DS6306_CaseStudy02

Rudranil Mitra

8/14/2020

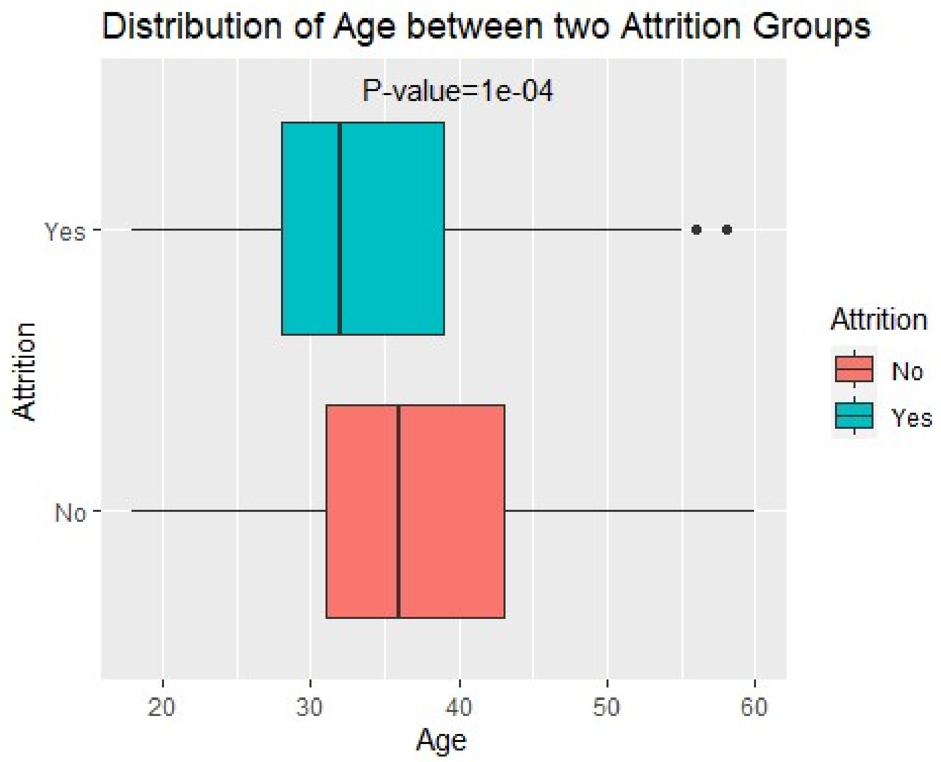
R Markdown

```
library(ggplot2)
library(dplyr)

##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union

library(tidyverse)

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

## v tibble  3.0.3     v purrr   0.3.4
## v tidyr   1.1.1     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.5.0

## -- Conflicts -------------------------------------------------------------
-------------------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()

library(caret)

## Loading required package: lattice

##
## Attaching package: 'caret'

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

library(class)
library(dplyr)
library(e1071)
library(FNN)
```

```
##
## Attaching package: 'FNN'

## The following objects are masked from 'package:class':
##
##     knn, knn.cv

library(gmodels)
library(psych)

##
## Attaching package: 'psych'

## The following objects are masked from 'package:ggplot2':
##
##     %+%, alpha

library(epiR)

## Loading required package: survival

##
## Attaching package: 'survival'

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

## Package epiR 1.0-15 is loaded

## Type help(epi.about) for summary information

## Type browseVignettes(package = 'epiR') to learn how to use epiR for
applied epidemiological analyses

##

library(DMwR)

## Loading required package: grid

## Registered S3 method overwritten by 'quantmod':
##   method            from
##   as.zoo.data.frame zoo

##
## Attaching package: 'DMwR'

## The following object is masked from 'package:psych':
##
##     crossValidation

# Load attrition dataset with labels
attrition_dataset = read.csv('D:\\SMU_MSDS\\MSDS_6306_Doing-Data-
Science\\Unit 14 and 15 Case Study 2\\CaseStudy2-data.csv')
# Load attrition dataset without labels
attrition_dataset_wl = read.csv('D:\\SMU_MSDS\\MSDS_6306_Doing-Data-
Science\\Unit 14 and 15 Case Study 2\\CaseStudy2CompSet No Attrition.csv')
#
attrition_dataset_lm = attrition_dataset
```

Does mean age differ significantly among Attrition Groups

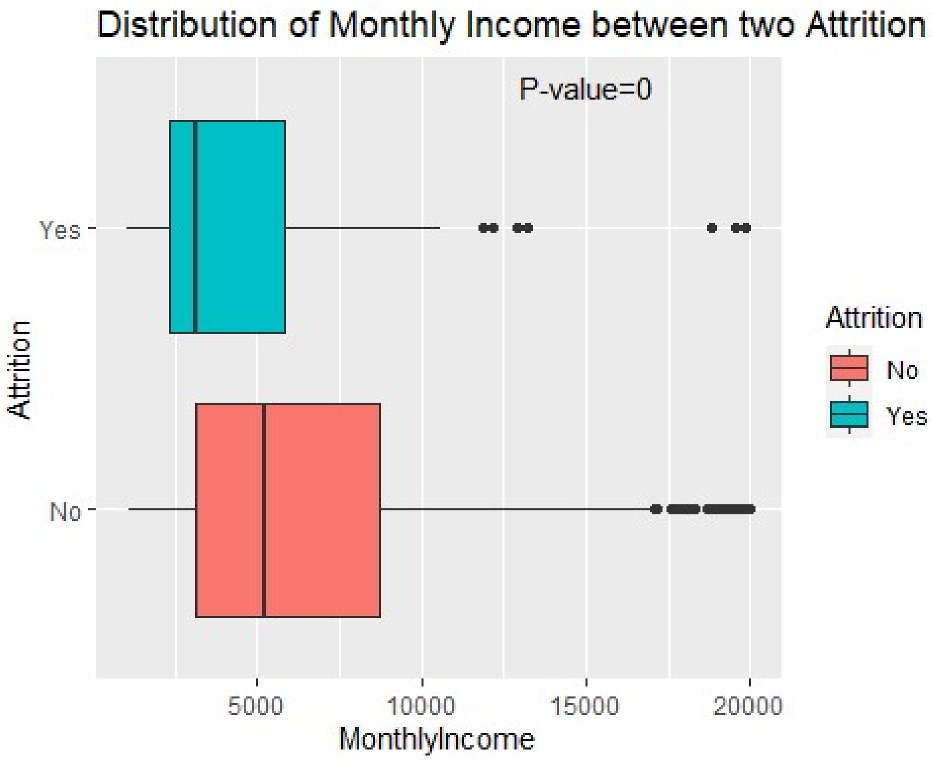
```
# Execute the t-test between two groups, we are assuming that the variances
are not equal
fit = t.test(
  attrition_dataset[attrition_dataset['Attrition']=='No',]$Age,
  attrition_dataset[attrition_dataset['Attrition']=='Yes',]$Age,
  var.equal = FALSE)
# P-value is less than the significance level(0.05), so we can reject the
null hypothesis and say that mean age is different between two groups
fit$p.value

## [1] 5.049764e-05

# Looking at the confidence intervals we can suggest that individuals who are
not leaving the company have higher mean age than the individuals who leave
the company
fit$conf.int

## [1] 1.902905 5.350324
## attr(,"conf.level")
## [1] 0.95

attrition_dataset %>% ggplot(mapping=aes(x=Age,y=Attrition,fill=Attrition)) +
  geom_boxplot() +
  annotate("text",x=40,y=2.5,label=paste0('P-value=',round(fit$p.value,4))) +
  ggtitle('Distribution of Age between two Attrition Groups')
```



### Does median monthly income differ significantly among Attrition Groups

```
# Execute the t-test between two groups, we are assuming that the variances
are not equal
fit = t.test(

log(attrition_dataset[attrition_dataset['Attrition']=='No',]$MonthlyIncome),

log(attrition_dataset[attrition_dataset['Attrition']=='Yes',]$MonthlyIncome),
  var.equal = FALSE)
# P-value is less than the significance level(0.05), so we can reject the
null hypothesis and say that median monthly income is different between two
groups
fit$p.value

## [1] 1.159977e-08

# Looking at the confidence intervals(confidence intervals are in the log
scale) we can suggest that individuals who are  not leaving the company have
higher median monthly income than the individuals who leave the company. So
monthly income is a significant reason for leaving the company
# The lower bound of the increase between two groups is 27%
round(exp(fit$conf.int[1])-1,2)

## [1] 0.27

# The upper bound of the increase between two groups is 60%
round(exp(fit$conf.int[2])-1,2)
```

```
## [1] 0.61

attrition_dataset %>%
ggplot(mapping=aes(x=MonthlyIncome,y=Attrition,fill=Attrition)) +
  geom_boxplot() +
  annotate("text",x=15000,y=2.5,label=paste0('P-
value=',round(fit$p.value,4))) +
  ggtitle('Distribution of Monthly Income between two Attrition Groups')
```

