HR_Analytics_A20392859_A20392402.R

mohan

Sat Nov 25 03:48:54 2017

```
####################### HR Analytics (A20392859) (A20392402) ############################
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.4.2
```

```
library(corrplot)
```

```
## Warning: package 'corrplot' was built under R version 3.4.2
```

```
## corrplot 0.84 loaded
```

```
library(magrittr)
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(leaps)
```

```
## Warning: package 'leaps' was built under R version 3.4.2
```

```
library(lars)
```

```
## Loaded lars 1.2
```

```
library(glmnet)
```

```
## Warning: package 'glmnet' was built under R version 3.4.2
```

```
## Loading required package: Matrix
```

```
## Loading required package: foreach
```

```
## Warning: package 'foreach' was built under R version 3.4.2
```

```
## Loaded glmnet 2.0-13
```

```
library(caret)
```

```
## Warning: package 'caret' was built under R version 3.4.2
```

```
## Loading required package: lattice
```

```
library(ROCR)
```

```
## Warning: package 'ROCR' was built under R version 3.4.2
```

```
## Loading required package: gplots
```

```
## Warning: package 'gplots' was built under R version 3.4.2

##
## Attaching package: 'gplots'

## The following object is masked from 'package:stats':
##
##     lowess
```

```r
library(rpart)
```

```
## Warning: package 'rpart' was built under R version 3.4.2
```

```r
library(randomForest)
```

```
## Warning: package 'randomForest' was built under R version 3.4.2

## randomForest 4.6-12

## Type rfNews() to see new features/changes/bug fixes.

##
## Attaching package: 'randomForest'

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

## The following object is masked from 'package:ggplot2':
##
##     margin
```

```r
library(pROC)
```

```
## Warning: package 'pROC' was built under R version 3.4.2

## Type 'citation("pROC")' for a citation.

##
## Attaching package: 'pROC'

## The following object is masked from 'package:glmnet':
##
##     auc

## The following objects are masked from 'package:stats':
##
##     cov, smooth, var
```

```r
library(e1071)
```

```
## Warning: package 'e1071' was built under R version 3.4.2
```

```r
# Step1: Loading the Data

HR_comma_sep <- read.csv("C:/Mohan/IITC/Fall 2017/CS584/Project/data/HR_comma_sep.csv")
HR_comma_sep<-data.frame(HR_comma_sep)


# Step2: Data Cleaning

## (2a). Renaming the variables names for irrelevant columns
```

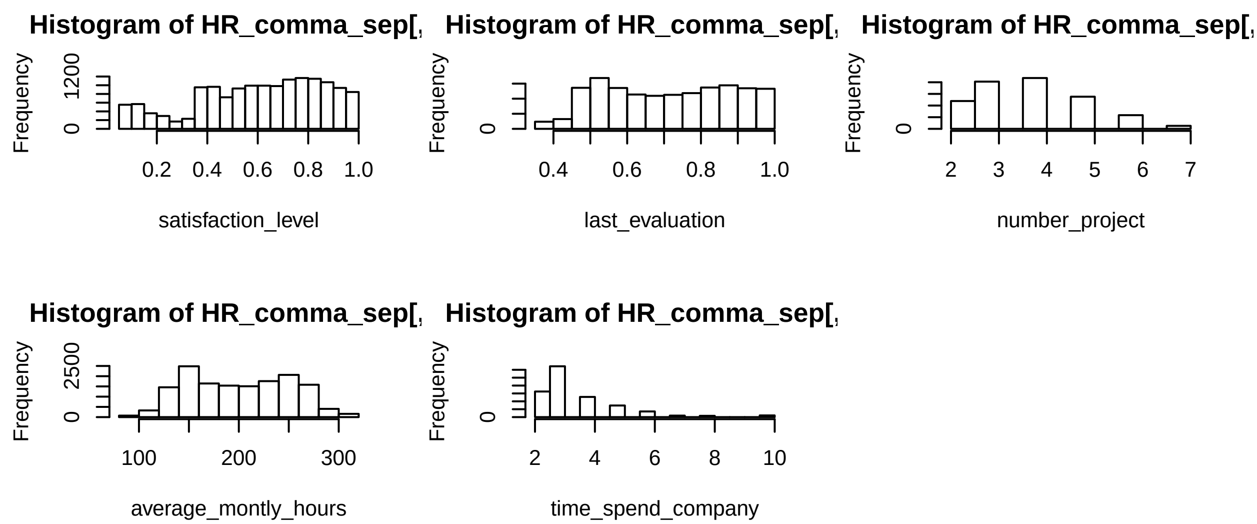
```r
colnames(HR_comma_sep)[9]<-"Department"

## (2b). Adding unique identifier for each employee

HR_comma_sep["ID"]<-seq.int(nrow(HR_comma_sep))
length(HR_comma_sep)
```

```
## [1] 11
```

```r
HR_comma_sep<-HR_comma_sep[colnames(HR_comma_sep)[c(11,1:10)]]

## (2c). Finding the NA values in the table

sum(is.na(HR_comma_sep))
```

```
## [1] 0
```

```r
# Step3: Exploring the Data

## (3a). Converting the variables to proper data type

HR_comma_sep$left=as.factor(HR_comma_sep$left)
HR_comma_sep$salary<-as.factor(HR_comma_sep$salary)
HR_comma_sep$Work_accident<-as.factor(HR_comma_sep$Work_accident)
HR_comma_sep$Department<-as.factor(HR_comma_sep$Department)
HR_comma_sep$promotion_last_5years<-as.factor(HR_comma_sep$promotion_last_5years)

## (3b). Converting the salary to ordinal variable

HR_comma_sep$salary<-ordered(HR_comma_sep$salary,levels=c("low","medium","high"))

## (3c). Finding the distribution for numeric variables

par(mfrow=c(3,3))
for(i in c(2:6)){hist(HR_comma_sep[,i],xlab=names(HR_comma_sep)[i])}
par(mfrow=c(1,1))
```

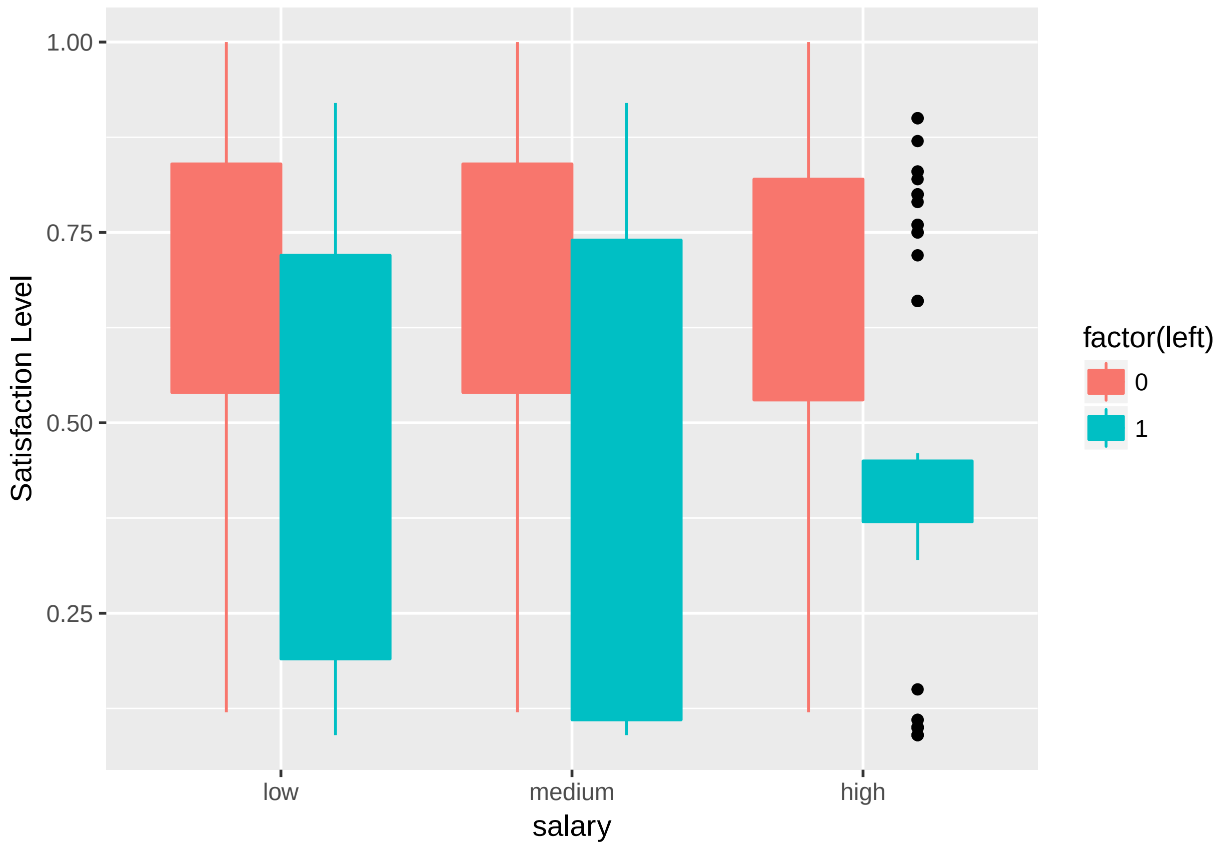



```r
## (3d). finding the descriptive statistics

summary(HR_comma_sep)
```

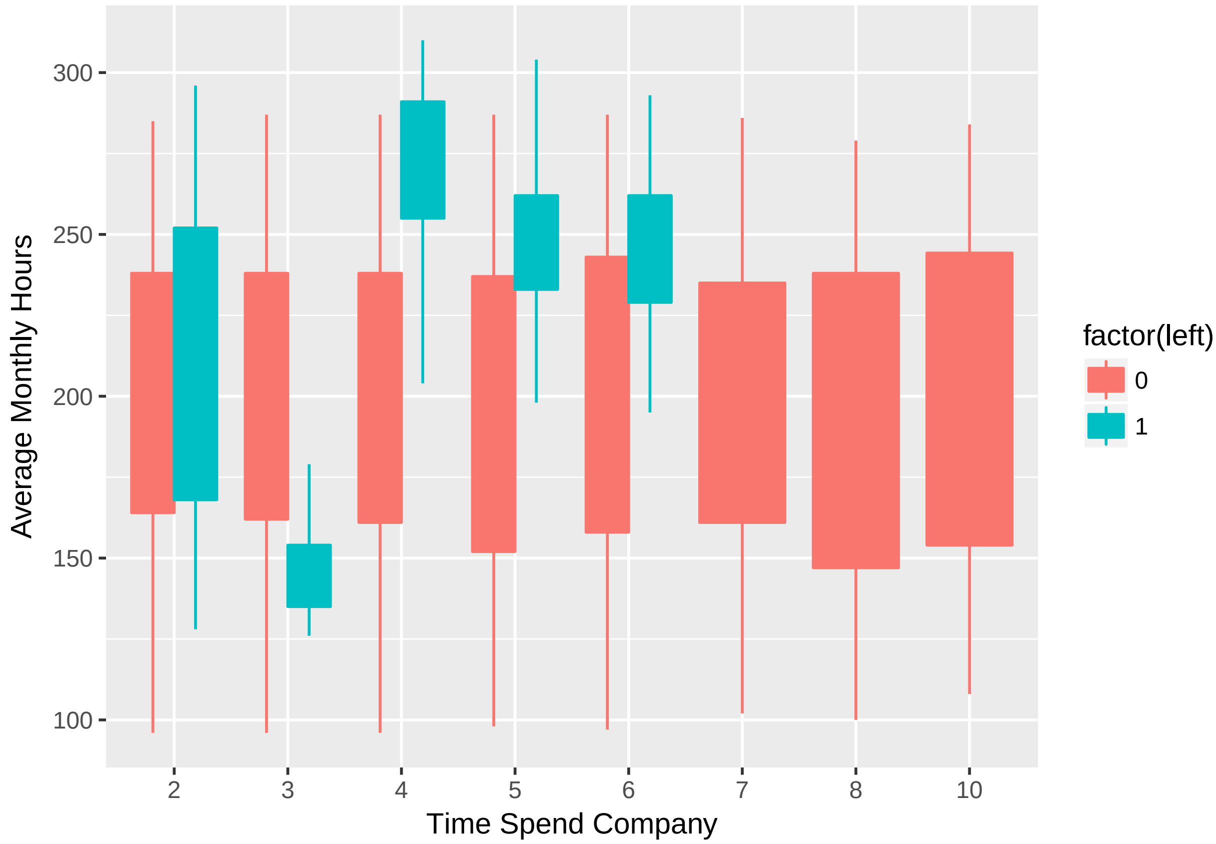
```
##        ID        satisfaction_level last_evaluation  number_project
##  Min.   :    1   Min.   :0.0900     Min.   :0.3600   Min.   :2.000
##  1st Qu.: 3750   1st Qu.:0.4400     1st Qu.:0.5600   1st Qu.:3.000
##  Median : 7500   Median :0.6400     Median :0.7200   Median :4.000
##  Mean   : 7500   Mean   :0.6128     Mean   :0.7161   Mean   :3.803
##  3rd Qu.:11250   3rd Qu.:0.8200     3rd Qu.:0.8700   3rd Qu.:5.000
##  Max.   :14999   Max.   :1.0000     Max.   :1.0000   Max.   :7.000
##
##  average_montly_hours time_spend_company Work_accident left
##  Min.   : 96.0        Min.   : 2.000     0:12830       0:11428
##  1st Qu.:156.0        1st Qu.: 3.000     1: 2169       1: 3571
##  Median :200.0        Median : 3.000
##  Mean   :201.1        Mean   : 3.498
##  3rd Qu.:245.0        3rd Qu.: 4.000
##  Max.   :310.0        Max.   :10.000
##
##  promotion_last_5years       Department      salary
##  0:14680                sales      :4140   low   :7316
##  1:  319                technical  :2720   medium:6446
##                         support    :2229   high  :1237
##                         IT         :1227
##                         product_mng: 902
```

```
##                          marketing  : 858
##                          (Other)    :2923
```

