R\_Project

Radhika Gedela

2023-12-10

# Load dataset and set working directory  
setwd("~/Downloads") # Set the working directory to where the CSV file is stored  
data <- read.csv("HR-Employee-Attrition.csv") # Read the CSV file into a data frame

# Remove unnecessary columns  
columns\_to\_remove <- c("DailyRate","EmployeeNumber", "HourlyRate", "MonthlyRate", "StandardHours", "StockOptionLevel", "EmployeeCount")   
  
data <- data[, !(names(data) %in% columns\_to\_remove)] # Drop the unnecessary columns

# View the first few rows of the dataset  
head(data)

## Age Attrition BusinessTravel Department DistanceFromHome  
## 1 41 Yes Travel\_Rarely Sales 1  
## 2 49 No Travel\_Frequently Research & Development 8  
## 3 37 Yes Travel\_Rarely Research & Development 2  
## 4 33 No Travel\_Frequently Research & Development 3  
## 5 27 No Travel\_Rarely Research & Development 2  
## 6 32 No Travel\_Frequently Research & Development 2  
## Education EducationField EnvironmentSatisfaction Gender JobInvolvement  
## 1 2 Life Sciences 2 Female 3  
## 2 1 Life Sciences 3 Male 2  
## 3 2 Other 4 Male 2  
## 4 4 Life Sciences 4 Female 3  
## 5 1 Medical 1 Male 3  
## 6 2 Life Sciences 4 Male 3  
## JobLevel JobRole JobSatisfaction MaritalStatus MonthlyIncome  
## 1 2 Sales Executive 4 Single 5993  
## 2 2 Research Scientist 2 Married 5130  
## 3 1 Laboratory Technician 3 Single 2090  
## 4 1 Research Scientist 3 Married 2909  
## 5 1 Laboratory Technician 2 Married 3468  
## 6 1 Laboratory Technician 4 Single 3068  
## NumCompaniesWorked Over18 OverTime PercentSalaryHike PerformanceRating  
## 1 8 Y Yes 11 3  
## 2 1 Y No 23 4  
## 3 6 Y Yes 15 3  
## 4 1 Y Yes 11 3  
## 5 9 Y No 12 3  
## 6 0 Y No 13 3  
## RelationshipSatisfaction TotalWorkingYears TrainingTimesLastYear  
## 1 1 8 0  
## 2 4 10 3  
## 3 2 7 3  
## 4 3 8 3  
## 5 4 6 3  
## 6 3 8 2  
## WorkLifeBalance YearsAtCompany YearsInCurrentRole YearsSinceLastPromotion  
## 1 1 6 4 0  
## 2 3 10 7 1  
## 3 3 0 0 0  
## 4 3 8 7 3  
## 5 3 2 2 2  
## 6 2 7 7 3  
## YearsWithCurrManager  
## 1 5  
## 2 7  
## 3 0  
## 4 0  
## 5 2  
## 6 6

# Get summary statistics for each column  
summary(data)

## Age Attrition BusinessTravel Department   
## Min. :18.00 Length:1470 Length:1470 Length:1470   
## 1st Qu.:30.00 Class :character Class :character Class :character   
## Median :36.00 Mode :character Mode :character Mode :character   
## Mean :36.92   
## 3rd Qu.:43.00   
## Max. :60.00   
## DistanceFromHome Education EducationField EnvironmentSatisfaction  
## Min. : 1.000 Min. :1.000 Length:1470 Min. :1.000   
## 1st Qu.: 2.000 1st Qu.:2.000 Class :character 1st Qu.:2.000   
## Median : 7.000 Median :3.000 Mode :character Median :3.000   
## Mean : 9.193 Mean :2.913 Mean :2.722   
## 3rd Qu.:14.000 3rd Qu.:4.000 3rd Qu.:4.000   
## Max. :29.000 Max. :5.000 Max. :4.000   
## Gender JobInvolvement JobLevel JobRole   
## Length:1470 Min. :1.00 Min. :1.000 Length:1470   
## Class :character 1st Qu.:2.00 1st Qu.:1.000 Class :character   
## Mode :character Median :3.00 Median :2.000 Mode :character   
## Mean :2.73 Mean :2.064   
## 3rd Qu.:3.00 3rd Qu.:3.000   
## Max. :4.00 Max. :5.000   
## JobSatisfaction MaritalStatus MonthlyIncome NumCompaniesWorked  
## Min. :1.000 Length:1470 Min. : 1009 Min. :0.000   
## 1st Qu.:2.000 Class :character 1st Qu.: 2911 1st Qu.:1.000   
## Median :3.000 Mode :character Median : 4919 Median :2.000   
## Mean :2.729 Mean : 6503 Mean :2.693   
## 3rd Qu.:4.000 3rd Qu.: 8379 3rd Qu.:4.000   
## Max. :4.000 Max. :19999 Max. :9.000   
## Over18 OverTime PercentSalaryHike PerformanceRating  
## Length:1470 Length:1470 Min. :11.00 Min. :3.000   
## Class :character Class :character 1st Qu.:12.00 1st Qu.:3.000   
## Mode :character Mode :character Median :14.00 Median :3.000   
## Mean :15.21 Mean :3.154   
## 3rd Qu.:18.00 3rd Qu.:3.000   
## Max. :25.00 Max. :4.000   
## RelationshipSatisfaction TotalWorkingYears TrainingTimesLastYear  
## Min. :1.000 Min. : 0.00 Min. :0.000   
## 1st Qu.:2.000 1st Qu.: 6.00 1st Qu.:2.000   
## Median :3.000 Median :10.00 Median :3.000   
## Mean :2.712 Mean :11.28 Mean :2.799   
## 3rd Qu.:4.000 3rd Qu.:15.00 3rd Qu.:3.000   
## Max. :4.000 Max. :40.00 Max. :6.000   
## WorkLifeBalance YearsAtCompany YearsInCurrentRole YearsSinceLastPromotion  
## Min. :1.000 Min. : 0.000 Min. : 0.000 Min. : 0.000   
## 1st Qu.:2.000 1st Qu.: 3.000 1st Qu.: 2.000 1st Qu.: 0.000   
## Median :3.000 Median : 5.000 Median : 3.000 Median : 1.000   
## Mean :2.761 Mean : 7.008 Mean : 4.229 Mean : 2.188   
## 3rd Qu.:3.000 3rd Qu.: 9.000 3rd Qu.: 7.000 3rd Qu.: 3.000   
## Max. :4.000 Max. :40.000 Max. :18.000 Max. :15.000   
## YearsWithCurrManager  
## Min. : 0.000   
## 1st Qu.: 2.000   
## Median : 3.000   
## Mean : 4.123   
## 3rd Qu.: 7.000   
## Max. :17.000

# Returns the number of rows and columns  
dim(data)

## [1] 1470 28

# Provides the structure of the dataset including data types  
str(data)

## 'data.frame': 1470 obs. of 28 variables:  
## $ Age : int 41 49 37 33 27 32 59 30 38 36 ...  
## $ Attrition : chr "Yes" "No" "Yes" "No" ...  
## $ BusinessTravel : chr "Travel\_Rarely" "Travel\_Frequently" "Travel\_Rarely" "Travel\_Frequently" ...  
## $ Department : chr "Sales" "Research & Development" "Research & Development" "Research & Development" ...  
## $ DistanceFromHome : int 1 8 2 3 2 2 3 24 23 27 ...  
## $ Education : int 2 1 2 4 1 2 3 1 3 3 ...  
## $ EducationField : chr "Life Sciences" "Life Sciences" "Other" "Life Sciences" ...  
## $ EnvironmentSatisfaction : int 2 3 4 4 1 4 3 4 4 3 ...  
## $ Gender : chr "Female" "Male" "Male" "Female" ...  
## $ JobInvolvement : int 3 2 2 3 3 3 4 3 2 3 ...  
## $ JobLevel : int 2 2 1 1 1 1 1 1 3 2 ...  
## $ JobRole : chr "Sales Executive" "Research Scientist" "Laboratory Technician" "Research Scientist" ...  
## $ JobSatisfaction : int 4 2 3 3 2 4 1 3 3 3 ...  
## $ MaritalStatus : chr "Single" "Married" "Single" "Married" ...  
## $ MonthlyIncome : int 5993 5130 2090 2909 3468 3068 2670 2693 9526 5237 ...  
## $ NumCompaniesWorked : int 8 1 6 1 9 0 4 1 0 6 ...  
## $ Over18 : chr "Y" "Y" "Y" "Y" ...  
## $ OverTime : chr "Yes" "No" "Yes" "Yes" ...  
## $ PercentSalaryHike : int 11 23 15 11 12 13 20 22 21 13 ...  
## $ PerformanceRating : int 3 4 3 3 3 3 4 4 4 3 ...  
## $ RelationshipSatisfaction: int 1 4 2 3 4 3 1 2 2 2 ...  
## $ TotalWorkingYears : int 8 10 7 8 6 8 12 1 10 17 ...  
## $ TrainingTimesLastYear : int 0 3 3 3 3 2 3 2 2 3 ...  
## $ WorkLifeBalance : int 1 3 3 3 3 2 2 3 3 2 ...  
## $ YearsAtCompany : int 6 10 0 8 2 7 1 1 9 7 ...  
## $ YearsInCurrentRole : int 4 7 0 7 2 7 0 0 7 7 ...  
## $ YearsSinceLastPromotion : int 0 1 0 3 2 3 0 0 1 7 ...  
## $ YearsWithCurrManager : int 5 7 0 0 2 6 0 0 8 7 ...

# Check for Null values in the entire data frame  
if (any(is.na(data))) {  
 print("There are NA values in the data frame.")  
} else {  
 print("There are no NA values in the data frame.")  
}

## [1] "There are no NA values in the data frame."

# Check for duplicate records  
any(duplicated(data))

## [1] FALSE

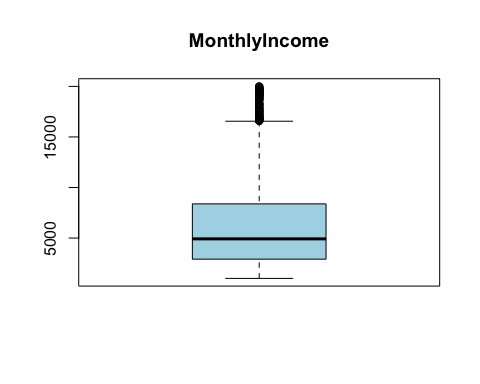
# Check for outliers in 'Age' column using IQR method  
variable <- data$Age  
Q1 <- quantile(variable, 0.25) # First quartile (25th percentile)  
Q3 <- quantile(variable, 0.75) # Third quartile (75th percentile)  
IQR <- Q3 - Q1  
  
# Identify potential outliers  
potential\_outliers <- variable < (Q1 - 1.5 \* IQR) | variable > (Q3 + 1.5 \* IQR)  
  
# Print the result  
print(any(potential\_outliers))

## [1] FALSE

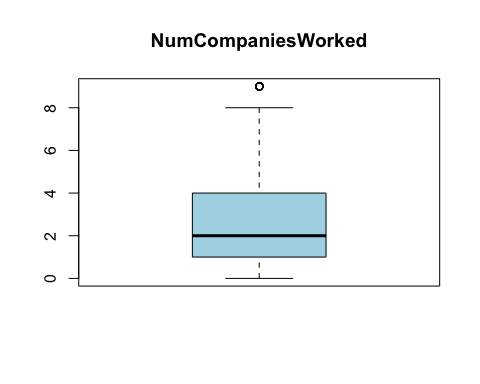
# Function to detect outliers using IQR method  
detect\_outliers <- function(variable) {  
 Q1 <- quantile(variable, 0.25)  
 Q3 <- quantile(variable, 0.75)  
 IQR <- Q3 - Q1  
   
 # Identify potential outliers  
 potential\_outliers <- variable < (Q1 - 1.5 \* IQR) | variable > (Q3 + 1.5 \* IQR)  
   
 return(potential\_outliers)  
}  
  
# Columns to check for outliers  
columns\_to\_check <- c(  
 "DistanceFromHome",  
 "MonthlyIncome",  
 "NumCompaniesWorked",  
 "PercentSalaryHike",  
 "TotalWorkingYears",  
 "YearsAtCompany",  
 "YearsInCurrentRole",  
 "YearsSinceLastPromotion",  
 "YearsWithCurrManager"  
)  
  
# Check for outliers in each column  
for (col in columns\_to\_check) {  
 variable <- data[[col]]  
 outliers <- detect\_outliers(variable)  
 print(paste("Outliers in", col, ":", any(outliers)))  
}

## [1] "Outliers in DistanceFromHome : FALSE"  
## [1] "Outliers in MonthlyIncome : TRUE"  
## [1] "Outliers in NumCompaniesWorked : TRUE"  
## [1] "Outliers in PercentSalaryHike : FALSE"  
## [1] "Outliers in TotalWorkingYears : TRUE"  
## [1] "Outliers in YearsAtCompany : TRUE"  
## [1] "Outliers in YearsInCurrentRole : TRUE"  
## [1] "Outliers in YearsSinceLastPromotion : TRUE"  
## [1] "Outliers in YearsWithCurrManager : TRUE"

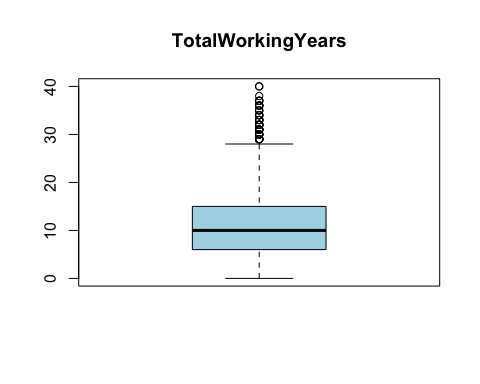
# Boxplots before outlier treatment  
boxplot(data$MonthlyIncome, main = "MonthlyIncome", col = "lightblue", border = "black", notch = FALSE)



