Assignment 2

Sebastian Paglia & Camila Perez
24/12/2021

#Clear plots and workspace

```
if(!is.null(dev.list())) dev.off()
rm(list=ls())
```

#Set working directory and load the dataframe

```
library(readr)
library(car)
setwd("C:/Users/sebas/OneDrive/Escritorio/Subjects/SIM/Assignment 2 -Description and Data-2021120
3")
filepath<-"C:/Users/sebas/OneDrive/Escritorio/Subjects/SIM/Assignment 2 -Description and Data-202
11203/"
df <- read_csv("aug_train.csv")
## Rows: 19158 Columns: 14
```

#Setting a random sample of 5000 observations as our df

```
### Use birthday of 1 member of the group as random seed:
set.seed(950524)
# Random selection of x registers:
sam<-as.vector(sort(sample(1:nrow(df),5000)))
head(df)  #Taking a look to the first rows/instances (6 rows)
df<-df[sam,]  # Subset of rows _ It will be my sample
summary(df)
##   enrollee_id        city           city_development_index    gender
##  Min.   :    1   Length:5000        Min.   :0.4480         Length:5000
##  1st Qu.: 8588   Class :character   1st Qu.:0.7400         Class :character
##  Median :17035   Mode  :character   Median :0.9030         Mode  :character
##  Mean   :16891                      Mean   :0.8301
##  3rd Qu.:25113                      3rd Qu.:0.9200
##  Max.   :33374                      Max.   :0.9490
##  relevent_experience enrolled_university education_level    major_discipline
##  Length:5000         Length:5000         Length:5000        Length:5000
##  Class :character    Class :character    Class :character   Class :character
##  Mode  :character    Mode  :character    Mode  :character   Mode  :character
##
##   experience         company_size       company_type       last_new_job
##  Length:5000        Length:5000        Length:5000        Length:5000
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##  training_hours      target
##  Min.   :  1.0   Min.   :0.0000
##  1st Qu.: 24.0   1st Qu.:0.0000
##  Median : 48.0   Median :0.0000
##  Mean   : 65.8   Mean   :0.2404
##  3rd Qu.: 89.0   3rd Qu.:0.0000
##  Max.   :336.0   Max.   :1.0000
save(list = c("df"),file="DatasetSample.RData")
```

#Clean workspace again and load our new df with 5000 observations

```
rm(list=ls())
filepath<-"C:/Users/sebas/OneDrive/Escritorio/Subjects/SIM/Assignment 2 -Description and Data-202
11203/"
load(paste0(filepath, "DatasetSample.RData"))
```

#Useful functions:

```
calcQ <- function(x) {
  s.x <- summary(x)
  iqr<-s.x[5]-s.x[2]
  list(souti=s.x[2]-3*iqr, mouti=s.x[2]-1.5*iqr, min=s.x[1], q1=s.x[2], q2=s.x[3],
       q3=s.x[5], max=s.x[6], mouts=s.x[5]+1.5*iqr, souts=s.x[5]+3*iqr ) }
countNA <- function(x) {
  mis_x <- NULL
  for (j in 1:ncol(x)) {mis_x[j] <- sum(is.na(x[,j])) }
  mis_x <- as.data.frame(mis_x)
  rownames(mis_x) <- names(x)
  mis_i <- rep(0,nrow(x))
  for (j in 1:ncol(x)) {mis_i <- mis_i + as.numeric(is.na(x[,j])) }
  list(mis_col=mis_x,mis_ind=mis_i) }
countX <- function(x,X) {
  n_x <- NULL
  for (j in 1:ncol(x)) {n_x[j] <- sum(x[,j]==X) }
  n_x <- as.data.frame(n_x)
  rownames(n_x) <- names(x)
  nx_i <- rep(0,nrow(x))
  for (j in 1:ncol(x)) {nx_i <- nx_i + as.numeric(x[,j]==X) }
  list(nx_col=n_x,nx_ind=nx_i) }
```

#Useful functions for packages treatment:

```
# Introduce required packages:
requiredPackages <- c("effects","FactoMineR","car", "factoextra","RColorBrewer","ggplot2","dplyr"
,"ggmap","ggthemes","knitr")

#use this function to check if each package is on the local machine
#if a package is installed, it will be loaded
#if any are not, the missing package(s) will be installed and loaded
package.check <- lapply(requiredPackages, FUN = function(x) {
  if (!require(x, character.only = TRUE)) {
    install.packages(x, dependencies = TRUE)
    library(x, character.only = TRUE)
  }
})
```

#Checking df

```
str(df)
## tibble [5,000 x 14] (S3: tbl_df/tbl/data.frame)
##  $ enrollee_id           : num [1:5000] 28806 27107 29452 4167 31972 ...
##  $ city                  : chr [1:5000] "city_160" "city_103" "city_21" "city_103" ...
##  $ city_development_index: num [1:5000] 0.92 0.92 0.624 0.92 0.843 0.855 0.624 0.92 0.92 0.884
...
##  $ gender                : chr [1:5000] "Male" "Male" NA NA ...
##  $ relevent_experience   : chr [1:5000] "Has relevent experience" "Has relevent experience" "N
o relevent experience" "Has relevent experience" ...
##  $ enrolled_university   : chr [1:5000] "no_enrollment" "no_enrollment" "Full time course" "no
_enrollment" ...
##  $ education_level       : chr [1:5000] "High School" "Graduate" "High School" "Graduate" ...
##  $ major_discipline      : chr [1:5000] NA "STEM" NA "STEM" ...
##  $ experience            : chr [1:5000] "5" "7" "2" "1" ...
##  $ company_size          : chr [1:5000] "50-99" "50-99" NA "50-99" ...
##  $ company_type          : chr [1:5000] "Funded Startup" "Pvt Ltd" NA "Pvt Ltd" ...
##  $ last_new_job          : chr [1:5000] "1" "1" "never" "never" ...
##  $ training_hours        : num [1:5000] 24 46 32 106 68 22 148 72 50 106 ...
##  $ target                : num [1:5000] 0 1 1 0 0 0 1 0 0 0 ...
names(df)
##  [1] "enrollee_id"            "city"                   "city_development_index"
##  [4] "gender"                 "relevent_experience"    "enrolled_university"
```

```
##  [7] "education_level"        "major_discipline"       "experience"
## [10] "company_size"           "company_type"           "last_new_job"
## [13] "training_hours"         "target"
##Duplicated obs
sum(duplicated(df))
## [1] 0
#No duplicated observation
```

#Setting as factors and numerics

```
df<-df[,-c(1)] #remove enrollee_id (not significant variable)
df$city = as.factor(df$city)
df$training_hours = as.numeric(df$training_hours)
df$gender = as.factor(df$gender)
df$relevent_experience = as.factor(df$relevent_experience)
df$enrolled_university = as.factor(df$enrolled_university)
df$education_level = as.factor(df$education_level)
df$major_discipline = as.factor(df$major_discipline)
df$experience = as.factor(df$experience)
df$company_size = as.factor(df$company_size)
df$company_type = as.factor(df$company_type)
df$last_new_job = as.factor(df$last_new_job)
df$training_hours = as.numeric(df$training_hours)
df$target = as.factor(df$target)
```

#Explore NA's

```
NAs=sapply(df, function(y) round((sum(length(which(is.na(y))))/nrow(df))*100.00,2))
data.frame(NAs)
missings=countNA(df)
sum(missings$mis_col)
## [1] 5475
# There are 5475 missings observations before starting to clean our dataset
```

#Reducing "City" levels into "Standard_city"(100-199), "Big_city"(>200) and "Small_city"(<100).

```
head(summary(df$city))
## city_103  city_21  city_16 city_114 city_160 city_136
##     1110      689      400      344      228      154
# No missings values

plot(sort(table(df$city), decreasing=TRUE)[1:20],type='h', xlab ="", cex.axis = 0.8, las=2, main
='Frecuency of city', ylab = 'Frecuency')
```

See plot Appendant (1)

```
tab <- c(table(df$city))
citynames <- setNames(names(tab), names(tab))
citynames[tab >= 100 ] <- "Standard_city"
citynames[tab > 200] <- "Big_city"
citynames[tab < 100] <- "Small_city"

#Taking into account "city" is a factor with 119 levels, we decided to create groups to reduce th
e amounts of levels, which will allow us to create more adequate and efficient models.

#Create new factor with proper labels
df$city_group<- factor(citynames[as.character(df$city)])
tab1<-prop.table(table(df$city,df$target))
tab2<-prop.table(table(df$city_group,df$target));tab2
##                      0      1
##   Big_city      0.4036 0.1506
##   Small_city    0.3078 0.0842
##   Standard_city 0.0482 0.0056
par(mfrow=c(1,2))
barplot(tab1, main = 'Contingency Table City - Target',xlab = 'Target', ylab = 'City')
```

```
barplot(tab2, legend.text = T, main = 'Contingency Table Group City - Target',xlab = 'Target', yl
ab = 'Group City')
```

See plot Appendant (2)

```
par(mfrow=c(1,1))
#With the contingency table we can observe that the proportions remain similar to the original on
es.
```

#Cleaning factors: “Gender” and “Relevant Experience”; reducing levels, and setting NAs from factors as “No Indicated”.

```
##Gender
summary(df$gender)
## Female   Male  Other   NA's
##    304   3481     52   1163
#1163 missing values
plot(df$gender, main = 'Factor - Gender', ylab = 'Frequency')
```

See plot Appendant (3)

```
levels(df$gender) <- c("Female", "Male", "Other", "No Indicated")
df$gender[which(is.na(df$gender))]<-"No Indicated"
summary(df$gender)
##       Female          Male         Other No Indicated
##          304          3481            52         1163
df$gender<-factor(df$gender, labels = c('Female','Male','No Indicated','No Indicated'))
summary(df$gender)
##       Female          Male No Indicated
##          304          3481         1215
#It's a very unbalanced factor, "Male" represent the level with more frequency.
#1163 missing values, plus 52 "Other". Total of 1215 as "No Indicated"

##Relevant experience
#Replace relevent for relevant
df$relevant_experience<-df$relevent_experience
df<-df[,-c(4)]
summary(df$relevant_experience)
## Has relevent experience  No relevent experience
##                    3601                    1399
levels(df$relevant_experience) <- c("Yes", "No")
plot(df$relevant_experience, main = 'Factor - Relevant experience', ylab = 'Frequency')
```

See plot Appendant (4)

```
#It's an unbalanced factor, "Yes" represent the level with more frequency.
#No missing values
```

#Cleaning factor “Enrolled university”, reducing levels, and setting NAs from factors as “No Indicated”.

```
##Enrolled university
summary(df$enrolled_university)
## Full time course    no_enrollment Part time course             NA's
##              975             3594              324              107
levels(df$enrolled_university) <- c("Full time course", "No enrollment", "Part time course", "No
Indicated")
df$enrolled_university[which(is.na(df$enrolled_university))]<-"No Indicated"
summary(df$enrolled_university)
plot(df$enrolled_university, main = 'Factor - Enrolled university', ylab = 'Frequency')
```

See plot Appendant (5)

```
#It's a very unbalanced factor, "No enrollment" represent the level with more frequency.
#107 missing values as "No Indicated"

##Reducing levels Enrolled university into "Yes", "No", "No Indicated"
```

```
df$group_enrolled_university<-factor(df$enrolled_university, labels = c('Yes','No','Yes','No Indi
cated'))
summary(df$group_enrolled_university)
##          Yes           No No Indicated
##         1299         3594          107
##contingency table
tab3<-prop.table(table(df$enrolled_university,df$target)); tab3
##                         0      1
##   Full time course 0.1228 0.0722
##   No enrollment    0.5732 0.1456
##   Part time course 0.0488 0.0160
##   No Indicated     0.0148 0.0066
tab4<-prop.table(table(df$group_enrolled_university,df$target)); tab4
##                     0      1
##   Yes          0.1716 0.0882
##   No           0.5732 0.1456
##   No Indicated 0.0148 0.0066
par(mfrow=c(1,2))
barplot(tab3, main = ' Contingency Table Enrolled university - Target',xlab = 'Target', ylab = 'C
ity')
barplot(tab4, legend.text = T, main = ' Contingency Table Group Enrolled University - Target',xla
b = 'Target', ylab = 'Group City')
```

See plot Appendant (6)

```
par(mfrow=c(1,1))

#With the contingency table we can observe that the proportions remain similar to the original on
es.
```