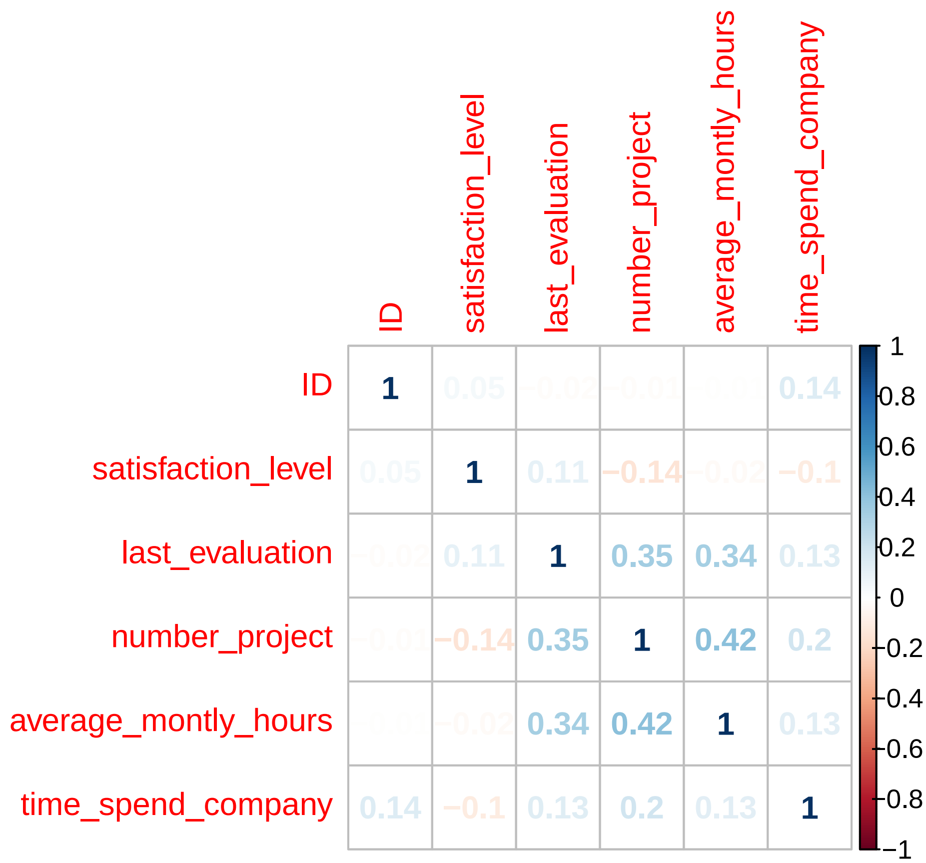
```
## (3e). Finding distributions for variables

ggplot(HR_comma_sep,aes(x=salary,y=satisfaction_level,fill=factor(left),
                        colour=factor(left)))+geom_boxplot(outlier.colour = "black")+
  xlab("salary")+ylab("Satisfaction Level")
```



