ...  
## $ MonthlyIncome : int 3407 3600 2377 16659 8620 7484 3482 5377 6812 9824 ...  
## $ MonthlyRate : int 25348 8429 19373 23258 23757 25796 19788 3835 17198 22908 ...  
## $ NumCompaniesWorked : int 1 1 1 2 1 3 2 2 1 3 ...  
## $ OverTime : chr "No" "No" "No" "Yes" ...  
## $ PercentSalaryHike : int 17 13 20 13 14 20 15 13 19 12 ...  
## $ PerformanceRating : Factor w/ 2 levels "Excellent","Outstanding": 1 1 2 1 1 2 1 1 1 1 ...  
## $ RelationshipSatisfaction: Factor w/ 4 levels "High","Low","Medium",..: 4 4 1 1 1 1 3 4 3 2 ...  
## $ TotalWorkingYears : int 10 5 1 30 10 23 16 10 10 12 ...  
## $ TrainingTimesLastYear : int 3 2 0 2 3 1 3 3 2 2 ...  
## $ WorkLifeBalance : Factor w/ 4 levels "Bad","Best","Better",..: 4 3 4 3 3 4 4 3 3 3 ...  
## $ YearsAtCompany : int 10 5 1 5 10 13 9 7 10 1 ...  
## $ YearsInCurrentRole : int 9 4 1 4 7 12 8 7 9 0 ...  
## $ YearsSinceLastPromotion : int 6 1 0 1 0 12 0 7 1 0 ...  
## $ YearsWithCurrManager : int 8 4 0 2 4 8 0 7 8 0 ...  
## $ MonthlyIncomeFact : Factor w/ 4 levels "Low","Average",..: 2 2 1 4 3 3 2 2 3 3 ...

#one hot encoding for all except continuous varibales and the unique ID column.  
dmy <- dummyVars(~., data = no.attrition\_df\_recode[-1])  
trsf.new <- data.frame(predict(dmy, newdata = no.attrition\_df\_recode[-1]))  
  
#many of the predictors are skewed so we transform the data set to minimize the skewness  
trsf.new <- trsf.new %>%  
 mutate(Age = log(Age + 1)  
 ,DailyRate = log(DailyRate + 1)  
 ,DistanceFromHome = log(DistanceFromHome + 1)  
 ,HourlyRate = log(HourlyRate + 1)  
 ,MonthlyIncome = log(MonthlyIncome + 1)  
 ,MonthlyRate = log(MonthlyRate + 1)  
 ,NumCompaniesWorked = log(NumCompaniesWorked + 1)  
 ,PercentSalaryHike = log(PercentSalaryHike + 1)  
 ,TotalWorkingYears = log(TotalWorkingYears + 1)  
 ,TrainingTimesLastYear = log(TrainingTimesLastYear + 1)  
 ,YearsAtCompany = log(YearsAtCompany +1)  
 ,YearsInCurrentRole = log(YearsInCurrentRole + 1)  
 ,YearsSinceLastPromotion = log(YearsSinceLastPromotion + 1)  
 ,YearsWithCurrManager = log(YearsWithCurrManager + 1)  
 )  
#standardize dataset  
prep\_num = preProcess(trsf.new, method=c("center", "scale"))  
final\_dataset1.new = predict(prep\_num, trsf.new)  
final\_dataset1.new = cbind(trsf.new, no.attrition\_df\_recode[1]) #Unknown is ready for prediction but we need to tag the observations with the unique ID  
  
set.seed(200)  
pred\_knn.new <- predict(knnFit, newdata=final\_dataset1.new, type ="raw")  
  
#I cannot build a confusion MAtrix because I don't have an expected label for the dataset.  
#confusionMatrix(table(pred\_knn.new,as.factor(training$Attrition)))  
summary(pred\_knn.new)

## 0 1   
## 284 16

#Just checking the dataset  
  
par(mfrow = c(2, 1))  
hist(trsf.new$MonthlyIncome)  
hist(final\_dataset1.new$MonthlyIncome)



par(mfrow = c(1, 1))  
  
  
prop.table(table(no.attrition\_df.pred$pred\_Attrition))

## Error in table(no.attrition\_df.pred$pred\_Attrition): object 'no.attrition\_df.pred' not found

str(no.attrition\_df.pred)

## Error in str(no.attrition\_df.pred): object 'no.attrition\_df.pred' not found

head(no.attrition\_df.pred)

## Error in h(simpleError(msg, call)): error in evaluating the argument 'x' in selecting a method for function 'head': object 'no.attrition\_df.pred' not found

no.attrition\_df\_recode%>%filter (ID == 1171) %>% dplyr::select(MonthlyRate)

## MonthlyRate  
## 1 25348

#combine the predicted attrition data with the unknown dataset and add unique ID

no.attrition\_df.pred <- cbind(final\_dataset.new, pred\_Attrition=pred\_knn.new)

## Error in cbind(final\_dataset.new, pred\_Attrition = pred\_knn.new): object 'final\_dataset.new' not found

no.attrition\_df.pred <- no.attrition\_df.pred %>%  
 mutate(pred\_Attrition = as.factor(if\_else(pred\_Attrition == 0,"No", "Yes")))

## Error in eval(lhs, parent, parent): object 'no.attrition\_df.pred' not found

no.attrition\_df.pred1 <- no.attrition\_df.pred %>% dplyr::select(ID, pred\_Attrition)

## Error in eval(lhs, parent, parent): object 'no.attrition\_df.pred' not found

no.attrition\_df.pred\_new <- left\_join(no.attrition\_df\_recode, no.attrition\_df.pred1, by = "ID")

## Error in is.data.frame(y): object 'no.attrition\_df.pred1' not found

str(no.attrition\_df.pred\_new)

## Error in str(no.attrition\_df.pred\_new): object 'no.attrition\_df.pred\_new' not found

#Write the result to a csv file and move it to github  
write.csv(no.attrition\_df.pred\_new, 'C:\\Users\\olani\\OneDrive\\Documents\\Data Science\\SMU-Data Science\\Doing Data Science\\MSDS\_6306\_DDS\\Unit 14 and 15 Case Study 2\\predicted attrition.csv')

## Error in is.data.frame(x): object 'no.attrition\_df.pred\_new' not found

###Result evaluation

Employee’s department and other factors should agree with the top factors selected by linear the model and trend in our training data. This is incorrect.