#Cleaning factor "Major discipline", reducing levels, and setting NAs from factors as "No Indicated".

```
##major_discipline
summary(df$major_discipline)
##            Arts Business Degree      Humanities        No Major           Other
##              69               88             163              47              89
##            STEM             NA's
##            3785              759
levels(df$major_discipline) <- c(levels(df$major_discipline), 'Not Apply', 'No Indicated')
df$major_discipline[which(df$education_level=='High School')] ='Not Apply'
df$major_discipline[which(df$education_level=='Primary School')] ='Not Apply'
df$major_discipline[which(is.na(df$major_discipline))]<-"No Indicated"
df$major_discipline<-factor(df$major_discipline, labels = c('Arts&humanities','Business Degree','
Arts&humanities','No Indicated','No Indicated','STEM', 'No Indicated', 'No Indicated'))
summary(df$major_discipline)
## Arts&humanities Business Degree    No Indicated            STEM
##             232              88             895            3785
plot(df$major_discipline, main = 'Factor - Major discipline', ylab = 'Frequency')
```

See plot Appendant (7)

```
#It's a very unbalanced factor, "STEM" represent the level with more frequency.)
# We grouped some disciplines in order to reduce levels of a factor variable, taking into account
that we consider relevant to group Arts with Humanities, and, No Major, Other and NAs as No Indic
ated
```

#Cleaning factor "education_level";grouping and setting NAs from factors as "No Indicated".

```
##education_level
summary(df$education_level)
##       Graduate    High School        Masters            Phd Primary School
##           3000            540           1147            102             85
##           NA's
##            126
```

```r
library(forcats)
df$education_level <- fct_collapse(df$education_level,
  Pre_graduate= c('High School', 'Primary School'),
  Post_graduate = c('Masters', 'Phd'),
  Graduate = 'Graduate')
summary(df$education_level)
##      Graduate  Pre_graduate Post_graduate           NA's
##          3000           625          1249            126
levels(df$education_level) <- c("Pre_graduate", "Post_graduate", "Graduate", "No Indicated")
df$education_level[which(is.na(df$education_level))]<-"No Indicated"
summary(df$education_level)
##  Pre_graduate Post_graduate      Graduate  No Indicated
##          3000           625          1249           126
plot(df$education_level, main = 'Factor - Education level', ylab = 'Frequency')
```

See plot Appendant (8)

```r
# It's a very unbalanced factor, "Pre-graduate" represent the level with more frequency.
# We grouped some categories in order to reduce levels according to the education level.
#126 missing values as "No Indicated"

##city_development_index
summary(df$city_development_index)
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.4480  0.7400  0.9030  0.8301  0.9200  0.9490
hist(df$city_development_index, main = 'City Development Index', ylab = 'Frequency', xlab='Index'
)
```

See plot Appendant (9)

```r
#No missing values and it's doesn't seem normally distributed.
```

#Cleaning variable “Experience”, create a factor version and setting NAs from factors as “No Indicated”.

```r
##Experience

#numeric
summary(df$experience) #in years
##   <1  >20    1   10   11   12   13   14   15   16   17   18   19    2   20    3
##  143  844  145  265  164  120  108  152  187  135   96   79   78  309   37  373
##    4    5    6    7    8    9 NA's
##  344  371  299  287  195  255   14
#Has 14 missing values that we are going to treat afterwards
sorted_labels<-suppressWarnings(paste(sort(as.integer(levels(df$experience)))))
levels(df$experience) <- c("0", "21", "1", "10", "11", "12", "13", "14", "15", "16", "17", "18",
"19", "2", "20", "3", "4", "5", "6", "7", "8", "9")
sorted_labels<-paste(sort(as.integer(levels(df$experience))))
df$experience<-factor(df$experience, levels = sorted_labels)
summary(df$experience)
##    0    1    2    3    4    5    6    7    8    9   10   11   12   13   14   15
##  143  145  309  373  344  371  299  287  195  255  265  164  120  108  152  187
##   16   17   18   19   20   21 NA's
##  135   96   79   78   37  844   14
df$experience <- as.numeric(as.character(df$experience))
table(df$experience)
##   0   1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16  17  18  19
## 143 145 309 373 344 371 299 287 195 255 265 164 120 108 152 187 135  96  79  78
##  20  21
##  37 844
summary(df$experience)
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    0.00    4.00    9.00   10.06   16.00   21.00      14
hist(df$experience, main = 'Experience', ylab = 'Frequency')
```

See plot Appendant (10)

```
#No missing values and it's doesn't seem normally distributed.

#Experience as a factor
df$f.experience <-df$experience
#Grouping it in a new factor with 5 intervals
table(df$f.experience)
##   0   1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16  17  18  19
## 143 145 309 373 344 371 299 287 195 255 265 164 120 108 152 187 135  96  79  78
##  20  21
##  37 844
df$f.experience[df$experience<=5]<-"0-5"
df$f.experience[df$experience > 5 & df$experience <= 10]<-"6-10"
df$f.experience[df$experience > 10 & df$experience <= 15] <- "11-15"
df$f.experience[df$experience > 15 & df$experience <= 20]<- "16-20"
df$f.experience[df$experience == 21] <- ">20"
df$f.experience = as.factor(df$f.experience)
levels(df$f.experience) <- c(levels(df$f.experience), 'No Indicated')
df$f.experience[which(is.na(df$f.experience))]<-"No Indicated"
summary(df$f.experience)
##          >20          0-5        11-15        16-20         6-10 No Indicated
##          844         1685          731          425         1301           14
plot(df$f.experience, main = 'Factor - Experience', ylab = 'Frequency')
```

See plot Appendant (11)

```
#14 missing values as "No Indicated"
```

##Cleaning factors: “Company size” and “Company Type”, reducing levels, and setting NAs from factors as “No Indicated”.

```
##company_size
summary(df$company_size)
##       <10     10/49   100-500 1000-4999    10000+     50-99   500-999 5000-9999
##       336       378       671       340       527       806       243       126
##      NA's
##      1573
df$company_size <- factor(df$company_size, labels = c("SME", "SME", "SME","Big", "Big","SME", "Bi
g", "Big"))
levels(df$company_size) <- c(levels(df$company_size),"No Indicated")
df$company_size[which(is.na(df$company_size))]<-"No Indicated"
plot(df$company_size,cex.axis = 0.8, main = 'Company size', ylab = 'Frequency')
```

See plot Appendant (12)

```
summary(df$company_size)
##          SME          Big No Indicated
##         2191         1236         1573
#Regroup into Small and Medium-sized Enterprises or Big companies
#1573 missing values as "No Indicated"

##company_type
summary(df$company_type)
## Early Stage Startup      Funded Startup                 NGO               Other
##                 141                 251                 135                  25
##       Public Sector             Pvt Ltd                NA's
##                 263                2546                1639
levels(df$company_type) <- c(levels(df$company_type),'No Indicated')
df$company_type[which(is.na(df$company_type))]<-"No Indicated"
df$company_type <- factor(df$company_type, labels = c("Startup", "Startup", "NGO", "No Indicated"
, "Public Sector","Private Limited Company", "No Indicated"))
plot(df$company_type, cex.axis = 0.8, main = 'Factor : Company type', ylab = 'Frequency')
```

See plot Appendant (13)

```
summary(df$company_type)
##                 Startup                     NGO            No Indicated
##                     392                     135                    1664
##           Public Sector Private Limited Company
##                     263                    2546
#Unify Startups, and Other with "No Indicated"
# It's a very unbalanced factor, "Private Limited Company" represent the level with more frequenc
y.
#1639 missing values as no indicated
```

##Cleaning factor: “Last new job” reducing levels, and setting NAs from factors as “No Indicated”.

```
##last_new_job
summary(df$last_new_job)
##    >4     1     2     3     4 never  NA's
##   831  2092   763   271   272   677    94
df$last_new_job <- factor(df$last_new_job, labels = c(">4", "1", "2","3", "4","None"))
levels(df$last_new_job) <- c(levels(df$last_new_job),'No Indicated')
df$last_new_job[which(is.na(df$last_new_job))]<-"No Indicated"
summary(df$last_new_job)
##           >4            1            2            3            4         None
##          831         2092          763          271          272          677
## No Indicated
##           94
plot(df$last_new_job, cex.axis = 0.8, main = 'Factor : Last new Job', ylab = 'Frequency')
```

See plot Appendant (14)

```
#Unbalanced factor, "1" represent the level with more frequency.
#94 missing values as "No Indicated"

##Group them into greater than cero, never or no indicated
df$group_last_new_job<-factor(df$last_new_job, labels = c('>0','>0','>0','>0','>0','None','No Ind
icated'))
summary(df$group_last_new_job)
##           >0         None No Indicated
##         4229          677           94
plot(df$group_last_new_job, cex.axis = 0.8, main = 'Factor : Last new Job', ylab = 'Frequency')
```

See plot Appendant (15)

```
##contingency table
tab5<-prop.table(table(df$last_new_job,df$target))
tab5
##                    0      1
##   >4          0.1378 0.0284
##   1           0.3100 0.1084
##   2           0.1196 0.0330
##   3           0.0428 0.0114
##   4           0.0422 0.0122
##   None        0.0958 0.0396
##   No Indicated 0.0114 0.0074
tab6<-prop.table(table(df$group_last_new_job,df$target))
tab6
##                    0      1
##   >0          0.6524 0.1934
##   None        0.0958 0.0396
##   No Indicated 0.0114 0.0074
par(mfrow=c(1,2))
barplot(tab5, main = 'Contingency Table Last new job - Target',xlab = 'Target', ylab = 'City')
```

```
barplot(tab6, legend.text = T, main = 'Contingency Table Group Last new job - Target',xlab = 'Target', ylab = 'Group City')
```
See plot Appendant (16)

```
par(mfrow=c(1,1))

#With the contingency table we can observe that the proportions remain equivalent to the original ones.
```

##Inspection of the variables "training hours" and "target"

```
##training_hours
summary(df$training_hours)
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     1.0    24.0    48.0    65.8    89.0   336.0
hist(df$training_hours)
```
See plot Appendant (17)

```
#No missing values and it's doesn't seem normally distributed.

##Target
summary(df$target) #0: Not looking for a job change. 1: Looking for a job change
##    0    1
## 3798 1202
plot(df$target, main='Target')
```
See plot Appendant (18)

```
#No missing values
#Very unbalanced factor, "0" represent the level with more frequency.

summary(df)
```

#Checking if any new duplicated observation was generated.

```
sum(duplicated(df))
## [1] 7
df <- df[-c(which(duplicated(df))),]
dim(df)
## [1] 4993   17
summary(df)
#We found 7 observations that were duplicated, so we removed them from our dataframe.
```

#CountNAs

```
missings=countNA(df)
sum(missings$mis_col)
## [1] 14
summary(df)
#So far we have 14 missing values from experience (numerical) that we haven't treated yet
```

#Treating Univariate Outliers

```
str(df)
names(df)
par(mfrow=c(1,1))

##experience
Boxplot(df$experience, main="Boxplot Experience", ylab='frequency') #No outliers
```
See plot Appendant (19)

```
##city_development_index
Boxplot(df$city_development_index, main="Boxplot City Development Index", ylab='frequency')
## [1] 1221 1446 4615
upsevout<-quantile(df$city_development_index,0.75, na.rm = T)+3*(quantile(df$city_development_index,0.75, na.rm = T)-quantile(df$city_development_index,0.25, na.rm = T))
```

```r
abline(h=upsevout,col="red",lwd=2)
uploutse<-which(df$city_development_index>upsevout[1]);length(uploutse)
## [1] 0
losevout<-quantile(df$city_development_index,0.25, na.rm = T)-3*(quantile(df$city_development_ind
ex,0.75, na.rm = T)-quantile(df$city_development_index,0.25, na.rm = T))
abline(h=losevout,col="red",lwd=2)
```

See plot Appendant (20)

```r
loloutse<-which(df$city_development_index<losevout[1]);length(loloutse)
## [1] 0
#No severe outliers


##training hours
Boxplot(df$training_hours, main="Boxplot Traning Hours", ylab='frequency')
##  [1] 4065 4805 4448   27  817 1730 2359 1924 3169 4212
upsevout<-quantile(df$training_hours,0.75, na.rm = T)+3*(quantile(df$training_hours,0.75, na.rm =
T)-quantile(df$training_hours,0.25, na.rm = T))
abline(h=upsevout,col="red",lwd=2)
uploutse<-which(df$training_hours>upsevout[1]);length(uploutse)
## [1] 57
#Upper threshold that identifies 57 severe outliers

losevout<-quantile(df$training_hours,0.25, na.rm = T)-3*(quantile(df$training_hours,0.75, na.rm =
T)-quantile(df$training_hours,0.25, na.rm = T))
abline(h=losevout,col="red",lwd=2)
```