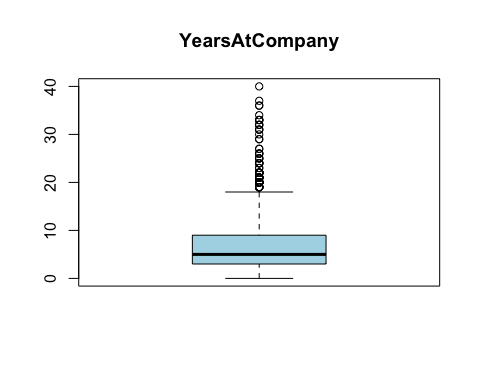
boxplot(data$NumCompaniesWorked, main = "NumCompaniesWorked", col = "lightblue", border = "black", notch = FALSE)



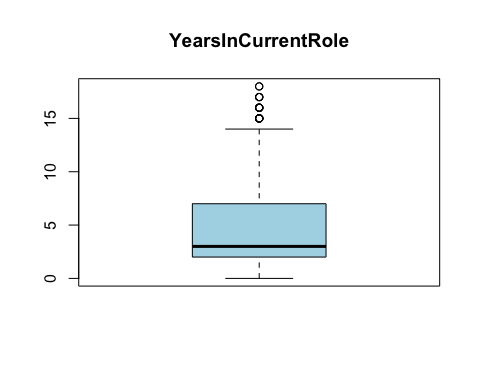
boxplot(data$TotalWorkingYears, main = "TotalWorkingYears", col = "lightblue", border = "black", notch = FALSE)



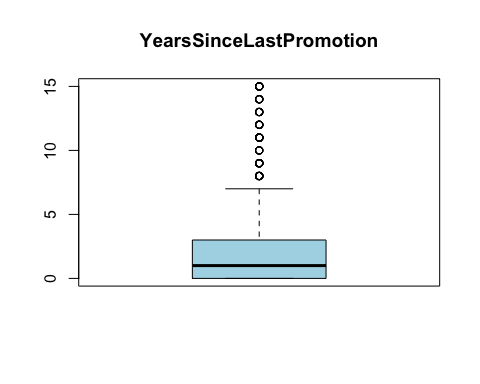
boxplot(data$YearsAtCompany, main = "YearsAtCompany", col = "lightblue", border = "black", notch = FALSE)



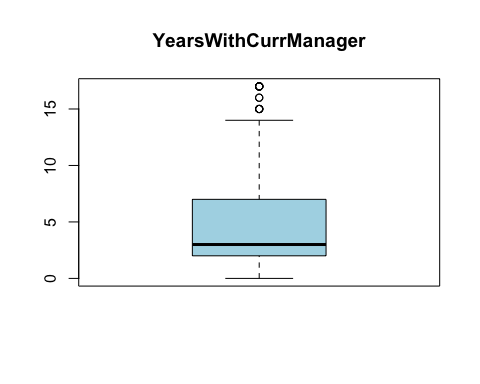
boxplot(data$YearsInCurrentRole, main = "YearsInCurrentRole", col = "lightblue", border = "black", notch = FALSE)



boxplot(data$YearsSinceLastPromotion, main = "YearsSinceLastPromotion", col = "lightblue", border = "black", notch = FALSE)

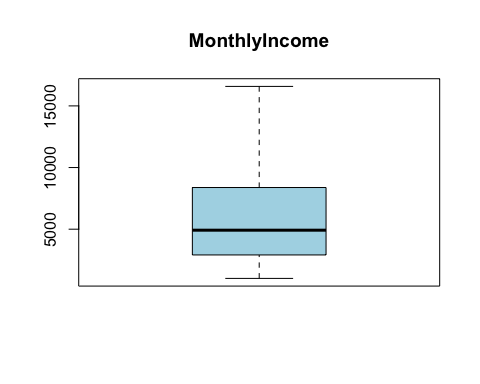


boxplot(data$YearsWithCurrManager, main = "YearsWithCurrManager", col = "lightblue", border = "black", notch = FALSE)

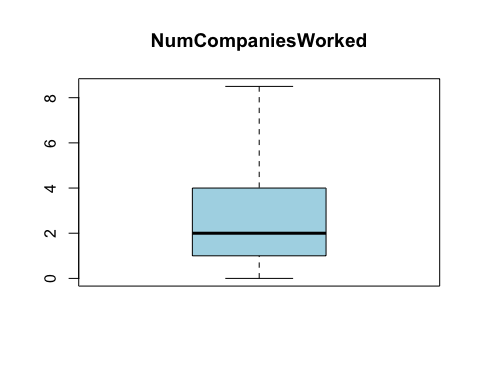


# Function to clip (cap) outliers based on IQR method  
clip\_outliers <- function(variable) {  
 Q1 <- quantile(variable, 0.25)  
 Q3 <- quantile(variable, 0.75)  
 IQR <- Q3 - Q1  
   
 # Set the clipping threshold  
 threshold <- 1.5  
   
 # Clip (cap) values beyond the threshold  
 variable[variable < (Q1 - threshold \* IQR)] <- (Q1 - threshold \* IQR)  
 variable[variable > (Q3 + threshold \* IQR)] <- (Q3 + threshold \* IQR)  
   
 return(variable)  
}  
  
# Columns to clip (cap) outliers  
columns\_to\_clip <- c(  
 "DistanceFromHome",  
 "MonthlyIncome",  
 "NumCompaniesWorked",  
 "PercentSalaryHike",  
 "TotalWorkingYears",  
 "YearsAtCompany",  
 "YearsInCurrentRole",  
 "YearsSinceLastPromotion",  
 "YearsWithCurrManager"  
)  
  
# Clip (cap) outliers in each column  
for (col in columns\_to\_clip) {  
 variable <- data[[col]]  
 data[[col]] <- clip\_outliers(variable)  
}

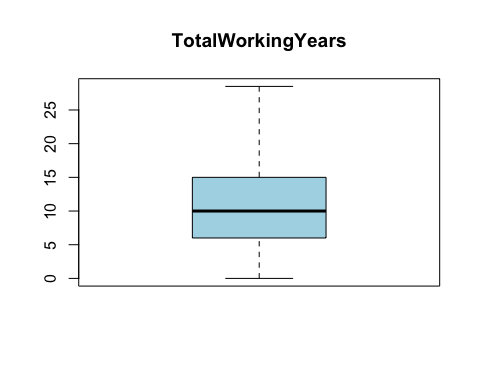
# Boxplots after outlier treatment  
boxplot(data$MonthlyIncome, main = "MonthlyIncome", col = "lightblue", border = "black", notch = FALSE)



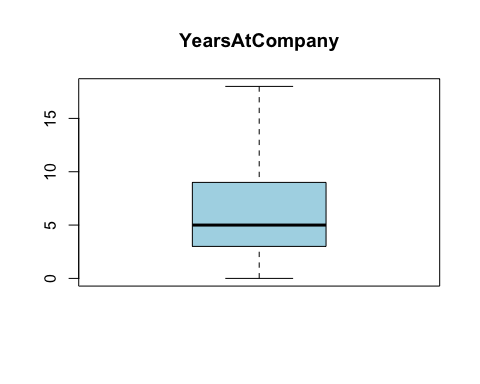
boxplot(data$NumCompaniesWorked, main = "NumCompaniesWorked", col = "lightblue", border = "black", notch = FALSE)



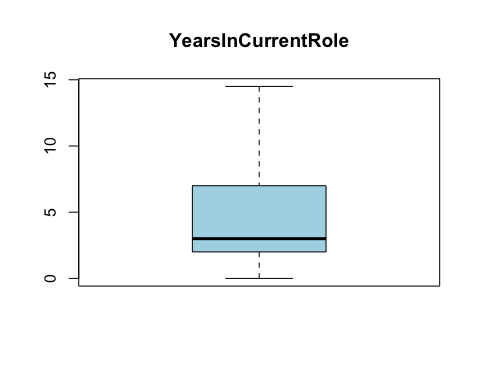
boxplot(data$TotalWorkingYears, main = "TotalWorkingYears", col = "lightblue", border = "black", notch = FALSE)



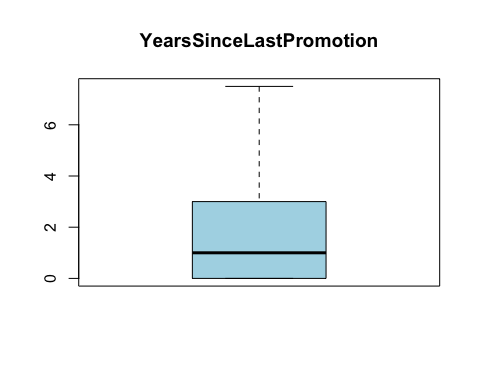
boxplot(data$YearsAtCompany, main = "YearsAtCompany", col = "lightblue", border = "black", notch = FALSE)



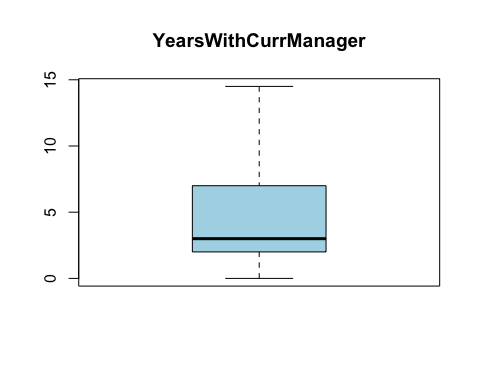
boxplot(data$YearsInCurrentRole, main = "YearsInCurrentRole", col = "lightblue", border = "black", notch = FALSE)



boxplot(data$YearsSinceLastPromotion, main = "YearsSinceLastPromotion", col = "lightblue", border = "black", notch = FALSE)



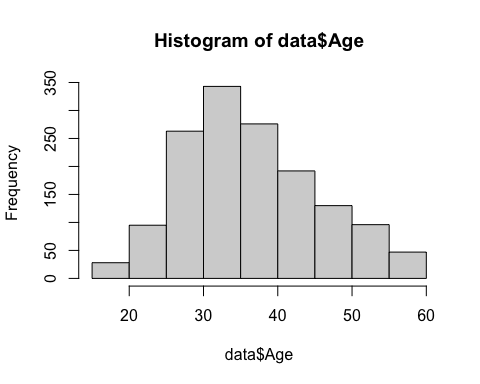
boxplot(data$YearsWithCurrManager, main = "YearsWithCurrManager", col = "lightblue", border = "black", notch = FALSE)



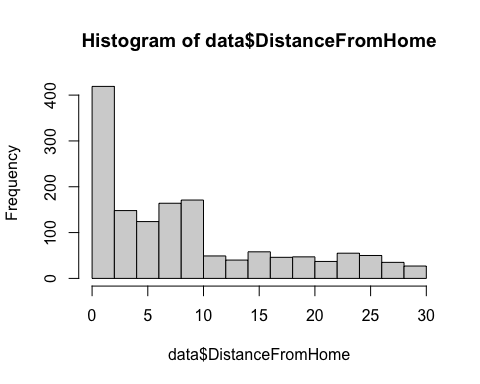
# Summary of dataset  
summary(data)

## Age Attrition BusinessTravel Department   
## Min. :18.00 Length:1470 Length:1470 Length:1470   
## 1st Qu.:30.00 Class :character Class :character Class :character   
## Median :36.00 Mode :character Mode :character Mode :character   
## Mean :36.92   
## 3rd Qu.:43.00   
## Max. :60.00   
## DistanceFromHome Education EducationField EnvironmentSatisfaction  
## Min. : 1.000 Min. :1.000 Length:1470 Min. :1.000   
## 1st Qu.: 2.000 1st Qu.:2.000 Class :character 1st Qu.:2.000   
## Median : 7.000 Median :3.000 Mode :character Median :3.000   
## Mean : 9.193 Mean :2.913 Mean :2.722   
## 3rd Qu.:14.000 3rd Qu.:4.000 3rd Qu.:4.000   
## Max. :29.000 Max. :5.000 Max. :4.000   
## Gender JobInvolvement JobLevel JobRole   
## Length:1470 Min. :1.00 Min. :1.000 Length:1470   
## Class :character 1st Qu.:2.00 1st Qu.:1.000 Class :character   
## Mode :character Median :3.00 Median :2.000 Mode :character   
## Mean :2.73 Mean :2.064   
## 3rd Qu.:3.00 3rd Qu.:3.000   
## Max. :4.00 Max. :5.000   
## JobSatisfaction MaritalStatus MonthlyIncome NumCompaniesWorked  
## Min. :1.000 Length:1470 Min. : 1009 Min. :0.000   
## 1st Qu.:2.000 Class :character 1st Qu.: 2911 1st Qu.:1.000   
## Median :3.000 Mode :character Median : 4919 Median :2.000   
## Mean :2.729 Mean : 6362 Mean :2.676   
## 3rd Qu.:4.000 3rd Qu.: 8379 3rd Qu.:4.000   
## Max. :4.000 Max. :16581 Max. :8.500   
## Over18 OverTime PercentSalaryHike PerformanceRating  
## Length:1470 Length:1470 Min. :11.00 Min. :3.000   
## Class :character Class :character 1st Qu.:12.00 1st Qu.:3.000   
## Mode :character Mode :character Median :14.00 Median :3.000   
## Mean :15.21 Mean :3.154   
## 3rd Qu.:18.00 3rd Qu.:3.000   
## Max. :25.00 Max. :4.000   
## RelationshipSatisfaction TotalWorkingYears TrainingTimesLastYear  
## Min. :1.000 Min. : 0.0 Min. :0.000   
## 1st Qu.:2.000 1st Qu.: 6.0 1st Qu.:2.000   
## Median :3.000 Median :10.0 Median :3.000   
## Mean :2.712 Mean :11.1 Mean :2.799   
## 3rd Qu.:4.000 3rd Qu.:15.0 3rd Qu.:3.000   
## Max. :4.000 Max. :28.5 Max. :6.000   
## WorkLifeBalance YearsAtCompany YearsInCurrentRole YearsSinceLastPromotion  
## Min. :1.000 Min. : 0.000 Min. : 0.000 Min. :0.000   
## 1st Qu.:2.000 1st Qu.: 3.000 1st Qu.: 2.000 1st Qu.:0.000   
## Median :3.000 Median : 5.000 Median : 3.000 Median :1.000   
## Mean :2.761 Mean : 6.618 Mean : 4.208 Mean :1.923   
## 3rd Qu.:3.000 3rd Qu.: 9.000 3rd Qu.: 7.000 3rd Qu.:3.000   
## Max. :4.000 Max. :18.000 Max. :14.500 Max. :7.500   
## YearsWithCurrManager  
## Min. : 0.000   
## 1st Qu.: 2.000   
## Median : 3.000   
## Mean : 4.107   
## 3rd Qu.: 7.000   
## Max. :14.500