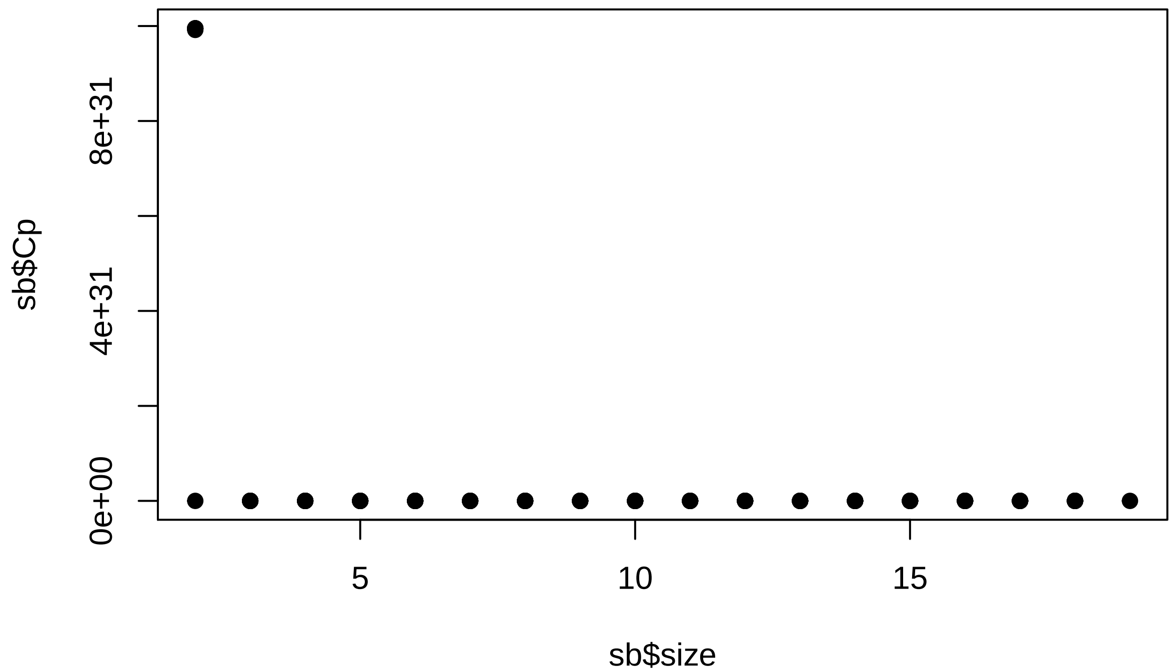

```
ggplot(HR_comma_sep,aes(x=factor(time_spend_company),y=average_montly_hours,
                        fill=factor(left),colour=factor(left)))+
  geom_boxplot(outlier.colour = NA)+xlab("Time Spend Company")+
  ylab("Average Monthly Hours")
```




```
## (3f). Finding the correlation between variables

nums<-sapply(HR_comma_sep,is.numeric)
cor_matrix<-cor(HR_comma_sep[,nums])
corrplot(cor_matrix,method = 'number')
```




```
HR_Corr<-HR_comma_sep %>% select(satisfaction_level:promotion_last_5years)

# 4. Model selection and Fitting:

## (4a). Model Selection using Cp

model.mat<-model.matrix(left~satisfaction_level+last_evaluation+number_project+
                          average_montly_hours+time_spend_company+Work_accident+
                          promotion_last_5years+Department+salary,data=HR_comma_sep)
sb<-leaps(x=model.mat[,2:19],y=HR_comma_sep[,7],method = 'Cp')
plot(sb$size,sb$Cp,pch=19)
```




```r
sb$which[which(sb$Cp==min(sb$Cp)),]
```

```
##     1     2     3     4     5     6     7     8     9     A     B     C
##  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE FALSE FALSE  TRUE FALSE  TRUE
##     D     E     F     G     H     I
##  TRUE FALSE FALSE FALSE  TRUE  TRUE
```

```r
## (4b). Forward selection and Backward Selection

fit.forward = regsubsets(left~satisfaction_level+last_evaluation+number_project
                         +average_montly_hours+time_spend_company+Work_accident
                         +promotion_last_5years+Department+salary,
                         data = HR_comma_sep,nvmax = 18,method = "forward")
summary(fit.forward)
```

```
## Subset selection object
## Call: regsubsets.formula(left ~ satisfaction_level + last_evaluation +
##     number_project + average_montly_hours + time_spend_company +
##     Work_accident + promotion_last_5years + Department + salary,
##     data = HR_comma_sep, nvmax = 18, method = "forward")
## 18 Variables  (and intercept)
##                          Forced in Forced out
## satisfaction_level           FALSE      FALSE
## last_evaluation              FALSE      FALSE
## number_project               FALSE      FALSE
## average_montly_hours         FALSE      FALSE
## time_spend_company           FALSE      FALSE
```

```
## Work_accident1            FALSE      FALSE
## promotion_last_5years1    FALSE      FALSE
## Departmenthr              FALSE      FALSE
## DepartmentIT              FALSE      FALSE
## Departmentmanagement      FALSE      FALSE
## Departmentmarketing       FALSE      FALSE
## Departmentproduct_mng     FALSE      FALSE
## DepartmentRandD           FALSE      FALSE
## Departmentsales           FALSE      FALSE
## Departmentsupport         FALSE      FALSE
## Departmenttechnical       FALSE      FALSE
## salary.L                  FALSE      FALSE
## salary.Q                  FALSE      FALSE
## 1 subsets of each size up to 18
## Selection Algorithm: forward
##           satisfaction_level last_evaluation number_project
## 1  ( 1 )  "*"                " "             " "
## 2  ( 1 )  "*"                " "             " "
## 3  ( 1 )  "*"                " "             " "
## 4  ( 1 )  "*"                " "             " "
## 5  ( 1 )  "*"                " "             "*"
## 6  ( 1 )  "*"                " "             "*"
## 7  ( 1 )  "*"                " "             "*"
## 8  ( 1 )  "*"                " "             "*"
## 9  ( 1 )  "*"                "*"             "*"
## 10  ( 1 ) "*"                "*"             "*"
## 11  ( 1 ) "*"                "*"             "*"
## 12  ( 1 ) "*"                "*"             "*"
## 13  ( 1 ) "*"                "*"             "*"
## 14  ( 1 ) "*"                "*"             "*"
## 15  ( 1 ) "*"                "*"             "*"
## 16  ( 1 ) "*"                "*"             "*"
## 17  ( 1 ) "*"                "*"             "*"
## 18  ( 1 ) "*"                "*"             "*"
##           average_montly_hours time_spend_company Work_accident1
## 1  ( 1 )  " "                  " "                " "
## 2  ( 1 )  " "                  " "                " "
## 3  ( 1 )  " "                  " "                "*"
## 4  ( 1 )  " "                  "*"                "*"
## 5  ( 1 )  " "                  "*"                "*"
## 6  ( 1 )  "*"                  "*"                "*"
## 7  ( 1 )  "*"                  "*"                "*"
## 8  ( 1 )  "*"                  "*"                "*"
## 9  ( 1 )  "*"                  "*"                "*"
## 10  ( 1 ) "*"                  "*"                "*"
## 11  ( 1 ) "*"                  "*"                "*"
## 12  ( 1 ) "*"                  "*"                "*"
## 13  ( 1 ) "*"                  "*"                "*"
## 14  ( 1 ) "*"                  "*"                "*"
## 15  ( 1 ) "*"                  "*"                "*"
## 16  ( 1 ) "*"                  "*"                "*"
## 17  ( 1 ) "*"                  "*"                "*"
## 18  ( 1 ) "*"                  "*"                "*"
##           promotion_last_5years1 Departmenthr DepartmentIT
```

```
## 1  ( 1 )  " "                   " "          " "
## 2  ( 1 )  " "                   " "          " "
## 3  ( 1 )  " "                   " "          " "
## 4  ( 1 )  " "                   " "          " "
## 5  ( 1 )  " "                   " "          " "
## 6  ( 1 )  " "                   " "          " "
## 7  ( 1 )  "*"                   " "          " "
## 8  ( 1 )  "*"                   " "          " "
## 9  ( 1 )  "*"                   " "          " "
## 10  ( 1 ) "*"                   " "          " "
## 11  ( 1 ) "*"                   " "          " "
## 12  ( 1 ) "*"                   "*"          " "
## 13  ( 1 ) "*"                   "*"          "*"
## 14  ( 1 ) "*"                   "*"          "*"
## 15  ( 1 ) "*"                   "*"          "*"
## 16  ( 1 ) "*"                   "*"          "*"
## 17  ( 1 ) "*"                   "*"          "*"
## 18  ( 1 ) "*"                   "*"          "*"
##           Departmentmanagement Departmentmarketing Departmentproduct_mng
## 1  ( 1 )  " "                  " "                 " "
## 2  ( 1 )  " "                  " "                 " "
## 3  ( 1 )  " "                  " "                 " "
## 4  ( 1 )  " "                  " "                 " "
## 5  ( 1 )  " "                  " "                 " "
## 6  ( 1 )  " "                  " "                 " "
## 7  ( 1 )  " "                  " "                 " "
## 8  ( 1 )  " "                  " "                 " "
## 9  ( 1 )  " "                  " "                 " "
## 10  ( 1 ) "*"                  " "                 " "
## 11  ( 1 ) "*"                  " "                 " "
## 12  ( 1 ) "*"                  " "                 " "
## 13  ( 1 ) "*"                  " "                 " "
## 14  ( 1 ) "*"                  " "                 "*"
## 15  ( 1 ) "*"                  " "                 "*"
## 16  ( 1 ) "*"                  "*"                 "*"
## 17  ( 1 ) "*"                  "*"                 "*"
## 18  ( 1 ) "*"                  "*"                 "*"
##           DepartmentRandD Departmentsales Departmentsupport
## 1  ( 1 )  " "             " "             " "
## 2  ( 1 )  " "             " "             " "
## 3  ( 1 )  " "             " "             " "
## 4  ( 1 )  " "             " "             " "
## 5  ( 1 )  " "             " "             " "
## 6  ( 1 )  " "             " "             " "
## 7  ( 1 )  " "             " "             " "
## 8  ( 1 )  "*"             " "             " "
## 9  ( 1 )  "*"             " "             " "
## 10  ( 1 ) "*"             " "             " "
## 11  ( 1 ) "*"             " "             " "
## 12  ( 1 ) "*"             " "             " "
## 13  ( 1 ) "*"             " "             " "
## 14  ( 1 ) "*"             " "             " "
## 15  ( 1 ) "*"             "*"             " "
## 16  ( 1 ) "*"             "*"             " "
```

```
## 17  ( 1 ) "*"              "*"              " "
## 18  ( 1 ) "*"              "*"              "*"
##           Departmenttechnical salary.L salary.Q
## 1  ( 1 )  " "                 " "      " "
## 2  ( 1 )  " "                 "*"      " "
## 3  ( 1 )  " "                 "*"      " "
## 4  ( 1 )  " "                 "*"      " "
## 5  ( 1 )  " "                 "*"      " "
## 6  ( 1 )  " "                 "*"      " "
## 7  ( 1 )  " "                 "*"      " "
## 8  ( 1 )  " "                 "*"      " "
## 9  ( 1 )  " "                 "*"      " "
## 10  ( 1 ) " "                 "*"      " "
## 11  ( 1 ) " "                 "*"      "*"
## 12  ( 1 ) " "                 "*"      "*"
## 13  ( 1 ) " "                 "*"      "*"
## 14  ( 1 ) " "                 "*"      "*"
## 15  ( 1 ) " "                 "*"      "*"
## 16  ( 1 ) " "                 "*"      "*"
## 17  ( 1 ) "*"                 "*"      "*"
## 18  ( 1 ) "*"                 "*"      "*"
```

```r
summary(fit.forward)$adjr2
```

```
##  [1] 0.1507785 0.1699466 0.1871023 0.2002490 0.2030084 0.2087386 0.2104291
##  [8] 0.2117506 0.2127275 0.2136280 0.2139914 0.2143088 0.2144996 0.2146386
## [15] 0.2146772 0.2146431 0.2145951 0.2145500
```

```r
which.max(summary(fit.forward)$adjr2)
```

```
## [1] 15
```

```r
coef(fit.forward,15)
```

```
##           (Intercept)     satisfaction_level        last_evaluation
##          1.4381089897          -0.6438647266           0.0872696609
##        number_project   average_montly_hours     time_spend_company
##         -0.0339856816           0.0006413777           0.0363956678
##         Work_accident1 promotion_last_5years1           Departmenthr
##         -0.1554263092          -0.1128482683           0.0298928079
##          DepartmentIT   Departmentmanagement  Departmentproduct_mng
##         -0.0301952167          -0.0668969528          -0.0288295235
##       DepartmentRandD        Departmentsales               salary.L
##         -0.0799823978          -0.0098817230          -0.1408300181
##             salary.Q
##         -0.0171257302
```

```r
fit.backward = regsubsets(left~satisfaction_level+last_evaluation+number_project+
                            average_montly_hours+time_spend_company+Work_accident+
                            promotion_last_5years+Department+salary,
                          data = HR_comma_sep,nvmax = 18,method = "backward")
summary(fit.backward)
```

```
## Subset selection object
## Call: regsubsets.formula(left ~ satisfaction_level + last_evaluation +
##     number_project + average_montly_hours + time_spend_company +
##     Work_accident + promotion_last_5years + Department + salary,
```

```
##     data = HR_comma_sep, nvmax = 18, method = "backward")
## 18 Variables  (and intercept)
##                        Forced in Forced out
## satisfaction_level         FALSE      FALSE
## last_evaluation            FALSE      FALSE
## number_project             FALSE      FALSE
## average_montly_hours       FALSE      FALSE
## time_spend_company         FALSE      FALSE
## Work_accident1             FALSE      FALSE
## promotion_last_5years1     FALSE      FALSE
## Departmenthr               FALSE      FALSE
## DepartmentIT               FALSE      FALSE
## Departmentmanagement       FALSE      FALSE
## Departmentmarketing        FALSE      FALSE
## Departmentproduct_mng      FALSE      FALSE
## DepartmentRandD            FALSE      FALSE
## Departmentsales            FALSE      FALSE
## Departmentsupport          FALSE      FALSE
## Departmenttechnical        FALSE      FALSE
## salary.L                   FALSE      FALSE
## salary.Q                   FALSE      FALSE
## 1 subsets of each size up to 18
## Selection Algorithm: backward
##           satisfaction_level last_evaluation number_project
## 1  ( 1 )  "*"                " "             " "
## 2  ( 1 )  "*"                " "             " "
## 3  ( 1 )  "*"                " "             " "
## 4  ( 1 )  "*"                " "             " "
## 5  ( 1 )  "*"                " "             "*"
## 6  ( 1 )  "*"                " "             "*"
## 7  ( 1 )  "*"                " "             "*"
## 8  ( 1 )  "*"                " "             "*"
## 9  ( 1 )  "*"                "*"             "*"
## 10  ( 1 ) "*"                "*"             "*"
## 11  ( 1 ) "*"                "*"             "*"
## 12  ( 1 ) "*"                "*"             "*"
## 13  ( 1 ) "*"                "*"             "*"
## 14  ( 1 ) "*"                "*"             "*"
## 15  ( 1 ) "*"                "*"             "*"
## 16  ( 1 ) "*"                "*"             "*"
## 17  ( 1 ) "*"                "*"             "*"
## 18  ( 1 ) "*"                "*"             "*"
##           average_montly_hours time_spend_company Work_accident1
## 1  ( 1 )  " "                  " "                " "
## 2  ( 1 )  " "                  " "                " "
## 3  ( 1 )  " "                  " "                "*"
## 4  ( 1 )  " "                  "*"                "*"
## 5  ( 1 )  " "                  "*"                "*"
## 6  ( 1 )  "*"                  "*"                "*"
## 7  ( 1 )  "*"                  "*"                "*"
## 8  ( 1 )  "*"                  "*"                "*"
## 9  ( 1 )  "*"                  "*"                "*"
## 10  ( 1 ) "*"                  "*"                "*"
## 11  ( 1 ) "*"                  "*"                "*"
```

```
## 12  ( 1 ) "*"                  "*"               "*"
## 13  ( 1 ) "*"                  "*"               "*"
## 14  ( 1 ) "*"                  "*"               "*"
## 15  ( 1 ) "*"                  "*"               "*"
## 16  ( 1 ) "*"                  "*"               "*"
## 17  ( 1 ) "*"                  "*"               "*"
## 18  ( 1 ) "*"                  "*"               "*"
##           promotion_last_5years1 Departmenthr DepartmentIT
## 1  ( 1 )  " "                    " "          " "
## 2  ( 1 )  " "                    " "          " "
## 3  ( 1 )  " "                    " "          " "
## 4  ( 1 )  " "                    " "          " "
## 5  ( 1 )  " "                    " "          " "
## 6  ( 1 )  " "                    " "          " "
## 7  ( 1 )  "*"                    " "          " "
## 8  ( 1 )  "*"                    " "          " "
## 9  ( 1 )  "*"                    " "          " "
## 10  ( 1 ) "*"                    " "          " "
## 11  ( 1 ) "*"                    " "          " "
## 12  ( 1 ) "*"                    "*"          " "
## 13  ( 1 ) "*"                    "*"          "*"
## 14  ( 1 ) "*"                    "*"          "*"
## 15  ( 1 ) "*"                    "*"          "*"
## 16  ( 1 ) "*"                    "*"          "*"
## 17  ( 1 ) "*"                    "*"          "*"
## 18  ( 1 ) "*"                    "*"          "*"
##           Departmentmanagement Departmentmarketing Departmentproduct_mng
## 1  ( 1 )  " "                  " "                 " "
## 2  ( 1 )  " "                  " "                 " "
## 3  ( 1 )  " "                  " "                 " "
## 4  ( 1 )  " "                  " "                 " "
## 5  ( 1 )  " "                  " "                 " "
## 6  ( 1 )  " "                  " "                 " "
## 7  ( 1 )  " "                  " "                 " "
## 8  ( 1 )  " "                  " "                 " "
## 9  ( 1 )  " "                  " "                 " "
## 10  ( 1 ) "*"                  " "                 " "
## 11  ( 1 ) "*"                  " "                 " "
## 12  ( 1 ) "*"                  " "                 " "
## 13  ( 1 ) "*"                  " "                 " "
## 14  ( 1 ) "*"                  " "                 "*"
## 15  ( 1 ) "*"                  " "                 "*"
## 16  ( 1 ) "*"                  " "                 "*"
## 17  ( 1 ) "*"                  " "                 "*"
## 18  ( 1 ) "*"                  "*"                 "*"
##           DepartmentRandD Departmentsales Departmentsupport
## 1  ( 1 )  " "             " "             " "
## 2  ( 1 )  " "             " "             " "
## 3  ( 1 )  " "             " "             " "
## 4  ( 1 )  " "             " "             " "
## 5  ( 1 )  " "             " "             " "
## 6  ( 1 )  " "             " "             " "
## 7  ( 1 )  " "             " "             " "
## 8  ( 1 )  "*"             " "             " "
```

```
## 9  ( 1 )  "*"              " "              " "
## 10  ( 1 ) "*"              " "              " "
## 11  ( 1 ) "*"              " "              " "
## 12  ( 1 ) "*"              " "              " "
## 13  ( 1 ) "*"              " "              " "
## 14  ( 1 ) "*"              " "              " "
## 15  ( 1 ) "*"              " "              " "
## 16  ( 1 ) "*"              " "              "*"
## 17  ( 1 ) "*"              "*"              "*"
## 18  ( 1 ) "*"              "*"              "*"
##           Departmenttechnical salary.L salary.Q
## 1  ( 1 )  " "                 " "      " "
## 2  ( 1 )  " "                 "*"      " "
## 3  ( 1 )  " "                 "*"      " "
## 4  ( 1 )  " "                 "*"      " "
## 5  ( 1 )  " "                 "*"      " "
## 6  ( 1 )  " "                 "*"      " "
## 7  ( 1 )  " "                 "*"      " "
## 8  ( 1 )  " "                 "*"      " "
## 9  ( 1 )  " "                 "*"      " "
## 10  ( 1 ) " "                 "*"      " "
## 11  ( 1 ) " "                 "*"      "*"
## 12  ( 1 ) " "                 "*"      "*"
## 13  ( 1 ) " "                 "*"      "*"
## 14  ( 1 ) " "                 "*"      "*"
## 15  ( 1 ) "*"                 "*"      "*"
## 16  ( 1 ) "*"                 "*"      "*"
## 17  ( 1 ) "*"                 "*"      "*"
## 18  ( 1 ) "*"                 "*"      "*"
```

```
summary(fit.backward)$adjr2
```

```
##  [1] 0.1507785 0.1699466 0.1871023 0.2002490 0.2030084 0.2087386 0.2104291
##  [8] 0.2117506 0.2127275 0.2136280 0.2139914 0.2143088 0.2144996 0.2146386
## [15] 0.2146409 0.2146475 0.2146016 0.2145500
```

```
which.max(summary(fit.backward)$adjr2)
```

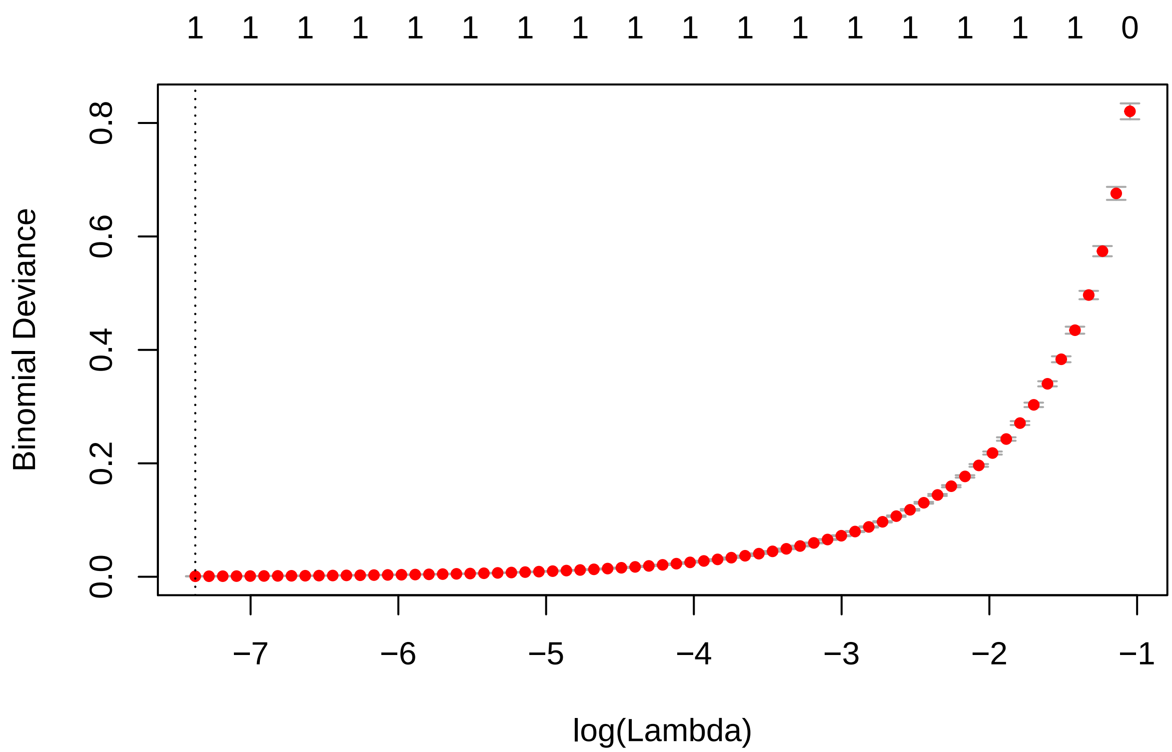
```
## [1] 16
```

```
coef(fit.backward,16)
```

```
##           (Intercept)     satisfaction_level       last_evaluation
##          1.4295574288          -0.6440923484          0.0874306340
##         number_project   average_montly_hours     time_spend_company
##         -0.0340556799           0.0006411873          0.0364409236
##         Work_accident1 promotion_last_5years1          Departmenthr
##         -0.1554294194          -0.1119518480          0.0386741109
##          DepartmentIT   Departmentmanagement  Departmentproduct_mng
##         -0.0213856069          -0.0583151959         -0.0200201300
##        DepartmentRandD      Departmentsupport    Departmenttechnical
##         -0.0711955983           0.0100103919          0.0113851306
##              salary.L               salary.Q
##         -0.1405758352          -0.0170558409
```