See plot Appendant (21)

```r
loloutse<-which(df$training_hours<losevout[1]);length(loloutse)
## [1] 0
##Setting outliers as NAs and also as "No indicated", the variable created to compute the sum of
error and unknown observations.
df$No_Indicated = 0
df$No_Indicated[which(df$training_hours>upsevout)] = 1
df$training_hours[which(df$training_hours>upsevout)] = NA
summary(df$training_hours)
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    1.00   23.00   47.00   62.96   87.00  288.00      57
#57 severe outliers as NAs

table(df$No_Indicated)
##    0    1
## 4936   57
```

#CountNAs

```r
missings=countNA(df)
sum(missings$mis_col)
## [1] 71
##total of 71  missing values (14 from experience + 57 from training hours)
```

#NAs Imputation

```r
library(missMDA)
nb <- estim_ncpPCA(df$training_hours,method.cv = "Kfold", verbose = FALSE) # estimate
nb$ncp
## [1] 0
names(df)
res.pca<-imputePCA(df[,c(11,7)], ncp=0) #”experience” and “training_hours”

summary(df[,c(11,7)])# original
##  training_hours     experience
##  Min.   :  1.00   Min.   : 0.00
```

```
##  1st Qu.: 23.00   1st Qu.: 4.00
##  Median : 47.00   Median : 9.00
##  Mean   : 62.96   Mean   :10.05
##  3rd Qu.: 87.00   3rd Qu.:16.00
##  Max.   :288.00   Max.   :21.00
##  NA's   :57       NA's   :14
summary(res.pca$completeObs)# imputed
##  training_hours     experience
##  Min.   :  1.00   Min.   : 0.00
##  1st Qu.: 24.00   1st Qu.: 4.00
##  Median : 48.00   Median : 9.00
##  Mean   : 62.96   Mean   :10.05
##  3rd Qu.: 86.00   3rd Qu.:16.00
##  Max.   :288.00   Max.   :21.00
#No significant variation between original and imputed, therefore, imputation was done correctly.

##replace imputed observations
df[,c(11,7)]<-res.pca$completeObs
missings=countNA(df)
sum(missings$mis_col)
## [1] 0
```

#Multivariate outliers

```
library(chemometrics)
names(df)
res.mout <- Moutlier( df[,c(2,11)], quantile = 0.999, plot=F) # “city_development_index” and “training_hours”
par(mfrow=c(1,1))
plot( res.mout$md, res.mout$rd, main= 'Multivariate Outliers', ylab='frequency' )
abline( h=res.mout$cutoff, lwd=2, col="red")
abline( v=res.mout$cutoff, lwd=2, col="red")
```

See plot Appendant (22)

```
mout_out <- which((res.mout$md > res.mout$cutoff ) & (res.mout$rd > res.mout$cutoff) );mout_out
str(mout_out)
summary(df)
mout=df[c(mout_out),];mout
summary(mout)
##        city    city_development_index          gender
##  city_21 :8    Min.    :0.5270         Female      : 0
##  city_11 :2    1st Qu.:0.5790          Male        :14
##  city_128:2    Median :0.6240          No Indicated: 7
##  city_67 :2    Mean    :0.6483
##  city_101:1    3rd Qu.:0.6980
##  city_102:1    Max.    :0.8550
##  (Other) :5
##        enrolled_university      education_level        major_discipline
##  Full time course: 5       Pre_graduate :18     Arts&humanities: 0
##  No enrollment   :16       Post_graduate: 0     Business Degree: 1
##  Part time course: 0       Graduate     : 2     No Indicated   : 2
##  No Indicated    : 0       No Indicated : 1     STEM           :18
##
##    experience           company_size                   company_type
##  Min.   : 1.00   SME          :11    Startup                 : 0
##  1st Qu.: 3.00   Big          : 4    NGO                     : 1
##  Median : 5.00   No Indicated: 6     No Indicated            : 5
##  Mean   : 6.19                       Public Sector           : 1
##  3rd Qu.: 9.00                       Private Limited Company:14
##  Max.   :18.00
##
##        last_new_job training_hours  target         city_group
```

```
##  >4          : 0    Min.   :228.0   0:11   Big_city     : 8
##  1           :12    1st Qu.:260.0   1:10   Small_city   :11
##  2           : 5    Median :268.0          Standard_city: 2
##  3           : 1    Mean   :266.3
##  4           : 1    3rd Qu.:278.0
##  None        : 2    Max.   :286.0
##  No Indicated: 0
##  relevant_experience group_enrolled_university       f.experience
##  Yes:16              Yes         : 5           >20         : 0
##  No : 5              No          :16           0-5         :11
##                      No Indicated: 0           11-15       : 2
##                                                16-20       : 1
##                                                6-10        : 7
##                                                No Indicated: 0
##     group_last_new_job  No_Indicated
##  >0          :19       Min.   :0
##  None        : 2       1st Qu.:0
##  No Indicated: 0       Median :0
##                        Mean   :0
##                        3rd Qu.:0
##                        Max.   :0
#Taking into account the numerical variables (city_development_index and training_hours), we can
observe 21 possible multivariate outliers (using 99.9% CI). Relevant characteristics in common:
#- None of this observations has Female as gender. However, the df is unbalanced so it is not rar
e that they share this characteristic. The same happens with No enrollment (enrolled_university),
STEM (major_discipline), SME (company_size), Private Limited Company (company_type), last_new_job
, relevant_experience, No enrolled to a university, f.experience and group_last_new_job;
#- Most of them are Pre-graduate 86%, while in the df the proportion of Pre-graduate is 60%;
#- Equity distributed on target variable, whereas, in the df the proportions are 76% related to 1
(look for a job change) and 34% to 0 (not looking for a job change);
#- 52% of the multivariate outliers live in small cities and 38% in big ones. But in the df the r
elation is inverted, 39% live in small cities and 55% in big ones.
#However, in this case and based on the plot we decided to maintain those observations because we
don't see very clear that they should be treated as outliers or atypical observations among our d
ataset.
```

#Unknown, errors and NA's variable

```
names(df)
df_group<-df[,-c(1,4,10,16)] # Remove “city”, “enrolled_university”, “last_new_job” and “f.experi
ence”
df$No_Indicated = (rowSums(df_group[,c(1:13)] == "No Indicated") + df$No_Indicated)
df_group$No_Indicated <- as.numeric(as.character(df$No_Indicated))
table(df_group$No_Indicated)
##    0    1    2    3    4    5    6    7
## 2216 1013  942  542  217   44   16    3
sum(df_group$No_Indicated>0)
#There are 2777 observations as "No Indicated"
```

#Correlation with “No Indicated” variable

```
library(FactoMineR)
names(df_group)
##  [1] "city_development_index"    "gender"
##  [3] "education_level"           "major_discipline"
##  [5] "experience"                "company_size"
##  [7] "company_type"              "training_hours"
##  [9] "target"                    "city_group"
## [11] "relevant_experience"       "group_enrolled_university"
## [13] "group_last_new_job"        "No_Indicated"
res.con <- condes(df_group, num.var=14, proba = 0.01 )
res.con$quanti
```

```
##                         correlation      p.value
## city_development_index  -0.1341615 1.720125e-21
## experience              -0.2539130 2.626110e-74
res.con$quali
##                                   R2       p.value
## education_level           0.338643127  0.000000e+00
## major_discipline          0.324098245  0.000000e+00
## company_size              0.640141382  0.000000e+00
## company_type              0.632309466  0.000000e+00
## gender                    0.213227394 1.392608e-260
## group_last_new_job        0.211356038 5.224577e-258
## relevant_experience       0.208306095 1.749399e-255
## group_enrolled_university 0.125659528 3.112088e-146
## target                    0.036766254  1.480066e-42
## city_group                0.007317961  1.099916e-08
res.con$category
##                                          Estimate       p.value
## company_type=No Indicated               1.71358972  0.000000e+00
## company_size=No Indicated               1.49576417  0.000000e+00
## major_discipline=No Indicated           1.40486206  0.000000e+00
## gender=No Indicated                     0.96539488 1.925096e-261
## relevant_experience=No                  0.65841320 1.749399e-255
## education_level=Post_graduate           0.53827555 4.222590e-224
## group_last_new_job=None                 0.23154303 1.584565e-179
## education_level=No Indicated            1.98890412 1.766235e-155
## group_enrolled_university=No Indicated  1.47162018  5.381045e-81
## group_last_new_job=No Indicated         1.05820321  6.187053e-58
## target=1                                0.29064607  1.480066e-42
## city_group=Small_city                   0.14048606  1.572041e-09
## major_discipline=Arts&humanities       -0.45880575  5.966932e-04
## city_group=Big_city                    -0.08921505  2.055727e-08
## company_type=Public Sector             -0.22959471  3.867507e-10
## gender=Female                          -0.54895952  1.575426e-10
## company_type=NGO                       -0.46767783  4.016251e-11
## education_level=Graduate               -1.32249234  1.406756e-34
## company_type=Startup                   -0.53184379  1.245752e-36
## target=0                               -0.29064607  1.480066e-42
## group_enrolled_university=Yes          -0.38673960  1.653353e-51
## education_level=Pre_graduate           -1.20468733  9.852167e-75
## group_enrolled_university=No           -1.08488058  2.472346e-98
## company_size=Big                       -0.79425641 2.190480e-132
## gender=Male                            -0.41643536 2.006641e-174
## group_last_new_job=>0                  -1.28974624 1.866360e-250
## company_size=SME                       -0.70150776 3.912663e-255
## relevant_experience=Yes                -0.65841320 1.749399e-255
## major_discipline=STEM                  -0.52350598 2.279536e-282
## company_type=Private Limited Company   -0.48447339  0.000000e+00
#The numerical variables are not highly correlated with our new variable. Between categorical var
iables, the most important and with higher R2 are "company_size" and "company_type" with 64% and
63%, followed by education_level and major_discipline with 33% and 32%.
```

#Train - Test - Split our dataset into train and test with 75% and 25% respectively.

```
names(df)
##  [1] "city"                    "city_development_index"
##  [3] "gender"                  "enrolled_university"
##  [5] "education_level"         "major_discipline"
##  [7] "experience"              "company_size"
##  [9] "company_type"            "last_new_job"
## [11] "training_hours"          "target"
## [13] "city_group"              "relevant_experience"
```

```
## [15] "group_enrolled_university" "f.experience"
## [17] "group_last_new_job"        "No_Indicated"
new_df<-df[,-c(1,4,10,18)]
set.seed(950524)
ind <- sample(2, nrow(new_df), replace = T, prob = c(0.75, 0.25))
train <- new_df[ind == 1,]
test <- new_df[ind == 2,]
```

#Balance review

```
names(train)
dim(train)
## [1] 3712   14
plot(df$target, main="Target", ylab='frequency')
```

See plot Appendant (23)

```
plot(df$city_group, main="City Group", ylab='frequency')
```

See plot Appendant (24)

```
plot(df$gender, main="Gender", ylab='frequency')
```

See plot Appendant (25)

```
plot(df$education_level, main="Education level", ylab='frequency')
```

See plot Appendant (26)

```
plot(df$major_discipline, main="Major Discipline", ylab='frequency')
```

See plot Appendant (27)

```
plot(df$company_size, main="Company Size", ylab='frequency')
```

See plot Appendant (28)

```
plot(df$company_type, main="Company Type", ylab='frequency')
```

See plot Appendant (29)

```
plot(df$group_last_new_job, main="Last New Job", ylab='frequency')
```

See plot Appendant (30)

```
plot(df$relevant_experience, main="Relevant Experience", ylab='frequency')
```

See plot Appendant (31)

```
plot(df$group_enrolled_university, main="Group Enrolled University", ylab='frequency')
```

See plot Appendant (32)

```
plot(df$f.experience, main="Factor Experience", ylab='frequency')
```

See plot Appendant (33)

```
#We can see, as we checked above, that our dataset is unbalanced in many variables but we are not
going to treat this issue in the project.
```

#Building models

```
#Total model
m0 <- glm(target ~ .,data=train,family=binomial)
summary(m0) # AIC: 3358.9 - reference
BIC(m0) # BIC: 3545.5
# We will check AIC and BIC for every model, taking into account that BIC is a variant of AIC wit
h a stronger penalty for including additional variables to the model.
```