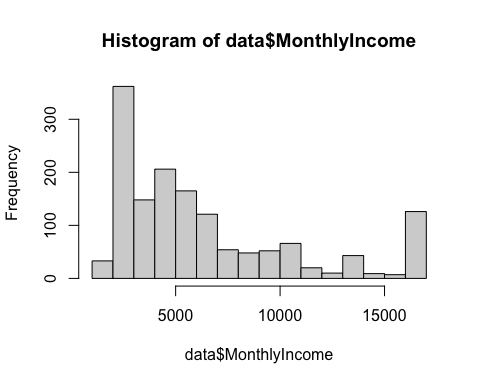
# Histograms for selected columns  
hist(data$Age)



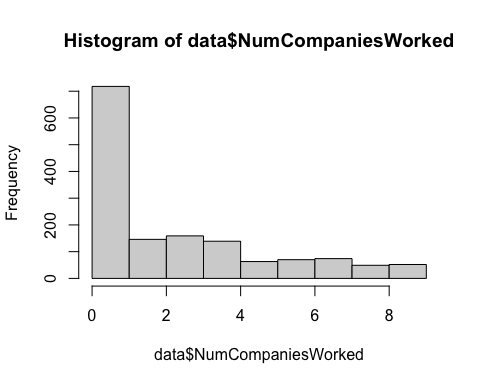
hist(data$DistanceFromHome)



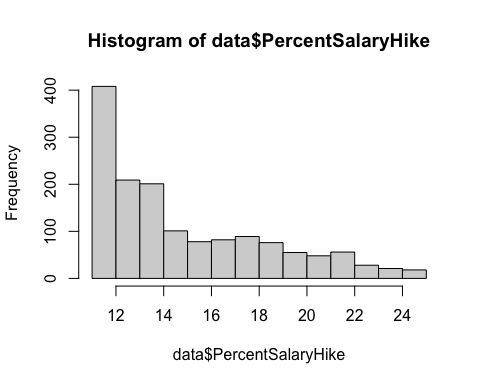
hist(data$MonthlyIncome)



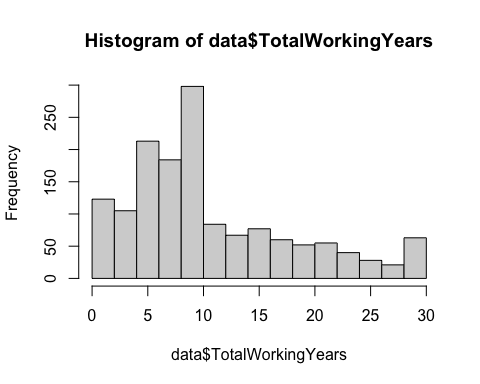
hist(data$NumCompaniesWorked)



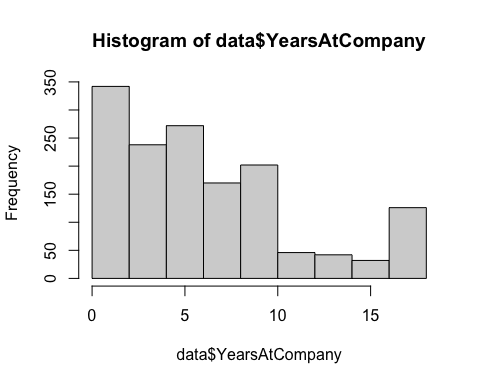
hist(data$PercentSalaryHike)



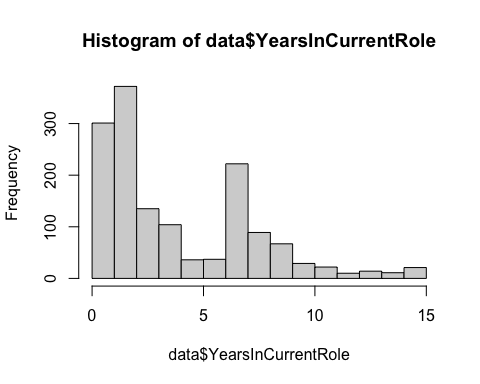
hist(data$TotalWorkingYears)



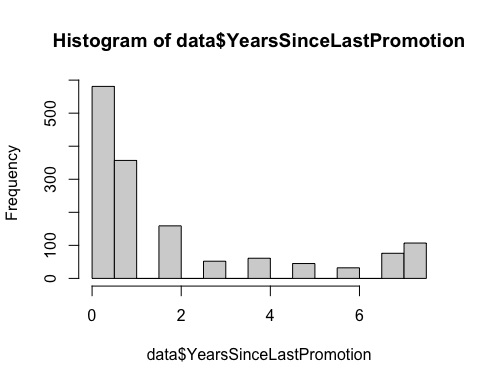
hist(data$YearsAtCompany)



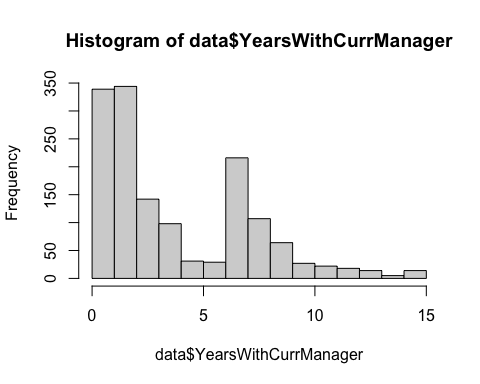
hist(data$YearsInCurrentRole)



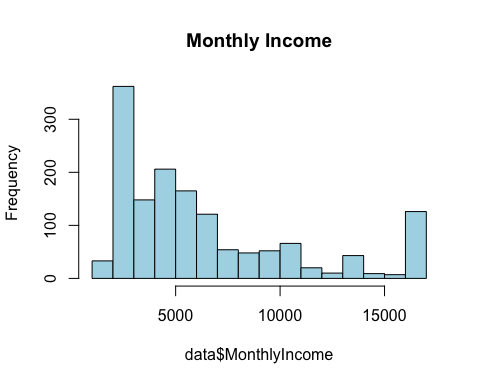
hist(data$YearsSinceLastPromotion)



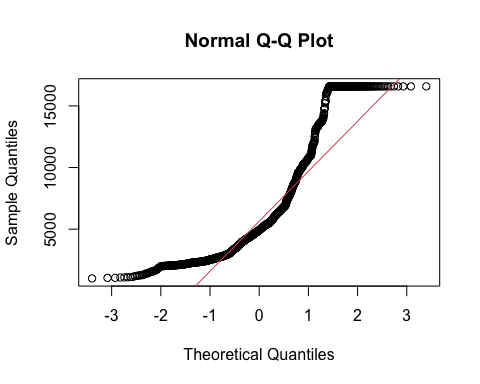
hist(data$YearsWithCurrManager)



# Histogram  
hist(data$MonthlyIncome, main="Monthly Income", col="lightblue", border="black")



# Q-Q Plot  
qqnorm(data$MonthlyIncome)  
qqline(data$MonthlyIncome, col = 2)

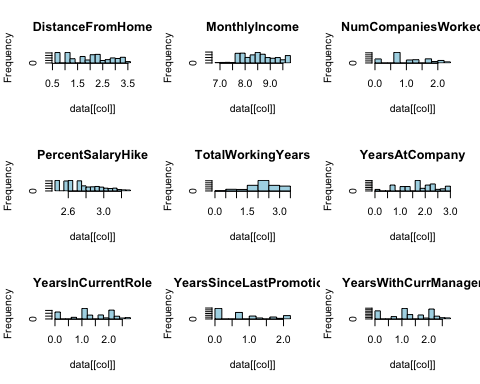


# Shapiro-Wilk Test for Normality  
shapiro.test(data$MonthlyIncome)

##   
## Shapiro-Wilk normality test  
##   
## data: data$MonthlyIncome  
## W = 0.84037, p-value < 2.2e-16

#columns\_to\_remove <- c("Over18")  
  
#data <- data[, !(names(data) %in% columns\_to\_remove)]

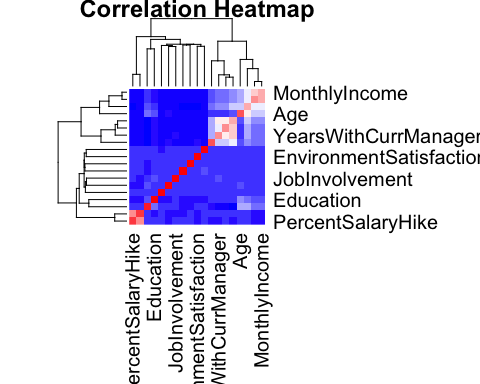
columns\_to\_transform <- c(  
 "DistanceFromHome",  
 "MonthlyIncome",  
 "NumCompaniesWorked",  
 "PercentSalaryHike",  
 "TotalWorkingYears",  
 "YearsAtCompany",  
 "YearsInCurrentRole",  
 "YearsSinceLastPromotion",  
 "YearsWithCurrManager"  
)  
  
for (col in columns\_to\_transform) {  
 data[[col]] <- ifelse(data[[col]] > 0, log(data[[col]] + 1), 0)  
}  
  
par(mfrow = c(3, 3))  
for (col in columns\_to\_transform) {  
 hist(data[[col]], main = col, col = "lightblue", border = "black")  
}