Attrition.Yes <- no.attrition\_df.pred\_new %>%  
 filter(pred\_Attrition == "Yes")

## Error in eval(lhs, parent, parent): object 'no.attrition\_df.pred\_new' not found

head(Attrition.Yes, n=20)

## Error in h(simpleError(msg, call)): error in evaluating the argument 'x' in selecting a method for function 'head': object 'Attrition.Yes' not found

###Examine the employees predicted to leave the company. Firstly,K-NN predicted result reveals that most of the employees predicted to leave the company work as Lab Tech or sales rep. They rarely travel on business and have low monthly income. This category of employees also have either low or very high total working years suggesting early career with high mobility and late career departing due to retirement. Some of these factors are not included in the top factors identified by the linear model. We should also note that the linear model was only able to explain about 16% (r = 0.411) of attrition.

###Predicting salary from another “unknown” dataset ###This appears to be a regression problem since we are not anticipating a categorical response. To minimize redundant, I ll remove variables that are perceived to be directly associated with salary: HourlyRate, MonthlyRate, Overtime, PercentSlaryHike, MonthlyIncomeFact.

no.salary\_df <- read.csv("C:/Users/olani/OneDrive/Documents/Data Science/SMU-Data Science/Doing Data Science/MSDS\_6306\_DDS/Unit 14 and 15 Case Study 2//CaseStudy2CompSet No Salary.csv")  
  
  
no.salary\_recode <- no.salary\_df %>%  
 mutate(Education = as.factor(if\_else(Education == 1,"Below College", if\_else(Education == 2, "College", if\_else(Education == 3, "Bachelor", if\_else(Education == 4, "Master","Doctor"))))),  
 EnvironmentSatisfaction = as.factor(if\_else(EnvironmentSatisfaction == 1,"Low",if\_else(EnvironmentSatisfaction == 2, "Medium", if\_else(EnvironmentSatisfaction == 3, "High", "Very High")))),  
 JobInvolvement = as.factor(if\_else(JobInvolvement == 1,"Low",if\_else(JobInvolvement == 2, "Medium",if\_else(JobInvolvement == 3, "High", "Very High")))),  
 JobSatisfaction = as.factor(if\_else(JobSatisfaction == 1, "Low",if\_else(JobSatisfaction == 2, "Medium",if\_else(JobSatisfaction == 3, "High","Very High")))),  
 PerformanceRating = as.factor(if\_else(PerformanceRating == 1, "Low",if\_else(PerformanceRating == 2, "Good", if\_else(PerformanceRating == 3, "Excellent", "Outstanding")))),  
 RelationshipSatisfaction = as.factor(if\_else(RelationshipSatisfaction == 1, "Low",if\_else(RelationshipSatisfaction == 2, "Medium", if\_else(RelationshipSatisfaction == 3, "High", "Very High")))),   
 WorkLifeBalance = as.factor(if\_else(WorkLifeBalance == 1, "Bad",if\_else(WorkLifeBalance == 2, "Good", if\_else(WorkLifeBalance == 3, "Better", "Best")))),  
 JobLevel = as.factor(JobLevel),  
 JobRole = as.factor(JobRole)) %>%  
 dplyr::select(-EmployeeCount, -EmployeeNumber, -Over18, -StandardHours, -EmployeeCount, -StockOptionLevel, -HourlyRate, -MonthlyRate, -OverTime, -PercentSalaryHike)

#check if those features are now in levels  
str(no.salary\_recode)

## 'data.frame': 300 obs. of 26 variables:  
## $ ï..ID : int 871 872 873 874 875 876 877 878 879 880 ...  
## $ Age : int 43 33 55 36 27 39 33 21 30 51 ...  
## $ Attrition : chr "No" "No" "Yes" "No" ...  
## $ BusinessTravel : chr "Travel\_Frequently" "Travel\_Rarely" "Travel\_Rarely" "Non-Travel" ...  
## $ DailyRate : int 1422 461 267 1351 1302 895 750 251 1312 1405 ...  
## $ Department : chr "Sales" "Research & Development" "Sales" "Research & Development" ...  
## $ DistanceFromHome : int 2 13 13 9 19 5 22 10 23 11 ...  
## $ Education : Factor w/ 5 levels "Bachelor","Below College",..: 5 2 5 5 1 1 3 3 1 3 ...  
## $ EducationField : chr "Life Sciences" "Life Sciences" "Marketing" "Life Sciences" ...  
## $ EnvironmentSatisfaction : Factor w/ 4 levels "High","Low","Medium",..: 2 3 2 2 4 4 1 2 2 4 ...  
## $ Gender : chr "Male" "Female" "Male" "Male" ...  
## $ JobInvolvement : Factor w/ 4 levels "High","Low","Medium",..: 1 1 4 4 3 1 1 3 2 3 ...  
## $ JobLevel : Factor w/ 5 levels "1","2","3","4",..: 2 1 4 1 1 2 2 1 1 4 ...  
## $ JobRole : Factor w/ 9 levels "Healthcare Representative",..: 8 7 8 3 3 9 8 3 7 5 ...  
## $ JobSatisfaction : Factor w/ 4 levels "High","Low","Medium",..: 4 4 1 3 2 4 3 1 1 3 ...  
## $ MaritalStatus : chr "Married" "Single" "Single" "Married" ...  
## $ NumCompaniesWorked : int 1 3 6 1 1 3 0 1 1 3 ...  
## $ PerformanceRating : Factor w/ 2 levels "Excellent","Outstanding": 2 1 1 2 1 1 1 2 2 1 ...  
## $ RelationshipSatisfaction: Factor w/ 4 levels "High","Low","Medium",..: 1 2 1 3 2 1 2 1 1 3 ...  
## $ TotalWorkingYears : int 7 5 24 5 7 19 8 2 10 29 ...  
## $ TrainingTimesLastYear : int 5 4 2 3 3 6 2 2 2 1 ...  
## $ WorkLifeBalance : Factor w/ 4 levels "Bad","Best","Better",..: 3 3 4 3 3 2 2 1 4 4 ...  
## $ YearsAtCompany : int 7 3 19 5 7 1 7 2 10 5 ...  
## $ YearsInCurrentRole : int 7 2 7 4 7 0 7 2 7 2 ...  
## $ YearsSinceLastPromotion : int 7 0 3 0 0 0 0 2 0 0 ...  
## $ YearsWithCurrManager : int 7 2 8 2 7 0 7 2 9 3 ...