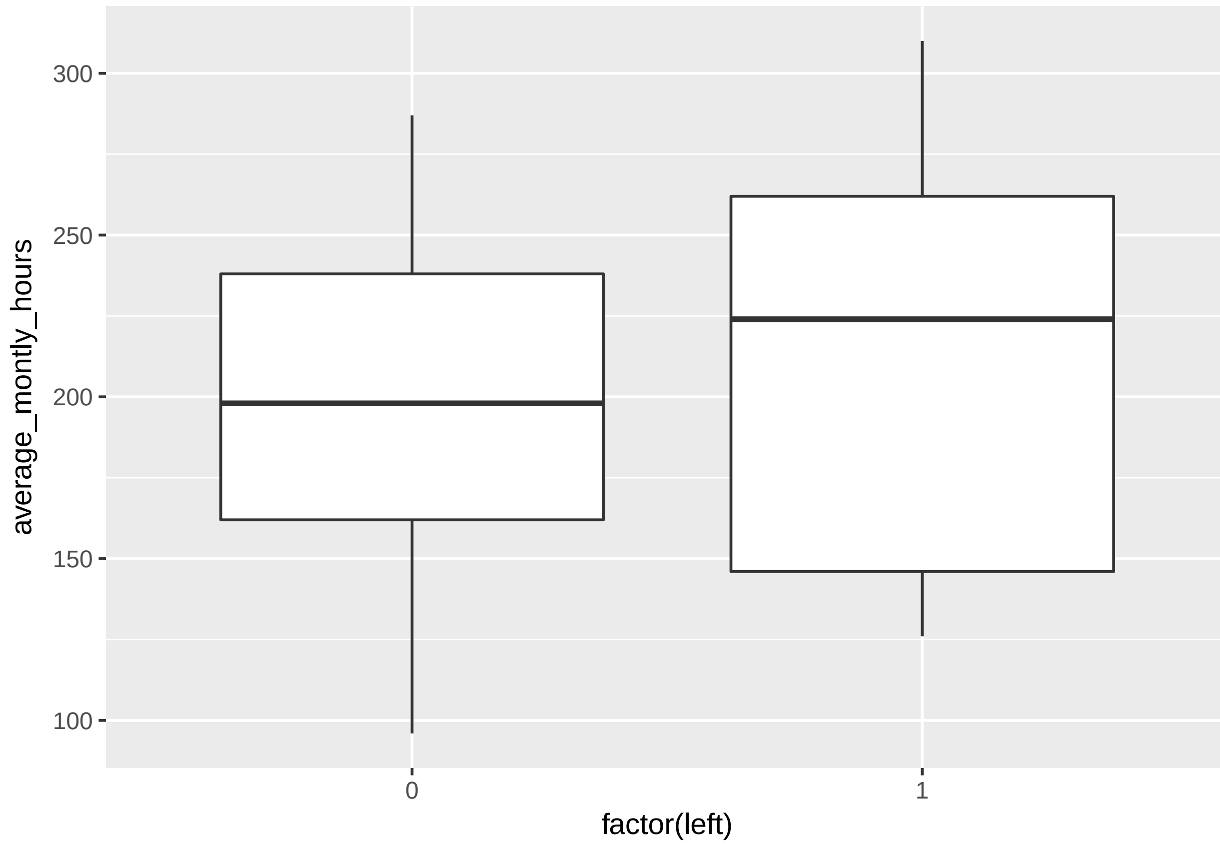
```
## (4c). LASSO
Xvars = model.matrix(left~satisfaction_level+last_evaluation+number_project+
```

```
                          average_montly_hours+time_spend_company+Work_accident+
                          promotion_last_5years+Department+salary,
                        data = HR_comma_sep)[,-1]
Yvars = HR_comma_sep[,7]


set.seed(1)
train = sample(1:nrow(Xvars),nrow(Xvars)/2)
test = -train
Yvars.test = Yvars[test]
grid =seq (0,10^10, length =10)
lasso.mod =glmnet(Xvars[train,],as.factor(Yvars[train]),alpha =1,
                  lambda =grid, family = "binomial")
cv.out = cv.glmnet(Xvars[train,],as.factor(Yvars[train]),
                   alpha =1, family = "binomial")
plot(cv.out)
```

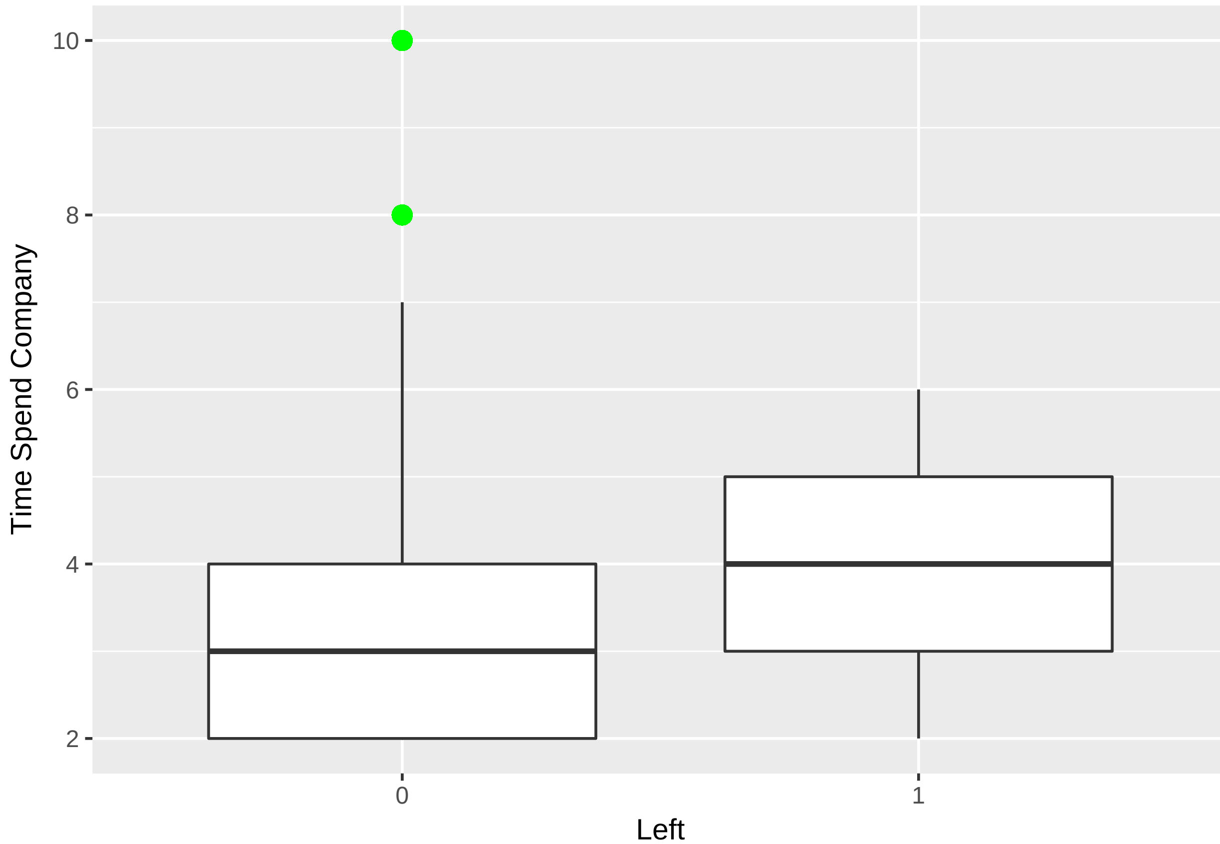



```
##  Number of projects and average monthly hours are correlated.
##  So find average time for spending time on single project

##  Creating new column for average hourly projects
HR_comma_sep['avg_hr_prj']<-
  (HR_comma_sep['average_montly_hours'] * 12)/HR_comma_sep['number_project']

##  Dividing the variable into 3 parts
HR_comma_sep['avg_hr_prj_range']<-cut(HR_comma_sep$avg_hr_prj,3)
```

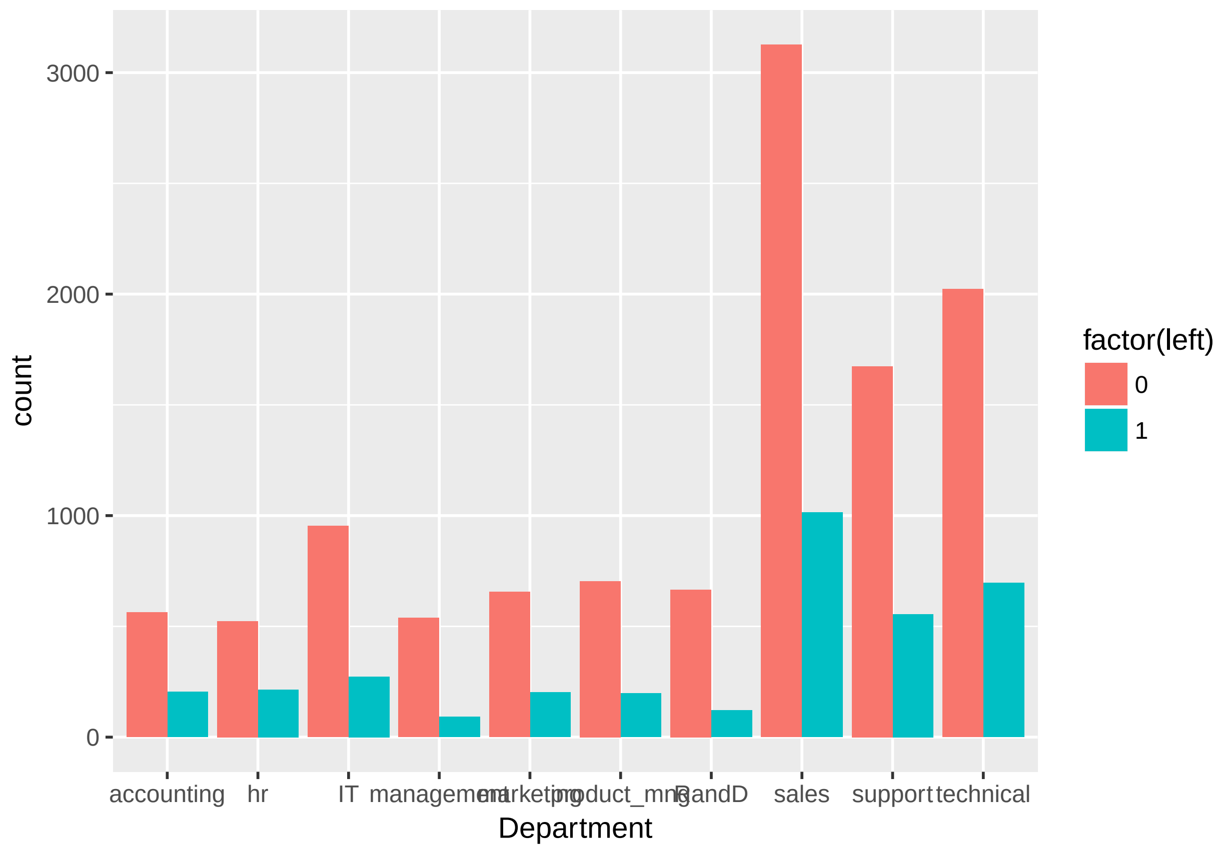
```
##  Assigning a variable with labels 0, 1, 2 according
##  to monthly hours spent range
HR_comma_sep['HR_Cat']<-cut(HR_comma_sep$avg_hr_prj,3,labels = c(0:2))

## Plotting for Observations
ggplot(HR_comma_sep,aes(factor(left),average_montly_hours))+
  geom_boxplot(outlier.colour = "green", outlier.size = 3)
```