# Correlation matrix  
numeric\_data <- data[, sapply(data, is.numeric)]  
  
# Calculate the correlation matrix  
correlation\_matrix <- cor(numeric\_data)  
  
# Print the correlation matrix  
print(correlation\_matrix)

## Age DistanceFromHome Education  
## Age 1.000000000 -0.021953855 0.208033731  
## DistanceFromHome -0.021953855 1.000000000 0.019414538  
## Education 0.208033731 0.019414538 1.000000000  
## EnvironmentSatisfaction 0.010146428 -0.009394644 -0.027128313  
## JobInvolvement 0.029819959 0.031489190 0.042437634  
## JobLevel 0.509604228 -0.004163625 0.101588886  
## JobSatisfaction -0.004891877 -0.011575089 -0.011296117  
## MonthlyIncome 0.491101313 -0.003100256 0.123132296  
## NumCompaniesWorked 0.329479594 -0.007007570 0.134197648  
## PercentSalaryHike 0.004219954 0.032099747 -0.006973840  
## PerformanceRating 0.001903896 0.011243042 -0.024538791  
## RelationshipSatisfaction 0.053534720 0.009517637 -0.009118377  
## TotalWorkingYears 0.648820781 -0.002108873 0.172696984  
## TrainingTimesLastYear -0.019620819 -0.017957375 -0.025100241  
## WorkLifeBalance -0.021490028 -0.023031058 0.009819189  
## YearsAtCompany 0.249425723 0.009945118 0.058541047  
## YearsInCurrentRole 0.184896347 0.016578402 0.056683882  
## YearsSinceLastPromotion 0.180329599 -0.002057297 0.039439521  
## YearsWithCurrManager 0.174974607 0.003337656 0.051920142  
## EnvironmentSatisfaction JobInvolvement JobLevel  
## Age 0.0101464279 0.029819959 0.509604228  
## DistanceFromHome -0.0093946436 0.031489190 -0.004163625  
## Education -0.0271283133 0.042437634 0.101588886  
## EnvironmentSatisfaction 1.0000000000 -0.008277598 0.001211699  
## JobInvolvement -0.0082775982 1.000000000 -0.012629883  
## JobLevel 0.0012116994 -0.012629883 1.000000000  
## JobSatisfaction -0.0067843526 -0.021475910 -0.001943708  
## MonthlyIncome -0.0170130875 -0.016791320 0.910295009  
## NumCompaniesWorked 0.0042947187 0.010960082 0.161525586  
## PercentSalaryHike -0.0311518054 -0.018065938 -0.038179810  
## PerformanceRating -0.0295479523 -0.029071333 -0.021222082  
## RelationshipSatisfaction 0.0076653835 0.034296821 0.021641511  
## TotalWorkingYears -0.0248540877 0.015371371 0.699664841  
## TrainingTimesLastYear -0.0193593083 -0.015337826 -0.018190550  
## WorkLifeBalance 0.0276272955 -0.014616593 0.037817746  
## YearsAtCompany 0.0095999614 0.019737471 0.428004073  
## YearsInCurrentRole 0.0177477983 0.020918916 0.328248355  
## YearsSinceLastPromotion 0.0267488930 -0.008943892 0.290561031  
## YearsWithCurrManager 0.0006257353 0.052569246 0.318352664  
## JobSatisfaction MonthlyIncome NumCompaniesWorked  
## Age -0.0048918771 0.491101313 0.329479594  
## DistanceFromHome -0.0115750892 -0.003100256 -0.007007570  
## Education -0.0112961167 0.123132296 0.134197648  
## EnvironmentSatisfaction -0.0067843526 -0.017013088 0.004294719  
## JobInvolvement -0.0214759103 -0.016791320 0.010960082  
## JobLevel -0.0019437080 0.910295009 0.161525586  
## JobSatisfaction 1.0000000000 -0.004709971 -0.050550281  
## MonthlyIncome -0.0047099713 1.000000000 0.181323836  
## NumCompaniesWorked -0.0505502812 0.181323836 1.000000000  
## PercentSalaryHike 0.0202742838 -0.029636034 -0.006927221  
## PerformanceRating 0.0022971971 -0.024053131 -0.006085083  
## RelationshipSatisfaction -0.0124535932 0.006049574 0.047612803  
## TotalWorkingYears -0.0213530642 0.729311449 0.303678405  
## TrainingTimesLastYear -0.0057793350 -0.032830834 -0.057376274  
## WorkLifeBalance -0.0194587102 0.033389828 0.003888189  
## YearsAtCompany 0.0098506535 0.460257229 -0.152960869  
## YearsInCurrentRole 0.0004203887 0.379218393 -0.113330170  
## YearsSinceLastPromotion 0.0003238095 0.290952764 -0.067096605  
## YearsWithCurrManager -0.0163095256 0.351363999 -0.132376325  
## PercentSalaryHike PerformanceRating  
## Age 0.004219954 0.001903896  
## DistanceFromHome 0.032099747 0.011243042  
## Education -0.006973840 -0.024538791  
## EnvironmentSatisfaction -0.031151805 -0.029547952  
## JobInvolvement -0.018065938 -0.029071333  
## JobLevel -0.038179810 -0.021222082  
## JobSatisfaction 0.020274284 0.002297197  
## MonthlyIncome -0.029636034 -0.024053131  
## NumCompaniesWorked -0.006927221 -0.006085083  
## PercentSalaryHike 1.000000000 0.725712189  
## PerformanceRating 0.725712189 1.000000000  
## RelationshipSatisfaction -0.039901401 -0.031351455  
## TotalWorkingYears -0.022035934 0.008214099  
## TrainingTimesLastYear -0.006839233 -0.015578882  
## WorkLifeBalance -0.004613474 0.002572361  
## YearsAtCompany -0.046350725 0.013126661  
## YearsInCurrentRole -0.021899921 0.024818382  
## YearsSinceLastPromotion -0.044746329 -0.000421025  
## YearsWithCurrManager -0.026543040 0.014621483  
## RelationshipSatisfaction TotalWorkingYears  
## Age 0.0535347197 0.648820781  
## DistanceFromHome 0.0095176374 -0.002108873  
## Education -0.0091183767 0.172696984  
## EnvironmentSatisfaction 0.0076653835 -0.024854088  
## JobInvolvement 0.0342968206 0.015371371  
## JobLevel 0.0216415105 0.699664841  
## JobSatisfaction -0.0124535932 -0.021353064  
## MonthlyIncome 0.0060495738 0.729311449  
## NumCompaniesWorked 0.0476128030 0.303678405  
## PercentSalaryHike -0.0399014012 -0.022035934  
## PerformanceRating -0.0313514554 0.008214099  
## RelationshipSatisfaction 1.0000000000 -0.002243549  
## TotalWorkingYears -0.0022435493 1.000000000  
## TrainingTimesLastYear 0.0024965264 -0.024971703  
## WorkLifeBalance 0.0196044057 0.004964684  
## YearsAtCompany -0.0074956580 0.609458913  
## YearsInCurrentRole -0.0153592659 0.500897860  
## YearsSinceLastPromotion 0.0259582307 0.355128695  
## YearsWithCurrManager 0.0003347191 0.501464349  
## TrainingTimesLastYear WorkLifeBalance YearsAtCompany  
## Age -0.019620819 -0.021490028 0.249425723  
## DistanceFromHome -0.017957375 -0.023031058 0.009945118  
## Education -0.025100241 0.009819189 0.058541047  
## EnvironmentSatisfaction -0.019359308 0.027627295 0.009599961  
## JobInvolvement -0.015337826 -0.014616593 0.019737471  
## JobLevel -0.018190550 0.037817746 0.428004073  
## JobSatisfaction -0.005779335 -0.019458710 0.009850654  
## MonthlyIncome -0.032830834 0.033389828 0.460257229  
## NumCompaniesWorked -0.057376274 0.003888189 -0.152960869  
## PercentSalaryHike -0.006839233 -0.004613474 -0.046350725  
## PerformanceRating -0.015578882 0.002572361 0.013126661  
## RelationshipSatisfaction 0.002496526 0.019604406 -0.007495658  
## TotalWorkingYears -0.024971703 0.004964684 0.609458913  
## TrainingTimesLastYear 1.000000000 0.028072207 -0.013795777  
## WorkLifeBalance 0.028072207 1.000000000 0.010549362  
## YearsAtCompany -0.013795777 0.010549362 1.000000000  
## YearsInCurrentRole -0.013718344 0.025722530 0.846560664  
## YearsSinceLastPromotion 0.009660688 0.011785458 0.543584433  
## YearsWithCurrManager -0.019920856 -0.006772938 0.834248045  
## YearsInCurrentRole YearsSinceLastPromotion  
## Age 0.1848963466 0.1803295992  
## DistanceFromHome 0.0165784020 -0.0020572970  
## Education 0.0566838822 0.0394395213  
## EnvironmentSatisfaction 0.0177477983 0.0267488930  
## JobInvolvement 0.0209189156 -0.0089438921  
## JobLevel 0.3282483548 0.2905610306  
## JobSatisfaction 0.0004203887 0.0003238095  
## MonthlyIncome 0.3792183927 0.2909527637  
## NumCompaniesWorked -0.1133301704 -0.0670966045  
## PercentSalaryHike -0.0218999214 -0.0447463295  
## PerformanceRating 0.0248183820 -0.0004210250  
## RelationshipSatisfaction -0.0153592659 0.0259582307  
## TotalWorkingYears 0.5008978600 0.3551286948  
## TrainingTimesLastYear -0.0137183438 0.0096606877  
## WorkLifeBalance 0.0257225296 0.0117854584  
## YearsAtCompany 0.8465606645 0.5435844327  
## YearsInCurrentRole 1.0000000000 0.5246673131  
## YearsSinceLastPromotion 0.5246673131 1.0000000000  
## YearsWithCurrManager 0.7318547779 0.4939325400  
## YearsWithCurrManager  
## Age 0.1749746074  
## DistanceFromHome 0.0033376563  
## Education 0.0519201424  
## EnvironmentSatisfaction 0.0006257353  
## JobInvolvement 0.0525692463  
## JobLevel 0.3183526641  
## JobSatisfaction -0.0163095256  
## MonthlyIncome 0.3513639990  
## NumCompaniesWorked -0.1323763248  
## PercentSalaryHike -0.0265430404  
## PerformanceRating 0.0146214829  
## RelationshipSatisfaction 0.0003347191  
## TotalWorkingYears 0.5014643494  
## TrainingTimesLastYear -0.0199208557  
## WorkLifeBalance -0.0067729380  
## YearsAtCompany 0.8342480449  
## YearsInCurrentRole 0.7318547779  
## YearsSinceLastPromotion 0.4939325400  
## YearsWithCurrManager 1.0000000000

par(mar = c(5, 5, 2, 2)) # Adjust the margins  
heatmap(correlation\_matrix,   
 col = colorRampPalette(c("blue", "white", "red"))(20),  
 main = "Correlation Heatmap",  
 cexRow = 1.5, cexCol = 1.5, # Adjust label size  
 margins = c(10, 10)) # Adjust margins in the heatmap



# Fisher's Exact Test for categorical associations  
your\_data <- lapply(data, as.factor)  
your\_data <- as.data.frame(your\_data)  
  
  
variable\_pairs <- combn(names(your\_data), 2, simplify = TRUE)  
  
associations <- list()  
  
# Perform Fisher's Exact Test for each pair  
for (i in seq(ncol(variable\_pairs))) {  
 # Create contingency table  
 contingency\_table <- table(your\_data[, variable\_pairs[1, i]], your\_data[, variable\_pairs[2, i]])  
   
   
 if (all(dim(contingency\_table) >= 2)) {  
 test\_result <- fisher.test(contingency\_table, simulate.p.value = TRUE)  
   
 # Check if the p-value is below a significance threshold (e.g., 0.05)  
 if (test\_result$p.value < 0.05) {  
 associations[[paste(variable\_pairs[1, i], variable\_pairs[2, i], sep = "\_")]] <- test\_result  
 }  
 } else {  
 cat("Insufficient data for Fisher's Exact Test for", variable\_pairs[1, i], "and", variable\_pairs[2, i], "\n")  
 }  
}

## Insufficient data for Fisher's Exact Test for Age and Over18   
## Insufficient data for Fisher's Exact Test for Attrition and Over18   
## Insufficient data for Fisher's Exact Test for BusinessTravel and Over18   
## Insufficient data for Fisher's Exact Test for Department and Over18   
## Insufficient data for Fisher's Exact Test for DistanceFromHome and Over18   
## Insufficient data for Fisher's Exact Test for Education and Over18   
## Insufficient data for Fisher's Exact Test for EducationField and Over18   
## Insufficient data for Fisher's Exact Test for EnvironmentSatisfaction and Over18   
## Insufficient data for Fisher's Exact Test for Gender and Over18   
## Insufficient data for Fisher's Exact Test for JobInvolvement and Over18   
## Insufficient data for Fisher's Exact Test for JobLevel and Over18   
## Insufficient data for Fisher's Exact Test for JobRole and Over18   
## Insufficient data for Fisher's Exact Test for JobSatisfaction and Over18   
## Insufficient data for Fisher's Exact Test for MaritalStatus and Over18   
## Insufficient data for Fisher's Exact Test for MonthlyIncome and Over18   
## Insufficient data for Fisher's Exact Test for NumCompaniesWorked and Over18   
## Insufficient data for Fisher's Exact Test for Over18 and OverTime   
## Insufficient data for Fisher's Exact Test for Over18 and PercentSalaryHike   
## Insufficient data for Fisher's Exact Test for Over18 and PerformanceRating   
## Insufficient data for Fisher's Exact Test for Over18 and RelationshipSatisfaction   
## Insufficient data for Fisher's Exact Test for Over18 and TotalWorkingYears   
## Insufficient data for Fisher's Exact Test for Over18 and TrainingTimesLastYear   
## Insufficient data for Fisher's Exact Test for Over18 and WorkLifeBalance   
## Insufficient data for Fisher's Exact Test for Over18 and YearsAtCompany   
## Insufficient data for Fisher's Exact Test for Over18 and YearsInCurrentRole   
## Insufficient data for Fisher's Exact Test for Over18 and YearsSinceLastPromotion   
## Insufficient data for Fisher's Exact Test for Over18 and YearsWithCurrManager

# Print the list of associations  
cat("List of associations:\n")

## List of associations:

print(associations)