#one hot encoding for all except continuous varibales and the unique ID column.  
dmy <- dummyVars(~., data = no.salary\_recode[-1])  
trsf.new1 <- data.frame(predict(dmy, newdata = no.salary\_recode[-1]))  
  
#many of the predictors are skewed so we transform the data set to minimize the skewness  
trsf.new1 <- trsf.new1 %>%  
 mutate(Age = log(Age + 1)  
 ,DailyRate = log(DailyRate + 1)  
 ,DistanceFromHome = log(DistanceFromHome + 1)  
 ,NumCompaniesWorked = log(NumCompaniesWorked + 1)  
 ,TotalWorkingYears = log(TotalWorkingYears + 1)  
 ,TrainingTimesLastYear = log(TrainingTimesLastYear + 1)  
 ,YearsAtCompany = log(YearsAtCompany +1)  
 ,YearsInCurrentRole = log(YearsInCurrentRole + 1)  
 ,YearsSinceLastPromotion = log(YearsSinceLastPromotion + 1)  
 ,YearsWithCurrManager = log(YearsWithCurrManager + 1)  
 )  
#standardize dataset  
prep\_num = preProcess(trsf.new1, method=c("center", "scale"))  
final\_dataset.new1 = predict(prep\_num, trsf.new1)  
final\_dataset.new1 = cbind(trsf.new, no.salary\_recode[-1]) #Unknown is ready for prediction but we need to tag the observations with the unique ID

###Splitting for the new Salary modelling

#Splitting the dataset  
set.seed(200)  
  
Train <- createDataPartition(final\_dataset$Attrition, p=0.7, list=FALSE)  
training.new <- final\_dataset.new1[ Train, ]  
testing.new <- final\_dataset.new1[ -Train, ]

#Training the data on salary.  
describe(training.new)