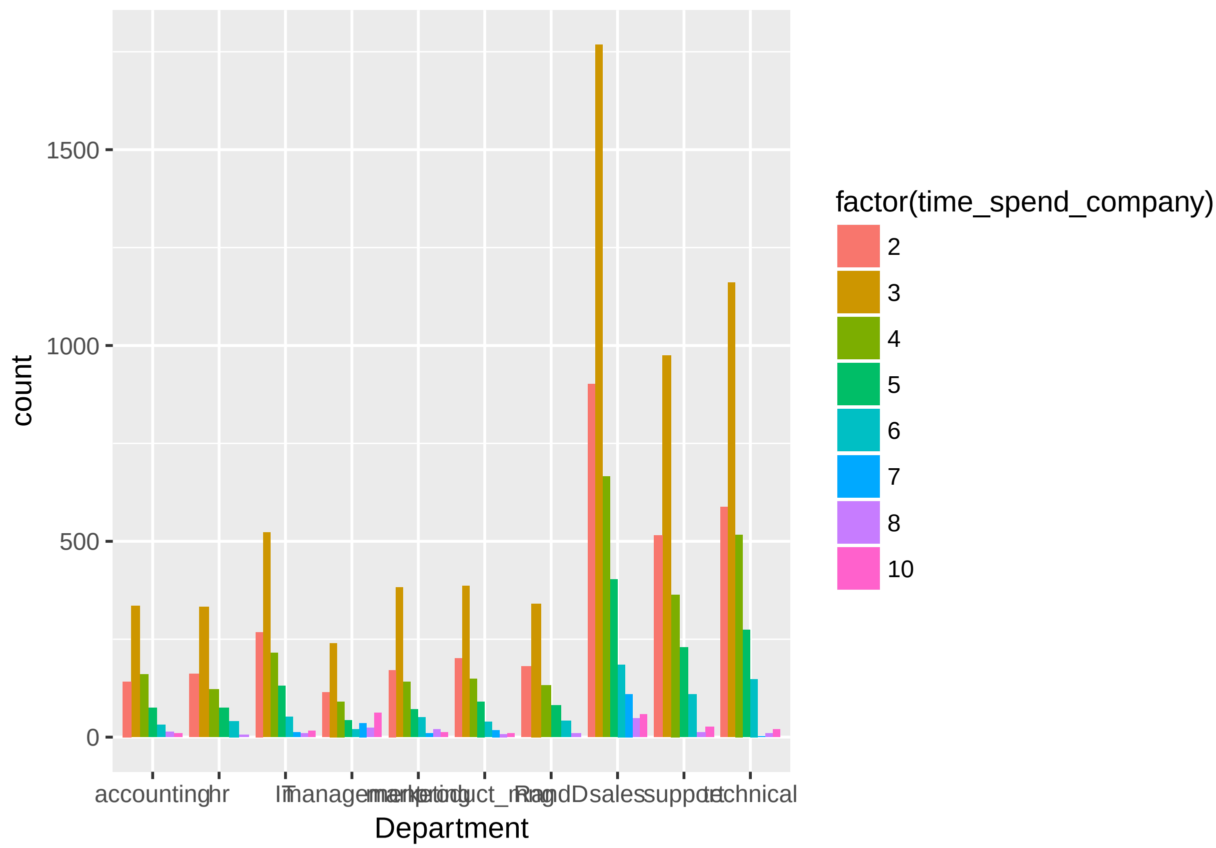

```
ggplot(HR_comma_sep,aes(factor(left),time_spend_company))+
  geom_boxplot(outlier.colour = "green",
               outlier.size = 3)+xlab("Left")+ylab("Time Spend Company")
```



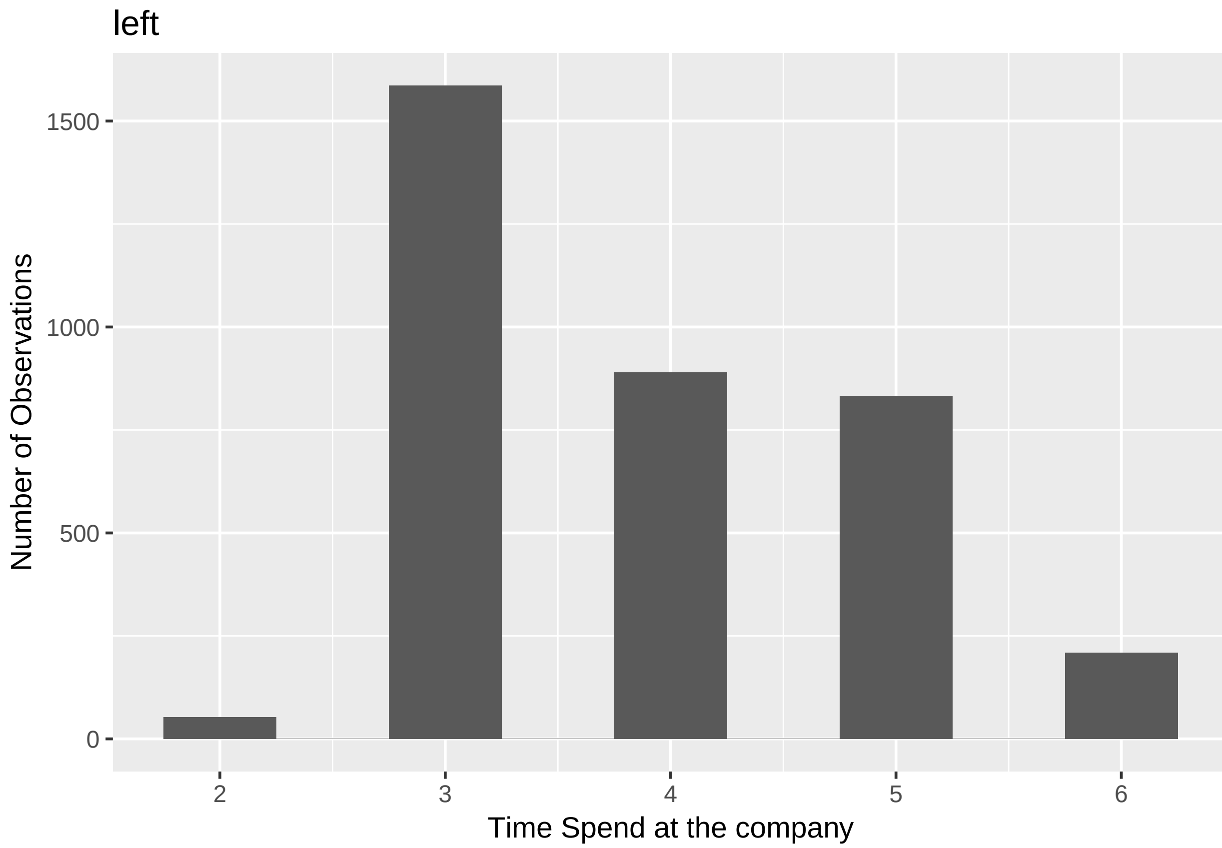

```
ggplot(HR_comma_sep,aes(Department))+
  geom_bar(aes(fill=factor(left)),position='dodge')
```



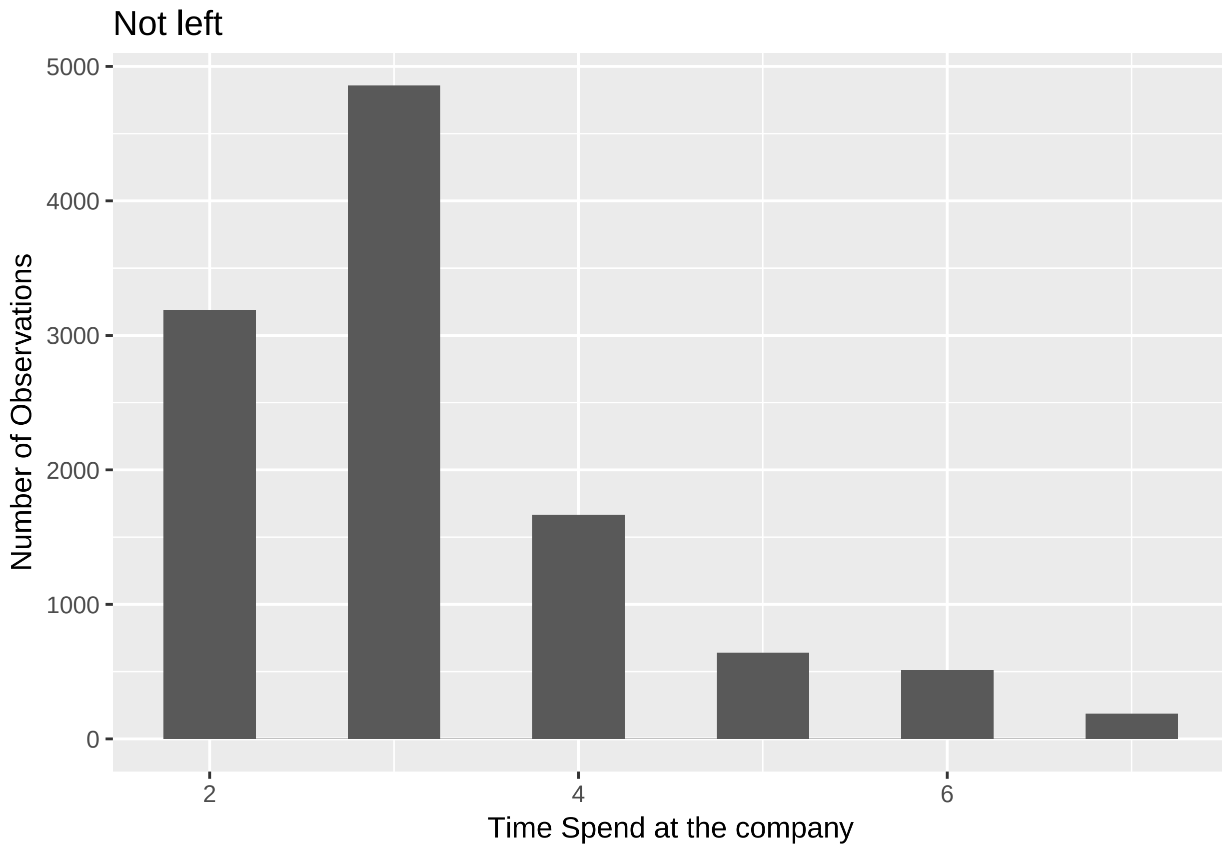

```
##  We observe that the highest employees left from the company belong
##  to departments 'Management' and 'RandD'

ggplot(HR_comma_sep,aes(Department))+
  geom_bar(aes(fill=factor(time_spend_company)),position='dodge')
```



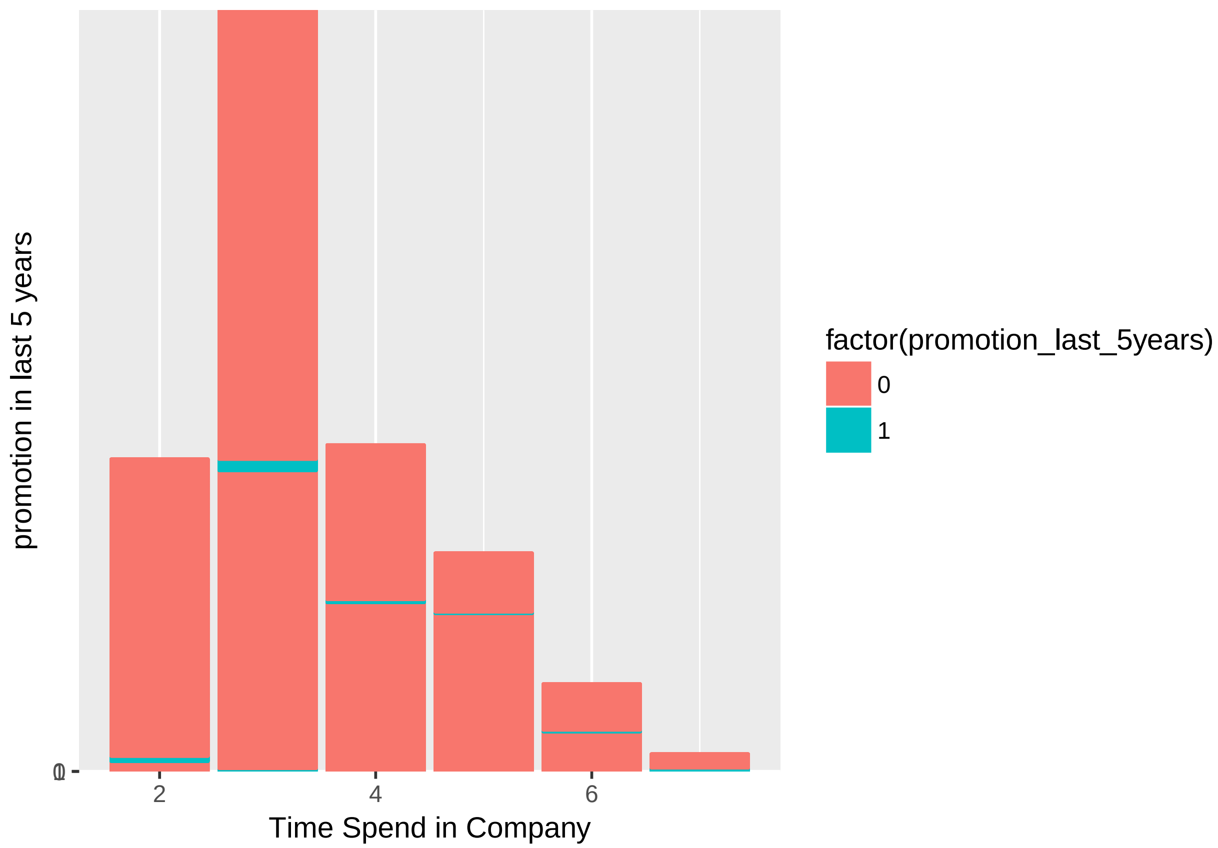

```
##  More number of employee from Management and sales are spending more
##  than 8 years in the company compared to other departments. So we cannot
##  remove outliers.

##  There are few outliers in the data set.
##  So we cannot ignore these observations because more
dropdata<-subset(HR_comma_sep,time_spend_company<8)
HR_comma_sep1<-dropdata
left=dropdata[(dropdata$left==1),]
non_left=dropdata[(dropdata$left==0),]
ggplot(left,aes(time_spend_company))+
  geom_histogram(binwidth = 0.5)+xlab("Time Spend at the company")+
  ylab("Number of Observations")+ggtitle("left")
```