## $Age\_Attrition  
##   
## Fisher's Exact Test for Count Data with simulated p-value (based on  
## 2000 replicates)  
##   
## data: contingency\_table  
## p-value = 0.0004998  
## alternative hypothesis: two.sided  
##   
##   
## $Age\_Education  
##   
## Fisher's Exact Test for Count Data with simulated p-value (based on  
## 2000 replicates)  
##   
## data: contingency\_table  
## p-value = 0.0004998  
## alternative hypothesis: two.sided  
##   
##   
## $Age\_JobLevel  
##   
## Fisher's Exact Test for Count Data with simulated p-value (based on  
## 2000 replicates)  
##   
## data: contingency\_table  
## p-value = 0.0004998  
## alternative hypothesis: two.sided  
##   
##   
## $Age\_JobRole  
##   
## Fisher's Exact Test for Count Data with simulated p-value (based on  
## 2000 replicates)  
##   
## data: contingency\_table  
## p-value = 0.0004998  
## alternative hypothesis: two.sided  
##   
##   
## $Age\_MaritalStatus  
##   
## Fisher's Exact Test for Count Data with simulated p-value (based on  
## 2000 replicates)  
##   
## data: contingency\_table  
## p-value = 0.0009995  
## alternative hypothesis: two.sided  
##   
##   
## $Age\_MonthlyIncome  
##   
## Fisher's Exact Test for Count Data with simulated p-value (based on  
## 2000 replicates)  
##   
## data: contingency\_table  
## p-value = 0.0004998  
## alternative hypothesis: two.sided  
##   
##   
## $Age\_NumCompaniesWorked  
##   
## Fisher's Exact Test for Count Data with simulated p-value (based on  
## 2000 replicates)  
##   
## data: contingency\_table  
## p-value = 0.0004998  
## alternative hypothesis: two.sided  
##   
##   
## $Age\_TotalWorkingYears  
##   
## Fisher's Exact Test for Count Data with simulated p-value (based on  
## 2000 replicates)  
##   
## data: contingency\_table  
## p-value = 0.0004998  
## alternative hypothesis: two.sided  
##   
##   
## $Age\_TrainingTimesLastYear  
##   
## Fisher's Exact Test for Count Data with simulated p-value (based on  
## 2000 replicates)  
##   
## data: contingency\_table  
## p-value = 0.02449  
## alternative hypothesis: two.sided  
##   
##   
## $Age\_YearsAtCompany  
##   
## Fisher's Exact Test for Count Data with simulated p-value (based on  
## 2000 replicates)  
##   
## data: contingency\_table  
## p-value = 0.0004998  
## alternative hypothesis: two.sided  
##   
##   
## $Age\_YearsInCurrentRole  
##   
## Fisher's Exact Test for Count Data with simulated p-value (based on  
## 2000 replicates)  
##   
## data: contingency\_table  
## p-value = 0.0004998  
## alternative hypothesis: two.sided  
##   
##   
## $Age\_YearsSinceLastPromotion  
##   
## Fisher's Exact Test for Count Data with simulated p-value (based on  
## 2000 replicates)  
##   
## data: contingency\_table  
## p-value = 0.001499  
## alternative hypothesis: two.sided  
##   
##   
## $Age\_YearsWithCurrManager  
##   
## Fisher's Exact Test for Count Data with simulated p-value (based on  
## 2000 replicates)  
##   
## data: contingency\_table  
## p-value = 0.002999  
## alternative hypothesis: two.sided  
##   
##   
## $Attrition\_BusinessTravel  
##   
## Fisher's Exact Test for Count Data with simulated p-value (based on  
## 2000 replicates)  
##   
## data: contingency\_table  
## p-value = 0.0004998  
## alternative hypothesis: two.sided  
##   
##   
## $Attrition\_Department  
##   
## Fisher's Exact Test for Count Data with simulated p-value (based on  
## 2000 replicates)  
##   
## data: contingency\_table  
## p-value = 0.002499  
## alternative hypothesis: two.sided  
##   
##   
## $Attrition\_EducationField  
##   
## Fisher's Exact Test for Count Data with simulated p-value (based on  
## 2000 replicates)  
##   
## data: contingency\_table  
## p-value = 0.008496  
## alternative hypothesis: two.sided  
##   
##   
## $Attrition\_EnvironmentSatisfaction  
##   
## Fisher's Exact Test for Count Data with simulated p-value (based on  
## 2000 replicates)  
##   
## data: contingency\_table  
## p-value = 0.0004998  
## alternative hypothesis: two.sided  
##   
##   
## $Attrition\_JobInvolvement  
##   
## Fisher's Exact Test for Count Data with simulated p-value (based on  
## 2000 replicates)  
##   
## data: contingency\_table  
## p-value = 0.0004998  
## alternative hypothesis: two.sided  
##   
##   
## $Attrition\_JobLevel  
##   
## Fisher's Exact Test for Count Data with simulated p-value (based on  
## 2000 replicates)  
##   
## data: contingency\_table  
## p-value = 0.0004998  
## alternative hypothesis: two.sided  
##   
##   
## $Attrition\_JobRole  
##   
## Fisher's Exact Test for Count Data with simulated p-value (based on  
## 2000 replicates)  
##   
## data: contingency\_table  
## p-value = 0.0004998  
## alternative hypothesis: two.sided  
##   
##   
## $Attrition\_JobSatisfaction  
##   
## Fisher's Exact Test for Count Data with simulated p-value (based on  
## 2000 replicates)  
##   
## data: contingency\_table  
## p-value = 0.0009995  
## alternative hypothesis: two.sided  
##   
##   
## $Attrition\_MaritalStatus  
##   
## Fisher's Exact Test for Count Data with simulated p-value (based on  
## 2000 replicates)  
##   
## data: contingency\_table  
## p-value = 0.0004998  
## alternative hypothesis: two.sided  
##   
##   
## $Attrition\_MonthlyIncome  
##   
## Fisher's Exact Test for Count Data with simulated p-value (based on  
## 2000 replicates)  
##   
## data: contingency\_table  
## p-value = 0.001499  
## alternative hypothesis: two.sided  
##   
##   
## $Attrition\_NumCompaniesWorked  
##   
## Fisher's Exact Test for Count Data with simulated p-value (based on  
## 2000 replicates)  
##   
## data: contingency\_table  
## p-value = 0.0004998  
## alternative hypothesis: two.sided  
##   
##   
## $Attrition\_OverTime  
##   
## Fisher's Exact Test for Count Data  
##   
## data: contingency\_table  
## p-value < 2.2e-16  
## alternative hypothesis: true odds ratio is not equal to 1  
## 95 percent confidence interval:  
## 2.799096 5.078460  
## sample estimates:  
## odds ratio   
## 3.767353   
##   
##   
## $Attrition\_TotalWorkingYears  
##   
## Fisher's Exact Test for Count Data with simulated p-value (based on  
## 2000 replicates)  
##   
## data: contingency\_table  
## p-value = 0.0004998  
## alternative hypothesis: two.sided  
##   
##   
## $Attrition\_TrainingTimesLastYear  
##   
## Fisher's Exact Test for Count Data with simulated p-value (based on  
## 2000 replicates)  
##   
## data: contingency\_table  
## p-value = 0.02599  
## alternative hypothesis: two.sided  
##   
##   
## $Attrition\_WorkLifeBalance  
##   
## Fisher's Exact Test for Count Data with simulated p-value (based on  
## 2000 replicates)  
##   
## data: contingency\_table  
## p-value = 0.001999  
## alternative hypothesis: two.sided  
##   
##   
## $Attrition\_YearsAtCompany  
##   
## Fisher's Exact Test for Count Data with simulated p-value (based on  
## 2000 replicates)  
##   
## data: contingency\_table  
## p-value = 0.0004998  
## alternative hypothesis: two.sided  
##   
##   
## $Attrition\_YearsInCurrentRole  
##   
## Fisher's Exact Test for Count Data with simulated p-value (based on  
## 2000 replicates)  
##   
## data: contingency\_table  
## p-value = 0.0004998  
## alternative hypothesis: two.sided  
##   
##   
## $Attrition\_YearsSinceLastPromotion  
##   
## Fisher's Exact Test for Count Data with simulated p-value (based on  
## 2000 replicates)  
##   
## data: contingency\_table  
## p-value = 0.03948  
## alternative hypothesis: two.sided  
##   
##   
## $Attrition\_YearsWithCurrManager  
##   
## Fisher's Exact Test for Count Data with simulated p-value (based on  
## 2000 replicates)  
##   
## data: contingency\_table  
## p-value = 0.0004998  
## alternative hypothesis: two.sided  
##   
##   
## $BusinessTravel\_PercentSalaryHike  
##   
## Fisher's Exact Test for Count Data with simulated p-value (based on  
## 2000 replicates)  
##   
## data: contingency\_table  
## p-value = 0.007496  
## alternative hypothesis: two.sided  
##   
##   
## $Department\_EducationField  
##   
## Fisher's Exact Test for Count Data with simulated p-value (based on  
## 2000 replicates)  
##   
## data: contingency\_table  
## p-value = 0.0004998  
## alternative hypothesis: two.sided  
##   
##   
## $Department\_JobLevel  
##   
## Fisher's Exact Test for Count Data with simulated p-value (based on  
## 2000 replicates)  
##   
## data: contingency\_table  
## p-value = 0.0004998  
## alternative hypothesis: two.sided  
##   
##   
## $Department\_JobRole  
##   
## Fisher's Exact Test for Count Data with simulated p-value (based on  
## 2000 replicates)  
##   
## data: contingency\_table  
## p-value = 0.0004998  
## alternative hypothesis: two.sided  
##   
##   
## $Department\_WorkLifeBalance  
##   
## Fisher's Exact Test for Count Data with simulated p-value (based on  
## 2000 replicates)  
##   
## data: contingency\_table  
## p-value = 0.03798  
## alternative hypothesis: two.sided  
##   
##   
## $DistanceFromHome\_Gender  
##   
## Fisher's Exact Test for Count Data with simulated p-value (based on  
## 2000 replicates)  
##   
## data: contingency\_table  
## p-value = 0.03198  
## alternative hypothesis: two.sided  
##   
##   
## $DistanceFromHome\_JobInvolvement  
##   
## Fisher's Exact Test for Count Data with simulated p-value (based on  
## 2000 replicates)  
##   
## data: contingency\_table  
## p-value = 0.04548  
## alternative hypothesis: two.sided  
##   
##   
## $DistanceFromHome\_MonthlyIncome  
##   
## Fisher's Exact Test for Count Data with simulated p-value (based on  
## 2000 replicates)  
##   
## data: contingency\_table  
## p-value = 0.002499  
## alternative hypothesis: two.sided  
##   
##   
## $DistanceFromHome\_YearsSinceLastPromotion  
##   
## Fisher's Exact Test for Count Data with simulated p-value (based on  
## 2000 replicates)  
##   
## data: contingency\_table  
## p-value = 0.03498  
## alternative hypothesis: two.sided  
##   
##   
## $Education\_EducationField  
##   
## Fisher's Exact Test for Count Data with simulated p-value (based on  
## 2000 replicates)  
##   
## data: contingency\_table  
## p-value = 0.01099  
## alternative hypothesis: two.sided  
##   
##   
## $Education\_JobLevel  
##   
## Fisher's Exact Test for Count Data with simulated p-value (based on  
## 2000 replicates)  
##   
## data: contingency\_table  
## p-value = 0.0004998  
## alternative hypothesis: two.sided  
##   
##   
## $Education\_NumCompaniesWorked  
##   
## Fisher's Exact Test for Count Data with simulated p-value (based on  
## 2000 replicates)  
##   
## data: contingency\_table  
## p-value = 0.0004998  
## alternative hypothesis: two.sided  
##   
##   
## $Education\_TotalWorkingYears  
##   
## Fisher's Exact Test for Count Data with simulated p-value (based on  
## 2000 replicates)  
##   
## data: contingency\_table  
## p-value = 0.0004998  
## alternative hypothesis: two.sided  
##   
##   
## $EducationField\_JobLevel  
##   
## Fisher's Exact Test for Count Data with simulated p-value (based on  
## 2000 replicates)  
##   
## data: contingency\_table  
## p-value = 0.0004998  
## alternative hypothesis: two.sided  
##   
##   
## $EducationField\_JobRole  
##   
## Fisher's Exact Test for Count Data with simulated p-value (based on  
## 2000 replicates)  
##   
## data: contingency\_table  
## p-value = 0.0004998  
## alternative hypothesis: two.sided  
##   
##   
## $EducationField\_NumCompaniesWorked  
##   
## Fisher's Exact Test for Count Data with simulated p-value (based on  
## 2000 replicates)  
##   
## data: contingency\_table  
## p-value = 0.005997  
## alternative hypothesis: two.sided  
##   
##   
## $EducationField\_TrainingTimesLastYear  
##   
## Fisher's Exact Test for Count Data with simulated p-value (based on  
## 2000 replicates)  
##   
## data: contingency\_table  
## p-value = 0.04498  
## alternative hypothesis: two.sided  
##   
##   
## $EnvironmentSatisfaction\_OverTime  
##   
## Fisher's Exact Test for Count Data with simulated p-value (based on  
## 2000 replicates)  
##   
## data: contingency\_table  
## p-value = 0.03948  
## alternative hypothesis: two.sided  
##   
##   
## $Gender\_JobRole  
##   
## Fisher's Exact Test for Count Data with simulated p-value (based on  
## 2000 replicates)  
##   
## data: contingency\_table  
## p-value = 0.04148  
## alternative hypothesis: two.sided  
##   
##   
## $JobLevel\_JobRole  
##   
## Fisher's Exact Test for Count Data with simulated p-value (based on  
## 2000 replicates)  
##   
## data: contingency\_table  
## p-value = 0.0004998  
## alternative hypothesis: two.sided  
##   
##   
## $JobLevel\_MonthlyIncome  
##   
## Fisher's Exact Test for Count Data with simulated p-value (based on  
## 2000 replicates)  
##   
## data: contingency\_table  
## p-value = 0.0004998  
## alternative hypothesis: two.sided  
##   
##   
## $JobLevel\_NumCompaniesWorked  
##   
## Fisher's Exact Test for Count Data with simulated p-value (based on  
## 2000 replicates)  
##   
## data: contingency\_table  
## p-value = 0.0004998  
## alternative hypothesis: two.sided  
##   
##   
## $JobLevel\_TotalWorkingYears  
##   
## Fisher's Exact Test for Count Data with simulated p-value (based on  
## 2000 replicates)  
##   
## data: contingency\_table  
## p-value = 0.0004998  
## alternative hypothesis: two.sided  
##   
##   
## $JobLevel\_YearsAtCompany  
##   
## Fisher's Exact Test for Count Data with simulated p-value (based on  
## 2000 replicates)  
##   
## data: contingency\_table  
## p-value = 0.0004998  
## alternative hypothesis: two.sided  
##   
##   
## $JobLevel\_YearsInCurrentRole  
##   
## Fisher's Exact Test for Count Data with simulated p-value (based on  
## 2000 replicates)  
##   
## data: contingency\_table  
## p-value = 0.0004998  
## alternative hypothesis: two.sided  
##   
##   
## $JobLevel\_YearsSinceLastPromotion  
##   
## Fisher's Exact Test for Count Data with simulated p-value (based on  
## 2000 replicates)  
##   
## data: contingency\_table  
## p-value = 0.0004998  
## alternative hypothesis: two.sided  
##   
##   
## $JobLevel\_YearsWithCurrManager  
##   
## Fisher's Exact Test for Count Data with simulated p-value (based on  
## 2000 replicates)  
##   
## data: contingency\_table  
## p-value = 0.0004998  
## alternative hypothesis: two.sided  
##   
##   
## $JobRole\_MaritalStatus  
##   
## Fisher's Exact Test for Count Data with simulated p-value (based on  
## 2000 replicates)  
##   
## data: contingency\_table  
## p-value = 0.04998  
## alternative hypothesis: two.sided  
##   
##   
## $JobRole\_MonthlyIncome  
##   
## Fisher's Exact Test for Count Data with simulated p-value (based on  
## 2000 replicates)  
##   
## data: contingency\_table  
## p-value = 0.0004998  
## alternative hypothesis: two.sided  
##   
##   
## $JobRole\_NumCompaniesWorked  
##   
## Fisher's Exact Test for Count Data with simulated p-value (based on  
## 2000 replicates)  
##   
## data: contingency\_table  
## p-value = 0.0004998  
## alternative hypothesis: two.sided  
##   
##   
## $JobRole\_TotalWorkingYears  
##   
## Fisher's Exact Test for Count Data with simulated p-value (based on  
## 2000 replicates)  
##   
## data: contingency\_table  
## p-value = 0.0004998  
## alternative hypothesis: two.sided  
##   
##   
## $JobRole\_YearsAtCompany  
##   
## Fisher's Exact Test for Count Data with simulated p-value (based on  
## 2000 replicates)  
##   
## data: contingency\_table  
## p-value = 0.0004998  
## alternative hypothesis: two.sided  
##   
##   
## $JobRole\_YearsInCurrentRole  
##   
## Fisher's Exact Test for Count Data with simulated p-value (based on  
## 2000 replicates)  
##   
## data: contingency\_table  
## p-value = 0.0004998  
## alternative hypothesis: two.sided  
##   
##   
## $JobRole\_YearsSinceLastPromotion  
##   
## Fisher's Exact Test for Count Data with simulated p-value (based on  
## 2000 replicates)  
##   
## data: contingency\_table  
## p-value = 0.0004998  
## alternative hypothesis: two.sided  
##   
##   
## $JobRole\_YearsWithCurrManager  
##   
## Fisher's Exact Test for Count Data with simulated p-value (based on  
## 2000 replicates)  
##   
## data: contingency\_table  
## p-value = 0.0004998  
## alternative hypothesis: two.sided  
##   
##   
## $MaritalStatus\_TotalWorkingYears  
##   
## Fisher's Exact Test for Count Data with simulated p-value (based on  
## 2000 replicates)  
##   
## data: contingency\_table  
## p-value = 0.01299  
## alternative hypothesis: two.sided  
##   
##   
## $MaritalStatus\_YearsAtCompany  
##   
## Fisher's Exact Test for Count Data with simulated p-value (based on  
## 2000 replicates)  
##   
## data: contingency\_table  
## p-value = 0.04148  
## alternative hypothesis: two.sided  
##   
##   
## $MonthlyIncome\_TotalWorkingYears  
##   
## Fisher's Exact Test for Count Data with simulated p-value (based on  
## 2000 replicates)  
##   
## data: contingency\_table  
## p-value = 0.0004998  
## alternative hypothesis: two.sided  
##   
##   
## $MonthlyIncome\_YearsAtCompany  
##   
## Fisher's Exact Test for Count Data with simulated p-value (based on  
## 2000 replicates)  
##   
## data: contingency\_table  
## p-value = 0.0004998  
## alternative hypothesis: two.sided  
##   
##   
## $NumCompaniesWorked\_TotalWorkingYears  
##   
## Fisher's Exact Test for Count Data with simulated p-value (based on  
## 2000 replicates)  
##   
## data: contingency\_table  
## p-value = 0.0004998  
## alternative hypothesis: two.sided  
##   
##   
## $NumCompaniesWorked\_WorkLifeBalance  
##   
## Fisher's Exact Test for Count Data with simulated p-value (based on  
## 2000 replicates)  
##   
## data: contingency\_table  
## p-value = 0.04748  
## alternative hypothesis: two.sided  
##   
##   
## $NumCompaniesWorked\_YearsAtCompany  
##   
## Fisher's Exact Test for Count Data with simulated p-value (based on  
## 2000 replicates)  
##   
## data: contingency\_table  
## p-value = 0.0004998  
## alternative hypothesis: two.sided  
##   
##   
## $NumCompaniesWorked\_YearsInCurrentRole  
##   
## Fisher's Exact Test for Count Data with simulated p-value (based on  
## 2000 replicates)  
##   
## data: contingency\_table  
## p-value = 0.0004998  
## alternative hypothesis: two.sided  
##   
##   
## $NumCompaniesWorked\_YearsWithCurrManager  
##   
## Fisher's Exact Test for Count Data with simulated p-value (based on  
## 2000 replicates)  
##   
## data: contingency\_table  
## p-value = 0.002499  
## alternative hypothesis: two.sided  
##   
##   
## $OverTime\_TrainingTimesLastYear  
##   
## Fisher's Exact Test for Count Data with simulated p-value (based on  
## 2000 replicates)  
##   
## data: contingency\_table  
## p-value = 0.001999  
## alternative hypothesis: two.sided  
##   
##   
## $PercentSalaryHike\_PerformanceRating  
##   
## Fisher's Exact Test for Count Data with simulated p-value (based on  
## 2000 replicates)  
##   
## data: contingency\_table  
## p-value = 0.0004998  
## alternative hypothesis: two.sided  
##   
##   
## $TotalWorkingYears\_YearsAtCompany  
##   
## Fisher's Exact Test for Count Data with simulated p-value (based on  
## 2000 replicates)  
##   
## data: contingency\_table  
## p-value = 0.0004998  
## alternative hypothesis: two.sided  
##   
##   
## $TotalWorkingYears\_YearsInCurrentRole  
##   
## Fisher's Exact Test for Count Data with simulated p-value (based on  
## 2000 replicates)  
##   
## data: contingency\_table  
## p-value = 0.0004998  
## alternative hypothesis: two.sided  
##   
##   
## $TotalWorkingYears\_YearsSinceLastPromotion  
##   
## Fisher's Exact Test for Count Data with simulated p-value (based on  
## 2000 replicates)  
##   
## data: contingency\_table  
## p-value = 0.0004998  
## alternative hypothesis: two.sided  
##   
##   
## $TotalWorkingYears\_YearsWithCurrManager  
##   
## Fisher's Exact Test for Count Data with simulated p-value (based on  
## 2000 replicates)  
##   
## data: contingency\_table  
## p-value = 0.0004998  
## alternative hypothesis: two.sided  
##   
##   
## $YearsAtCompany\_YearsInCurrentRole  
##   
## Fisher's Exact Test for Count Data with simulated p-value (based on  
## 2000 replicates)  
##   
## data: contingency\_table  
## p-value = 0.0004998  
## alternative hypothesis: two.sided  
##   
##   
## $YearsAtCompany\_YearsSinceLastPromotion  
##   
## Fisher's Exact Test for Count Data with simulated p-value (based on  
## 2000 replicates)  
##   
## data: contingency\_table  
## p-value = 0.0004998  
## alternative hypothesis: two.sided  
##   
##   
## $YearsAtCompany\_YearsWithCurrManager  
##   
## Fisher's Exact Test for Count Data with simulated p-value (based on  
## 2000 replicates)  
##   
## data: contingency\_table  
## p-value = 0.0004998  
## alternative hypothesis: two.sided  
##   
##   
## $YearsInCurrentRole\_YearsSinceLastPromotion  
##   
## Fisher's Exact Test for Count Data with simulated p-value (based on  
## 2000 replicates)  
##   
## data: contingency\_table  
## p-value = 0.0004998  
## alternative hypothesis: two.sided  
##   
##   
## $YearsInCurrentRole\_YearsWithCurrManager  
##   
## Fisher's Exact Test for Count Data with simulated p-value (based on  
## 2000 replicates)  
##   
## data: contingency\_table  
## p-value = 0.0004998  
## alternative hypothesis: two.sided  
##   
##   
## $YearsSinceLastPromotion\_YearsWithCurrManager  
##   
## Fisher's Exact Test for Count Data with simulated p-value (based on  
## 2000 replicates)  
##   
## data: contingency\_table  
## p-value = 0.0004998  
## alternative hypothesis: two.sided