## vars n mean sd median trimmed mad min max range skew  
## 1 1 213 3.62 0.26 3.61 3.63 0.26 3.04 4.08 1.03 -0.12  
## 2 2 213 0.11 0.31 0.00 0.01 0.00 0.00 1.00 1.00 2.51  
## 3 3 213 0.20 0.40 0.00 0.12 0.00 0.00 1.00 1.00 1.51  
## 4 4 213 0.69 0.46 1.00 0.74 0.00 0.00 1.00 1.00 -0.84  
## 5 5 213 6.49 0.68 6.69 6.58 0.59 4.63 7.29 2.66 -0.96  
## 6 6 213 0.02 0.15 0.00 0.00 0.00 0.00 1.00 1.00 6.25  
## 7 7 213 0.68 0.47 1.00 0.73 0.00 0.00 1.00 1.00 -0.77  
## 8 8 213 0.30 0.46 0.00 0.25 0.00 0.00 1.00 1.00 0.89  
## 9 9 213 1.95 0.89 2.08 1.94 1.12 0.69 3.40 2.71 0.03  
## 10 10 213 0.45 0.50 0.00 0.44 0.00 0.00 1.00 1.00 0.20  
## 11 11 213 0.11 0.31 0.00 0.01 0.00 0.00 1.00 1.00 2.51  
## 12 12 213 0.16 0.37 0.00 0.08 0.00 0.00 1.00 1.00 1.80  
## 13 13 213 0.03 0.18 0.00 0.00 0.00 0.00 1.00 1.00 5.20  
## 14 14 213 0.24 0.43 0.00 0.18 0.00 0.00 1.00 1.00 1.18  
## 15 15 213 0.01 0.10 0.00 0.00 0.00 0.00 1.00 1.00 10.10  
## 16 16 213 0.44 0.50 0.00 0.43 0.00 0.00 1.00 1.00 0.23  
## 17 17 213 0.10 0.31 0.00 0.01 0.00 0.00 1.00 1.00 2.59  
## 18 18 213 0.31 0.46 0.00 0.26 0.00 0.00 1.00 1.00 0.82  
## 19 19 213 0.03 0.17 0.00 0.00 0.00 0.00 1.00 1.00 5.66  
## 20 20 213 0.11 0.31 0.00 0.01 0.00 0.00 1.00 1.00 2.51  
## 21 21 213 0.34 0.48 0.00 0.30 0.00 0.00 1.00 1.00 0.66  
## 22 22 213 0.19 0.40 0.00 0.12 0.00 0.00 1.00 1.00 1.55  
## 23 23 213 0.19 0.40 0.00 0.12 0.00 0.00 1.00 1.00 1.55  
## 24 24 213 0.27 0.45 0.00 0.22 0.00 0.00 1.00 1.00 1.02  
## 25 25 213 0.32 0.47 0.00 0.28 0.00 0.00 1.00 1.00 0.75  
## 26 26 213 0.68 0.47 1.00 0.72 0.00 0.00 1.00 1.00 -0.75  
## 27 27 213 4.18 0.31 4.23 4.21 0.33 3.43 4.61 1.17 -0.63  
## 28 28 213 0.58 0.49 1.00 0.60 0.00 0.00 1.00 1.00 -0.33  
## 29 29 213 0.06 0.24 0.00 0.00 0.00 0.00 1.00 1.00 3.64  
## 30 30 213 0.26 0.44 0.00 0.20 0.00 0.00 1.00 1.00 1.07  
## 31 31 213 0.09 0.29 0.00 0.00 0.00 0.00 1.00 1.00 2.76  
## 32 32 213 0.37 0.48 0.00 0.33 0.00 0.00 1.00 1.00 0.55  
## 33 33 213 0.37 0.48 0.00 0.33 0.00 0.00 1.00 1.00 0.55  
## 34 34 213 0.10 0.30 0.00 0.00 0.00 0.00 1.00 1.00 2.67  
## 35 35 213 0.09 0.29 0.00 0.00 0.00 0.00 1.00 1.00 2.76  
## 36 36 213 0.08 0.26 0.00 0.00 0.00 0.00 1.00 1.00 3.20  
## 37 37 213 0.08 0.26 0.00 0.00 0.00 0.00 1.00 1.00 3.20  
## 38 38 213 0.01 0.12 0.00 0.00 0.00 0.00 1.00 1.00 8.19  
## 39 39 213 0.19 0.40 0.00 0.12 0.00 0.00 1.00 1.00 1.55  
## 40 40 213 0.11 0.32 0.00 0.02 0.00 0.00 1.00 1.00 2.43  
## 41 41 213 0.08 0.28 0.00 0.00 0.00 0.00 1.00 1.00 2.97  
## 42 42 213 0.05 0.22 0.00 0.00 0.00 0.00 1.00 1.00 4.02  
## 43 43 213 0.22 0.42 0.00 0.15 0.00 0.00 1.00 1.00 1.34  
## 44 44 213 0.21 0.41 0.00 0.13 0.00 0.00 1.00 1.00 1.44  
## 45 45 213 0.04 0.20 0.00 0.00 0.00 0.00 1.00 1.00 4.52  
## 46 46 213 0.31 0.46 0.00 0.26 0.00 0.00 1.00 1.00 0.82  
## 47 47 213 0.17 0.38 0.00 0.09 0.00 0.00 1.00 1.00 1.71  
## 48 48 213 0.22 0.41 0.00 0.15 0.00 0.00 1.00 1.00 1.37  
## 49 49 213 0.30 0.46 0.00 0.25 0.00 0.00 1.00 1.00 0.86  
## 50 50 213 0.22 0.41 0.00 0.15 0.00 0.00 1.00 1.00 1.37  
## 51 51 213 0.42 0.50 0.00 0.40 0.00 0.00 1.00 1.00 0.31  
## 52 52 213 0.36 0.48 0.00 0.33 0.00 0.00 1.00 1.00 0.57  
## 53 53 213 8.60 0.70 8.55 8.57 0.82 7.12 9.90 2.78 0.30  
## 54 54 213 9.38 0.67 9.61 9.46 0.59 7.65 10.19 2.54 -0.86  
## 55 55 213 1.04 0.63 1.10 1.03 0.60 0.00 2.30 2.30 0.10  
## 56 56 213 0.71 0.45 1.00 0.77 0.00 0.00 1.00 1.00 -0.94  
## 57 57 213 0.29 0.45 0.00 0.23 0.00 0.00 1.00 1.00 0.94  
## 58 58 213 2.76 0.22 2.71 2.74 0.21 2.48 3.26 0.77 0.46  
## 59 59 213 0.85 0.36 1.00 0.94 0.00 0.00 1.00 1.00 -1.94  
## 60 60 213 0.15 0.36 0.00 0.06 0.00 0.00 1.00 1.00 1.94  
## 61 61 213 0.36 0.48 0.00 0.32 0.00 0.00 1.00 1.00 0.59  
## 62 62 213 0.16 0.37 0.00 0.08 0.00 0.00 1.00 1.00 1.80  
## 63 63 213 0.18 0.38 0.00 0.10 0.00 0.00 1.00 1.00 1.67  
## 64 64 213 0.30 0.46 0.00 0.25 0.00 0.00 1.00 1.00 0.86  
## 65 65 213 2.35 0.75 2.40 2.40 0.73 0.00 3.66 3.66 -0.61  
## 66 66 213 1.24 0.39 1.10 1.26 0.43 0.00 1.95 1.95 -1.00  
## 67 67 213 0.05 0.21 0.00 0.00 0.00 0.00 1.00 1.00 4.25  
## 68 68 213 0.09 0.29 0.00 0.00 0.00 