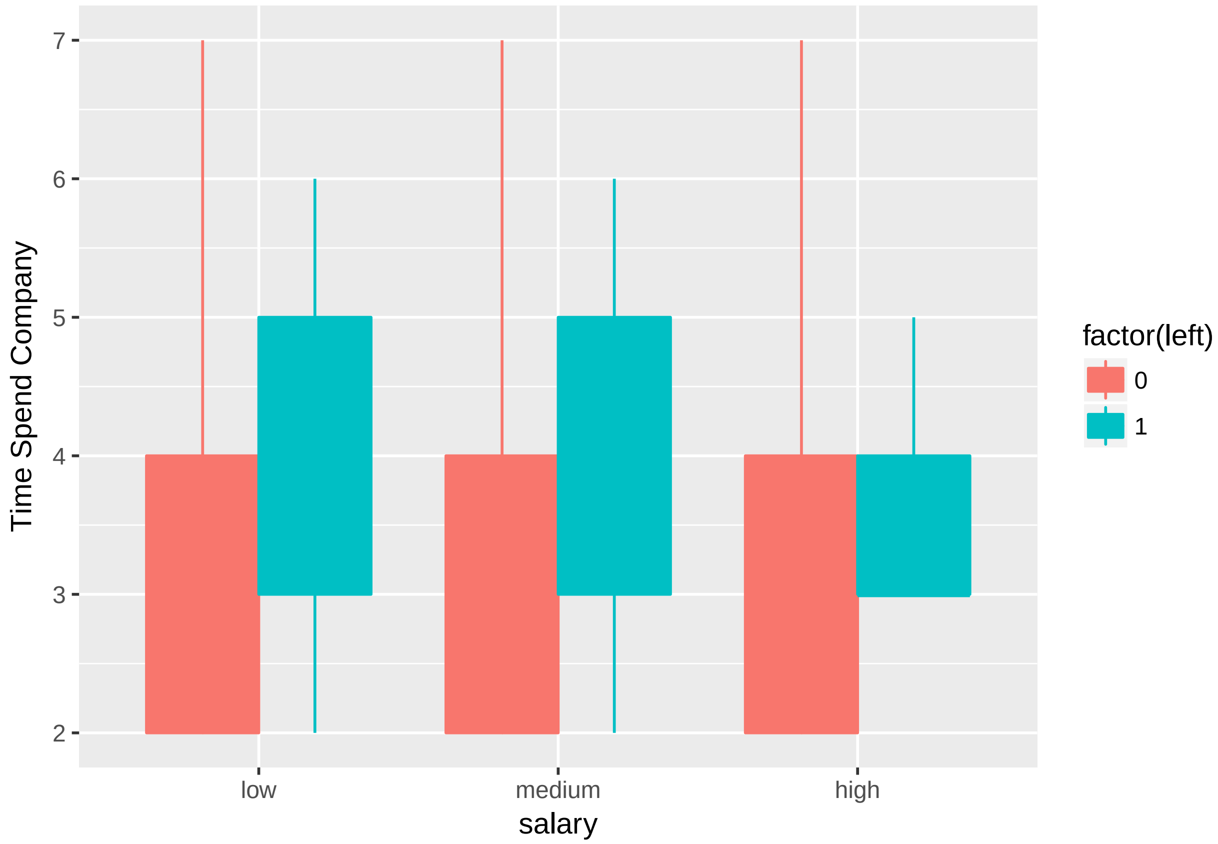

```
ggplot(non_left,aes(time_spend_company))+
  geom_histogram(binwidth = 0.5)+xlab("Time Spend at the company")+
  ylab("Number of Observations")+ggtitle("Not left")
```




```r
# Observations from above plots:

# a. From the above plots we can say that, people who work more than
#    6 years and who work for 2 years are less likely to leave
# b. People are more likely to leave when they spend 3 to 5 years
# c. People with 5-years are more likely to leave
# d. When the years people spent in the company lies in 3-5: the more
#    they've been here, the more likely they leave.

ggplot(dropdata,aes(x=time_spend_company,y=left,fill=
                      factor(promotion_last_5years),colour=
                      factor(promotion_last_5years)))+
  geom_bar(position='stack', stat='identity')+xlab("Time Spend in Company")+
  ylab("promotion in last 5 years")
```




```
# e. Very less people got promoted even though they are spending
#    more time in the office.

ggplot(dropdata,aes(x=salary,y=time_spend_company,fill=factor(left),
                    colour=factor(left)))+geom_boxplot(outlier.colour = NA)+
  xlab("salary")+ylab("Time Spend Company")
```




```
# f. The low and medium income people are leaving the company

HR_comma_sep1['avg_hr_prj']<-
  (HR_comma_sep1['average_montly_hours'] * 12)/HR_comma_sep1['number_project']
HR_comma_sep1['avg_hr_prj_range']<-cut(HR_comma_sep1$avg_hr_prj,3)

# who are valuable employess??

## The evaluation criteria and Monthly hours spend in the company are considered
## as valuable. Here we are not considering the promotion because very less
## people got promoted in last 5 years.

## For our analysis we are finding the average time an employee spent on each
## project. Then, we converted the variable into 3 levels.

## In general an employee must work for 160 hours per month. We have splitted
## this variable into 3 levels and then according to the level we have
## given categories as [0,1,2]

b1<-HR_comma_sep$last_evaluation > 0.5
b2<-HR_comma_sep$HR_Cat==1 | HR_comma_sep$HR_Cat==2
sum(b1 & b2)
```

```
## [1] 4386
```

```
# There are total of 4386 valuable employees
```

```r
# Decide who all are valuable employees
HR_comma_sep['valuedEmployee']<-0
head(HR_comma_sep)
```

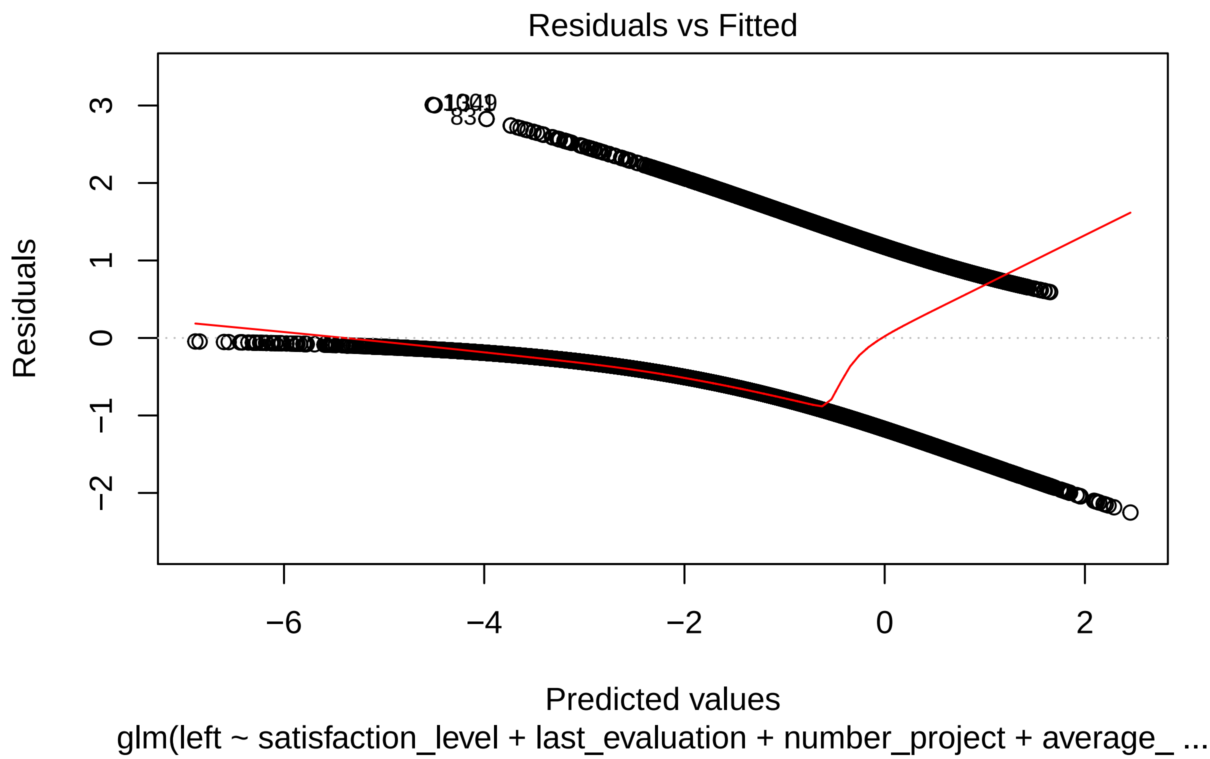
```
##   ID satisfaction_level last_evaluation number_project
## 1  1               0.38            0.53              2
## 2  2               0.80            0.86              5
## 3  3               0.11            0.88              7
## 4  4               0.72            0.87              5
## 5  5               0.37            0.52              2
## 6  6               0.41            0.50              2
##   average_montly_hours time_spend_company Work_accident left
## 1                  157                  3             0    1
## 2                  262                  6             0    1
## 3                  272                  4             0    1
## 4                  223                  5             0    1
## 5                  159                  3             0    1
## 6                  153                  3             0    1
##   promotion_last_5years Department salary avg_hr_prj avg_hr_prj_range
## 1                     0      sales    low   942.0000    (749,1.3e+03]
## 2                     0      sales medium   628.8000        (192,749]
## 3                     0      sales medium   466.2857        (192,749]
## 4                     0      sales    low   535.2000        (192,749]
## 5                     0      sales    low   954.0000    (749,1.3e+03]
## 6                     0      sales    low   918.0000    (749,1.3e+03]
##   HR_Cat valuedEmployee
## 1      1              0
## 2      0              0
## 3      0              0
## 4      0              0
## 5      1              0
## 6      1              0
```

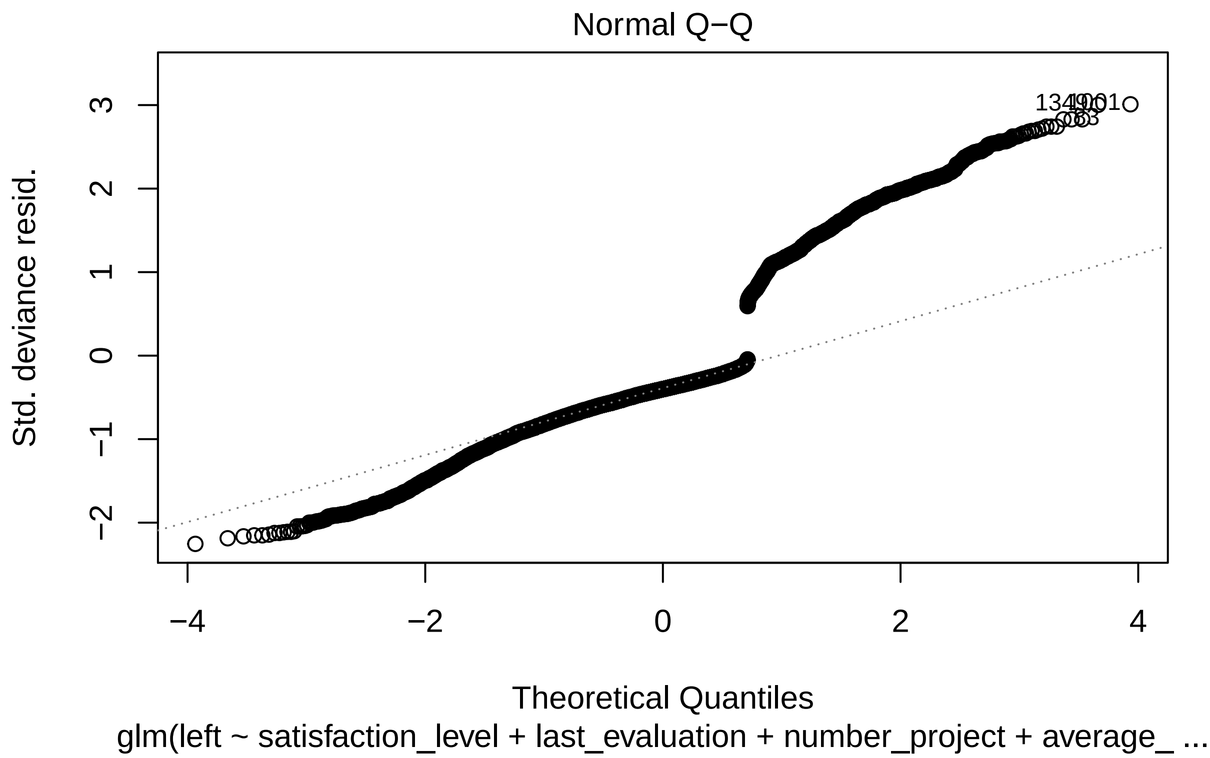
```r
for (i in (1: nrow(HR_comma_sep))){
  b1<-(HR_comma_sep[i,'last_evaluation'] > 0.5)
  b2<-((HR_comma_sep[i,'HR_Cat']==1) | (HR_comma_sep[i,'HR_Cat']==2))
  if(b1 & b2){
    HR_comma_sep[i,'valuedEmployee'] = 1
  }
}


# 5. Algorithms:

# (5a) Stratified sampling

xvars=c('satisfaction_level','last_evaluation','number_project',
        'average_montly_hours','time_spend_company','Work_accident',
        'promotion_last_5years','sales','salary')
yvars='left'
p1<-0.8
set.seed(12345)
inTrain<-createDataPartition(y=HR_comma_sep[,yvars],p=p1,list=FALSE)
train_HR<-HR_comma_sep[inTrain,]
```