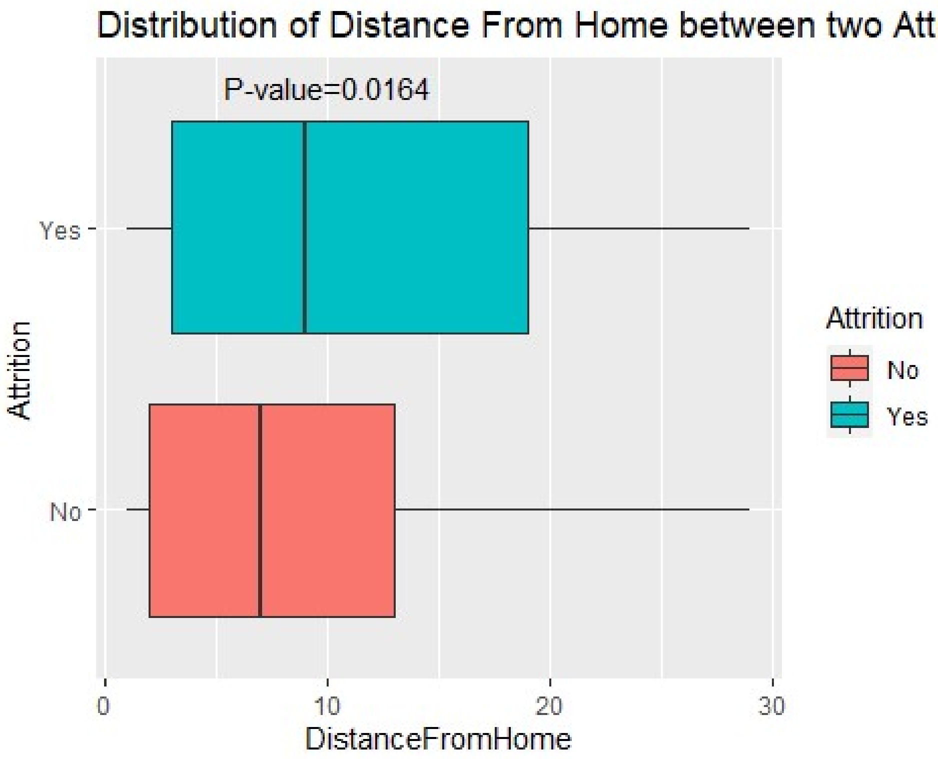


Does mean distance from home differ significantly among Attrition Groups

```
# Execute the t-test between two groups, we are assuming that the variances
are not equal
fit = t.test(
  attrition_dataset[attrition_dataset['Attrition']=='No',]$DistanceFromHome,
  attrition_dataset[attrition_dataset['Attrition']=='Yes',]$DistanceFromHome,
  var.equal = FALSE)
# P-value is less than the significance level(0.05), so we can reject the
null hypothesis and say that mean distance from home is different between two
groups
fit$p.value

## [1] 0.01640519

# Looking at the confidence intervals we can suggest that individuals who are
not leaving the company are staying close to home than the individuals who
are leaving the company
fit$conf.int
```

```
## [1] -3.4992554 -0.3574961
## attr(,"conf.level")
## [1] 0.95
```

```
attrition_dataset %>%
ggplot(mapping=aes(x=DistanceFromHome,y=Attrition,fill=Attrition)) +
  geom_boxplot() +
  annotate("text",x=10,y=2.5,label=paste0('P-value=',round(fit$p.value,4))) +
  ggtitle('Distribution of Distance From Home between two Attrition Groups')
```

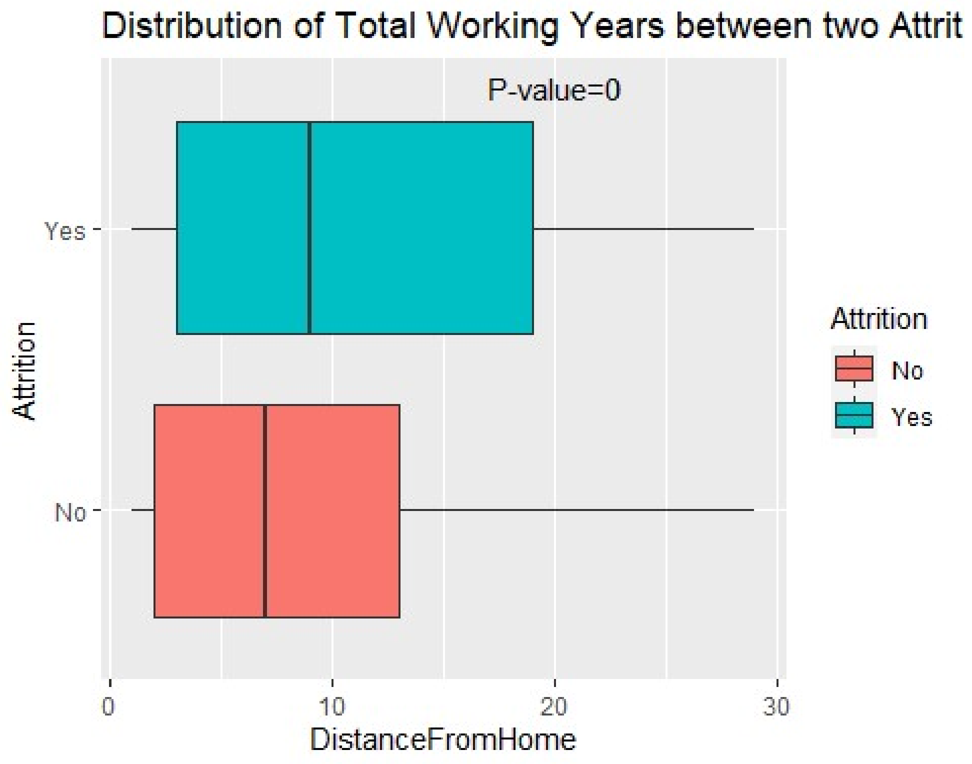


## Does mean total working years differ significantly among Attrition Groups

```
# Execute the t-test between two groups, we are assuming that the variances
are not equal
fit = t.test(
  attrition_dataset[attrition_dataset['Attrition']=='No',]$TotalWorkingYears,

attrition_dataset[attrition_dataset['Attrition']=='Yes',]$TotalWorkingYears,
  var.equal = FALSE)
# P-value is less than the significance level(0.05), so we can reject the
null hypothesis and say that mean total working years is different between
two groups
fit$p.value
```

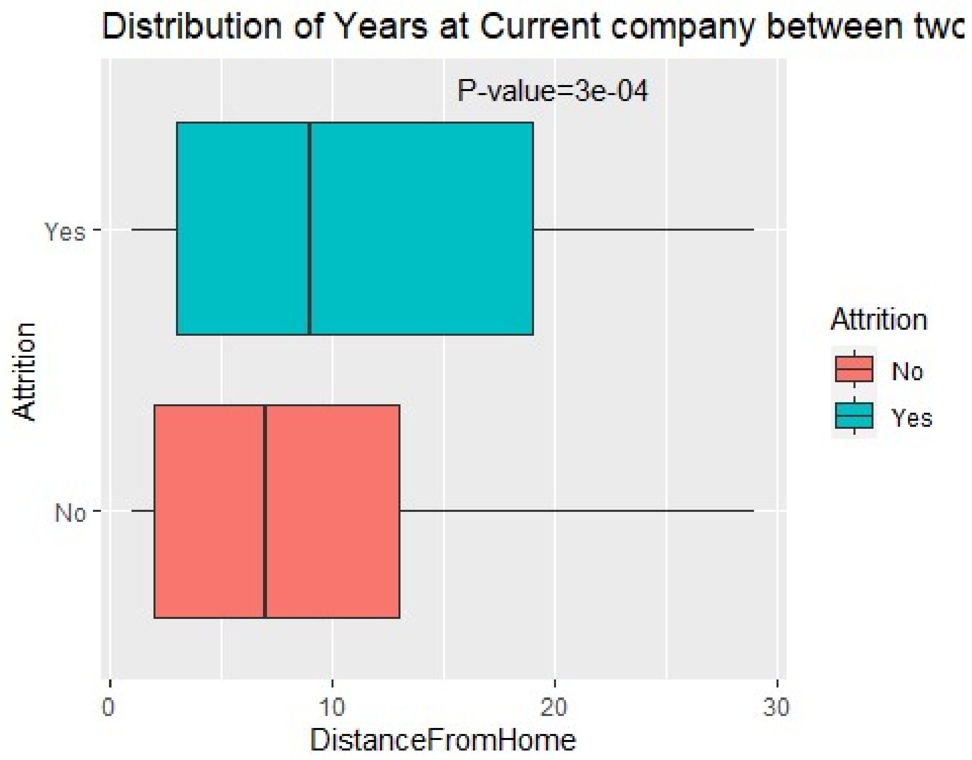
```
## [1] 6.595682e-07
```

```
attrition_dataset %>%
ggplot(mapping=aes(x=DistanceFromHome,y=Attrition,fill=Attrition)) +
  geom_boxplot() +
    annotate("text",x=20,y=2.5,label=paste0('P-value=',round(fit$p.value,4)))
+
  ggtitle('Distribution of Total Working Years between two Attrition Groups')
```