#Numeric model

```
mnumeric <- glm(target ~ city_development_index +  training_hours + experience ,data=train,family
=binomial )
summary(mnumeric) # AIC: 3661.7
```

```
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5376  -0.6305  -0.5533  -0.4791   2.1343
## Coefficients:
##                            Estimate Std. Error z value Pr(>|z|)
## (Intercept)               3.5928522  0.2574862  13.954  < 2e-16 ***
## city_development_index -5.5799985  0.3307861 -16.869  < 2e-16 ***
## training_hours         -0.0003374  0.0007651  -0.441 0.659230
## experience             -0.0238827  0.0067772  -3.524 0.000425 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4054.2  on 3711  degrees of freedom
## Residual deviance: 3653.7  on 3708  degrees of freedom
## AIC: 3661.7
## Number of Fisher Scoring iterations: 4
BIC(mnumeric) # BIC: 3686.6
mnumeric_2 <- glm(target ~ poly(city_development_index,2) +  training_hours + experience ,data=tr
ain,family=binomial)
summary(mnumeric_2) # AIC: 3632.1
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1524  -0.6280  -0.5719  -0.5088   2.0701
## Coefficients:
##                                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)                       -1.027e+00  9.070e-02 -11.322  < 2e-16 ***
## poly(city_development_index, 2)1 -4.015e+01  2.477e+00 -16.210  < 2e-16 ***
## poly(city_development_index, 2)2  1.354e+01  2.508e+00   5.398 6.73e-08 ***
## training_hours                    -3.852e-04  7.732e-04  -0.498 0.618373
## experience                        -2.415e-02  6.794e-03  -3.555 0.000379 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## (Dispersion parameter for binomial family taken to be 1)
##     Null deviance: 4054.2  on 3711  degrees of freedom
## Residual deviance: 3622.1  on 3707  degrees of freedom
## AIC: 3632.1
## Number of Fisher Scoring iterations: 4
BIC(mnumeric_2) # BIC: 3663.2
anova(mnumeric, mnumeric_2, test="Chisq")
## Analysis of Deviance Table
## Model 1: target ~ city_development_index + training_hours + experience
## Model 2: target ~ poly(city_development_index, 2) + training_hours + experience
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1      3708     3653.7
## 2      3707     3622.1  1   31.625 1.87e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#We reject H0 so we can say that both models are not equivalent and maintain order 2 on city_deve
lopment_index

mnumeric_3 <- glm(target ~ city_development_index +  poly(training_hours,2) + experience ,data=tr
ain,family=binomial)
summary(mnumeric_3) # AIC: 3661.9
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5417  -0.6330  -0.5515  -0.4717   2.2331
## Coefficients:
##                              Estimate Std. Error z value Pr(>|z|)
## (Intercept)                  3.574370   0.251668  14.203  < 2e-16 ***
```

```
## city_development_index   -5.584057   0.330867 -16.877  < 2e-16 ***
## poly(training_hours, 2)1 -1.170823   2.560322  -0.457 0.647459
## poly(training_hours, 2)2 -3.404728   2.591938  -1.314 0.188986
## experience               -0.023902   0.006778  -3.526 0.000421 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## (Dispersion parameter for binomial family taken to be 1)
##     Null deviance: 4054.2  on 3711  degrees of freedom
## Residual deviance: 3651.9  on 3707  degrees of freedom
## AIC: 3661.9
## Number of Fisher Scoring iterations: 4
BIC(mnumeric_3) # BIC: 3693
anova(mnumeric, mnumeric_3, test="Chisq")
## Analysis of Deviance Table
## Model 1: target ~ city_development_index + training_hours + experience
## Model 2: target ~ city_development_index + poly(training_hours, 2) + experience
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1      3708     3653.7
## 2      3707     3651.9  1   1.7615   0.1844
#We fail to reject H0 so we can say that both models are equivalent and maintain training_hours w
ithout transformation

mnumeric_4 <- glm(target ~ city_development_index + training_hours +  poly(experience,2) ,data=tr
ain,family=binomial)
summary(mnumeric_4) # AIC: 3663
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5514  -0.6338  -0.5457  -0.4899   2.1142
## Coefficients:
##                           Estimate Std. Error z value Pr(>|z|)
## (Intercept)              3.3472044  0.2739605  12.218  < 2e-16 ***
## city_development_index -5.5745473  0.3309224 -16.845  < 2e-16 ***
## training_hours         -0.0003332  0.0007652  -0.435 0.663254
## poly(experience, 2)1   -9.5973229  2.8138191  -3.411 0.000648 ***
## poly(experience, 2)2    2.0521156  2.5288327   0.811 0.417086
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## (Dispersion parameter for binomial family taken to be 1)
##     Null deviance: 4054.2  on 3711  degrees of freedom
## Residual deviance: 3653.0  on 3707  degrees of freedom
## AIC: 3663
## Number of Fisher Scoring iterations: 4
BIC(mnumeric_4) # BIC: 3694.1
anova(mnumeric, mnumeric_4, test="Chisq")
## Analysis of Deviance Table
## Model 1: target ~ city_development_index + training_hours + experience
## Model 2: target ~ city_development_index + training_hours + poly(experience,
##     2)
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1      3708     3653.7
## 2      3707     3653.0  1   0.6575   0.4174
#We fail to reject H0 so we can say that both models are equivalent and maintain experience witho
ut transformation

#The best numerical model obtained is mnumeric_2 (with second order over city_development_index)
```

#Factor o numeric for “experience”

```
#factor experience or numeric experience
m1 <- glm(formula = target ~ poly(city_development_index,2) + f.experience + training_hours, fami
```

```
ly = binomial, data = train)
summary(m1) # AIC: 3640.7
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.3990  -0.6182  -0.5936  -0.5005   2.1017
## Coefficients:
##                                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)                       -1.507e+00  1.267e-01 -11.893  < 2e-16 ***
## poly(city_development_index, 2)1 -4.084e+01  2.475e+00 -16.501  < 2e-16 ***
## poly(city_development_index, 2)2  1.374e+01  2.519e+00   5.455 4.91e-08 ***
## f.experience0-5                   3.640e-01  1.401e-01   2.598  0.00936 **
## f.experience11-15                 3.169e-01  1.606e-01   1.973  0.04852 *
## f.experience16-20                -6.968e-02  1.985e-01  -0.351  0.72559
## f.experience6-10                  2.825e-01  1.433e-01   1.971  0.04869 *
## f.experienceNo Indicated          9.348e-01  8.008e-01   1.167  0.24312
## training_hours                   -3.903e-04  7.737e-04  -0.504  0.61397
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## (Dispersion parameter for binomial family taken to be 1)
##     Null deviance: 4054.2  on 3711  degrees of freedom
## Residual deviance: 3622.7  on 3703  degrees of freedom
## AIC: 3640.7
## Number of Fisher Scoring iterations: 4
BIC(m1) # BIC: 3696.7
AIC(mnumeric_2,m1)
##            df      AIC
## mnumeric_2  5 3632.053
## m1          9 3640.713
BIC(mnumeric_2,m1)
##            df      BIC
## mnumeric_2  5 3663.150
## m1          9 3696.687
#mnumeric_2 has a lower AIC and BIC number, so we keep "experience" as numerical, according to Ak
aike test and Bayesian Information Criterion.
```

#Numeric model after treatment

```
step(mnumeric_2)
## Start:  AIC=3632.05
## target ~ poly(city_development_index, 2) + training_hours + experience
##
##                                   Df Deviance    AIC
## - training_hours                   1   3622.3 3630.3
## <none>                                 3622.1 3632.1
## - experience                       1   3634.8 3642.8
## - poly(city_development_index, 2)  2   3951.1 3957.1
##
## Step:  AIC=3630.3
## target ~ poly(city_development_index, 2) + experience
##                                   Df Deviance    AIC
## <none>                                 3622.3 3630.3
## - experience                       1   3635.2 3641.2
## - poly(city_development_index, 2)  2   3951.1 3955.1
## Call:  glm(formula = target ~ poly(city_development_index, 2) + experience,
##     family = binomial, data = train)
## Coefficients:
##                      (Intercept)  poly(city_development_index, 2)1
##                         -1.05058                         -40.10956
## poly(city_development_index, 2)2                        experience
##                         13.52601                          -0.02421
## Degrees of Freedom: 3711 Total (i.e. Null);  3708 Residual
```

```
## Null Deviance:       4054
## Residual Deviance: 3622  AIC: 3630
mnumeric_5<-glm(target ~ poly(city_development_index,2) + experience ,data=train,family=binomial)
summary(mnumeric_5) # AIC: 3630.3
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1680  -0.6284  -0.5713  -0.5128   2.0471
## Coefficients:
##                                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)                       -1.050581   0.077390 -13.575  < 2e-16 ***
## poly(city_development_index, 2)1 -40.109561   2.474991 -16.206  < 2e-16 ***
## poly(city_development_index, 2)2  13.526006   2.507626   5.394 6.89e-08 ***
## experience                        -0.024206   0.006793  -3.564 0.000366 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## (Dispersion parameter for binomial family taken to be 1)
##     Null deviance: 4054.2  on 3711  degrees of freedom
## Residual deviance: 3622.3  on 3708  degrees of freedom
## AIC: 3630.3
## Number of Fisher Scoring iterations: 4
BIC(mnumeric_5) # BIC: 3655.2
#After applying step function, we removed training_hours since it is not significant to our model
.
```

#Influential data

```
library(chemometrics)
influenceIndexPlot(mnumeric_5,id=c(method=abs(cooks.distance(mnumeric_5)), n=3))
```

See plot Appendant (34)

```
train[c("2289","2737","3007"),c(1,5,9)]
## # A tibble: 3 x 3
##   city_development_index experience target
##                    <dbl>      <dbl> <fct>
## 1                  0.487        5   0
## 2                  0.479       10.1 0
## 3                  0.479        2   0
summary(train[,c(1,5,9)])
##  city_development_index   experience    target
##  Min.   :0.4480         Min.   : 0.00   0:2837
##  1st Qu.:0.7430         1st Qu.: 4.00   1: 875
##  Median :0.9100         Median : 9.00
##  Mean   :0.8317         Mean   :10.12
##  3rd Qu.:0.9200         3rd Qu.:16.00
##  Max.   :0.9490         Max.   :21.00
Boxplot(cooks.distance(mnumeric_5))
```

See plot Appendant (35)

```
##  [1] 3007 2737 2289 3204 1053 1755 3345  231 3078 3686
#Taking into account the potentially highly influential observations, we can summarise that the t
hree of them belong to the first quartile of the variable city_development_index, below the mean
of experience and none of them would look for a job change. Therefore, we decided to remove them
from the train dataset.
train<-train[-c(2289,2737,3007),]
Boxplot(hatvalues(mnumeric_5))
```

See plot Appendant (36)

```
##  [1] 1906 1305 2737 3345 1093 2289  154 2876 3204 3440
summary(train[c(1906,1305),])
##  city_development_index          gender       education_level
##  Min.   :0.479          Female      :0   Pre_graduate :0
##  1st Qu.:0.481          Male        :1   Post_graduate:1
```

```
##  Median :0.483          No Indicated:1   Graduate     :1
##  Mean   :0.483                           No Indicated :0
##  3rd Qu.:0.485
##  Max.   :0.487
##         major_discipline   experience       company_size
##  Arts&humanities:0       Min.   :19   SME          :0
##  Business Degree:0       1st Qu.:19   Big          :0
##  No Indicated   :1       Median :19   No Indicated:2
##  STEM           :1       Mean   :19
##                          3rd Qu.:19
##                          Max.   :19
##                    company_type training_hours target         city_group
##  Startup                 :0     Min.   :52.0   0:0    Big_city     :0
##  NGO                     :0     1st Qu.:54.5   1:2    Small_city   :2
##  No Indicated            :1     Median :57.0          Standard_city:0
##  Public Sector           :0     Mean   :57.0
##  Private Limited Company:1     3rd Qu.:59.5
##                                 Max.   :62.0
##  relevant_experience group_enrolled_university       f.experience
##  Yes:2               Yes         :0            >20         :0
##  No :0               No          :2            0-5         :0
##                      No Indicated:0            11-15       :0
##                                                16-20       :2
##                                                6-10        :0
##                                                No Indicated:0
##     group_last_new_job
##  >0          :1
##  None        :1
##  No Indicated:0

#There are two observations with the most important hat value: obs 1906 and 1305. We can see that both of them belong to the group that look for a job change.

mnumericfinal<-glm(target ~ poly(city_development_index,2) + experience ,data=train,family=binomial)
summary(mnumericfinal) # AIC: 3616.4
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1159  -0.6261  -0.5710  -0.5107   2.0523
##
## Coefficients:
##                                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)                      -1.052569   0.077541 -13.574  < 2e-16 ***
## poly(city_development_index, 2)1 -40.356842   2.478811 -16.281  < 2e-16 ***
## poly(city_development_index, 2)2  15.076075   2.535747   5.945 2.76e-09 ***
## experience                       -0.023977   0.006803  -3.524 0.000425 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##     Null deviance: 4052.6  on 3708  degrees of freedom
## Residual deviance: 3608.4  on 3705  degrees of freedom
## AIC: 3616.4
## Number of Fisher Scoring iterations: 4
BIC(mnumericfinal) # BIC: 3641.2
influenceIndexPlot(mnumericfinal,id=c(method=abs(cooks.distance(mnumericfinal)), n=3))
```