0.00 1.00 1.00 2.76  
## 69 69 213 0.62 0.49 1.00 0.64 0.00 0.00 1.00 1.00 -0.47  
## 70 70 213 0.24 0.43 0.00 0.18 0.00 0.00 1.00 1.00 1.18  
## 71 71 213 1.85 0.78 1.79 1.84 0.76 0.00 3.64 3.64 -0.03  
## 72 72 213 1.40 0.78 1.39 1.44 1.03 0.00 2.94 2.94 -0.43  
## 73 73 213 0.73 0.79 0.69 0.62 1.03 0.00 2.77 2.77 0.85  
## 74 74 213 1.36 0.83 1.39 1.38 1.03 0.00 2.83 2.83 -0.32  
## 75 75 213 0.24 0.43 0.00 0.18 0.00 0.00 1.00 1.00 1.18  
## 76 76 213 0.37 0.48 0.00 0.33 0.00 0.00 1.00 1.00 0.55  
## 77 77 213 0.18 0.38 0.00 0.10 0.00 0.00 1.00 1.00 1.67  
## 78 78 213 0.21 0.41 0.00 0.14 0.00 0.00 1.00 1.00 1.40  
## 79 79 213 36.36 8.94 36.00 35.80 8.90 18.00 60.00 42.00 0.48  
## 80 80 213 1.14 0.35 1.00 1.05 0.00 1.00 2.00 1.00 2.05  
## 81 81 213 2.62 0.64 3.00 2.76 0.00 1.00 3.00 2.00 -1.45  
## 82 82 213 773.17 407.73 693.00 772.57 532.25 105.00 1490.00 1385.00 0.04  
## 83 83 213 2.23 0.55 2.00 2.23 0.00 1.00 3.00 2.00 0.08  
## 84 84 213 8.47 7.34 7.00 7.39 7.41 1.00 29.00 28.00 1.08  
## 85 85 213 2.48 1.59 2.00 2.36 1.48 1.00 5.00 4.00 0.57  
## 86 86 213 3.33 1.36 3.00 3.19 1.48 1.00 6.00 5.00 0.46  
## 87 87 213 2.48 1.24 2.00 2.48 1.48 1.00 4.00 3.00 0.02  
## 88 88 213 1.61 0.49 2.00 1.63 0.00 1.00 2.00 1.00 -0.43  
## 89 89 213 1.85 1.11 1.00 1.69 0.00 1.00 4.00 3.00 0.72  
## 90 90 213 2.02 1.00 2.00 1.88 1.48 1.00 5.00 4.00 0.99  
## 91 91 213 5.29 2.54 6.00 5.44 2.97 1.00 9.00 8.00 -0.30  
## 92 92 213 2.50 1.23 3.00 2.50 1.48 1.00 4.00 3.00 -0.01  
## 93 93 213 2.04 0.75 2.00 2.05 1.48 1.00 3.00 2.00 -0.07  
## 94 94 213 2.67 2.61 2.00 2.31 1.48 0.00 9.00 9.00 1.05  
## 95 95 213 1.15 0.36 1.00 1.06 0.00 1.00 2.00 1.00 1.94  
## 96 96 213 2.56 1.15 3.00 2.57 1.48 1.00 4.00 3.00 -0.14  
## 97 97 213 10.91 7.34 9.00 10.03 4.45 0.00 36.00 36.00 1.13  
## 98 98 213 2.80 1.38 3.00 2.74 1.48 0.00 6.00 6.00 0.45  
## 99 99 213 3.08 0.74 3.00 3.16 0.00 1.00 4.00 3.00 -0.97  
## 100 100 213 6.80 5.34 6.00 5.95 4.45 0.00 33.00 33.00 1.75  
## 101 101 213 4.32 3.49 3.00 3.99 2.97 0.00 16.00 16.00 0.84  
## 102 102 213 2.15 3.02 1.00 1.52 1.48 0.00 15.00 15.00 1.86  
## 103 103 213 3.89 3.08 3.00 3.65 2.97 0.00 13.00 13.00 0.66  
## kurtosis se  
## 1 -0.77 0.02  
## 2 4.31 0.02  
## 3 0.29 0.03  
## 4 -1.30 0.03  
## 5 0.00 0.05  
## 6 37.24 0.01  
## 7 -1.41 0.03  
## 8 -1.22 0.03  
## 9 -1.26 0.06  
## 10 -1.97 0.03  
## 11 4.31 0.02  
## 12 1.24 0.03  
## 13 25.20 0.01  
## 14 -0.60 0.03  
## 15 100.53 0.01  
## 16 -1.95 0.03  
## 17 4.72 0.02  
## 18 -1.34 0.03  
## 19 30.21 0.01  
## 20 4.31 0.02  
## 21 -1.57 0.03  
## 22 0.40 0.03  
## 23 0.40 0.03  
## 24 -0.97 0.03  
## 25 -1.45 0.03  
## 26 -1.45 0.03  
## 27 -0.53 0.02  
## 28 -1.90 0.03  
## 29 11.31 0.02  
## 30 -0.86 0.03  
## 31 5.67 0.02  
## 32 -1.70 0.03  
## 33 -1.70 0.03  
## 34 5.17 0.02  
## 35 5.67 0.02  
## 36 8.29 0.02  
## 37 8.29 0.02  
## 38 65.37 0.01  
## 39 0.40 0.03  
## 40 3.94 0.02  
## 41 6.83 0.02  
## 42 14.25 0.02  
## 43 -0.21 0.03  
## 44 0.07 0.03  
## 45 18.51 0.01  
## 46 -1.34 0.03  
## 47 0.93 0.03  
## 48 -0.12 0.03  
## 49 -1.26 0.03  
## 50 -0.12 0.03  
## 51 -1.91 0.03  
## 52 -1.68 0.03  
## 53 -0.87 0.05  
## 54 -0.39 0.05  
## 55 -0.76 0.04  
## 56 -1.12 0.03  
## 57 -1.12 0.03  
## 58 -0.90 0.01  
## 59 1.79 0.02  
## 60 1.79 0.02  
## 61 -1.66 0.03  
## 62 1.24 0.03  
## 63 0.79 0.03  
## 64 -1.26 0.03  
## 65 0.26 0.05  
## 66 2.56 0.03  
## 67 16.17 0.01  
## 68 5.67 0.02  
## 69 -1.79 0.03  
## 70 -0.60 0.03  
## 71 -0.29 0.05  
## 72 -0.72 0.05  
## 73 -0.30 0.05  
## 74 -0.94 0.06  
## 75 -0.60 0.03  
## 76 -1.70 0.03  
## 77 0.79 0.03  
## 78 -0.03 0.03  
## 79 -0.32 0.61  
## 80 2.21 0.02  
## 81 0.80 0.04  
## 82 -1.26 27.94  
## 83 -0.27 0.04  
## 84 0.25 0.50  
## 85 -1.24 0.11  
## 86 -0.87 0.09  
## 87 -1.62 0.09  
## 88 -1.82 0.03  
## 89 -1.14 0.08  
## 90 0.63 0.07  
## 91 -1.35 0.17  
## 92 -1.60 0.08  
## 93 -1.25 0.05  
## 94 0.00 0.18  
## 95 1.79 0.02  
## 96 -1.42 0.08  
## 97 0.80 0.50  
## 98 0.27 0.09  
## 99 1.49 0.05  
## 100 4.00 0.37  
## 101 0.20 0.24  
## 102 3.21 0.21  
## 103 -0.52 0.21