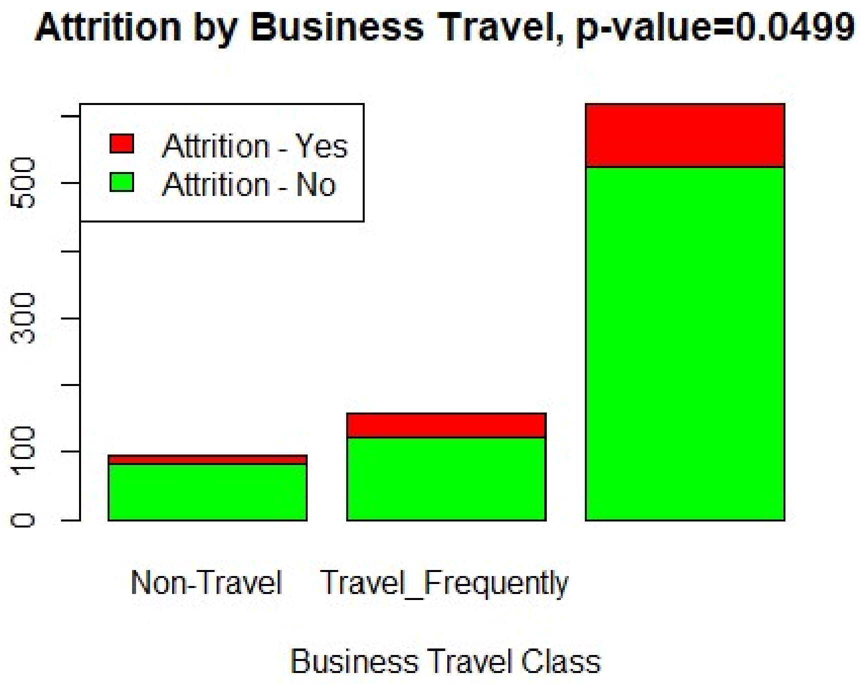
Does mean total years working at the company different significantly among Attrition Groups

```
# Execute the t-test between two groups, we are assuming that the variances
are not equal
fit = t.test(
  attrition_dataset[attrition_dataset['Attrition']=='No',]$YearsAtCompany,
  attrition_dataset[attrition_dataset['Attrition']=='Yes',]$YearsAtCompany,
  var.equal = FALSE)
# P-value is less than the significance level(0.05), so we can reject the
null hypothesis and say that mean years at current company is different
between two groups
fit$p.value

## [1] 0.0002563021

attrition_dataset %>%
ggplot(mapping=aes(x=DistanceFromHome,y=Attrition,fill=Attrition)) +
  geom_boxplot() +
  annotate("text",x=20,y=2.5,label=paste0('P-value=',round(fit$p.value,4))) +
  ggtitle('Distribution of Years at Current company between two Attrition
Groups')
```

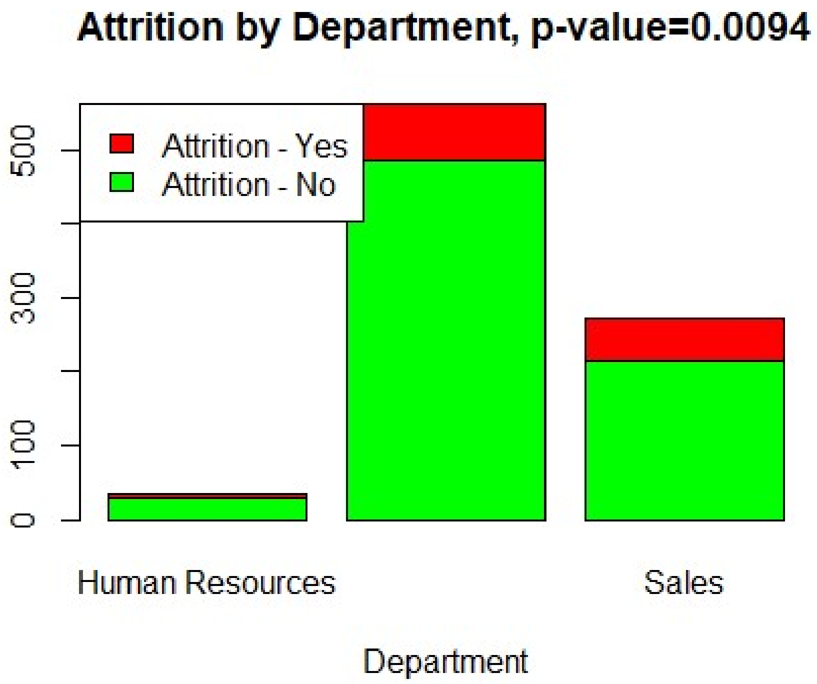


## Run categorical tests to check whether those variables are associated with attrition

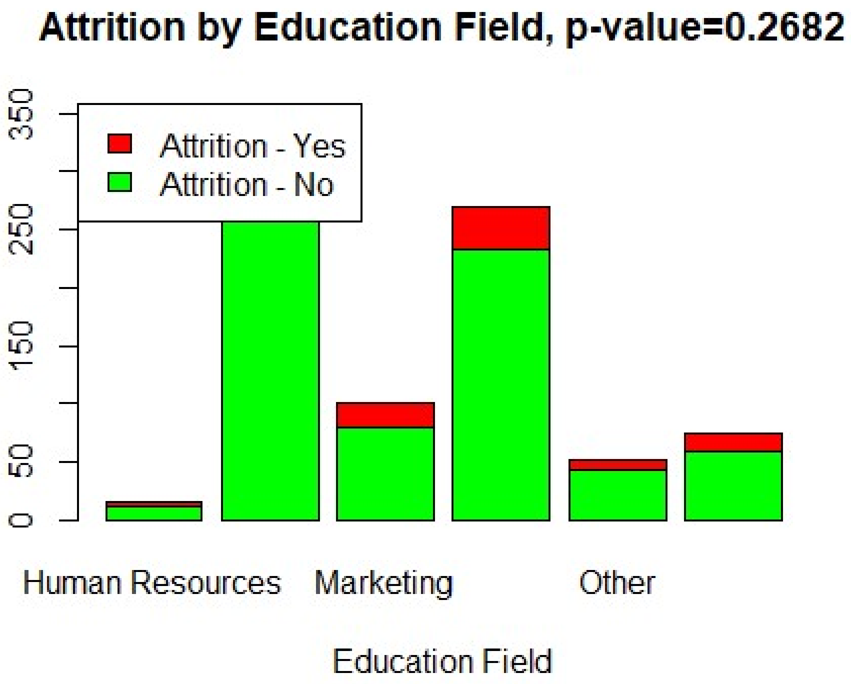
```
fit=chisq.test(table(attrition_dataset$BusinessTravel,attrition_dataset$Attri
tion))
barplot(table(attrition_dataset$Attrition,attrition_dataset$BusinessTravel),
        col = c("green","red"),
        main=paste0('Attrition by Business Travel, p-
value=',round(fit$p.value,4)),
        xlab='Business Travel Class')
legend("topleft",c("Attrition - Yes","Attrition - No"),fill =
c("red","green"))
```