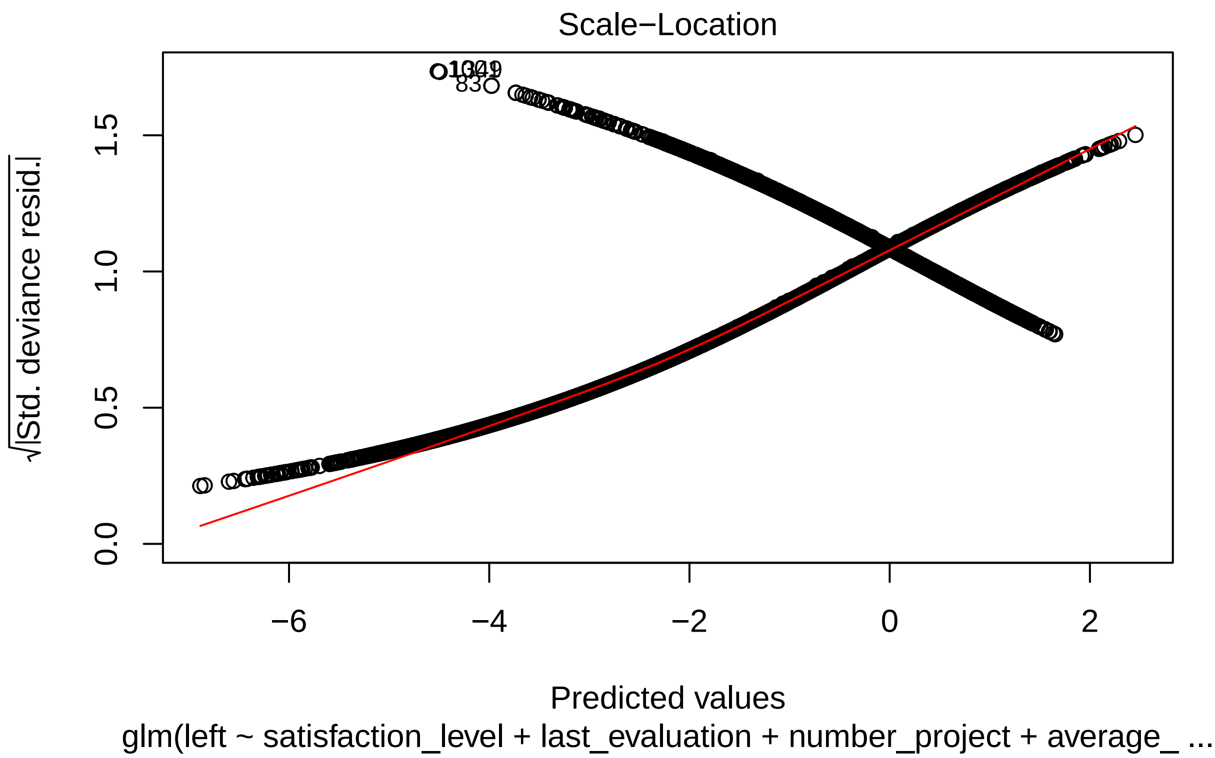
```r
test_HR<-HR_comma_sep[-inTrain,]
stopifnot(nrow(train_HR)+nrow(test_HR)==nrow(HR_comma_sep))

# (5b) Logistic Regression (Fitting GLM)

glm.fit<-glm(left~satisfaction_level+last_evaluation+number_project+
               average_montly_hours+time_spend_company+Work_accident+
               promotion_last_5years+Department+salary,
             data=train_HR,family = binomial(link="logit"))
summary(glm.fit)
```

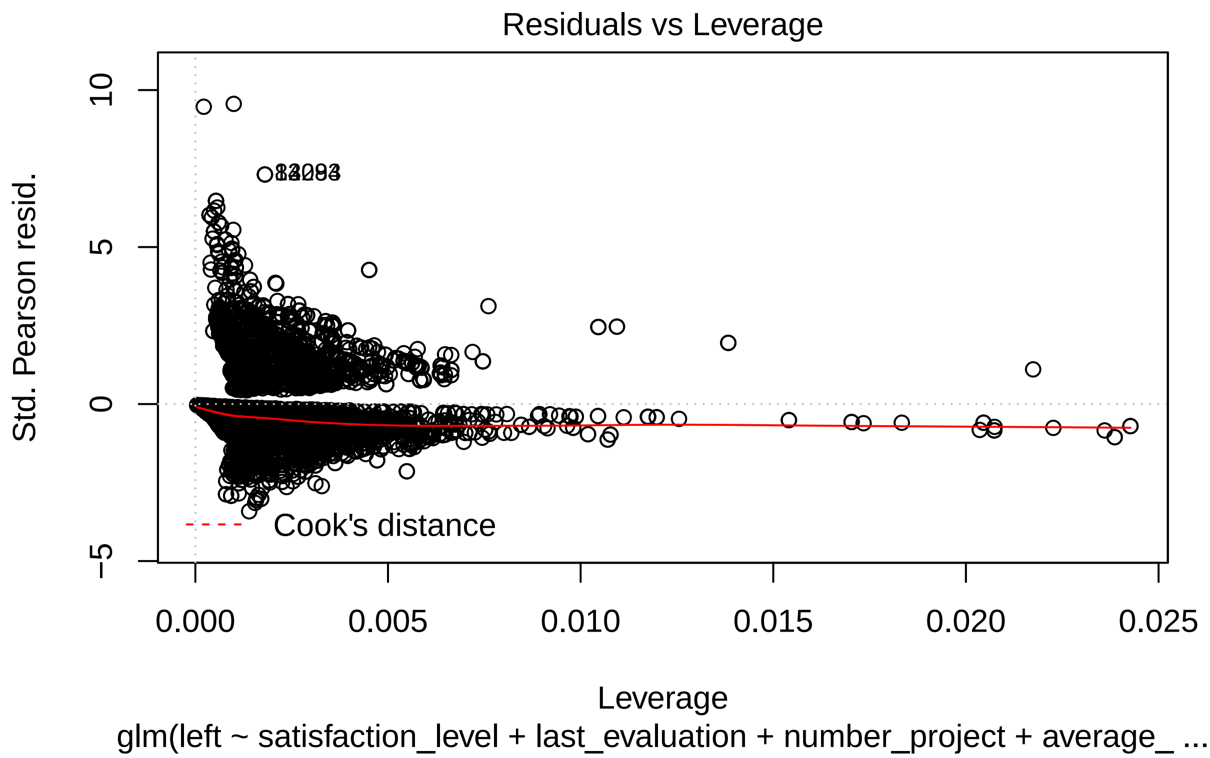
```
##
## Call:
## glm(formula = left ~ satisfaction_level + last_evaluation + number_project +
##     average_montly_hours + time_spend_company + Work_accident +
##     promotion_last_5years + Department + salary, family = binomial(link = "logit"),
##     data = train_HR)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2527  -0.6588  -0.4010  -0.1188   3.0084
##
## Coefficients:
##                        Estimate Std. Error z value Pr(>|z|)
## (Intercept)           -0.356444   0.171307  -2.081 0.037459 *
## satisfaction_level    -4.144288   0.109977 -37.683  < 2e-16 ***
## last_evaluation        0.775304   0.167395   4.632 3.63e-06 ***
## number_project        -0.326962   0.023895 -13.683  < 2e-16 ***
## average_montly_hours   0.004294   0.000578   7.430 1.09e-13 ***
## time_spend_company     0.284947   0.017576  16.212  < 2e-16 ***
## Work_accident1        -1.451776   0.097491 -14.891  < 2e-16 ***
## promotion_last_5years1 -1.319956   0.282791  -4.668 3.05e-06 ***
## Departmenthr           0.138999   0.149162   0.932 0.351406
## DepartmentIT          -0.196621   0.136715  -1.438 0.150381
## Departmentmanagement  -0.471916   0.179565  -2.628 0.008586 **
## Departmentmarketing   -0.037971   0.147277  -0.258 0.796546
## Departmentproduct_mng -0.216840   0.145833  -1.487 0.137039
## DepartmentRandD       -0.608230   0.162932  -3.733 0.000189 ***
## Departmentsales       -0.033144   0.114960  -0.288 0.773111
## Departmentsupport      0.041642   0.122869   0.339 0.734675
## Departmenttechnical    0.066998   0.119840   0.559 0.576120
## salary.L              -1.371530   0.099341 -13.806  < 2e-16 ***
## salary.Q              -0.309246   0.065104  -4.750 2.03e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 13173  on 11999  degrees of freedom
## Residual deviance: 10276  on 11981  degrees of freedom
## AIC: 10314
##
## Number of Fisher Scoring iterations: 5
```

```r
par(mfrow=c(1,1))
plot(glm.fit)
```




Normal Q−Q
Std. deviance resid.
−2
−1
0
1
2
3
−4
−2
0
2
4
Theoretical Quantiles
glm(left ~ satisfaction_level + last_evaluation + number_project + average_ ...


Scale−Location
√|Std. deviance resid.|
0.0
0.5
1.0
1.5
−6
−4
−2
0
2
Predicted values
glm(left ~ satisfaction_level + last_evaluation + number_project + average_ ...
1041
83

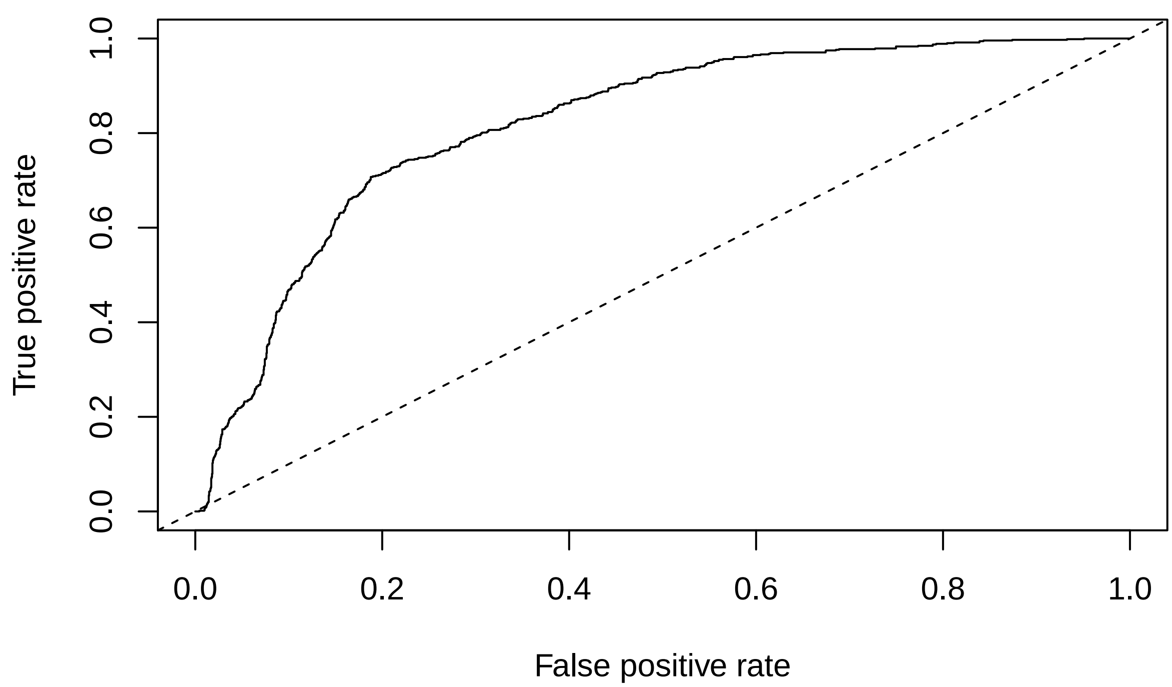



```
## confusion matrix

test_HR[,'Yhat']<-predict(glm.fit,newdata=test_HR)
fitted.values<-test_HR[,'Yhat']
test_HR$Yhat<-ifelse( test_HR$Yhat>0.5,1,0)
conf<-confusionMatrix(test_HR$Yhat,test_HR$left)
conf
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    0    1
##          0 2197  573
##          1   88  141
##
##                Accuracy : 0.7796
##                  95% CI : (0.7643, 0.7943)
##     No Information Rate : 0.7619
##     P-Value [Acc > NIR] : 0.0117
##
##                   Kappa : 0.2074
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 0.9615
##             Specificity : 0.1975
##          Pos Pred Value : 0.7931
##          Neg Pred Value : 0.6157
```

```
##               Prevalence : 0.7619
##           Detection Rate : 0.7326
##     Detection Prevalence : 0.9236
##        Balanced Accuracy : 0.5795
##
##         'Positive' Class : 0
##
```

```
## ROC Curve
fit_values<-prediction(fitted.values,test_HR$left)
p<-performance(fit_values,measure = 'tpr',x.measure = 'fpr')
plot(p)
abline(0, 1, lty = 2)
```




```
# (5c) Fitting random forest
fit_rf<-randomForest(as.factor(left)~.,data=train_HR,importance=TRUE,ntree=1000)
fit_rf$confusion
```

```
##      0    1  class.error
## 0 9139    4 0.0004374932
## 1   48 2809 0.0168008400
```

```
## confusion matrix for random forest
fitted.values.rf<-predict(fit_rf,newdata = test_HR,type='class')
fitted.values.rf1<-predict(fit_rf,newdata = test_HR,type='prob')
conf.rf<-confusionMatrix(fitted.values.rf,test_HR$left)
conf.rf
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    0    1
##          0 2285    5
##          1    0  709
##
##                Accuracy : 0.9983
##                  95% CI : (0.9961, 0.9995)
##     No Information Rate : 0.7619
##     P-Value [Acc > NIR] : < 2e-16
##
##                   Kappa : 0.9954
##  Mcnemar's Test P-Value : 0.07364
##
##             Sensitivity : 1.0000
##             Specificity : 0.9930
##          Pos Pred Value : 0.9978
##          Neg Pred Value : 1.0000
##              Prevalence : 0.7619
##          Detection Rate : 0.7619
##    Detection Prevalence : 0.7636
##       Balanced Accuracy : 0.9965
##
##        'Positive' Class : 0
##
```

```r
## ROC curve for random forest
HR.rf<-roc(test_HR$left, fitted.values.rf1[,2])
plot(HR.rf, print.auc=TRUE, auc.polygon=TRUE)

# (5d) Fitting SVM algorithm
svm_model<-svm(left~.,data=train_HR,type='C-classification')
svm_model1<-svm(left~.,data=train_HR,type='C-classification',probability = TRUE)
summary(svm_model)
```