See plot Appendant (37)

```
#influenceIndexPlot after removing the most influential data. We can observe new influential data but the scale is different than before, so we will keep this observations.
```

#Model: numeric + factors

```
#numeric + factors
m_num_fact <- glm(target ~ poly(city_development_index,2) + experience + gender + education_level
+ major_discipline + company_size + company_type + city_group + relevant_experience + group_enrol
led_university+ group_last_new_job, family = binomial, data = train)
summary(m_num_fact) # AIC: 3339.4
BIC(m_num_fact) # BIC: 3494.8
step(m_num_fact)
#step model
m2 <- glm(formula = target ~ poly(city_development_index, 2) + experience + education_level + com
pany_size + company_type + city_group + group_enrolled_university + group_last_new_job, family =
binomial, data = train)
summary(m2) # AIC: 3329.7
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.3412  -0.6027  -0.4613  -0.2611   2.7232
## Coefficients:
##                                         Estimate Std. Error z value Pr(>|z|)
## (Intercept)                            -1.022121   0.204331  -5.002 5.67e-07
## poly(city_development_index, 2)1      -43.280032   2.729077 -15.859  < 2e-16
## poly(city_development_index, 2)2       11.633558   2.877231   4.043 5.27e-05
## experience                             -0.021475   0.007829  -2.743 0.006089
## education_levelPost_graduate           -0.993291   0.159803  -6.216 5.11e-10
## education_levelGraduate                -0.131488   0.109903  -1.196 0.231538
## education_levelNo Indicated            -0.957548   0.305306  -3.136 0.001711
## company_sizeBig                         0.043070   0.126949   0.339 0.734406
## company_sizeNo Indicated                1.213831   0.168087   7.221 5.14e-13
## company_typeNGO                        -0.553437   0.383887  -1.442 0.149397
## company_typeNo Indicated                0.345468   0.234749   1.472 0.141116
## company_typePublic Sector               0.303769   0.266910   1.138 0.255081
## company_typePrivate Limited Company     0.012024   0.193772   0.062 0.950522
## city_groupSmall_city                   -0.565635   0.102313  -5.528 3.23e-08
## city_groupStandard_city                -0.845189   0.261673  -3.230 0.001238
## group_enrolled_universityNo            -0.293332   0.105599  -2.778 0.005473
## group_enrolled_universityNo Indicated  -0.178897   0.304668  -0.587 0.557078
## group_last_new_jobNone                 -0.553827   0.142527  -3.886 0.000102
## group_last_new_jobNo Indicated          0.108924   0.278286   0.391 0.695493
## (Intercept)                           ***
## poly(city_development_index, 2)1      ***
## poly(city_development_index, 2)2      ***
## experience                            **
## education_levelPost_graduate          ***
## education_levelGraduate
## education_levelNo Indicated           **
## company_sizeBig
## company_sizeNo Indicated              ***
## company_typeNGO
## company_typeNo Indicated
## company_typePublic Sector
## company_typePrivate Limited Company
## city_groupSmall_city                  ***
## city_groupStandard_city               **
## group_enrolled_universityNo           **
## group_enrolled_universityNo Indicated
## group_last_new_jobNone                ***
## group_last_new_jobNo Indicated
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## (Dispersion parameter for binomial family taken to be 1)
##     Null deviance: 4052.6  on 3708  degrees of freedom
```

```
## Residual deviance: 3291.7  on 3690  degrees of freedom
## AIC: 3329.7
## Number of Fisher Scoring iterations: 5
BIC(m2) # BIC: 3447.8
Anova(m2,test="LR")
## Analysis of Deviance Table (Type II tests)
## Response: target
##                                  LR Chisq Df Pr(>Chisq)
## poly(city_development_index, 2)  312.093  2  < 2.2e-16 ***
## experience                         7.567  1  0.0059433 **
## education_level                   47.830  3  2.314e-10 ***
## company_size                      56.640  2  5.022e-13 ***
## company_type                       9.252  4  0.0551095 .
## city_group                        36.957  2  9.441e-09 ***
## group_enrolled_university          7.659  2  0.0217208 *
## group_last_new_job                16.290  2  0.0002901 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
vif(m2)
##                                     GVIF Df GVIF^(1/(2*Df))
## poly(city_development_index, 2) 1.530984  2        1.112353
## experience                      1.418487  1        1.191002
## education_level                 1.433119  3        1.061811
## company_size                    3.723926  2        1.389154
## company_type                    3.486763  4        1.168968
## city_group                      1.289123  2        1.065549
## group_enrolled_university       1.333962  2        1.074696
## group_last_new_job              1.372135  2        1.082304
# company_size and company_type are correlated, so we will check a new model removing company_typ
e because is the one with less significance in our model (checked in Anova test)

m3<- glm(formula = target ~ poly(city_development_index, 2) + experience + education_level + comp
any_size + city_group + group_enrolled_university + group_last_new_job, family = binomial, data =
train)
summary(m3) # AIC: 3330.9
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.3155  -0.5916  -0.4660  -0.2713   2.7145
## Coefficients:
##                                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)                          -0.973342   0.127953  -7.607 2.80e-14
## poly(city_development_index, 2)1    -43.022652   2.710984 -15.870  < 2e-16
## poly(city_development_index, 2)2     11.318643   2.862705   3.954 7.69e-05
## experience                           -0.022743   0.007802  -2.915 0.003556
## education_levelPost_graduate         -0.985984   0.159688  -6.174 6.64e-10
## education_levelGraduate              -0.123233   0.109344  -1.127 0.259736
## education_levelNo Indicated          -0.950220   0.306639  -3.099 0.001943
## company_sizeBig                       0.045563   0.123010   0.370 0.711083
## company_sizeNo Indicated              1.496132   0.108914  13.737  < 2e-16
## city_groupSmall_city                 -0.570757   0.102095  -5.590 2.26e-08
## city_groupStandard_city              -0.863347   0.261777  -3.298 0.000974
## group_enrolled_universityNo          -0.289541   0.104746  -2.764 0.005706
## group_enrolled_universityNo Indicated  -0.157865   0.304356  -0.519 0.603979
## group_last_new_jobNone               -0.567187   0.142467  -3.981 6.86e-05
## group_last_new_jobNo Indicated        0.127506   0.278016   0.459 0.646501
## (Intercept)                         ***
## poly(city_development_index, 2)1    ***
## poly(city_development_index, 2)2    ***
## experience                          **
## education_levelPost_graduate        ***
```

```
## education_levelGraduate
## education_levelNo Indicated           **
## company_sizeBig
## company_sizeNo Indicated              ***
## city_groupSmall_city                  ***
## city_groupStandard_city               ***
## group_enrolled_universityNo           **
## group_enrolled_universityNo Indicated
## group_last_new_jobNone                ***
## group_last_new_jobNo Indicated
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## (Dispersion parameter for binomial family taken to be 1)
##     Null deviance: 4052.6  on 3708  degrees of freedom
## Residual deviance: 3300.9  on 3694  degrees of freedom
## AIC: 3330.9
##
## Number of Fisher Scoring iterations: 5
BIC(m3) # BIC: 3424.2
Anova(m3,test="LR")
## Analysis of Deviance Table (Type II tests)
## Response: target
##                                    LR Chisq Df Pr(>Chisq)
## poly(city_development_index, 2)  311.468  2  < 2.2e-16 ***
## experience                         8.552  1  0.0034520 **
## education_level                   47.077  3  3.348e-10 ***
## company_size                     228.084  2  < 2.2e-16 ***
## city_group                        37.989  2  5.633e-09 ***
## group_enrolled_university          7.585  2  0.0225431 *
## group_last_new_job                17.242  2  0.0001803 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
anova(m2,m3, test="Chisq")
## Analysis of Deviance Table
## Model 1: target ~ poly(city_development_index, 2) + experience + education_level +
##     company_size + company_type + city_group + group_enrolled_university +
##     group_last_new_job
## Model 2: target ~ poly(city_development_index, 2) + experience + education_level +
##     company_size + city_group + group_enrolled_university + group_last_new_job
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1      3690     3291.7
## 2      3694     3300.9 -4  -9.2516  0.05511 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
vif(m3)
##                                     GVIF Df GVIF^(1/(2*Df))
## poly(city_development_index, 2) 1.508264  2        1.108203
## experience                      1.411315  1        1.187988
## education_level                 1.424295  3        1.060718
## company_size                    1.316953  2        1.071254
## city_group                      1.285733  2        1.064848
## group_enrolled_university       1.314728  2        1.070801
## group_last_new_job              1.369522  2        1.081788
# Although the AIC is a little bit higher, the BIC is lower. Hence, we decided to remove company_
type taking into account the statistical p-value on anova test over 0.05, failing to reject the n
ull hypothesis, improving generalized VIF values and using less number of parameters as possible,
maintaining the most significant variables in our new best model so far, m3.
```

#Influential observation of best model

```
#Influential data
influenceIndexPlot(m3,id=c(method=abs(cooks.distance(m3)), n=3))
```

See plot Appendant (38)

```
summary(train[c("712"),])
##  city_development_index          gender       education_level
##  Min.   :0.897          Female      :0   Pre_graduate :0
##  1st Qu.:0.897          Male        :1   Post_graduate:0
##  Median :0.897          No Indicated:0   Graduate     :0
##  Mean   :0.897                           No Indicated :1
##  3rd Qu.:0.897
##  Max.   :0.897
##         major_discipline   experience       company_size
##  Arts&humanities:0       Min.   :7    SME         :1
##  Business Degree:0       1st Qu.:7    Big         :0
##  No Indicated   :1       Median :7    No Indicated:0
##  STEM           :0       Mean   :7
##                          3rd Qu.:7
##                          Max.   :7
##                     company_type training_hours target         city_group
##  Startup                  :0     Min.   :39     0:0    Big_city     :0
##  NGO                      :0     1st Qu.:39     1:1    Small_city   :0
##  No Indicated             :0     Median :39            Standard_city:1
##  Public Sector            :0     Mean   :39
##  Private Limited Company:1     3rd Qu.:39
##                                  Max.   :39
##  relevant_experience group_enrolled_university       f.experience
##  Yes:1               Yes         :0            >20         :0
##  No :0               No          :1            0-5         :0
##                      No Indicated:0            11-15       :0
##                                                16-20       :0
##                                                6-10        :1
##                                                No Indicated:0
##    group_last_new_job
##  >0          :1
##  None        :0
##  No Indicated:0
summary(train)
##  city_development_index          gender          education_level
##  Min.   :0.448          Female      : 229   Pre_graduate :2235
##  1st Qu.:0.743          Male        :2552   Post_graduate: 458
##  Median :0.910          No Indicated: 928   Graduate     : 918
##  Mean   :0.832                              No Indicated :  98
##  3rd Qu.:0.920
##  Max.   :0.949
##         major_discipline   experience          company_size
##  Arts&humanities: 179    Min.   : 0.00   SME         :1604
##  Business Degree:  61    1st Qu.: 4.00   Big         : 918
##  No Indicated   : 663    Median : 9.00   No Indicated:1187
##  STEM           :2806    Mean   :10.12
##                          3rd Qu.:16.00
##                          Max.   :21.00
##                     company_type  training_hours   target           city_group
##  Startup                  : 296   Min.   :  1.00   0:2834   Big_city     :2062
##  NGO                      : 107   1st Qu.: 24.00   1: 875   Small_city   :1450
##  No Indicated             :1254   Median : 48.00            Standard_city: 197
##  Public Sector            : 203   Mean   : 63.03
##  Private Limited Company:1849   3rd Qu.: 86.00
##                                   Max.   :288.00
##  relevant_experience group_enrolled_university       f.experience
##  Yes:2662            Yes         : 973         >20         : 643
##  No :1047            No          :2659         0-5         :1237
```

```
##                      No Indicated:  77         11-15       : 525
##                                                16-20       : 324
##                                                6-10        : 973
##                                                No Indicated:   7
##     group_last_new_job
##  >0          :3123
##  None        : 509
##  No Indicated:  77
Boxplot(cooks.distance(m3), main="Cooks Distance model numerical v. + factors")
```

See plot Appendant (39)

```
##  [1]  712 3472  493  875 1686 1034 3201 2334 2112  631
# In this case, the cooksdistance is not much greater than the rest of the observations and check
ing its value on the variables, compared to the rest of the dataset, we can see that in most of t
he parameters is not very influential, with typical values. Therefore, we decided to keep it in o
ur train dataset.
```