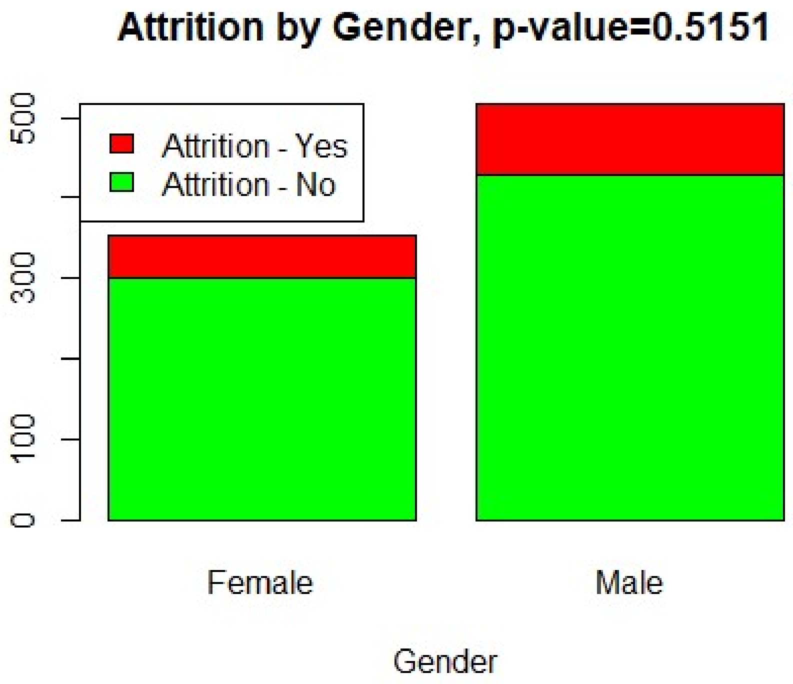
```
fit=chisq.test(table(attrition_dataset$Department,attrition_dataset$Attrition
))
barplot(table(attrition_dataset$Attrition,attrition_dataset$Department),
        col = c("green","red"),
        main=paste0('Attrition by Department, p-
value=',round(fit$p.value,4)),
        xlab='Department')
legend("topleft",c("Attrition - Yes","Attrition - No"),fill =
c("red","green"))
```



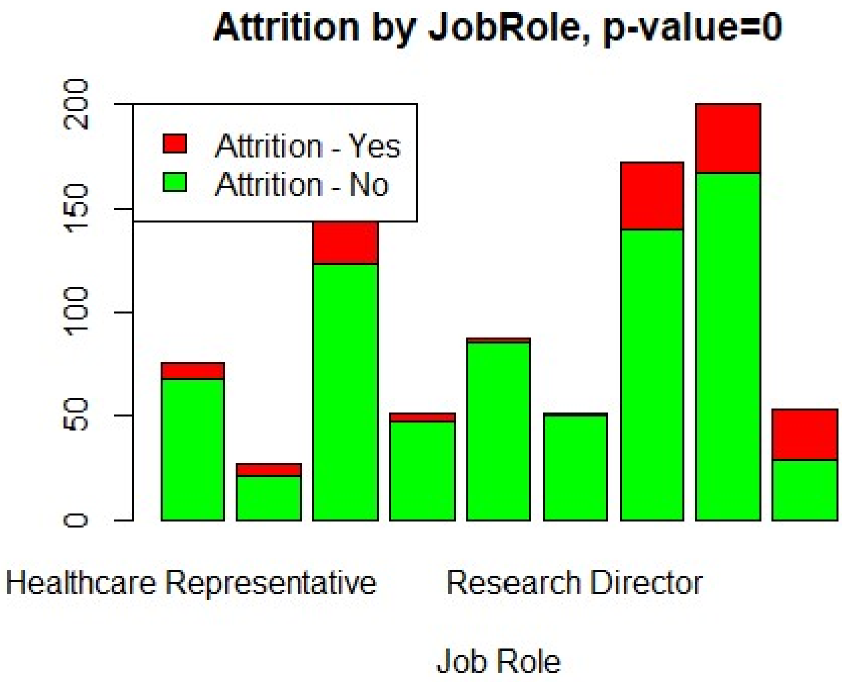
```
fit=chisq.test(table(attrition_dataset$EducationField,attrition_dataset$Attri
tion))

## Warning in chisq.test(table(attrition_dataset$EducationField,
## attrition_dataset$Attrition)): Chi-squared approximation may be incorrect

barplot(table(attrition_dataset$Attrition,attrition_dataset$EducationField),
        col = c("green","red"),
        main=paste0('Attrition by Education Field, p-
value=',round(fit$p.value,4)),
        xlab='Education Field')
legend("topleft",c("Attrition - Yes","Attrition - No"),fill =
c("red","green"))
```



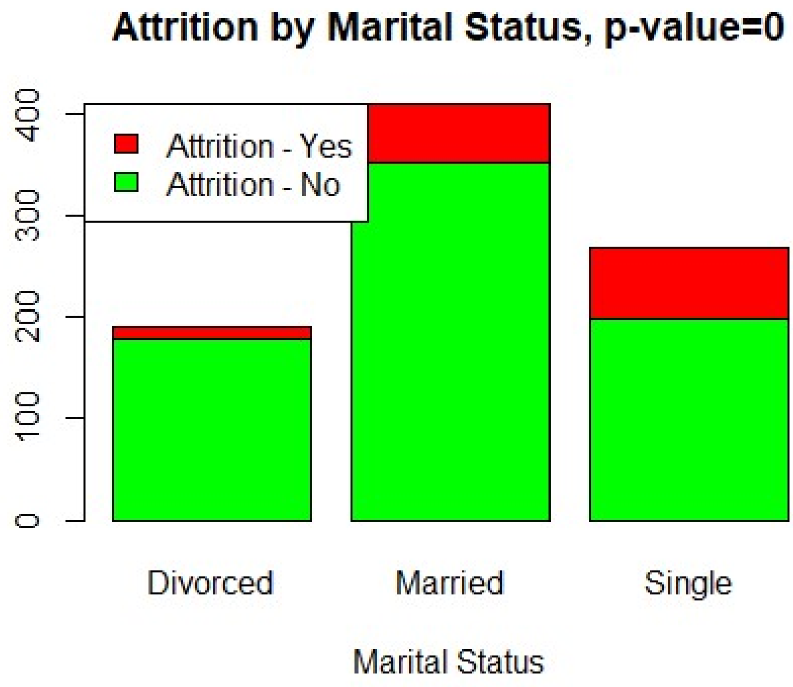
```
fit=chisq.test(table(attrition_dataset$Gender,attrition_dataset$Attrition))
barplot(table(attrition_dataset$Attrition,attrition_dataset$Gender),
        col = c("green","red"),
        main=paste0('Attrition by Gender, p-value=',round(fit$p.value,4)),
        xlab='Gender')
legend("topleft",c("Attrition - Yes","Attrition - No"),fill =
c("red","green"))
```



```
fit=chisq.test(table(attrition_dataset$JobRole,attrition_dataset$Attrition))

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

barplot(table(attrition_dataset$Attrition,attrition_dataset$JobRole),
        col = c("green","red"),
        main=paste0('Attrition by JobRole, p-value=',round(fit$p.value,4)),
        xlab='Job Role')
legend("topleft",c("Attrition - Yes","Attrition - No"),fill =
c("red","green"))
```



```r
fit=chisq.test(table(attrition_dataset$MaritalStatus,attrition_dataset$Attrition))
barplot(table(attrition_dataset$Attrition,attrition_dataset$MaritalStatus),
        col = c("green","red"),
        main=paste0('Attrition by Marital Status, p-value=',round(fit$p.value,4)),
        xlab='Marital Status')
legend("topleft",c("Attrition - Yes","Attrition - No"),fill = c("red","green"))
```