# Linear regression  
model = lm(MonthlyIncome ~ Age + Attrition + BusinessTravel + Department + DistanceFromHome + Education + EducationField + EnvironmentSatisfaction + Gender + JobInvolvement + JobLevel + JobRole + JobSatisfaction + MaritalStatus + NumCompaniesWorked + OverTime + PercentSalaryHike + PerformanceRating + RelationshipSatisfaction + TotalWorkingYears + TrainingTimesLastYear + WorkLifeBalance + YearsAtCompany + YearsInCurrentRole + YearsSinceLastPromotion + YearsWithCurrManager, data = data)  
  
summary(model)

##   
## Call:  
## lm(formula = MonthlyIncome ~ Age + Attrition + BusinessTravel +   
## Department + DistanceFromHome + Education + EducationField +   
## EnvironmentSatisfaction + Gender + JobInvolvement + JobLevel +   
## JobRole + JobSatisfaction + MaritalStatus + NumCompaniesWorked +   
## OverTime + PercentSalaryHike + PerformanceRating + RelationshipSatisfaction +   
## TotalWorkingYears + TrainingTimesLastYear + WorkLifeBalance +   
## YearsAtCompany + YearsInCurrentRole + YearsSinceLastPromotion +   
## YearsWithCurrManager, data = data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.82940 -0.15394 0.00715 0.15135 0.64647   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 7.5944838 0.1329497 57.123 < 2e-16 \*\*\*  
## Age -0.0017825 0.0009128 -1.953 0.051052 .   
## AttritionYes -0.0471153 0.0188418 -2.501 0.012510 \*   
## BusinessTravelTravel\_Frequently 0.0320385 0.0235285 1.362 0.173511   
## BusinessTravelTravel\_Rarely 0.0270216 0.0201515 1.341 0.180157   
## DepartmentResearch & Development 0.0878774 0.0821742 1.069 0.285068   
## DepartmentSales 0.0825627 0.0852733 0.968 0.333102   
## DistanceFromHome 0.0017354 0.0070188 0.247 0.804756   
## Education 0.0056146 0.0060195 0.933 0.351114   
## EducationFieldLife Sciences 0.0264577 0.0590486 0.448 0.654173   
## EducationFieldMarketing 0.0315339 0.0629184 0.501 0.616316   
## EducationFieldMedical 0.0218582 0.0592614 0.369 0.712299   
## EducationFieldOther 0.0344217 0.0634257 0.543 0.587415   
## EducationFieldTechnical Degree 0.0406051 0.0616258 0.659 0.510068   
## EnvironmentSatisfaction -0.0097805 0.0055617 -1.759 0.078870 .   
## GenderMale 0.0014889 0.0122997 0.121 0.903665   
## JobInvolvement -0.0155101 0.0085084 -1.823 0.068525 .   
## JobLevel 0.3212808 0.0126460 25.406 < 2e-16 \*\*\*  
## JobRoleHuman Resources -0.1475279 0.0861107 -1.713 0.086886 .   
## JobRoleLaboratory Technician -0.3372791 0.0280668 -12.017 < 2e-16 \*\*\*  
## JobRoleManager 0.1481949 0.0423067 3.503 0.000474 \*\*\*  
## JobRoleManufacturing Director -0.0136659 0.0276279 -0.495 0.620929   
## JobRoleResearch Director 0.1909328 0.0367447 5.196 2.33e-07 \*\*\*  
## JobRoleResearch Scientist -0.3414362 0.0276990 -12.327 < 2e-16 \*\*\*  
## JobRoleSales Executive 0.0020522 0.0543952 0.038 0.969910   
## JobRoleSales Representative -0.3992807 0.0609997 -6.546 8.24e-11 \*\*\*  
## JobSatisfaction -0.0023358 0.0054663 -0.427 0.669224   
## MaritalStatusMarried 0.0150356 0.0154695 0.972 0.331239   
## MaritalStatusSingle 0.0127142 0.0168287 0.756 0.450071   
## NumCompaniesWorked 0.0154414 0.0107852 1.432 0.152444   
## OverTimeYes 0.0295194 0.0139528 2.116 0.034546 \*   
## PercentSalaryHike 0.0648837 0.0408374 1.589 0.112320   
## PerformanceRating -0.0463788 0.0241347 -1.922 0.054847 .   
## RelationshipSatisfaction -0.0055853 0.0055766 -1.002 0.316723   
## TotalWorkingYears 0.1557096 0.0178543 8.721 < 2e-16 \*\*\*  
## TrainingTimesLastYear -0.0078878 0.0046799 -1.685 0.092119 .   
## WorkLifeBalance -0.0013517 0.0085090 -0.159 0.873806   
## YearsAtCompany 0.0044703 0.0215360 0.208 0.835592   
## YearsInCurrentRole 0.0381616 0.0144802 2.635 0.008494 \*\*   
## YearsSinceLastPromotion -0.0015368 0.0096380 -0.159 0.873338   
## YearsWithCurrManager -0.0205978 0.0137027 -1.503 0.133011   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.2271 on 1429 degrees of freedom  
## Multiple R-squared: 0.881, Adjusted R-squared: 0.8776   
## F-statistic: 264.4 on 40 and 1429 DF, p-value: < 2.2e-16

model1 = lm(MonthlyIncome ~ JobLevel + JobRole, data = data)  
  
summary(model1)

##   
## Call:  
## lm(formula = MonthlyIncome ~ JobLevel + JobRole, data = data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.02421 -0.16832 0.00512 0.16726 0.61059   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 7.88687 0.03548 222.319 < 2e-16 \*\*\*  
## JobLevel 0.39818 0.01153 34.545 < 2e-16 \*\*\*  
## JobRoleHuman Resources -0.26281 0.04123 -6.375 2.45e-10 \*\*\*  
## JobRoleLaboratory Technician -0.35655 0.02956 -12.064 < 2e-16 \*\*\*  
## JobRoleManager 0.06921 0.03826 1.809 0.070643 .   
## JobRoleManufacturing Director -0.02931 0.02913 -1.006 0.314590   
## JobRoleResearch Director 0.14232 0.03841 3.705 0.000219 \*\*\*  
## JobRoleResearch Scientist -0.34314 0.02933 -11.698 < 2e-16 \*\*\*  
## JobRoleSales Executive -0.02419 0.02505 -0.966 0.334410   
## JobRoleSales Representative -0.49401 0.03749 -13.176 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.2417 on 1460 degrees of freedom  
## Multiple R-squared: 0.8624, Adjusted R-squared: 0.8615   
## F-statistic: 1016 on 9 and 1460 DF, p-value: < 2.2e-16

#ANOVA   
summary(aov(MonthlyIncome ~ JobRole , data = data))

## Df Sum Sq Mean Sq F value Pr(>F)   
## JobRole 8 464.5 58.06 547.5 <2e-16 \*\*\*  
## Residuals 1461 155.0 0.11   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