#Interactions

#Interactions of poly(city_development_index,2)

```
m4 <- glm(target ~ poly(city_development_index,2) * experience + education_level + company_size +
city_group + group_enrolled_university + group_last_new_job, data=train,family=binomial)
summary(m4)#AIC: 3332.1
BIC(m4) # BIC: 3437.9
m4.1<- glm(target ~experience  + poly(city_development_index,2) * education_level + company_size
+ city_group + group_enrolled_university  + group_last_new_job,data=train,family=binomial )
summary(m4.1)#AIC: 3329.1
BIC(m4.1) # BIC: 3459.6
m4.2<-  glm(target ~experience  +  education_level + poly(city_development_index,2) * company_siz
e + city_group + group_enrolled_university  + group_last_new_job,data=train,family=binomial)
summary(m4.2)#AIC: 3258.6
BIC(m4.2) # BIC: 3376.8
m4.3<- glm(target ~experience  +  education_level +  company_size + poly(city_development_index,2
) * city_group + group_enrolled_university + group_last_new_job,data=train,family=binomial )
summary(m4.3)#AIC: 3331.6
BIC(m4.3) # BIC: 3443.5
m4.4<- glm(target ~experience  +  education_level  +  company_size + city_group + poly(city_devel
opment_index,2) * group_enrolled_university + group_last_new_job,data=train,family=binomial )
summary(m4.4)#AIC:3326.7
BIC(m4.4) # BIC: 3444.9
m4.5<- glm(target ~experience  +  education_level +  company_size + city_group + group_enrolled_u
niversity + poly(city_development_index,2) * group_last_new_job, data=train,family=binomial )
summary(m4.5)#AIC: 3329.9
BIC(m4.5) # BIC: 3448
Anova(m4.2, test = "LR")
## Analysis of Deviance Table (Type II tests)
## Response: target
##                                              LR Chisq Df Pr(>Chisq)
## experience                                      7.404  1   0.006508 **
## education_level                                40.459  3  8.515e-09 ***
## poly(city_development_index, 2)               311.468  2  < 2.2e-16 ***
## company_size                                  228.084  2  < 2.2e-16 ***
## city_group                                     33.353  2  5.722e-08 ***
## group_enrolled_university                      10.837  2   0.004433 **
## group_last_new_job                             11.287  2   0.003540 **
## poly(city_development_index, 2):company_size   80.303  4  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
anova(m3, m4.2, test = "Chisq")
## Analysis of Deviance Table
## Model 1: target ~ poly(city_development_index, 2) + experience + education_level +
```

```
##     company_size + city_group + group_enrolled_university + group_last_new_job
## Model 2: target ~ experience + education_level + poly(city_development_index,
##     2) * company_size + city_group + group_enrolled_university +
##     group_last_new_job
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1      3694     3300.9
## 2      3690     3220.6  4   80.303 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##Best model obtained so far is m4.2 with AIC: 3258.6 and BIC: 3376.8. We have statistical argume
nts to reject the null hypothesis and confirm that it is not equal to m3.
```

#Interactions of group_enrolled_university

```
m5<- glm(target ~group_enrolled_university * experience  +  education_level + poly(city_developme
nt_index,2) + company_size + city_group + group_last_new_job,data=train,family=binomial)
summary(m5) # AIC: 3331
BIC(m5) # BIC: 3436.7
m5.1<- glm(target ~ experience  + group_enrolled_university * education_level + poly(city_develop
ment_index,2) + company_size + city_group+ group_last_new_job,data=train,family=binomial)
summary(m5.1) # AIC: 3335.3
BIC(m5.1) # BIC: 3465.9
m5.2<- glm(target ~ experience  +  education_level + poly(city_development_index,2) + group_enrol
led_university * company_size + city_group+ group_last_new_job,data=train,family=binomial)
summary(m5.2) #AIC: 3338.3
BIC(m5.2) # BIC: 3456.4
m5.3<- glm(target ~ experience  +  education_level + poly(city_development_index,2) + company_siz
e + group_enrolled_university * city_group+ group_last_new_job,data=train,family=binomial)
summary(m5.3) #AIC: 3337.3
BIC(m5.3) # BIC: 3455.5
m5.4<- glm(target ~ experience  +  education_level + poly(city_development_index,2) + company_siz
e + city_group + group_enrolled_university * group_last_new_job,data=train,family=binomial)
summary(m5.4) #AIC: 3333.1
BIC(m5.4) # BIC: 3451.3
##Best model so far is m4.2 with AIC: 3258.6 and BIC: 3376.8.
```

#Interactions of educational_level

```
m6<-  glm(target ~ education_level * experience + poly(city_development_index,2) + company_size +
city_group + group_enrolled_university + group_last_new_job, data=train, family=binomial)
summary(m6)  # AIC: 3331.7
BIC(m6) # BIC: 3443.6
m6.1<- glm(target ~ experience + poly(city_development_index,2) + education_level * company_size
+ city_group + group_enrolled_university + group_last_new_job, data=train, family=binomial)
summary(m6.1) # AIC: 3332.9
BIC(m6.1) # BIC: 3463.5
m6.2<- glm(target ~ experience + poly(city_development_index,2) + company_size + education_level
* city_group + group_enrolled_university + group_last_new_job, data=train, family=binomial)
summary(m6.2) # AIC: 3340.1
BIC(m6.2) # BIC: 3470.7
m6.3<- glm(target ~ experience + poly(city_development_index,2) + company_size + city_group + gro
up_enrolled_university + education_level * group_last_new_job, data=train, family=binomial)
summary(m6.3) # AIC: 3330.3
BIC(m6.3) # BIC: 3460.9
##Best model so far is m4.2 with AIC: 3258.6 and BIC: 3376.8.
```

#Interactions of experience

```
m7<- glm(target ~ education_level + poly(city_development_index,2) + experience * company_size +
city_group + group_enrolled_university + group_last_new_job, data=train, family=binomial)
summary(m7)  # AIC: 3323.9
BIC(m7) # BIC: 3429.6
```

```r
m7.1<- glm(target ~ education_level + poly(city_development_index,2) + company_size + experience
* city_group + group_enrolled_university + group_last_new_job, data=train, family=binomial)
summary(m7.1)  # AIC: 3334.8
BIC(m7.1) # BIC: 3440.5
m7.2<- glm(target ~ education_level + poly(city_development_index,2) + company_size + city_group
+ group_enrolled_university + experience * group_last_new_job, data=train, family=binomial)
summary(m7.2)  # AIC: 33327.1
BIC(m7.2) # BIC: 3432.9
##Best model so far is m4.2 with AIC: 3258.6 and BIC: 3376.8.
```

#Interactions of company_size

```r
m8<- glm(target ~  experience + education_level + poly(city_development_index,2) + company_size *
city_group + group_enrolled_university + group_last_new_job, data=train, family=binomial)
summary(m8)  # AIC: 3334.5
BIC(m8) # BIC: 3452.7
m8.1<-  glm(target ~  experience + education_level + poly(city_development_index,2) + city_group
+ group_enrolled_university + company_size * group_last_new_job, data=train, family=binomial)
summary(m8.1)  # AIC: 3327.4
BIC(m8.1) # BIC: 3445.6
##Best model so far is m4.2 with AIC: 3258.6 and BIC: 3376.8.
```

#Interactions of city_group and group_last_new_job

```r
m9<- glm(target ~  experience + education_level + poly(city_development_index,2) + company_size +
group_enrolled_university + city_group * group_last_new_job, data=train, family=binomial)
summary(m9)  # AIC: 3335
BIC(m9) # BIC: 3453.2
##Best model so far is m4.2 with AIC: 3258.6 and BIC: 3376.8.
```

#Analize Best Model

```r
summary(m4.2)#AIC: 3258.6
## Call:
## glm(formula = target ~ experience + education_level + poly(city_development_index,
##     2) * company_size + city_group + group_enrolled_university +
##     group_last_new_job, family = binomial, data = train)
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.0894  -0.6051  -0.4178  -0.2425   2.8334
## Coefficients:
##                                                               Estimate Std. Error
## (Intercept)                                                  -1.133219   0.135617
## experience                                                   -0.021537   0.007937
## education_levelPost_graduate                                 -0.890605   0.153614
## education_levelGraduate                                      -0.118870   0.111340
## education_levelNo Indicated                                  -0.848711   0.295720
## poly(city_development_index, 2)1                            -59.358262   4.293624
## poly(city_development_index, 2)2                             22.603639   4.923966
## company_sizeBig                                               0.178641   0.136101
## company_sizeNo Indicated                                      1.658936   0.115085
## city_groupSmall_city                                         -0.525813   0.103196
## city_groupStandard_city                                      -0.904980   0.267395
## group_enrolled_universityNo                                  -0.347390   0.105247
## group_enrolled_universityNo Indicated                        -0.195785   0.295556
## group_last_new_jobNone                                       -0.436511   0.136502
## group_last_new_jobNo Indicated                                0.142400   0.276427
## poly(city_development_index, 2)1:company_sizeBig              3.585883   6.965724
## poly(city_development_index, 2)2:company_sizeBig            -17.137980   8.938993
## poly(city_development_index, 2)1:company_sizeNo Indicated   40.631497   5.551824
## poly(city_development_index, 2)2:company_sizeNo Indicated -18.125671   5.889856
##                                                             z value Pr(>|z|)
## (Intercept)                                                  -8.356  < 2e-16 ***
## experience                                                   -2.714 0.006655 **
```

```
## education_levelPost_graduate                               -5.798 6.72e-09 ***
## education_levelGraduate                                     -1.068 0.285688
## education_levelNo Indicated                                 -2.870 0.004105 **
## poly(city_development_index, 2)1                           -13.825  < 2e-16 ***
## poly(city_development_index, 2)2                             4.591 4.42e-06 ***
## company_sizeBig                                              1.313 0.189330
## company_sizeNo Indicated                                    14.415  < 2e-16 ***
## city_groupSmall_city                                        -5.095 3.48e-07 ***
## city_groupStandard_city                                     -3.384 0.000713 ***
## group_enrolled_universityNo                                 -3.301 0.000964 ***
## group_enrolled_universityNo Indicated                       -0.662 0.507696
## group_last_new_jobNone                                      -3.198 0.001385 **
## group_last_new_jobNo Indicated                               0.515 0.606451
## poly(city_development_index, 2)1:company_sizeBig             0.515 0.606700
## poly(city_development_index, 2)2:company_sizeBig            -1.917 0.055211 .
## poly(city_development_index, 2)1:company_sizeNo Indicated   7.319 2.51e-13 ***
## poly(city_development_index, 2)2:company_sizeNo Indicated  -3.077 0.002088 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## (Dispersion parameter for binomial family taken to be 1)
##     Null deviance: 4052.6  on 3708  degrees of freedom
## Residual deviance: 3220.6  on 3690  degrees of freedom
## AIC: 3258.6
## Number of Fisher Scoring iterations: 5
BIC(m4.2) # BIC: 3376.8
Anova(m4.2, test = "LR")
## Analysis of Deviance Table (Type II tests)
## Response: target
##                                              LR Chisq Df Pr(>Chisq)
## experience                                      7.404  1   0.006508 **
## education_level                                40.459  3  8.515e-09 ***
## poly(city_development_index, 2)               311.468  2  < 2.2e-16 ***
## company_size                                  228.084  2  < 2.2e-16 ***
## city_group                                     33.353  2  5.722e-08 ***
## group_enrolled_university                      10.837  2   0.004433 **
## group_last_new_job                             11.287  2   0.003540 **
## poly(city_development_index, 2):company_size   80.303  4  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
# Test hosmer-Lemeshow
library(ResourceSelection)
hl_test <- hoslem.test(m4.2$y, fitted(m4.2));hl_test
##  Hosmer and Lemeshow goodness of fit (GOF) test
## X-squared = 45.267, df = 8, p-value = 3.276e-07
# Although small values with large p-values on Hosmer-Lemeshow test indicate a good fit while lar
ge values with p-values below 0.05 indicate a poor fit, searching for more information regarding
this test, we found that for larger datasets (>1000 observations) it’s highly likely that it will
fail. Therefore we'll evaluate our model with more tools.

#Residual Analisis Best Model
library(effects)
summary(m4.2)
plot(allEffects(m4.2),ask=FALSE)
```

See plot Appendant (40)

```
# The plots show how the target variable respond to variability among the different parameters. F
or example, as greater the experience, less is the probability that a person looks for a job chan
ge. This is validated also with the second plot due to is more probable that people more experien
ced are those with a post graduate education and also whose are less probable to look for a job c
hange. In addition, people living in a big city are more susceptible to look for a job change, th
```

```
an people living in small cities.

influenceIndexPlot(m4.2,id=c(method=abs(cooks.distance(m4.2)), n=5))
```
See plot Appendant (41)

```
# We can see that the observation 712 is still the one with more Cook's distance. However, as we
analyzed before, taking into account the characteristics that it has, we still decide to keep it
into our data.

marginalModelPlots(m4.2,id=list(method=abs(cooks.distance(m4.2)), n=5))
```
See plot Appendant (42)

```
# As we can see from the plots above, the model follows the same pattern, so we have significant
evidence to affirm that the model we reached fits well.

#For binary targets with many factors variables into the model (with interaction between some of
them and with many levels for each), the Added-Variable Plot does not deliver much valuable infor
mation to the analysis, so we decided not to approach it in our analysis.
```