Convert categorical variables by using dummy encoding

```
clean_wrangle_dataset = function(df,test){
  ##
  input_dataset = df
  ##
  if(test==0){
    input_dataset$Attrition = as.factor(input_dataset$Attrition)
  }
  ##
  input_dataset$BusinessTravel = as.factor(input_dataset$BusinessTravel)
  input_dataset$Department = as.factor(input_dataset$Department)
  input_dataset$EducationField = as.factor(input_dataset$EducationField)
  input_dataset$Gender = as.factor(input_dataset$Gender)
  input_dataset$JobRole = as.factor(input_dataset$JobRole)
  input_dataset$MaritalStatus = as.factor(input_dataset$MaritalStatus)
  input_dataset$Over18 = as.factor(input_dataset$Over18)
  input_dataset$OverTime = as.factor(input_dataset$OverTime)

  # Shorten Department Names by replacing the values

  input_dataset$Department = str_replace(input_dataset$Department, "Research
& Development", "RnD")
  input_dataset$Department = str_replace(input_dataset$Department, "Human
Resources", "HR")
```

```
  # Change variables with 2 levels to 1 and 0
  input_dataset$Gender <- ifelse(input_dataset$Gender == "Female", 1, 0)
  input_dataset$OverTime <- ifelse(input_dataset$OverTime == "Yes", 1, 0)

  # Remove Variables not needed in the model
  input_dataset = input_dataset %>% select(-Over18)
  input_dataset = input_dataset %>% select(-EmployeeCount)
  input_dataset = input_dataset %>% select(-EmployeeNumber)
  input_dataset = input_dataset %>% select(-ID)
  input_dataset = input_dataset %>% select(-StandardHours)
  input_dataset = input_dataset %>% select(-DailyRate)
  input_dataset = input_dataset %>% select(-HourlyRate)
  input_dataset = input_dataset %>% select(-MonthlyRate)



  # Scale numeric variables
  input_dataset[,
c("Age","DistanceFromHome","Education","EnvironmentSatisfaction","JobInvolvem
ent","JobLevel","JobSatisfaction","MonthlyIncome","NumCompaniesWorked","Perce
ntSalaryHike","PerformanceRating","RelationshipSatisfaction","StockOptionLeve
l","TotalWorkingYears","TrainingTimesLastYear","WorkLifeBalance","YearsAtComp
any","YearsInCurrentRole","YearsSinceLastPromotion","YearsWithCurrManager")]
=  scale(input_dataset[,
c("Age","DistanceFromHome","Education","EnvironmentSatisfaction","JobInvolvem
ent","JobLevel","JobSatisfaction","MonthlyIncome","NumCompaniesWorked","Perce
ntSalaryHike","PerformanceRating","RelationshipSatisfaction","StockOptionLeve
l","TotalWorkingYears","TrainingTimesLastYear","WorkLifeBalance","YearsAtComp
any","YearsInCurrentRole","YearsSinceLastPromotion","YearsWithCurrManager")])


  # Add feature names and replace spaces

  input_dataset$BusinessTravel =
paste("BT_",str_replace(input_dataset$BusinessTravel," ","_"),sep="")
  input_dataset$Department =
paste("DP_",str_replace(input_dataset$Department," ","_"),sep="")
  input_dataset$EducationField =
paste("EF_",str_replace(input_dataset$EducationField," ","_"),sep="")
  input_dataset$JobRole = paste("JR_",str_replace(input_dataset$JobRole,"
","_"),sep="")
  input_dataset$MaritalStatus =
paste("MS_",str_replace(input_dataset$MaritalStatus," ","_"),sep="")


  # Dummy code categorical variables having 3 or more levels
  BusinessTravel = as.data.frame(dummy.code(input_dataset$BusinessTravel))
  Department = as.data.frame(dummy.code(input_dataset$Department))
  EducationField = as.data.frame(dummy.code(input_dataset$EducationField))
  JobRole = as.data.frame(dummy.code(input_dataset$JobRole))
  MaritalStatus = as.data.frame(dummy.code(input_dataset$MaritalStatus))


  # Add the dummy codes to the dataset
  input_dataset =
cbind(input_dataset,BusinessTravel,Department,EducationField,JobRole,MaritalS
tatus)

  # Remove the original categorical variables
  input_dataset = input_dataset %>% select(-one_of(c("BusinessTravel",
"Department", "EducationField","JobRole","MaritalStatus")))

  # Return the dataset
  return(input_dataset)
}
```

Clean and make the datasets ready for KNN algorithm

```
attrition_dataset = clean_wrangle_dataset(attrition_dataset,0)
#Move outcome to the last column
attrition_dataset$Outcome = attrition_dataset$Attrition
attrition_dataset = attrition_dataset %>% select(-Attrition)

attrition_dataset_wl = clean_wrangle_dataset(attrition_dataset_wl,1)
```

Run KNN on training set to check specificity and sensitivity of the model

```
#set.seed(1243) # set the seed to make the partition reproducible

# 80% of the sample size
smp_size <- floor(0.8 * nrow(attrition_dataset))
train_ind <- sample(seq_len(nrow(attrition_dataset)), size = smp_size)

train_df <- attrition_dataset[train_ind, ]
test_df <- attrition_dataset[-train_ind, ]
# Use SMOTE to oversample the Attrition = Yes Observations

train_df = SMOTE(Outcome~.,train_df,perc.over = 600,perc.under=100,k=10)
prop.table(table(train_df$Outcome))

##
##        No       Yes
## 0.4615385 0.5384615

classifications = knn(train_df[,c(1:46)], test_df[,c(1:46)],
train_df$Outcome,prob = T,k=10)
confusionMatrix(table(test_df$Outcome,classifications,dnn=c('Predicted','Actu
al')))
```

```
## Confusion Matrix and Statistics
## 
##           Actual
## Predicted No Yes
##       No  70  73
##       Yes 12  19
##                                           
##                 Accuracy : 0.5115         
##                   95% CI : (0.4347, 0.5879)
##      No Information Rate : 0.5287         
##      P-Value [Acc > NIR] : 0.7028         
##                                           
##                    Kappa : 0.0578         
##                                           
##  Mcnemar's Test P-Value : 7.62e-11       
##                                           
##              Sensitivity : 0.8537         
##              Specificity : 0.2065         
##           Pos Pred Value : 0.4895         
##           Neg Pred Value : 0.6129         
##               Prevalence : 0.4713         
##           Detection Rate : 0.4023         
##     Detection Prevalence : 0.8218         
##        Balanced Accuracy : 0.5301         
##                                           
##         'Positive' Class : No             
## 
```

Predict classifications on the test data set

```
# attrition_dataset_train = SMOTE(Outcome~.,attrition_dataset,perc.over =
600,perc.under=100,k=10)
# classifications = knn(attrition_dataset_train[,c(1:46)],
attrition_dataset_wl[,c(1:46)], attrition_dataset_train$Outcome,prob = F,k=5)
# write.csv(x=classifications, file='D:\\SMU_MSDS\\MSDS_6306_Doing-Data-
Science\\Unit 14 and 15 Case Study 2\\attrition_results.csv',row.names = F)
```