# Pairwise t test  
pairwise.t.test(data$MonthlyIncome, data$JobRole, p.adj = "none")

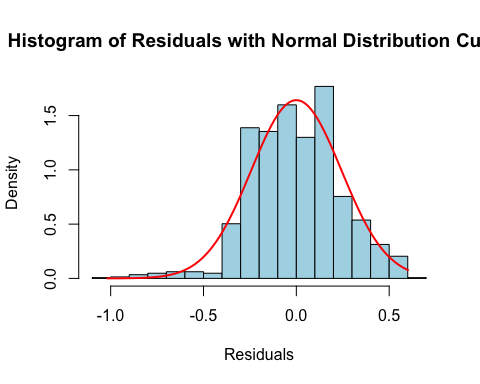
##   
## Pairwise comparisons using t tests with pooled SD   
##   
## data: data$MonthlyIncome and data$JobRole   
##   
## Healthcare Representative Human Resources  
## Human Resources < 2e-16 -   
## Laboratory Technician < 2e-16 0.00013   
## Manager < 2e-16 < 2e-16   
## Manufacturing Director 0.31740 < 2e-16   
## Research Director < 2e-16 < 2e-16   
## Research Scientist < 2e-16 9.9e-05   
## Sales Executive 0.01511 < 2e-16   
## Sales Representative < 2e-16 2.1e-11   
## Laboratory Technician Manager Manufacturing Director  
## Human Resources - - -   
## Laboratory Technician - - -   
## Manager < 2e-16 - -   
## Manufacturing Director < 2e-16 < 2e-16 -   
## Research Director < 2e-16 0.23440 < 2e-16   
## Research Scientist 0.95828 < 2e-16 < 2e-16   
## Sales Executive < 2e-16 < 2e-16 0.18931   
## Sales Representative 1.4e-06 < 2e-16 < 2e-16   
## Research Director Research Scientist Sales Executive  
## Human Resources - - -   
## Laboratory Technician - - -   
## Manager - - -   
## Manufacturing Director - - -   
## Research Director - - -   
## Research Scientist < 2e-16 - -   
## Sales Executive < 2e-16 < 2e-16 -   
## Sales Representative < 2e-16 1.2e-06 < 2e-16   
##   
## P value adjustment method: none

# Encoding categorical variable:  
  
encoded\_data <- cbind(data, model.matrix(~ JobRole - 1, data = data))  
  
encoded\_data <- encoded\_data[, -which(names(encoded\_data) %in% c("JobRole"))]  
  
colnames(encoded\_data)[colnames(encoded\_data) == "JobRoleHuman Resources"] <- "JobRoleHumanResources"  
colnames(encoded\_data)[colnames(encoded\_data) == "JobRoleLaboratory Technician"] <- "JobRoleLaboratoryTechnician"  
colnames(encoded\_data)[colnames(encoded\_data) == "JobRoleManufacturing Director"] <- "JobRoleManufacturingDirector"  
colnames(encoded\_data)[colnames(encoded\_data) == "JobRoleResearch Scientist"] <- "JobRoleResearchScientist"  
colnames(encoded\_data)[colnames(encoded\_data) == "JobRoleSales Executive"] <- "JobRoleSalesExecutive"  
colnames(encoded\_data)[colnames(encoded\_data) == "JobRoleSales Representative"] <- "JobRoleSalesRepresentative"  
colnames(encoded\_data)[colnames(encoded\_data) == "JobRoleResearch Director"] <- "JobRoleResearchDirector"

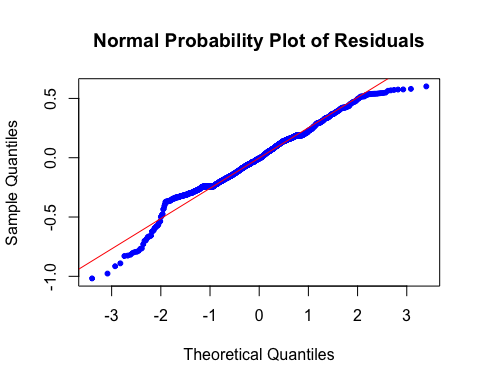
# Final regression model with encoded job roles  
model1 = lm(MonthlyIncome ~ JobLevel + JobRoleHumanResources + JobRoleLaboratoryTechnician + JobRoleResearchScientist + JobRoleSalesRepresentative , data = encoded\_data)  
  
summary(model1)

##   
## Call:  
## lm(formula = MonthlyIncome ~ JobLevel + JobRoleHumanResources +   
## JobRoleLaboratoryTechnician + JobRoleResearchScientist +   
## JobRoleSalesRepresentative, data = encoded\_data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.0175 -0.1747 -0.0049 0.1699 0.6013   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 7.802519 0.024490 318.599 < 2e-16 \*\*\*  
## JobLevel 0.431186 0.008178 52.725 < 2e-16 \*\*\*  
## JobRoleHumanResources -0.227332 0.036474 -6.233 5.98e-10 \*\*\*  
## JobRoleLaboratoryTechnician -0.313102 0.021609 -14.490 < 2e-16 \*\*\*  
## JobRoleResearchScientist -0.298462 0.021189 -14.086 < 2e-16 \*\*\*  
## JobRoleSalesRepresentative -0.445451 0.031394 -14.189 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.2433 on 1464 degrees of freedom  
## Multiple R-squared: 0.8601, Adjusted R-squared: 0.8596   
## F-statistic: 1800 on 5 and 1464 DF, p-value: < 2.2e-16

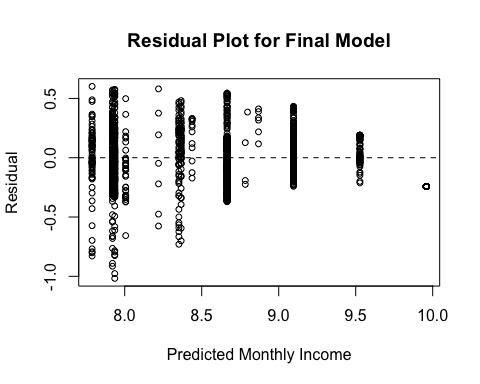
# Residual plots  
residuals <- residuals(model1)  
  
hist(residuals, main = "Histogram of Residuals with Normal Distribution Curve", col = "lightblue", border = "black", xlab = "Residuals", prob = TRUE)  
  
  
mu <- mean(residuals)  
sigma <- sd(residuals)  
x <- seq(min(residuals), max(residuals), length = 100)  
lines(x, dnorm(x, mean = mu, sd = sigma), col = "red", lwd = 2)



qqnorm(residuals, main = "Normal Probability Plot of Residuals", col = "blue", pch = 20)  
qqline(residuals, col = "red")



#residual plot  
plot(resid(model1) ~ fitted(model1),  
 xlab = "Predicted Monthly Income", ylab = "Residual",  
 main = "Residual Plot for Final Model",  
 pch = 21, cex = 0.8)  
abline(h = 0, lty = 2)



# Interaction model  
model2 = lm(MonthlyIncome ~ JobLevel + JobRole + JobLevel\*JobRole, data = data)  
  
summary(model2)

##   
## Call:  
## lm(formula = MonthlyIncome ~ JobLevel + JobRole + JobLevel \*   
## JobRole, data = data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.99371 -0.13786 -0.01523 0.13747 0.61013   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 7.783578 0.080978 96.119 < 2e-16  
## JobLevel 0.439946 0.031753 13.855 < 2e-16  
## JobRoleHuman Resources -0.553317 0.109648 -5.046 5.07e-07  
## JobRoleLaboratory Technician -0.288425 0.090722 -3.179 0.00151  
## JobRoleManager 1.404181 0.166781 8.419 < 2e-16  
## JobRoleManufacturing Director -0.122751 0.111311 -1.103 0.27031  
## JobRoleResearch Director 1.291956 0.148844 8.680 < 2e-16  
## JobRoleResearch Scientist -0.421273 0.090753 -4.642 3.76e-06  
## JobRoleSales Executive -0.038403 0.097396 -0.394 0.69342  
## JobRoleSales Representative -0.397669 0.128616 -3.092 0.00203  
## JobLevel:JobRoleHuman Resources 0.224176 0.055252 4.057 5.23e-05  
## JobLevel:JobRoleLaboratory Technician -0.013389 0.044374 -0.302 0.76290  
## JobLevel:JobRoleManager -0.327938 0.046139 -7.108 1.85e-12  
## JobLevel:JobRoleManufacturing Director 0.038594 0.043847 0.880 0.37889  
## JobLevel:JobRoleResearch Director -0.304993 0.044215 -6.898 7.85e-12  
## JobLevel:JobRoleResearch Scientist 0.109165 0.045265 2.412 0.01600  
## JobLevel:JobRoleSales Executive 0.008705 0.038982 0.223 0.82332  
## JobLevel:JobRoleSales Representative -0.035358 0.094745 -0.373 0.70906  
##   
## (Intercept) \*\*\*  
## JobLevel \*\*\*  
## JobRoleHuman Resources \*\*\*  
## JobRoleLaboratory Technician \*\*   
## JobRoleManager \*\*\*  
## JobRoleManufacturing Director   
## JobRoleResearch Director \*\*\*  
## JobRoleResearch Scientist \*\*\*  
## JobRoleSales Executive   
## JobRoleSales Representative \*\*   
## JobLevel:JobRoleHuman Resources \*\*\*  
## JobLevel:JobRoleLaboratory Technician   
## JobLevel:JobRoleManager \*\*\*  
## JobLevel:JobRoleManufacturing Director   
## JobLevel:JobRoleResearch Director \*\*\*  
## JobLevel:JobRoleResearch Scientist \*   
## JobLevel:JobRoleSales Executive   
## JobLevel:JobRoleSales Representative   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.226 on 1452 degrees of freedom  
## Multiple R-squared: 0.8803, Adjusted R-squared: 0.8789   
## F-statistic: 628 on 17 and 1452 DF, p-value: < 2.2e-16

# Attrition balancing  
minority\_indices <- which(data$Attrition == "Yes")  
table(data$Attrition)

##   
## No Yes   
## 1233 237

# Over sampling  
minority\_indices <- which(data$Attrition == "Yes")  
  
data\_oversampled <- data  
oversampled\_indices <- sample(minority\_indices, replace = TRUE, size = length(setdiff(1:nrow(data), minority\_indices)))  
data\_oversampled <- rbind(data\_oversampled, data[oversampled\_indices, ])  
  
table(data\_oversampled$Attrition)

##   
## No Yes   
## 1233 1470

data\_oversampled$Attrition <- as.factor(data\_oversampled$Attrition)

# Convert multiple columns to factors and check structure  
factor\_columns <- c("BusinessTravel", "Department", "Education", "Gender",   
 "JobInvolvement", "JobLevel", "JobRole", "JobSatisfaction",   
 "MaritalStatus", "NumCompaniesWorked", "OverTime")  
  
for (col in factor\_columns) {  
 data[[col]] <- as.factor(data[[col]])  
 cat("\nStructure of", col, ":\n")  
 str(data[[col]])  
}

##   
## Structure of BusinessTravel :  
## Factor w/ 3 levels "Non-Travel","Travel\_Frequently",..: 3 2 3 2 3 2 3 3 2 3 ...  
##   
## Structure of Department :  
## Factor w/ 3 levels "Human Resources",..: 3 2 2 2 2 2 2 2 2 2 ...  
##   
## Structure of Education :  
## Factor w/ 5 levels "1","2","3","4",..: 2 1 2 4 1 2 3 1 3 3 ...  
##   
## Structure of Gender :  
## Factor w/ 2 levels "Female","Male": 1 2 2 1 2 2 1 2 2 2 ...  
##   
## Structure of JobInvolvement :  
## Factor w/ 4 levels "1","2","3","4": 3 2 2 3 3 3 4 3 2 3 ...  
##   
## Structure of JobLevel :  
## Factor w/ 5 levels "1","2","3","4",..: 2 2 1 1 1 1 1 1 3 2 ...  
##   
## Structure of JobRole :  
## Factor w/ 9 levels "Healthcare Representative",..: 8 7 3 7 3 3 3 3 5 1 ...  
##   
## Structure of JobSatisfaction :  
## Factor w/ 4 levels "1","2","3","4": 4 2 3 3 2 4 1 3 3 3 ...  
##   
## Structure of MaritalStatus :  
## Factor w/ 3 levels "Divorced","Married",..: 3 2 3 2 2 3 2 1 3 2 ...  
##   
## Structure of NumCompaniesWorked :  
## Factor w/ 10 levels "0","0.693147180559945",..: 9 2 7 2 10 1 5 2 1 7 ...  
##   
## Structure of OverTime :  
## Factor w/ 2 levels "No","Yes": 2 1 2 2 1 1 2 1 1 1 ...

# Display the first few rows of the dataset  
cat("\nHead of the data:\n")

##   
## Head of the data:

head(data)

## Age Attrition BusinessTravel Department DistanceFromHome  
## 1 41 Yes Travel\_Rarely Sales 0.6931472  
## 2 49 No Travel\_Frequently Research & Development 2.1972246  
## 3 37 Yes Travel\_Rarely Research & Development 1.0986123  
## 4 33 No Travel\_Frequently Research & Development 1.3862944  
## 5 27 No Travel\_Rarely Research & Development 1.0986123  
## 6 32 No Travel\_Frequently Research & Development 1.0986123  
## Education EducationField EnvironmentSatisfaction Gender JobInvolvement  
## 1 2 Life Sciences 2 Female 3  
## 2 1 Life Sciences 3 Male 2  
## 3 2 Other 4 Male 2  
## 4 4 Life Sciences 4 Female 3  
## 5 1 Medical 1 Male 3  
## 6 2 Life Sciences 4 Male 3  
## JobLevel JobRole JobSatisfaction MaritalStatus MonthlyIncome  
## 1 2 Sales Executive 4 Single 8.698514  
## 2 2 Research Scientist 2 Married 8.543056  
## 3 1 Laboratory Technician 3 Single 7.645398  
## 4 1 Research Scientist 3 Married 7.975908  
## 5 1 Laboratory Technician 2 Married 8.151622  
## 6 1 Laboratory Technician 4 Single 8.029107  
## NumCompaniesWorked Over18 OverTime PercentSalaryHike PerformanceRating  
## 1 2.19722457733622 Y Yes 2.484907 3  
## 2 0.693147180559945 Y No 3.178054 4  
## 3 1.94591014905531 Y Yes 2.772589 3  
## 4 0.693147180559945 Y Yes 2.484907 3  
## 5 2.2512917986065 Y No 2.564949 3  
## 6 0 Y No 2.639057 3  
## RelationshipSatisfaction TotalWorkingYears TrainingTimesLastYear  
## 1 1 2.197225 0  
## 2 4 2.397895 3  
## 3 2 2.079442 3  
## 4 3 2.197225 3  
## 5 4 1.945910 3  
## 6 3 2.197225 2  
## WorkLifeBalance YearsAtCompany YearsInCurrentRole YearsSinceLastPromotion  
## 1 1 1.945910 1.609438 0.0000000  
## 2 3 2.397895 2.079442 0.6931472  
## 3 3 0.000000 0.000000 0.0000000  
## 4 3 2.197225 2.079442 1.3862944  
## 5 3 1.098612 1.098612 1.0986123  
## 6 2 2.079442 2.079442 1.3862944  
## YearsWithCurrManager  
## 1 1.791759  
## 2 2.079442  
## 3 0.000000  
## 4 0.000000  
## 5 1.098612  
## 6 1.945910