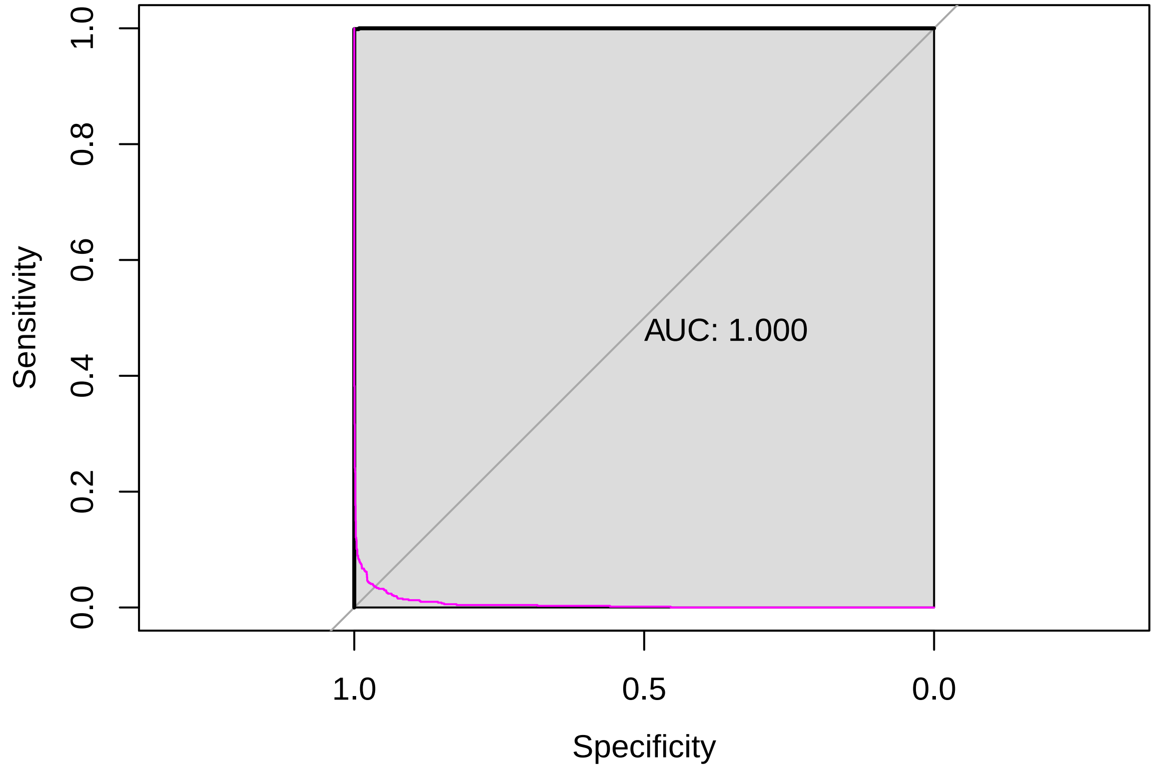
```
##
## Call:
## svm(formula = left ~ ., data = train_HR, type = "C-classification")
##
##
## Parameters:
##    SVM-Type:  C-classification
##  SVM-Kernel:  radial
##        cost:  1
##       gamma:  0.03846154
##
## Number of Support Vectors:  1523
##
##  ( 756 767 )
##
##
## Number of Classes:  2
##
## Levels:
```

```
##  0 1
```

```r
## predicting values and confusion matrix
pred<-predict(svm_model,newdata = test_HR)
pred.prob<-predict(svm_model1,newdata = test_HR,type='prob',probability = TRUE)
conf.svm<-confusionMatrix(pred,test_HR$left)
conf.svm
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    0    1
##          0 2255   49
##          1   30  665
##
##                Accuracy : 0.9737
##                  95% CI : (0.9673, 0.9791)
##     No Information Rate : 0.7619
##     P-Value [Acc > NIR] : < 2e-16
##
##                   Kappa : 0.9267
##  Mcnemar's Test P-Value : 0.04285
##
##             Sensitivity : 0.9869
##             Specificity : 0.9314
##          Pos Pred Value : 0.9787
##          Neg Pred Value : 0.9568
##              Prevalence : 0.7619
##          Detection Rate : 0.7519
##    Detection Prevalence : 0.7683
##       Balanced Accuracy : 0.9591
##
##        'Positive' Class : 0
##
```

```r
## ROC curve for SVM
p.svm<-prediction(attr(pred.prob,"probabilities")[,2],test_HR$left)
svm.perf<-performance(p.svm,measure = 'tpr',x.measure = 'fpr')
plot(svm.perf,add=TRUE,col=6)
```




```r
# (5e) CART implementation
cart.fit<-rpart(left~.,data=train_HR,method='class')
summary(cart.fit)
```

```
## Call:
## rpart(formula = left ~ ., data = train_HR, method = "class")
##   n= 12000
##
##           CP nsplit rel error      xerror         xstd
## 1 0.5582779      0 1.0000000 1.000000000 0.0163304675
## 2 0.2219111      1 0.4417221 0.442072104 0.0117663642
## 3 0.1099055      2 0.2198110 0.220861043 0.0085580526
## 4 0.0100000      4 0.0000000 0.001050053 0.0006061723
##
## Variable importance
##                    ID   satisfaction_level       number_project
##                    80                   10                    6
## average_montly_hours      last_evaluation
##                     3                    1
##
## Node number 1: 12000 observations,    complexity param=0.5582779
##   predicted class=0  expected loss=0.2380833  P(node) =1
##     class counts:  9143  2857
##    probabilities: 0.762 0.238
##   left son=2 (10405 obs) right son=3 (1595 obs)
##   Primary splits:
```

```
##       ID                  < 2000.5  to the right, improve=2135.7220, (0 missing)
##       satisfaction_level   < 0.465   to the right, improve=1230.8090, (0 missing)
##       number_project       < 2.5     to the right, improve= 813.6592, (0 missing)
##       time_spend_company   < 2.5     to the left,  improve= 333.4971, (0 missing)
##       average_montly_hours < 287.5   to the left,  improve= 318.2650, (0 missing)
##   Surrogate splits:
##       satisfaction_level   < 0.115   to the right, agree=0.876, adj=0.066, (0 split)
##       average_montly_hours < 287.5   to the left,  agree=0.870, adj=0.019, (0 split)
##       number_project       < 6.5     to the left,  agree=0.870, adj=0.018, (0 split)
##       avg_hr_prj           < 1749    to the left,  agree=0.867, adj=0.001, (0 split)
##
## Node number 2: 10405 observations,    complexity param=0.2219111
##   predicted class=0  expected loss=0.1212878  P(node) =0.8670833
##     class counts:  9143  1262
##    probabilities: 0.879 0.121
##   left son=4 (9771 obs) right son=5 (634 obs)
##   Primary splits:
##       ID                   < 14211.5 to the left,  improve=1042.59500, (0 missing)
##       satisfaction_level   < 0.115   to the right, improve= 478.67250, (0 missing)
##       number_project       < 2.5     to the right, improve= 335.62500, (0 missing)
##       average_montly_hours < 288     to the left,  improve= 185.89420, (0 missing)
##       time_spend_company   < 2.5     to the left,  improve=  90.52302, (0 missing)
##   Surrogate splits:
##       satisfaction_level   < 0.095   to the right, agree=0.940, adj=0.011, (0 split)
##       average_montly_hours < 289.5   to the left,  agree=0.939, adj=0.005, (0 split)
##
## Node number 3: 1595 observations
##   predicted class=1  expected loss=0  P(node) =0.1329167
##     class counts:     0  1595
##    probabilities: 0.000 1.000
##
## Node number 4: 9771 observations,    complexity param=0.1099055
##   predicted class=0  expected loss=0.06427182  P(node) =0.81425
##     class counts:  9143   628
##    probabilities: 0.936 0.064
##   left son=8 (8008 obs) right son=9 (1763 obs)
##   Primary splits:
##       ID                   < 12000.5 to the left,  improve=366.67560, (0 missing)
##       satisfaction_level   < 0.115   to the right, improve=264.96550, (0 missing)
##       number_project       < 2.5     to the right, improve=118.98290, (0 missing)
##       average_montly_hours < 288     to the left,  improve=103.94700, (0 missing)
##       time_spend_company   < 2.5     to the left,  improve= 25.38359, (0 missing)
##   Surrogate splits:
##       satisfaction_level   < 0.115   to the right, agree=0.835, adj=0.085, (0 split)
##       average_montly_hours < 288     to the left,  agree=0.826, adj=0.033, (0 split)
##       number_project       < 6.5     to the left,  agree=0.824, adj=0.024, (0 split)
##
## Node number 5: 634 observations
##   predicted class=1  expected loss=0  P(node) =0.05283333
##     class counts:     0   634
##    probabilities: 0.000 1.000
##
## Node number 8: 8008 observations
##   predicted class=0  expected loss=0  P(node) =0.6673333
```

```
##     class counts:  8008     0
##    probabilities: 1.000 0.000
##
## Node number 9: 1763 observations,    complexity param=0.1099055
##   predicted class=0  expected loss=0.356211  P(node) =0.1469167
##     class counts:  1135   628
##    probabilities: 0.644 0.356
##   left son=18 (1135 obs) right son=19 (628 obs)
##   Primary splits:
##       ID                   < 12784   to the right, improve=808.59900, (0 missing)
##       satisfaction_level   < 0.465   to the right, improve=265.06650, (0 missing)
##       number_project       < 2.5     to the right, improve=172.87420, (0 missing)
##       time_spend_company   < 2.5     to the left,  improve= 63.60336, (0 missing)
##       average_montly_hours < 275.5   to the left,  improve= 56.52522, (0 missing)
##   Surrogate splits:
##       satisfaction_level   < 0.465   to the right, agree=0.805, adj=0.454, (0 split)
##       number_project       < 2.5     to the right, agree=0.763, adj=0.336, (0 split)
##       average_montly_hours < 274.5   to the left,  agree=0.687, adj=0.123, (0 split)
##       last_evaluation      < 0.575   to the right, agree=0.663, adj=0.054, (0 split)
##       Department           splits as  LRLLLLLLLL, agree=0.653, adj=0.025, (0 split)
##
## Node number 18: 1135 observations
##   predicted class=0  expected loss=0  P(node) =0.09458333
##     class counts:  1135     0
##    probabilities: 1.000 0.000
##
## Node number 19: 628 observations
##   predicted class=1  expected loss=0  P(node) =0.05233333
##     class counts:     0   628
##    probabilities: 0.000 1.000
```

```
## Predicting using CART model
fit.values.cart<-predict(cart.fit,newdata = test_HR)
fit.val1<-ifelse(fit.values.cart[,1]>0.5,1,0)
fit.val2<-ifelse(fit.values.cart[,2]>0.5,1,0)

conf.cart<-confusionMatrix(fit.val2,test_HR$left)
conf.cart
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    0    1
##          0 2285    0
##          1    0  714
##
##                Accuracy : 1
##                  95% CI : (0.9988, 1)
##     No Information Rate : 0.7619
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 1
##  Mcnemar's Test P-Value : NA
##
##             Sensitivity : 1.0000
```

```
##             Specificity : 1.0000
##          Pos Pred Value : 1.0000
##          Neg Pred Value : 1.0000
##              Prevalence : 0.7619
##          Detection Rate : 0.7619
##    Detection Prevalence : 0.7619
##       Balanced Accuracy : 1.0000
##
##        'Positive' Class : 0
##
```

```r
p.cart<-prediction(fit.values.cart[,2],test_HR$left)
p.cart<-performance(p.cart,measure = 'tpr',x.measure = 'fpr')
plot(p.cart)
abline(0,1,lty=2)
```