Run linear regression

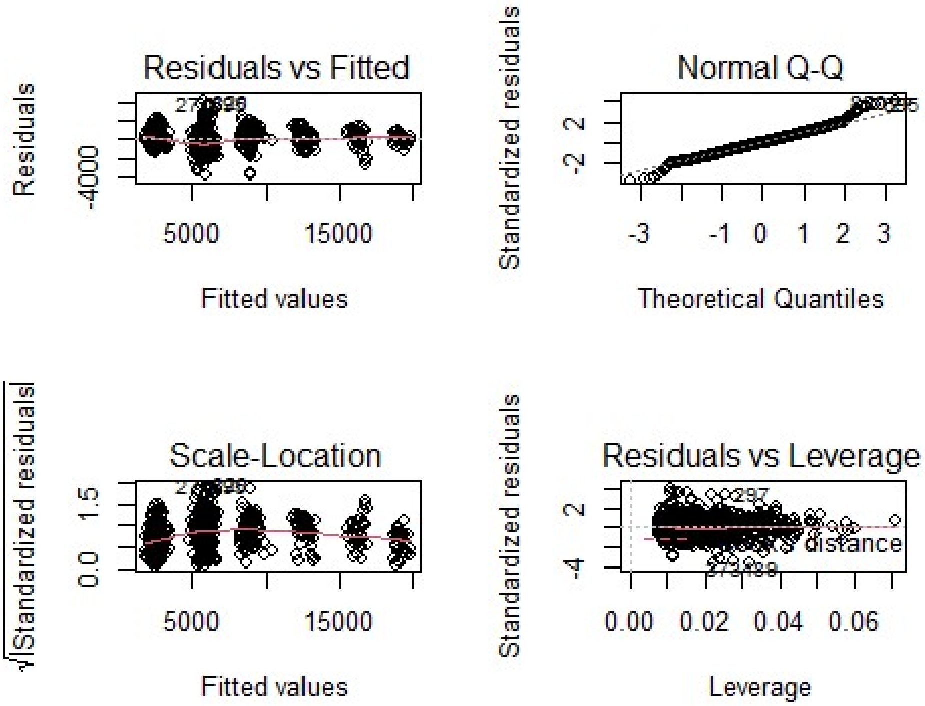
```
attrition_dataset_lm$BusinessTravel =
as.factor(attrition_dataset_lm$BusinessTravel)
attrition_dataset_lm$Department =as.factor(attrition_dataset_lm$Department)
attrition_dataset_lm$Gender = as.factor(attrition_dataset_lm$Gender)
attrition_dataset_lm$JobRole = as.factor(attrition_dataset_lm$JobRole)
attrition_dataset_lm$OverTime = as.factor(attrition_dataset_lm$OverTime)

model.full = lm(MonthlyIncome~Age+
                BusinessTravel+
                Department+
                Education+
                EducationField+
                EnvironmentSatisfaction+
                Gender+
                JobInvolvement+
                JobLevel+
                JobRole+
                JobSatisfaction+
                NumCompaniesWorked+
                OverTime+
                PercentSalaryHike+
                PerformanceRating+
                RelationshipSatisfaction+
                StockOptionLevel+
                TotalWorkingYears+
                TrainingTimesLastYear+
                WorkLifeBalance+
                YearsAtCompany+
                YearsInCurrentRole+
                YearsSinceLastPromotion+
                YearsWithCurrManager,data=attrition_dataset_lm)

model.aic.backward <- step(model.full, direction = "backward", trace = 1)
```

```
## Start:  AIC=12155.21
## MonthlyIncome ~ Age + BusinessTravel + Department + Education + 
##     EducationField + EnvironmentSatisfaction + Gender + JobInvolvement + 
##     JobLevel + JobRole + JobSatisfaction + NumCompaniesWorked + 
##     OverTime + PercentSalaryHike + PerformanceRating + 
RelationshipSatisfaction + 
##     StockOptionLevel + TotalWorkingYears + TrainingTimesLastYear + 
##     WorkLifeBalance + YearsAtCompany + YearsInCurrentRole + 
YearsSinceLastPromotion + 
##     YearsWithCurrManager
## 
##                             Df Sum of Sq       RSS   AIC
## - EducationField             5   2188361 934002638 12147
## - StockOptionLevel           1      3063 931817340 12153
## - OverTime                   1      5013 931819290 12153
## - YearsAtCompany             1     25354 931839631 12153
## - YearsInCurrentRole         1     84141 931898418 12153
## - EnvironmentSatisfaction    1    104984 931919261 12153
## - Age                        1    127371 931941648 12153
## - JobInvolvement             1    127908 931942184 12153
## - RelationshipSatisfaction   1    173320 931987597 12153
## - NumCompaniesWorked         1    458909 932273186 12154
## - JobSatisfaction            1    573107 932387383 12154
## - TrainingTimesLastYear      1    693826 932508103 12154
## - WorkLifeBalance            1    738527 932552804 12154
## - Education                  1   1079327 932893604 12154
## - Department                 2   3464790 935279067 12154
## <none>                                   931814277 12155
## - Gender                    1    2629318  934443595 12156
## - YearsWithCurrManager      1    2723629  934537906 12156
## - PercentSalaryHike         1    3093401  934907678 12156
## - YearsSinceLastPromotion   1    4468805  936283081 12157
## - PerformanceRating         1    4994816  936809093 12158
## - BusinessTravel            2   14309251  946123528 12164
## - TotalWorkingYears         1   23152293  954966570 12175
## - JobRole                   8  639534354 1571348630 12594
## - JobLevel                  1 1254725478 2186539755 12895
##
## Step:  AIC=12147.25
## MonthlyIncome ~ Age + BusinessTravel + Department + Education +
##     EnvironmentSatisfaction + Gender + JobInvolvement + JobLevel +
##     JobRole + JobSatisfaction + NumCompaniesWorked + OverTime +
##     PercentSalaryHike + PerformanceRating + RelationshipSatisfaction +
##     StockOptionLevel + TotalWorkingYears + TrainingTimesLastYear +
##     WorkLifeBalance + YearsAtCompany + YearsInCurrentRole +
YearsSinceLastPromotion +
##     YearsWithCurrManager
##
##                            Df  Sum of Sq        RSS   AIC
## - StockOptionLevel          1       1475  934004112 12145
## - OverTime                  1      11571  934014208 12145
## - YearsAtCompany            1      44460  934047098 12145
## - Age                       1      66006  934068644 12145
## - EnvironmentSatisfaction   1      93279  934095917 12145
## - RelationshipSatisfaction  1      94114  934096752 12145
## - YearsInCurrentRole        1     108600  934111237 12145
## - JobInvolvement            1     141435  934144073 12145
## - NumCompaniesWorked        1     455968  934458606 12146
## - TrainingTimesLastYear     1     474340  934476978 12146
## - JobSatisfaction           1     584562  934587200 12146
## - WorkLifeBalance           1     741738  934744376 12146
## - Education                 1    1025519  935028157 12146
## - Department                2    3219440  937222078 12146
## <none>                                   934002638 12147
## - YearsWithCurrManager      1    2637717  936640355 12148
## - Gender                    1    2684309  936686946 12148
## - PercentSalaryHike         1    2834550  936837188 12148
## - YearsSinceLastPromotion   1    4384582  938387219 12149
## - PerformanceRating         1    4665021  938667659 12150
## - BusinessTravel            2   15020987  949023625 12157
## - TotalWorkingYears         1   22511231  956513869 12166
## - JobRole                   8  638539302 1572541940 12584
## - JobLevel                  1 1263905298 2197907936 12890
##
## Step:  AIC=12145.25
## MonthlyIncome ~ Age + BusinessTravel + Department + Education +
##     EnvironmentSatisfaction + Gender + JobInvolvement + JobLevel +
##     JobRole + JobSatisfaction + NumCompaniesWorked + OverTime +
##     PercentSalaryHike + PerformanceRating + RelationshipSatisfaction +
##     TotalWorkingYears + TrainingTimesLastYear + WorkLifeBalance +
##     YearsAtCompany + YearsInCurrentRole + YearsSinceLastPromotion +
##     YearsWithCurrManager
##
##                            Df  Sum of Sq        RSS   AIC
## - OverTime                  1      11562  934015674 12143
## - YearsAtCompany            1      44894  934049006 12143
## - Age                       1      65574  934069686 12143
## - EnvironmentSatisfaction   1      92788  934096900 12143
## - RelationshipSatisfaction  1      93397  934097509 12143
## - YearsInCurrentRole        1     111499  934115611 12143
## - JobInvolvement            1     143909  934148021 12143
## - NumCompaniesWorked        1     457120  934461233 12144
## - TrainingTimesLastYear     1     476007  934480120 12144
## - JobSatisfaction           1     584445  934588558 12144
## - WorkLifeBalance           1     740271  934744383 12144
## - Education                 1    1024752  935028865 12144
## - Department                2    3218145  937222258 12144
## <none>                                    934004112 12145
## - YearsWithCurrManager      1    2639133  936643245 12146
## - Gender                    1    2689316  936693429 12146
## - PercentSalaryHike         1    2835993  936840105 12146
## - YearsSinceLastPromotion   1    4383213  938387326 12147
## - PerformanceRating         1    4670853  938674965 12148
## - BusinessTravel            2   15021284  949025396 12155
## - TotalWorkingYears         1   22511103  956515215 12164
## - JobRole                   8  638565271 1572569384 12582
## - JobLevel                  1 1264319112 2198323224 12888
##
## Step:  AIC=12143.26
## MonthlyIncome ~ Age + BusinessTravel + Department + Education +
##     EnvironmentSatisfaction + Gender + JobInvolvement + JobLevel +
##     JobRole + JobSatisfaction + NumCompaniesWorked + PercentSalaryHike +
##     PerformanceRating + RelationshipSatisfaction + TotalWorkingYears +
##     TrainingTimesLastYear + WorkLifeBalance + YearsAtCompany +
##     YearsInCurrentRole + YearsSinceLastPromotion + YearsWithCurrManager
##
##                            Df  Sum of Sq        RSS   AIC
## - YearsAtCompany            1      46383  934062057 12141
## - Age                       1      67583  934083257 12141
## - RelationshipSatisfaction  1      91890  934107564 12141
## - EnvironmentSatisfaction   1      97656  934113330 12141
## - YearsInCurrentRole        1     112579  934128253 12141
## - JobInvolvement            1     148008  934163682 12141
## - NumCompaniesWorked        1     456652  934472326 12142
## - TrainingTimesLastYear     1     484940  934500614 12142
## - JobSatisfaction           1     581620  934597294 12142
## - WorkLifeBalance           1     740094  934755768 12142
## - Education                 1    1018293  935033967 12142
## - Department                2    3212824  937228498 12142
## <none>                                    934015674 12143
## - YearsWithCurrManager      1    2630980  936646653 12144
## - Gender                    1    2689018  936704692 12144
## - PercentSalaryHike         1    2831337  936847011 12144
## - YearsSinceLastPromotion   1    4384873  938400547 12145
## - PerformanceRating         1    4663175  938678849 12146
## - BusinessTravel            2   15013172  949028846 12153
## - TotalWorkingYears         1   22560376  956576050 12162