income\_mod1 = glm(MonthlyIncome~., data = training[,c(-4, -11,-18,-21,-20,-31)])  
  
model\_summary <- summary(income\_mod1)  
  
step1<- stepAIC(income\_mod1,direction = "backward",trace=FALSE)  
summary(step1)

##   
## Call:  
## glm(formula = MonthlyIncome ~ BusinessTravelNon.Travel + DailyRate +   
## Education.Master + JobLevel.1 + JobLevel.3 + JobLevel.4 +   
## JobLevel.5 + JobRole.Healthcare.Representative + JobRole.Manager +   
## JobRole.Manufacturing.Director + JobRole.Research.Director +   
## JobRole.Research.Scientist + JobRole.Sales.Executive + TotalWorkingYears +   
## WorkLifeBalance.Better + MonthlyIncomeFact.Low + MonthlyIncomeFact.Average +   
## MonthlyIncomeFact.High, data = training[, c(-4, -11, -18,   
## -21, -20, -31)])  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -0.66014 -0.07899 0.00647 0.07461 0.38742   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 8.564028 0.080205 106.777 < 2e-16 \*\*\*  
## BusinessTravelNon.Travel -0.048380 0.017546 -2.757 0.006009 \*\*   
## DailyRate 0.016137 0.008811 1.831 0.067531 .   
## Education.Master 0.031318 0.012760 2.454 0.014401 \*   
## JobLevel.1 -0.168720 0.025390 -6.645 6.89e-11 \*\*\*  
## JobLevel.3 0.207525 0.026430 7.852 1.94e-14 \*\*\*  
## JobLevel.4 0.412475 0.041265 9.996 < 2e-16 \*\*\*  
## JobLevel.5 0.555799 0.048692 11.415 < 2e-16 \*\*\*  
## JobRole.Healthcare.Representative 0.076545 0.028337 2.701 0.007107 \*\*   
## JobRole.Manager 0.327759 0.044681 7.336 7.33e-13 \*\*\*  
## JobRole.Manufacturing.Director 0.099034 0.027470 3.605 0.000338 \*\*\*  
## JobRole.Research.Director 0.330610 0.040272 8.209 1.40e-15 \*\*\*  
## JobRole.Research.Scientist 0.038828 0.016377 2.371 0.018067 \*   
## JobRole.Sales.Executive 0.103285 0.024459 4.223 2.79e-05 \*\*\*  
## TotalWorkingYears 0.088240 0.012175 7.247 1.34e-12 \*\*\*  
## WorkLifeBalance.Better 0.016873 0.011425 1.477 0.140224   
## MonthlyIncomeFact.Low -0.924202 0.043158 -21.414 < 2e-16 \*\*\*  
## MonthlyIncomeFact.Average -0.484159 0.037672 -12.852 < 2e-16 \*\*\*  
## MonthlyIncomeFact.High -0.165334 0.030936 -5.344 1.30e-07 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for gaussian family taken to be 0.01830624)  
##   
## Null deviance: 262.322 on 608 degrees of freedom  
## Residual deviance: 10.801 on 590 degrees of freedom  
## AIC: -687.35  
##   
## Number of Fisher Scoring iterations: 2

step1$coefficients

## (Intercept) BusinessTravelNon.Travel   
## 8.56402751 -0.04838006   
## DailyRate Education.Master   
## 0.01613731 0.03131770   
## JobLevel.1 JobLevel.3   
## -0.16871983 0.20752462   
## JobLevel.4 JobLevel.5   
## 0.41247534 0.55579886   
## JobRole.Healthcare.Representative JobRole.Manager   
## 0.07654460 0.32775939   
## JobRole.Manufacturing.Director JobRole.Research.Director   
## 0.09903408 0.33060977   
## JobRole.Research.Scientist JobRole.Sales.Executive   
## 0.03882794 0.10328517   
## TotalWorkingYears WorkLifeBalance.Better   
## 0.08824014 0.01687346   
## MonthlyIncomeFact.Low MonthlyIncomeFact.Average   
## -0.92420242 -0.48415914   
## MonthlyIncomeFact.High   
## -0.16533395