# Logistic regression model  
model <- glm(Attrition ~ Age + MonthlyIncome + BusinessTravel + Department +   
 DistanceFromHome + Education + EducationField + EnvironmentSatisfaction +   
 Gender + JobInvolvement + JobLevel + JobRole + JobSatisfaction +   
 MaritalStatus + NumCompaniesWorked + OverTime + PercentSalaryHike +   
 PerformanceRating + RelationshipSatisfaction + TotalWorkingYears +   
 TrainingTimesLastYear + WorkLifeBalance + YearsAtCompany +   
 YearsInCurrentRole + YearsSinceLastPromotion + YearsWithCurrManager,   
 data = data\_oversampled, family = binomial(link = "logit"))  
  
summary(model)

##   
## Call:  
## glm(formula = Attrition ~ Age + MonthlyIncome + BusinessTravel +   
## Department + DistanceFromHome + Education + EducationField +   
## EnvironmentSatisfaction + Gender + JobInvolvement + JobLevel +   
## JobRole + JobSatisfaction + MaritalStatus + NumCompaniesWorked +   
## OverTime + PercentSalaryHike + PerformanceRating + RelationshipSatisfaction +   
## TotalWorkingYears + TrainingTimesLastYear + WorkLifeBalance +   
## YearsAtCompany + YearsInCurrentRole + YearsSinceLastPromotion +   
## YearsWithCurrManager, family = binomial(link = "logit"),   
## data = data\_oversampled)  
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -5.771845 375.487106 -0.015 0.987736   
## Age -0.011814 0.007376 -1.602 0.109256   
## MonthlyIncome -0.889969 0.221894 -4.011 6.05e-05 \*\*\*  
## BusinessTravelTravel\_Frequently 2.004240 0.233433 8.586 < 2e-16 \*\*\*  
## BusinessTravelTravel\_Rarely 1.277452 0.210817 6.060 1.37e-09 \*\*\*  
## DepartmentResearch & Development 14.895816 375.482454 0.040 0.968355   
## DepartmentSales 14.902341 375.482553 0.040 0.968341   
## DistanceFromHome 0.422953 0.062582 6.758 1.39e-11 \*\*\*  
## Education 0.035920 0.052275 0.687 0.491992   
## EducationFieldLife Sciences -1.054479 0.500331 -2.108 0.035069 \*   
## EducationFieldMarketing -0.735792 0.530198 -1.388 0.165208   
## EducationFieldMedical -0.855733 0.495714 -1.726 0.084300 .   
## EducationFieldOther -0.714123 0.530286 -1.347 0.178085   
## EducationFieldTechnical Degree 0.209786 0.515330 0.407 0.683942   
## EnvironmentSatisfaction -0.425195 0.048497 -8.767 < 2e-16 \*\*\*  
## GenderMale 0.397831 0.108115 3.680 0.000234 \*\*\*  
## JobInvolvement -0.451673 0.074152 -6.091 1.12e-09 \*\*\*  
## JobLevel 0.713756 0.138431 5.156 2.52e-07 \*\*\*  
## JobRoleHuman Resources 15.910734 375.482519 0.042 0.966200   
## JobRoleLaboratory Technician 1.325816 0.282507 4.693 2.69e-06 \*\*\*  
## JobRoleManager -0.370393 0.411053 -0.901 0.367545   
## JobRoleManufacturing Director 0.201479 0.278970 0.722 0.470155   
## JobRoleResearch Director -1.310336 0.440044 -2.978 0.002904 \*\*   
## JobRoleResearch Scientist 0.537608 0.283940 1.893 0.058306 .   
## JobRoleSales Executive 1.037598 0.582435 1.781 0.074834 .   
## JobRoleSales Representative 2.172146 0.631099 3.442 0.000578 \*\*\*  
## JobSatisfaction -0.340621 0.047621 -7.153 8.50e-13 \*\*\*  
## MaritalStatusMarried 0.727037 0.146678 4.957 7.17e-07 \*\*\*  
## MaritalStatusSingle 1.543149 0.155262 9.939 < 2e-16 \*\*\*  
## NumCompaniesWorked 0.802384 0.091930 8.728 < 2e-16 \*\*\*  
## OverTimeYes 1.799070 0.111659 16.112 < 2e-16 \*\*\*  
## PercentSalaryHike -1.085949 0.357864 -3.035 0.002409 \*\*   
## PerformanceRating 0.559446 0.214540 2.608 0.009117 \*\*   
## RelationshipSatisfaction -0.154953 0.046953 -3.300 0.000966 \*\*\*  
## TotalWorkingYears -0.774955 0.145275 -5.334 9.59e-08 \*\*\*  
## TrainingTimesLastYear -0.192613 0.040351 -4.773 1.81e-06 \*\*\*  
## WorkLifeBalance -0.266673 0.068323 -3.903 9.49e-05 \*\*\*  
## YearsAtCompany 0.618059 0.187039 3.304 0.000952 \*\*\*  
## YearsInCurrentRole -0.511560 0.128384 -3.985 6.76e-05 \*\*\*  
## YearsSinceLastPromotion 0.638694 0.088496 7.217 5.31e-13 \*\*\*  
## YearsWithCurrManager -0.486420 0.119229 -4.080 4.51e-05 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 3726.3 on 2702 degrees of freedom  
## Residual deviance: 2419.1 on 2662 degrees of freedom  
## AIC: 2501.1  
##   
## Number of Fisher Scoring iterations: 14

# Ordinary Least Square Regression  
  
#install.packages("ordinal")  
library(ordinal)  
  
data$PerformanceRating <- ordered(data$PerformanceRating)  
  
  
model3 <- clm(PerformanceRating ~ JobSatisfaction + EnvironmentSatisfaction + RelationshipSatisfaction +   
 WorkLifeBalance, data = data)  
  
summary(model3)

## formula:   
## PerformanceRating ~ JobSatisfaction + EnvironmentSatisfaction + RelationshipSatisfaction + WorkLifeBalance  
## data: data  
##   
## link threshold nobs logLik AIC niter max.grad cond.H   
## logit flexible 1470 -627.31 1268.62 5(0) 3.46e-11 8.7e+02  
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)  
## JobSatisfaction2 -0.09140 0.22928 -0.399 0.690  
## JobSatisfaction3 -0.32837 0.21350 -1.538 0.124  
## JobSatisfaction4 0.04191 0.20095 0.209 0.835  
## EnvironmentSatisfaction -0.08000 0.06603 -1.212 0.226  
## RelationshipSatisfaction -0.08069 0.06657 -1.212 0.225  
## WorkLifeBalance 0.02067 0.10272 0.201 0.841  
##   
## Threshold coefficients:  
## Estimate Std. Error z value  
## 3|4 1.2360 0.4058 3.046

model4 <- clm(PerformanceRating ~ JobSatisfaction + EnvironmentSatisfaction + RelationshipSatisfaction + WorkLifeBalance + JobInvolvement + DistanceFromHome + TotalWorkingYears + Education + JobLevel + NumCompaniesWorked + TotalWorkingYears + TrainingTimesLastYear + WorkLifeBalance + YearsAtCompany + YearsInCurrentRole + YearsSinceLastPromotion + YearsWithCurrManager , data = data)  
  
summary(model4)

## formula:   
## PerformanceRating ~ JobSatisfaction + EnvironmentSatisfaction + RelationshipSatisfaction + WorkLifeBalance + JobInvolvement + DistanceFromHome + TotalWorkingYears + Education + JobLevel + NumCompaniesWorked + TotalWorkingYears + TrainingTimesLastYear + WorkLifeBalance + YearsAtCompany + YearsInCurrentRole + YearsSinceLastPromotion + YearsWithCurrManager  
## data: data  
##   
## link threshold nobs logLik AIC niter max.grad cond.H   
## logit flexible 1470 -617.34 1302.67 5(0) 7.28e-09 6.2e+03  
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## JobSatisfaction2 -0.073446 0.233338 -0.315 0.7529   
## JobSatisfaction3 -0.333272 0.216481 -1.539 0.1237   
## JobSatisfaction4 0.018944 0.204497 0.093 0.9262   
## EnvironmentSatisfaction -0.086067 0.067189 -1.281 0.2002   
## RelationshipSatisfaction -0.075268 0.067592 -1.114 0.2655   
## WorkLifeBalance 0.020388 0.104567 0.195 0.8454   
## JobInvolvement2 -0.334485 0.313732 -1.066 0.2864   
## JobInvolvement3 -0.363261 0.293256 -1.239 0.2155   
## JobInvolvement4 -0.498298 0.370337 -1.346 0.1785   
## DistanceFromHome 0.015389 0.085504 0.180 0.8572   
## TotalWorkingYears 0.309281 0.224860 1.375 0.1690   
## Education2 -0.009394 0.262876 -0.036 0.9715   
## Education3 -0.250733 0.241821 -1.037 0.2998   
## Education4 -0.243209 0.257096 -0.946 0.3442   
## Education5 0.101946 0.434727 0.235 0.8146   
## JobLevel2 -0.256543 0.196116 -1.308 0.1908   
## JobLevel3 -0.277246 0.269172 -1.030 0.3030   
## JobLevel4 -0.193539 0.371272 -0.521 0.6022   
## JobLevel5 -0.859478 0.484726 -1.773 0.0762 .  
## NumCompaniesWorked0.693147180559945 0.311370 0.242138 1.286 0.1985   
## NumCompaniesWorked1.09861228866811 -0.286811 0.362964 -0.790 0.4294   
## NumCompaniesWorked1.38629436111989 0.201415 0.325680 0.618 0.5363   
## NumCompaniesWorked1.6094379124341 0.227587 0.336662 0.676 0.4990   
## NumCompaniesWorked1.79175946922805 -0.048148 0.438861 -0.110 0.9126   
## NumCompaniesWorked1.94591014905531 -0.107935 0.436880 -0.247 0.8049   
## NumCompaniesWorked2.07944154167984 -0.375589 0.445115 -0.844 0.3988   
## NumCompaniesWorked2.19722457733622 0.601352 0.422723 1.423 0.1549   
## NumCompaniesWorked2.2512917986065 -0.372075 0.528434 -0.704 0.4814   
## TrainingTimesLastYear -0.039483 0.057919 -0.682 0.4954   
## YearsAtCompany -0.252151 0.284663 -0.886 0.3757   
## YearsInCurrentRole 0.196936 0.187829 1.048 0.2944   
## YearsSinceLastPromotion -0.048774 0.116679 -0.418 0.6759   
## YearsWithCurrManager 0.052328 0.169210 0.309 0.7571   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Threshold coefficients:  
## Estimate Std. Error z value  
## 3|4 1.153 0.678 1.7

model4 <- clm(PerformanceRating ~ JobLevel + JobInvolvement + JobInvolvement\*JobLevel, data = data)

## Warning: (1) Hessian is numerically singular: parameters are not uniquely determined   
## In addition: Absolute convergence criterion was met, but relative criterion was not met

# Display summary of the model  
summary(model4)

## formula:   
## PerformanceRating ~ JobLevel + JobInvolvement + JobInvolvement \* JobLevel  
## data: data  
##   
## link threshold nobs logLik AIC niter max.grad cond.H   
## logit flexible 1470 -623.18 1286.35 19(0) 4.33e-09 7.0e+11  
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)  
## JobLevel2 0.39304 NA NA NA  
## JobLevel3 0.76214 NA NA NA  
## JobLevel4 -19.59346 NA NA NA  
## JobLevel5 0.22314 NA NA NA  
## JobInvolvement2 -0.09963 NA NA NA  
## JobInvolvement3 0.04462 NA NA NA  
## JobInvolvement4 -0.37648 NA NA NA  
## JobLevel2:JobInvolvement2 -0.49408 NA NA NA  
## JobLevel3:JobInvolvement2 -0.89890 NA NA NA  
## JobLevel4:JobInvolvement2 20.25271 NA NA NA  
## JobLevel5:JobInvolvement2 -0.05452 NA NA NA  
## JobLevel2:JobInvolvement3 -0.68015 NA NA NA  
## JobLevel3:JobInvolvement3 -0.88372 NA NA NA  
## JobLevel4:JobInvolvement3 19.50962 NA NA NA  
## JobLevel5:JobInvolvement3 -1.24859 NA NA NA  
## JobLevel2:JobInvolvement4 0.22957 NA NA NA  
## JobLevel3:JobInvolvement4 -1.34117 NA NA NA  
## JobLevel4:JobInvolvement4 19.01443 NA NA NA  
## JobLevel5:JobInvolvement4 -19.44012 NA NA NA  
##   
## Threshold coefficients:  
## Estimate Std. Error z value  
## 3|4 1.609 NA NA

response\_variable <- data$PerformanceRating  
  
# Check if it is a factor  
if (is.factor(response\_variable)) {  
 print("Response variable is a factor.")  
} else {  
 print("Response variable is not a factor.")  
}

## [1] "Response variable is a factor."

data$PerformanceRating <- as.factor(data$PerformanceRating)

table(data$PercentSalaryHike, data$PerformanceRating)

##   
## 3 4  
## 2.484906649788 210 0  
## 2.56494935746154 198 0  
## 2.63905732961526 209 0  
## 2.70805020110221 201 0  
## 2.77258872223978 101 0  
## 2.83321334405622 78 0  
## 2.89037175789616 82 0  
## 2.94443897916644 89 0  
## 2.99573227355399 76 0  
## 3.04452243772342 0 55  
## 3.09104245335832 0 48  
## 3.13549421592915 0 56  
## 3.17805383034795 0 28  
## 3.2188758248682 0 21  
## 3.25809653802148 0 18