#Predict the probability of a candidate will work for the company

```
prediction_table <- predict(m4.2, newdata=test,type="response")
probabtarget <- ifelse(prediction_table<0.5,0,1);probabtarget
cm <- table(probabtarget,test$target);cm
## probabtarget   0   1
##            0 877 188
##            1  78 138
library(cvAUC)
AUC(predict(m4.2, type="response"), train$target) #same calculation, but manually : accuracy <- s
um(cm[1], cm[4]) / sum(cm[1:4]);accuracy
## [1] 0.7978064
precision <- cm[4] / sum(cm[4], cm[2]); precision
## [1] 0.6388889
recall <- cm[4] / sum(cm[4], cm[3]); recall
## [1] 0.4233129
fscore <- (2 * (recall * precision))/(recall + precision); fscore
## [1] 0.5092251
#Taking into account our best model (m4.2), we obtained the following rates :
# - Accuracy : 79%, which indicates overall, how often is the classifier correct.
# - Precision : 64%, which indicates  when it predicts yes, how often is it correct.
# - Recall: 42%, which indicates when it's actually yes, how often does it predict yes. (true pos
itive rate)
# - Fscore: 51%. This is a weighted average of the recall and precision.

# We focus our analysis on accuracy rate, and since this value is between 70% and 80%, we can ind
icate that it's a good model considering that we didn't treat unbalance issue because it was out
of the project scope.

#ROC curve
roc<-prediction(predict(m4.2,type="response",newdata=test), test$target)
par(mfrow=c(1,1))
plot(performance(roc,"tpr","fpr",fpr.stop=0.05), col = "blue", main = "ROC Curve")
abline(0,1,lty=2,col='red')
```
See plot Appendant (43)

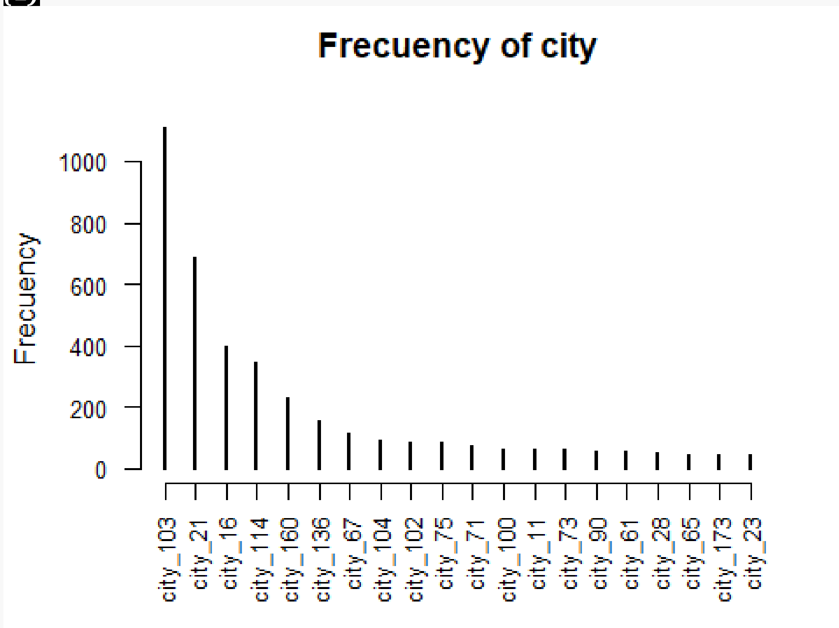
```
#Taking into account our AUC= 79%, ant it represents the area under the ROC curve, we can observe
that the curve approaches closer to the top-left corner, representing that model performance is g
ood. This ROC curve allow us to visualize the true positive rate (sensitivity) vs the false posit
ive rate (specificity)

#Probability of a candidate will look for a job change and potentially work for the company
prob_1 <- cm[4] / sum(cm[1:4]);prob_1*100
```

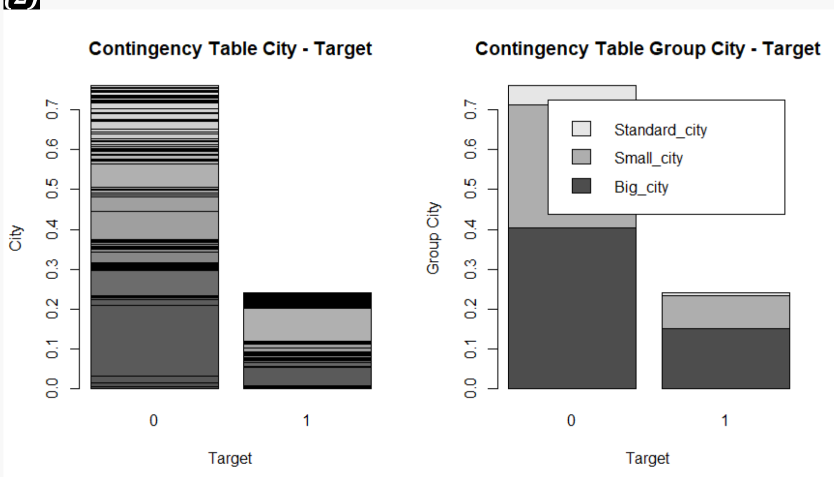
```
## [1] 10.77283
# There is 11% of probability that a candidate will look for a job change and potentially work fo
r the company.
```

Appendant

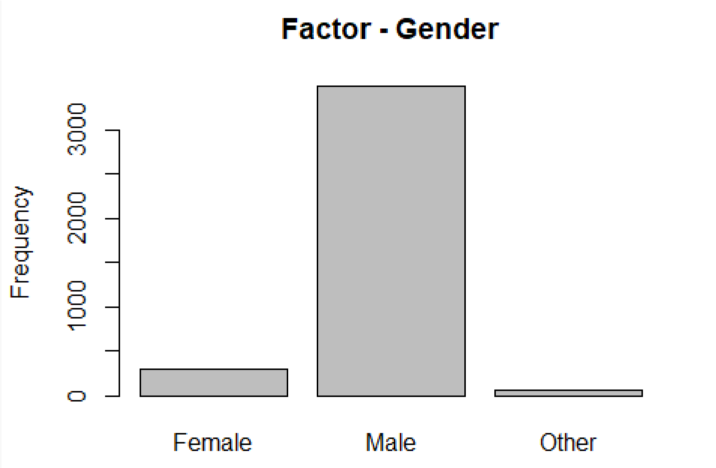
(1)



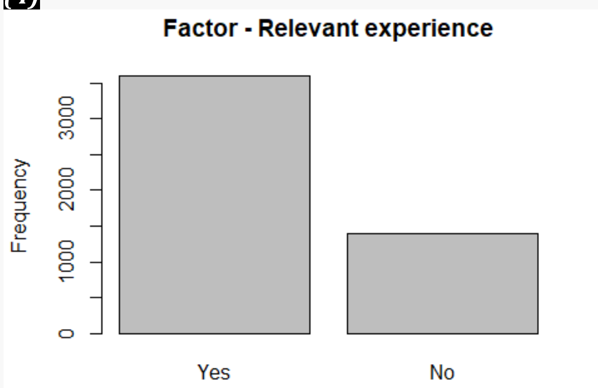
(2)



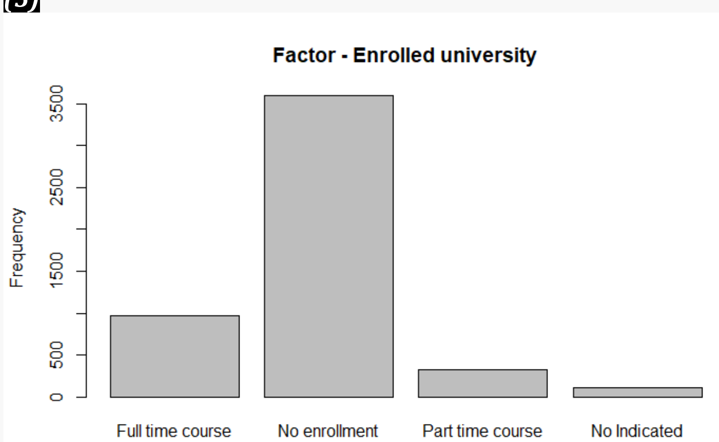
(3)


Factor - Gender
Frequency
0
1000
2000
3000
Female
Male
Other

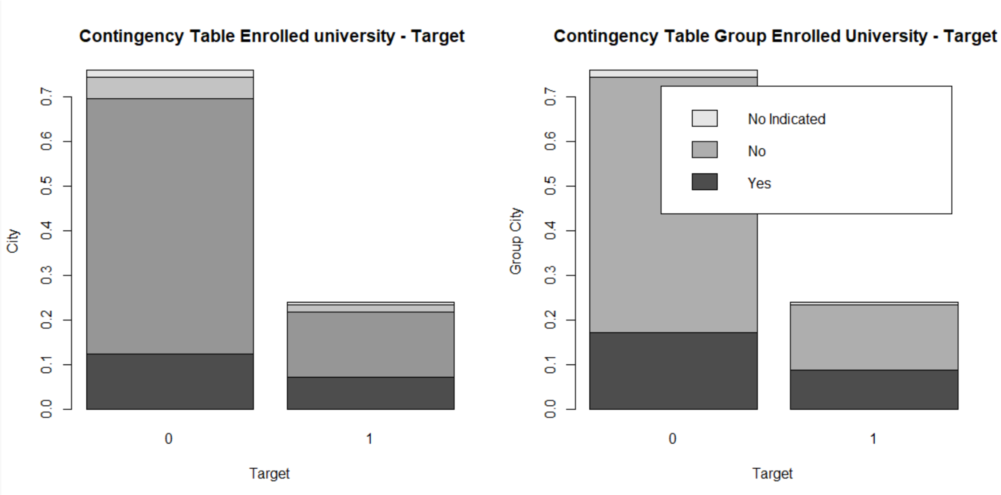
(4)


Factor - Relevant experience
Frequency
0
1000
2000
3000
Yes
No

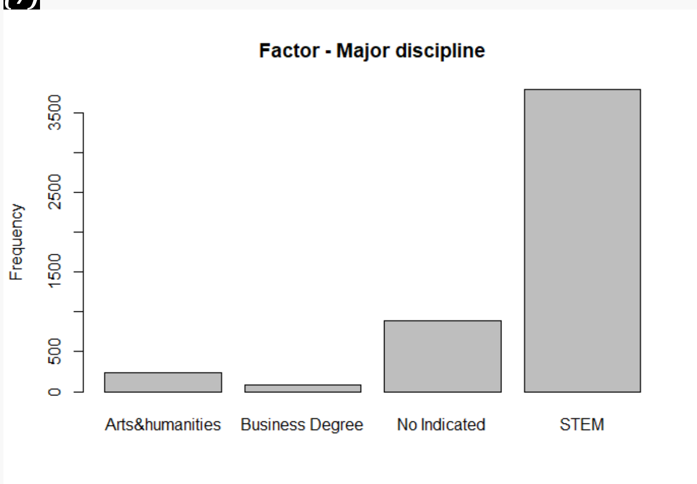
(5)


Factor - Enrolled university
Frequency
0
500
1500
2500
3500
Full time course
No enrollment
Part time course
No Indicated

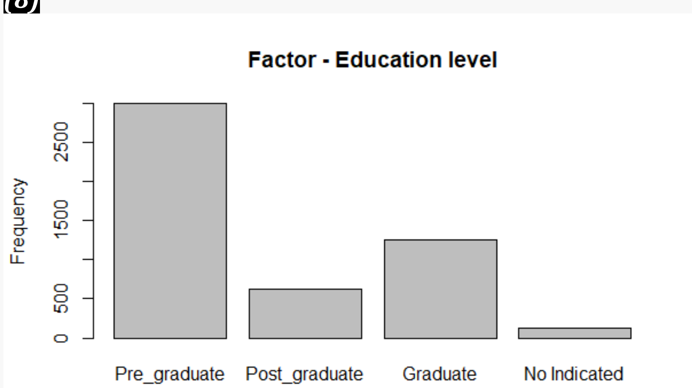
(6)


Contingency Table Enrolled university - Target
City
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0
1
Target
Contingency Table Group Enrolled University - Target
Group City
No Indicated
No
Yes
Target

(7)


Factor - Major discipline
Frequency
0
500
1500
2500
3500
Arts&humanities
Business Degree
No Indicated
STEM