step1$anova

## Stepwise Model Path   
## Analysis of Deviance Table  
##   
## Initial Model:  
## MonthlyIncome ~ Age + BusinessTravelNon.Travel + BusinessTravelTravel\_Frequently +   
## DailyRate + DepartmentHuman.Resources + DepartmentResearch...Development +   
## DistanceFromHome + Education.Bachelor + Education.Below.College +   
## Education.Doctor + Education.Master + EducationFieldHuman.Resources +   
## EducationFieldLife.Sciences + EducationFieldMarketing + EducationFieldMedical +   
## EducationFieldTechnical.Degree + EnvironmentSatisfaction.Medium +   
## EnvironmentSatisfaction.Very.High + GenderFemale + HourlyRate +   
## JobInvolvement.High + JobInvolvement.Low + JobInvolvement.Medium +   
## JobInvolvement.Very.High + JobLevel.1 + JobLevel.3 + JobLevel.4 +   
## JobLevel.5 + JobRole.Healthcare.Representative + JobRole.Laboratory.Technician +   
## JobRole.Manager + JobRole.Manufacturing.Director + JobRole.Research.Director +   
## JobRole.Research.Scientist + JobRole.Sales.Executive + JobRole.Sales.Representative +   
## JobSatisfaction.High + JobSatisfaction.Low + JobSatisfaction.Medium +   
## JobSatisfaction.Very.High + MaritalStatusDivorced + MaritalStatusMarried +   
## MaritalStatusSingle + MonthlyRate + NumCompaniesWorked +   
## OverTimeNo + PercentSalaryHike + PerformanceRating.Excellent +   
## RelationshipSatisfaction.High + RelationshipSatisfaction.Low +   
## RelationshipSatisfaction.Medium + RelationshipSatisfaction.Very.High +   
## TotalWorkingYears + TrainingTimesLastYear + WorkLifeBalance.Bad +   
## WorkLifeBalance.Best + WorkLifeBalance.Better + WorkLifeBalance.Good +   
## YearsAtCompany + YearsInCurrentRole + YearsSinceLastPromotion +   
## YearsWithCurrManager + MonthlyIncomeFact.Low + MonthlyIncomeFact.Average +   
## MonthlyIncomeFact.High + MonthlyIncomeFact.Very.High + Attrition +   
## ID  
##   
## Final Model:  
## MonthlyIncome ~ BusinessTravelNon.Travel + DailyRate + Education.Master +   
## JobLevel.1 + JobLevel.3 + JobLevel.4 + JobLevel.5 + JobRole.Healthcare.Representative +   
## JobRole.Manager + JobRole.Manufacturing.Director + JobRole.Research.Director +   
## JobRole.Research.Scientist + JobRole.Sales.Executive + TotalWorkingYears +   
## WorkLifeBalance.Better + MonthlyIncomeFact.Low + MonthlyIncomeFact.Average +   
## MonthlyIncomeFact.High  
##   
##   
## Step Df Deviance Resid. Df Resid. Dev  
## 1 546 10.38978  
## 2 - MonthlyIncomeFact.Very.High 0 0.000000e+00 546 10.38978  
## 3 - WorkLifeBalance.Good 0 0.000000e+00 546 10.38978  
## 4 - RelationshipSatisfaction.Very.High 0 0.000000e+00 546 10.38978  
## 5 - MaritalStatusSingle 0 0.000000e+00 546 10.38978  
## 6 - JobSatisfaction.Very.High 0 0.000000e+00 546 10.38978  
## 7 - JobInvolvement.Very.High 0 0.000000e+00 546 10.38978  
## 8 - JobSatisfaction.Low 1 3.186725e-06 547 10.38979  
## 9 - EducationFieldTechnical.Degree 1 4.815241e-06 548 10.38979  
## 10 - EnvironmentSatisfaction.Very.High 1 2.732764e-05 549 10.38982  
## 11 - JobRole.Laboratory.Technician 1 5.336607e-05 550 10.38987  
## 12 - TrainingTimesLastYear 1 7.214548e-05 551 10.38995  
## 13 - YearsWithCurrManager 1 1.117445e-04 552 10.39006  
## 14 - WorkLifeBalance.Bad 1 1.995956e-04 553 10.39026  
## 15 - YearsAtCompany 1 4.983113e-04 554 10.39075  
## 16 - JobInvolvement.Medium 1 7.219306e-04 555 10.39148  
## 17 - EducationFieldHuman.Resources 1 7.671272e-04 556 10.39224  
## 18 - WorkLifeBalance.Best 1 8.448582e-04 557 10.39309  
## 19 - JobInvolvement.Low 1 8.893834e-04 558 10.39398  
## 20 - ID 1 2.010484e-03 559 10.39599  
## 21 - DistanceFromHome 1 2.029542e-03 560 10.39802  
## 22 - NumCompaniesWorked 1 2.252648e-03 561 10.40027  
## 23 - JobInvolvement.High 1 3.132678e-03 562 10.40340  
## 24 - HourlyRate 1 3.667057e-03 563 10.40707  
## 25 - Attrition 1 3.523456e-03 564 10.41059  
## 26 - JobRole.Sales.Representative 1 3.602943e-03 565 10.41420  
## 27 - PercentSalaryHike 1 3.474824e-03 566 10.41767  
## 28 - RelationshipSatisfaction.High 1 3.942465e-03 567 10.42161  
## 29 - Education.Doctor 1 3.962831e-03 568 10.42558  
## 30 - DepartmentHuman.Resources 1 3.871713e-03 569 10.42945  
## 31 - DepartmentResearch...Development 1 2.097058e-03 570 10.43155  
## 32 - MonthlyRate 1 4.067479e-03 571 10.43561  
## 33 - EducationFieldLife.Sciences 1 4.694283e-03 572 10.44031  
## 34 - YearsSinceLastPromotion 1 4.614977e-03 573 10.44492  
## 35 - EducationFieldMedical 1 6.534616e-03 574 10.45146  
## 36 - Education.Below.College 1 1.502325e-02 575 10.46648  
## 37 - GenderFemale 1 1.476318e-02 576 10.48124  
## 38 - MaritalStatusMarried 1 1.529805e-02 577 10.49654  
## 39 - OverTimeNo 1 1.505295e-02 578 10.51159  
## 40 - MaritalStatusDivorced 1 1.731365e-02 579 10.52891  
## 41 - YearsInCurrentRole 1 1.811734e-02 580 10.54703  
## 42 - EducationFieldMarketing 1 2.340808e-02 581 10.57043  
## 43 - JobSatisfaction.Medium 1 2.019929e-02 582 10.59063  
## 44 - EnvironmentSatisfaction.Medium 1 2.387124e-02 583 10.61450  
## 45 - JobSatisfaction.High 1 2.068219e-02 584 10.63519  
## 46 - BusinessTravelTravel\_Frequently 1 2.341676e-02 585 10.65860  
## 47 - PerformanceRating.Excellent 1 2.587868e-02 586 10.68448  
## 48 - RelationshipSatisfaction.Low 1 2.714476e-02 587 10.71163  
## 49 - RelationshipSatisfaction.Medium 1 3.010493e-02 588 10.74173  
## 50 - Education.Bachelor 1 2.905192e-02 589 10.77078  
## 51 - Age 1 2.990040e-02 590 10.80068  
## AIC  
## 1 -622.9690  
## 2 -622.9690  
## 3 -622.9690  
## 4 -622.9690  
## 5 -622.9690  
## 6 -622.9690  
## 7 -622.9690  
## 8 -624.9688  
## 9 -626.9685  
## 10 -628.9669  
## 11 -630.9638  
## 12 -632.9595  
## 13 -634.9530  
## 14 -636.9413  
## 15 -638.9121  
## 16 -640.8698  
## 17 -642.8248  
## 18 -644.7753  
## 19 -646.7232  
## 20 -648.6054  
## 21 -650.4865  
## 22 -652.3546  
## 23 -654.1712  
## 24 -655.9566  
## 25 -657.7504  
## 26 -659.5397  
## 27 -661.3365  
## 28 -663.1061  
## 29 -664.8746  
## 30 -666.6484  
## 31 -668.5260  
## 32 -670.2886  
## 33 -672.0147  
## 34 -673.7456  
## 35 -675.3647  
## 36 -676.4899  
## 37 -677.6315  
## 38 -678.7433  
## 39 -679.8705  
## 40 -680.8683  
## 41 -681.8213  
## 42 -682.4711  
## 43 -683.3085  
## 44 -683.9374  
## 45 -684.7519  
## 46 -685.4125  
## 47 -685.9356  
## 48 -686.3904  
## 49 -686.6812  
## 50 -687.0363  
## 51 -687.3480