## - JobRole                   8  638556143 1572571817 12580
## - JobLevel                  1 1264447159 2198462833 12886
## 
## Step:  AIC=12141.31
## MonthlyIncome ~ Age + BusinessTravel + Department + Education + 
##     EnvironmentSatisfaction + Gender + JobInvolvement + JobLevel + 
##     JobRole + JobSatisfaction + NumCompaniesWorked + PercentSalaryHike + 
##     PerformanceRating + RelationshipSatisfaction + TotalWorkingYears + 
##     TrainingTimesLastYear + WorkLifeBalance + YearsInCurrentRole + 
##     YearsSinceLastPromotion + YearsWithCurrManager
## 
##                             Df  Sum of Sq        RSS   AIC
## - Age                        1      60206  934122263 12139
## - YearsInCurrentRole         1      73558  934135615 12139
## - RelationshipSatisfaction   1      86548  934148605 12139
## - EnvironmentSatisfaction    1      95411  934157468 12139
## - JobInvolvement             1     162641  934224698 12140
## - TrainingTimesLastYear      1     464219  934526276 12140
## - JobSatisfaction            1     559113  934621170 12140
## - NumCompaniesWorked         1     563109  934625166 12140
## - WorkLifeBalance            1     732744  934794801 12140
## - Education                  1    1008788  935070846 12140
## - Department                 2    3238870  937300927 12140
## <none>                                    934062057 12141
## - Gender                     1    2689800  936751857 12142
## - PercentSalaryHike          1    2821084  936883141 12142
## - YearsWithCurrManager       1    3542462  937604519 12143
## - YearsSinceLastPromotion    1    4507553  938569610 12144
## - PerformanceRating          1    4641401  938703458 12144
## - BusinessTravel             2   15016362  949078419 12151
## - TotalWorkingYears          1   23873634  957935692 12161
## - JobRole                    8  639224009 1573286066 12579
## - JobLevel                   1 1273330299 2207392356 12888
## 
## Step:  AIC=12139.36
## MonthlyIncome ~ BusinessTravel + Department + Education + 
EnvironmentSatisfaction + 
##     Gender + JobInvolvement + JobLevel + JobRole + JobSatisfaction + 
##     NumCompaniesWorked + PercentSalaryHike + PerformanceRating + 
##     RelationshipSatisfaction + TotalWorkingYears + TrainingTimesLastYear + 
##     WorkLifeBalance + YearsInCurrentRole + YearsSinceLastPromotion + 
##     YearsWithCurrManager
## 
##                            Df  Sum of Sq        RSS   AIC
## - YearsInCurrentRole        1      80707  934202970 12137
## - RelationshipSatisfaction  1      85611  934207874 12137
## - EnvironmentSatisfaction   1      97018  934219282 12138
## - JobInvolvement            1     156026  934278289 12138
## - TrainingTimesLastYear     1     464773  934587037 12138
## - NumCompaniesWorked        1     535391  934657654 12138
## - JobSatisfaction           1     551692  934673956 12138
## - WorkLifeBalance           1     721655  934843919 12138
## - Department                2    3204811  937327075 12138
## - Education                 1    1113579  935235843 12138
## <none>                                    934122263 12139
## - Gender                    1    2700660  936822924 12140
## - PercentSalaryHike         1    2797788  936920051 12140
## - YearsWithCurrManager      1    3490081  937612344 12141
## - YearsSinceLastPromotion   1    4536647  938658910 12142
## - PerformanceRating         1    4608768  938731031 12142
## - BusinessTravel            2   15107121  949229385 12149
## - TotalWorkingYears         1   29571597  963693860 12164
## - JobRole                   8  639614967 1573737230 12577
## - JobLevel                  1 1279471476 2213593740 12888
## 
## Step:  AIC=12137.44
## MonthlyIncome ~ BusinessTravel + Department + Education + 
EnvironmentSatisfaction + 
##     Gender + JobInvolvement + JobLevel + JobRole + JobSatisfaction + 
##     NumCompaniesWorked + PercentSalaryHike + PerformanceRating + 
##     RelationshipSatisfaction + TotalWorkingYears + TrainingTimesLastYear + 
##     WorkLifeBalance + YearsSinceLastPromotion + YearsWithCurrManager
## 
##                            Df  Sum of Sq        RSS   AIC
## - RelationshipSatisfaction  1      87624  934290595 12136
## - EnvironmentSatisfaction   1      89384  934292354 12136
## - JobInvolvement            1     162881  934365851 12136
## - TrainingTimesLastYear     1     461468  934664439 12136
## - NumCompaniesWorked        1     498728  934701699 12136
## - JobSatisfaction           1     555540  934758511 12136
## - WorkLifeBalance           1     684204  934887174 12136
## - Department                2    3197456  937400427 12136
## - Education                 1    1127285  935330256 12136
## <none>                                    934202970 12137
## - Gender                    1    2667998  936870968 12138
## - PercentSalaryHike         1    2813352  937016323 12138
## - YearsWithCurrManager      1    4053911  938256882 12139
## - PerformanceRating         1    4596343  938799314 12140
## - YearsSinceLastPromotion   1    5185810  939388781 12140
## - BusinessTravel            2   15257474  949460445 12148
## - TotalWorkingYears         1   30708719  964911689 12164
## - JobRole                  8  640216778 1574419748 12576
## - JobLevel                 1 1280125017 2214327987 12886
##
## Step:  AIC=12135.52
## MonthlyIncome ~ BusinessTravel + Department + Education + 
EnvironmentSatisfaction + 
##     Gender + JobInvolvement + JobLevel + JobRole + JobSatisfaction + 
##     NumCompaniesWorked + PercentSalaryHike + PerformanceRating + 
##     TotalWorkingYears + TrainingTimesLastYear + WorkLifeBalance + 
##     YearsSinceLastPromotion + YearsWithCurrManager
## 
##                            Df  Sum of Sq        RSS   AIC
## - EnvironmentSatisfaction  1      90448  934381043 12134
## - JobInvolvement           1     167861  934458455 12134
## - TrainingTimesLastYear    1     468392  934758987 12134
## - NumCompaniesWorked       1     524253  934814848 12134
## - JobSatisfaction          1     542238  934832833 12134
## - WorkLifeBalance          1     669703  934960298 12134
## - Department               2    3186222  937476817 12134
## - Education                1    1143383  935433977 12135
## <none>                                   934290595 12136
## - Gender                   1    2683039  936973634 12136
## - PercentSalaryHike        1    2788558  937079152 12136
## - YearsWithCurrManager     1    4074126  938364720 12137
## - PerformanceRating        1    4597944  938888539 12138
## - YearsSinceLastPromotion  1    5271089  939561683 12138
## - BusinessTravel           2   15183961  949474556 12146
## - TotalWorkingYears        1   30622610  964913204 12162
## - JobRole                  8  640255663 1574546257 12574
## - JobLevel                 1 1280577371 2214867966 12884
##
## Step:  AIC=12133.6
## MonthlyIncome ~ BusinessTravel + Department + Education + Gender + 
##     JobInvolvement + JobLevel + JobRole + JobSatisfaction + 
NumCompaniesWorked + 
##     PercentSalaryHike + PerformanceRating + TotalWorkingYears + 
##     TrainingTimesLastYear + WorkLifeBalance + YearsSinceLastPromotion + 
##     YearsWithCurrManager
## 
##                            Df  Sum of Sq        RSS   AIC
## - JobInvolvement           1     166174  934547216 12132
## - TrainingTimesLastYear    1     475544  934856586 12132
## - NumCompaniesWorked       1     513481  934894523 12132
## - JobSatisfaction          1     550060  934931103 12132
## - WorkLifeBalance          1     714729  935095772 12132
## - Department               2    3135929  937516972 12132
## - Education                1    1125593  935506636 12133
## <none>                                   934381043 12134
## - Gender                   1    2663265  937044307 12134
## - PercentSalaryHike        1    2775796  937156838 12134
## - YearsWithCurrManager     1    4049332  938430375 12135
## - PerformanceRating        1    4576288  938957331 12136
## - YearsSinceLastPromotion  1    5239083  939620125 12136
## - BusinessTravel           2   15176270  949557313 12144
## - TotalWorkingYears        1   30772196  965153239 12160
## - JobRole                  8  646149738 1580530780 12575
## - JobLevel                 1 1282553207 2216934250 12883
## 
## Step:  AIC=12131.76
## MonthlyIncome ~ BusinessTravel + Department + Education + Gender + 
##     JobLevel + JobRole + JobSatisfaction + NumCompaniesWorked + 
##     PercentSalaryHike + PerformanceRating + TotalWorkingYears + 
##     TrainingTimesLastYear + WorkLifeBalance + YearsSinceLastPromotion + 
##     YearsWithCurrManager
## 
##                            Df  Sum of Sq        RSS   AIC
## - TrainingTimesLastYear    1     462929  935010145 12130
## - NumCompaniesWorked       1     508433  935055649 12130
## - JobSatisfaction          1     515180  935062396 12130
## - WorkLifeBalance          1     705094  935252310 12130
## - Department               2    3157508  937704724 12131
## - Education                1    1107352  935654568 12131
## <none>                                  934547216 12132
## - Gender                   1    2699867  937247083 12132
## - PercentSalaryHike        1    2779368  937326584 12132
## - YearsWithCurrManager     1    3991098  938538315 12134
## - PerformanceRating        1    4572419  939119635 12134
## - YearsSinceLastPromotion  1    5182906  939730122 12135
## - BusinessTravel           2   15346441  949893658 12142
## - TotalWorkingYears        1   30732900  965280116 12158
## - JobRole                  8  649418679 1583965896 12575
## - JobLevel                 1 1284903930 2219451146 12882
## 
## Step:  AIC=12130.19
## MonthlyIncome ~ BusinessTravel + Department + Education + Gender + 
##     JobLevel + JobRole + JobSatisfaction + NumCompaniesWorked + 
##     PercentSalaryHike + PerformanceRating + TotalWorkingYears + 
##     WorkLifeBalance + YearsSinceLastPromotion + YearsWithCurrManager
## 
##                            Df  Sum of Sq        RSS   AIC
## - NumCompaniesWorked       1     455369  935465514 12129
## - JobSatisfaction          1     481044  935491189 12129
## - WorkLifeBalance          1     679194  935689339 12129
## - Department               2    3192432  938202577 12129
## - Education                1    1175777  936185922 12129
## <none>                                  935010145 12130
## - Gender                   1    2693652  937703797 12131
## - PercentSalaryHike        1    2838961  937849106 12131
## - YearsWithCurrManager     1    3917100  938927245 12132
## - PerformanceRating        1    4660379  939670524 12132
## - YearsSinceLastPromotion  1    5044764  940054909 12133
## - BusinessTravel           2   15390450  950400596 12140
## - TotalWorkingYears        1   30861379  965871524 12156
## - JobRole                  8  650046383 1585056528 12573
## - JobLevel                 1 1284675894 2219686039 12880
##
## Step:  AIC=12128.61
## MonthlyIncome ~ BusinessTravel + Department + Education + Gender +
##     JobLevel + JobRole + JobSatisfaction + PercentSalaryHike +
##     PerformanceRating + TotalWorkingYears + WorkLifeBalance +
##     YearsSinceLastPromotion + YearsWithCurrManager
##
##                           Df  Sum of Sq        RSS   AIC
## - JobSatisfaction          1     421477  935886991 12127
## - WorkLifeBalance          1     648770  936114284 12127
## - Department               2    3135967  938601481 12128
## - Education                1    1003975  936469489 12128
## <none>                                   935465514 12129
## - Gender                   1    2660866  938126380 12129
## - PercentSalaryHike        1    2814843  938280356 12129
## - PerformanceRating        1    4670907  940136420 12131
## - YearsSinceLastPromotion  1    4741665  940207179 12131
## - YearsWithCurrManager     1    4854286  940319800 12131
## - BusinessTravel           2   15137039  950602553 12139
## - TotalWorkingYears        1   37423188  972888702 12161
## - JobRole                  8  649602115 1585067629 12571
## - JobLevel                 1 1287418605 2222884119 12880
##
## Step:  AIC=12127.01
## MonthlyIncome ~ BusinessTravel + Department + Education + Gender +
##     JobLevel + JobRole + PercentSalaryHike + PerformanceRating +
##     TotalWorkingYears + WorkLifeBalance + YearsSinceLastPromotion +
##     YearsWithCurrManager
##
##                           Df  Sum of Sq        RSS   AIC
## - WorkLifeBalance          1     671894  936558885 12126
## - Department               2    3108168  938995159 12126
## - Education                1     981158  936868149 12126
## <none>                                   935886991 12127
## - Gender                   1    2710035  938597027 12128
## - PercentSalaryHike        1    2866513  938753504 12128
## - YearsSinceLastPromotion  1    4682723  940569714 12129
## - PerformanceRating        1    4741301  940628293 12129
## - YearsWithCurrManager     1    4760429  940647420 12129
## - BusinessTravel           2   14984008  950870999 12137
## - TotalWorkingYears        1   37261608  973148599 12159
## - JobRole                  8  649219309 1585106300 12569
## - JobLevel                 1 1287190977 2223077969 12878
##
## Step:  AIC=12125.63
## MonthlyIncome ~ BusinessTravel + Department + Education + Gender + 
##     JobLevel + JobRole + PercentSalaryHike + PerformanceRating + 
##     TotalWorkingYears + YearsSinceLastPromotion + YearsWithCurrManager
## 
##                           Df  Sum of Sq        RSS   AIC
## - Department               2    3044302  939603187 12124
## - Education                1     997785  937556671 12125
## <none>                                   936558885 12126
## - Gender                   1    2768792  939327677 12126
## - PercentSalaryHike        1    2873679  939432564 12126
## - YearsSinceLastPromotion  1    4549174  941108059 12128
## - YearsWithCurrManager     1    4723230  941282115 12128
## - PerformanceRating        1    4794127  941353012 12128
## - BusinessTravel           2   15152301  951711186 12136
## - TotalWorkingYears        1   37286211  973845096 12158
## - JobRole                  8  649097073 1585655959 12568
## - JobLevel                 1 1286772821 2223331707 12876
## 
## Step:  AIC=12124.45
## MonthlyIncome ~ BusinessTravel + Education + Gender + JobLevel + 
##     JobRole + PercentSalaryHike + PerformanceRating + TotalWorkingYears + 
##     YearsSinceLastPromotion + YearsWithCurrManager
## 
##                           Df  Sum of Sq        RSS   AIC
## - Education                1     964753  940567941 12123
## <none>                                   939603187 12124
## - PercentSalaryHike        1    2811590  942414777 12125
## - Gender                   1    2873416  942476604 12125
## - PerformanceRating        1    4624294  944227482 12127
## - YearsSinceLastPromotion  1    4795892  944399079 12127
## - YearsWithCurrManager     1    4901852  944505039 12127
## - BusinessTravel           2   14577582  954180769 12134
## - TotalWorkingYears        1   36362628  975965815 12156
## - JobRole                  8  678333126 1617936314 12581
## - JobLevel                 1 1297921805 2237524993 12877
## 
## Step:  AIC=12123.35
## MonthlyIncome ~ BusinessTravel + Gender + JobLevel + JobRole + 
##     PercentSalaryHike + PerformanceRating + TotalWorkingYears + 
##     YearsSinceLastPromotion + YearsWithCurrManager
## 
##                           Df  Sum of Sq        RSS   AIC
## <none>                                   940567941 12123
## - PercentSalaryHike        1    2694685  943262626 12124
## - Gender                   1    2883603  943451543 12124
## - PerformanceRating        1    4452833  945020774 12126
## - YearsSinceLastPromotion  1    4775586  945343526 12126
## - YearsWithCurrManager     1    4975294  945543234 12126
## - BusinessTravel           2   14477895  955045836 12133
## - TotalWorkingYears        1   35742372  976310313 12154
## - JobRole                   8  677398331 1617966271 12579
## - JobLevel                  1 1300663927 2241231868 12877
```

```
fit=lm(MonthlyIncome ~ BusinessTravel + Gender + JobLevel + JobRole +
  PercentSalaryHike + PerformanceRating + TotalWorkingYears +
  YearsSinceLastPromotion + YearsWithCurrManager,data=attrition_dataset_lm)
summary(fit)
```

```
##
## Call:
## lm(formula = MonthlyIncome ~ BusinessTravel + Gender + JobLevel +
##     JobRole + PercentSalaryHike + PerformanceRating + TotalWorkingYears +
##     YearsSinceLastPromotion + YearsWithCurrManager, data =
attrition_dataset_lm)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3728.8  -636.5   -11.9   627.8  4121.8
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                        209.57     424.91   0.493 0.621986
## BusinessTravelTravel_Frequently   226.20     138.64   1.632 0.103153
## BusinessTravelTravel_Rarely       396.70     117.09   3.388 0.000736 ***
## GenderMale                         118.50      73.32   1.616 0.106423
## JobLevel                          2785.86      81.16  34.325  < 2e-16 ***
## JobRoleHuman Resources            -367.63     249.43  -1.474 0.140882
## JobRoleLaboratory Technician      -606.90     167.79  -3.617 0.000315 ***
## JobRoleManager                    4016.47     228.33  17.591  < 2e-16 ***
## JobRoleManufacturing Director      157.44     166.13   0.948 0.343562
## JobRoleResearch Director          4007.65     214.07  18.721  < 2e-16 ***
## JobRoleResearch Scientist         -356.06     167.77  -2.122 0.034101 *
## JobRoleSales Executive             -47.75     143.02  -0.334 0.738557
## JobRoleSales Representative       -450.59     211.05  -2.135 0.033044 *
## PercentSalaryHike                   24.28      15.54   1.562 0.118577
## PerformanceRating                 -319.71     159.19  -2.008 0.044920 *
## TotalWorkingYears                   48.25       8.48   5.690 1.75e-08 ***
## YearsSinceLastPromotion             28.53      13.71   2.080 0.037835 *
## YearsWithCurrManager               -26.27      12.38  -2.123 0.034048 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1051 on 852 degrees of freedom
## Multiple R-squared:  0.9488, Adjusted R-squared:  0.9478
## F-statistic: 928.7 on 17 and 852 DF,  p-value: < 2.2e-16
```

```
par(mfrow=c(2,2))
plot(fit)
```



Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.