# #The final model selected based on the above.  
# training\_sal <- training %>% dplyr::select(-HourlyRate, -MonthlyRate, -Overtime, -PercentSlaryHike, -MonthlyIncomeFact)  
# Final.model.sal <- glm(MonthlyIncome ~ BusinessTravel + Education + Gender + JobInvolvement +   
# JobLevel + JobRole + TotalWorkingYears, data = training[,c(-4, -11,-18,-21,-20,-31)])  
  
step\_select.r1 <- with(summary(income\_mod1), 1 - deviance/null.deviance) #r =0.96, r^2 = 0.81  
  
  
  
  
#Adding the predicted salary to the test data and also back converting the log transformed salary response   
salary.pred\_df <- cbind(testing, pred\_Monthly\_income.log=pred\_lr.sal, pred\_monthlyincome = (exp(pred\_lr.sal)-1))

## Error in data.frame(..., check.names = FALSE): object 'pred\_lr.sal' not found

str(salary.pred\_df)

## Error in str(salary.pred\_df): object 'salary.pred\_df' not found

#Write the result to a csv file and move it to github  
write.csv(salary.pred\_df, 'C:\\Users\\olani\\OneDrive\\Documents\\Data Science\\SMU-Data Science\\Doing Data Science\\MSDS\_6306\_DDS\\Unit 14 and 15 Case Study 2\\predicted monthly income.csv')

## Error in is.data.frame(x): object 'salary.pred\_df' not found

#Residual plots and Cook's D plots to check for assumptions.  
  
par(mfrow = c(2, 1))  
p\_r1 <- plot(Final.model.sal$fitted.values,Final.model.sal$residuals, main = "Residual Plot for salary prediction")

## Error in h(simpleError(msg, call)): error in evaluating the argument 'x' in selecting a method for function 'plot': object 'Final.model.sal' not found

p\_c1 <- plot(cooks.distance(Final.model.sal), main = "Cooks' D for Salary prediction")

## Error in h(simpleError(msg, call)): error in evaluating the argument 'x' in selecting a method for function 'plot': object 'Final.model.sal' not found

par(mfrow = c(1, 1))  
  
#These are no random clouds of residuals around -2 to +2.  
#This shows that the linear model satisyfies major assumption and good to classify the response.  
#More so, the Cooks'D plot did not show any strongly high leverage point,so there are no extreme outlier in the plotted data.  
  
  
#calculating the RMSE  
library(qpcR)  
qpcR::RMSE(Final.model.sal)

## Error in residuals(object): object 'Final.model.sal' not found

#Convert to actual salary value  
exp(qpcR::RMSE(Final.model.sal))

## Error in residuals(object): object 'Final.model.sal' not found