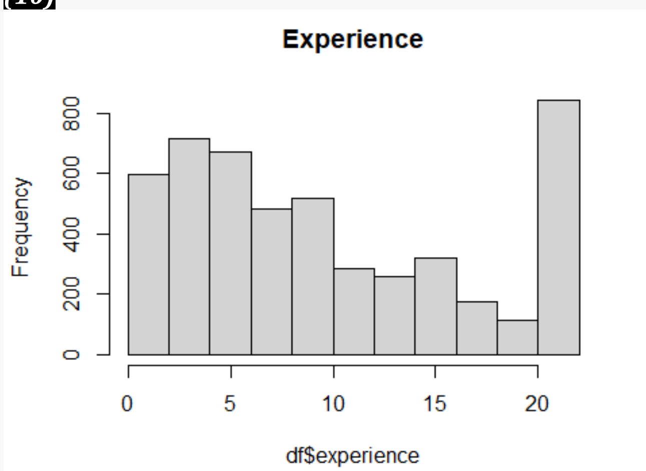
(8)


Factor - Education level
Frequency
0
500
1500
2500
Pre_graduate
Post_graduate
Graduate
No Indicated

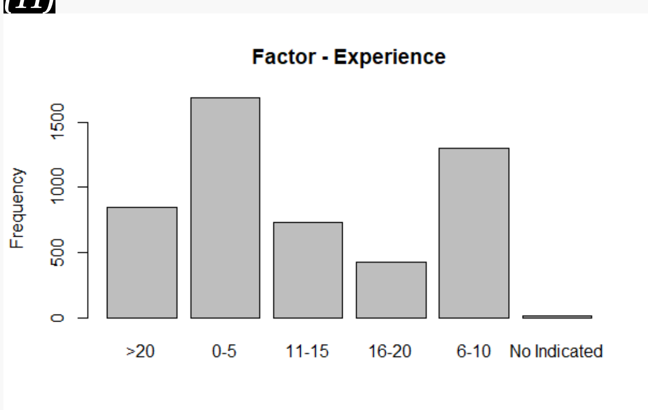
(9)


City Development Index
Frequency
0
500
1500
2500
0.4
0.5
0.6
0.7
0.8
0.9
Index

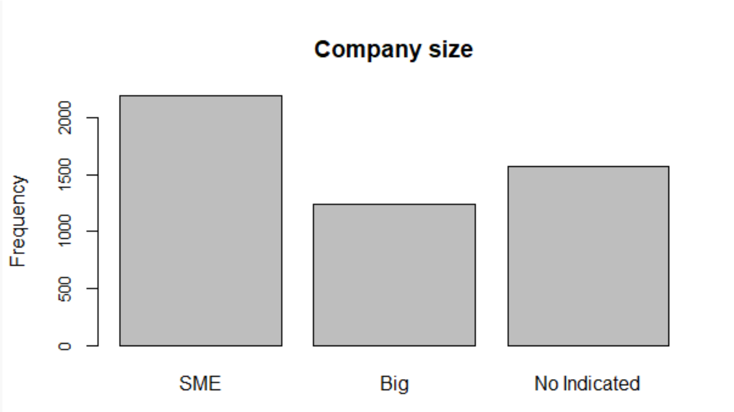
(10)


Experience
Frequency
0
200
400
600
800
0
5
10
15
20
df$experience

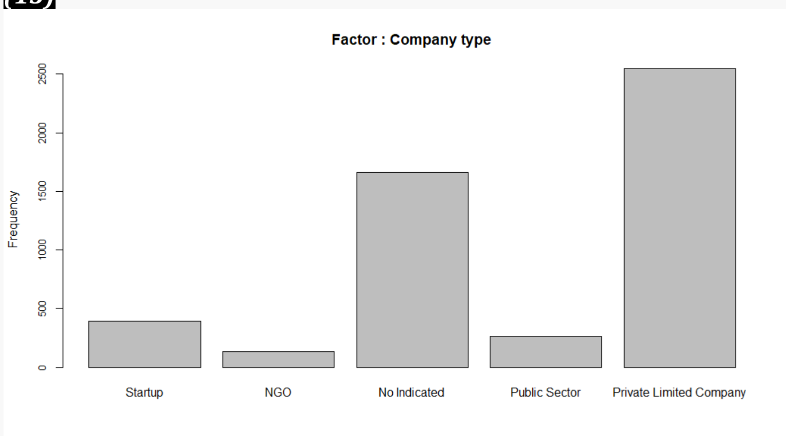
(11)


Factor - Experience
Frequency
0
500
1000
1500
>20
0-5
11-15
16-20
6-10
No Indicated

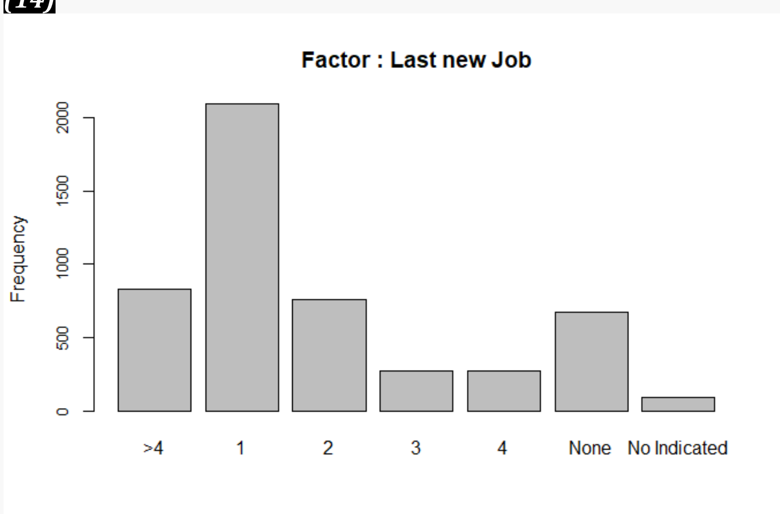
(12)


Company size
Frequency
0
500
1000
1500
2000
SME
Big
No Indicated

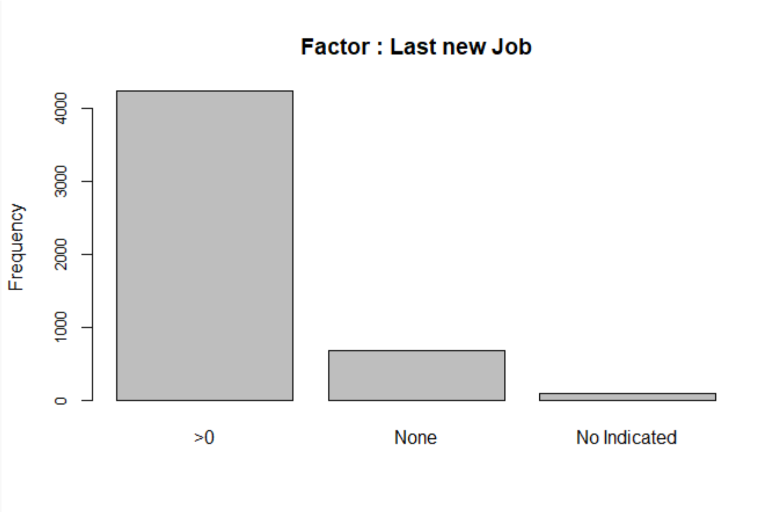
(13)


Factor : Company type
Frequency
Startup
NGO
No Indicated
Public Sector
Private Limited Company

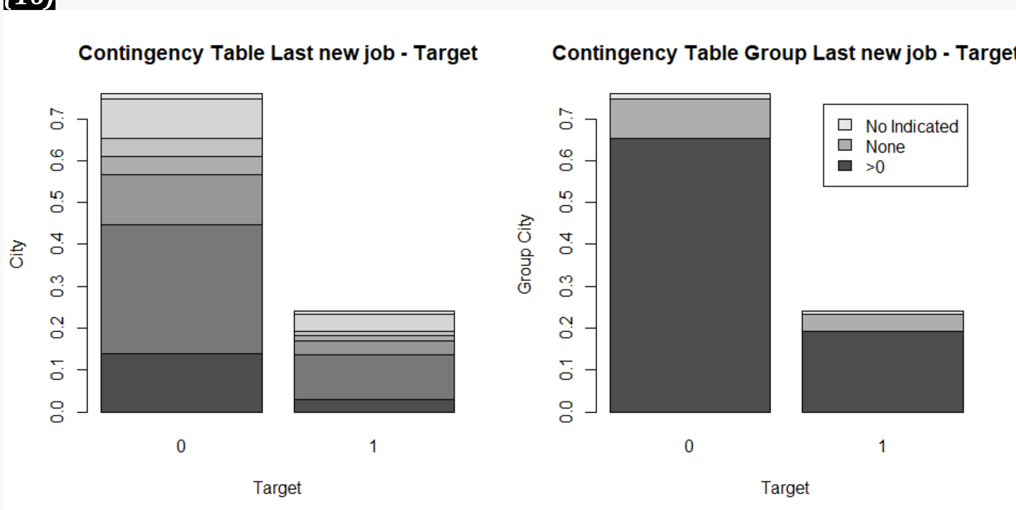
(14)


Factor : Last new Job
Frequency
0
500
1000
1500
2000
>4
1
2
3
4
None
No Indicated

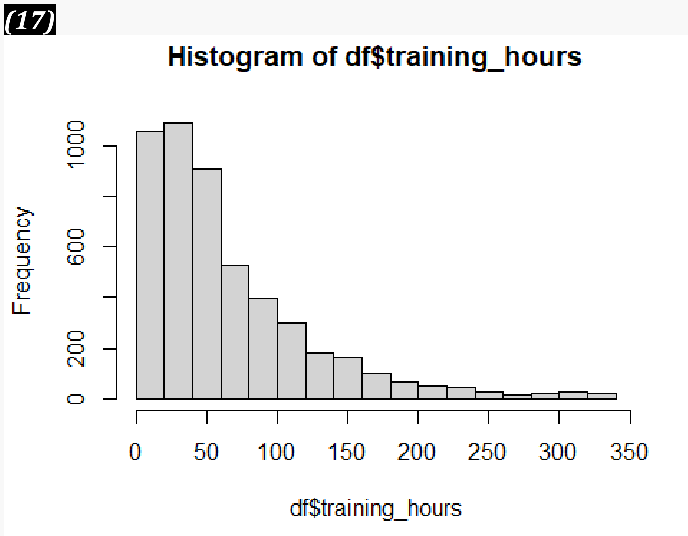
(15)


Factor : Last new Job
Frequency
0
1000
2000
3000
4000
>0
None
No Indicated

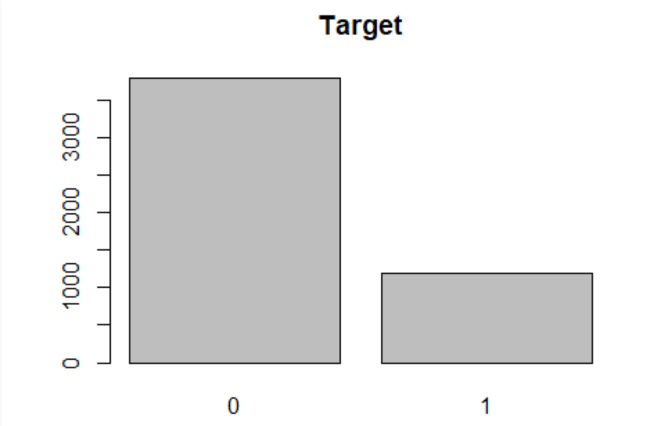
(16)


Contingency Table Last new job - Target
Contingency Table Group Last new job - Target
City
Group City
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0
1
Target
No Indicated
None
>0

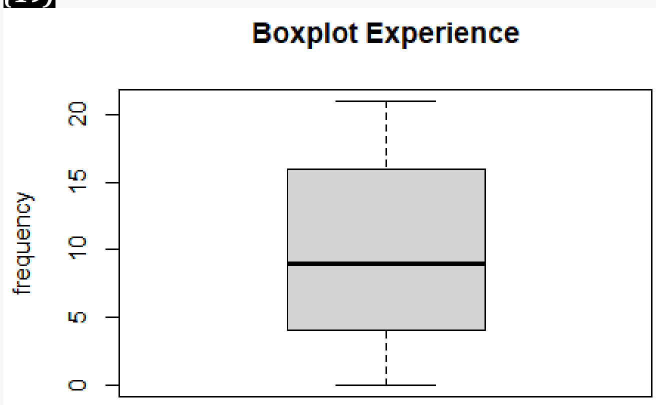
(17)


Histogram of df$training_hours
Frequency
0
200
600
1000
0
50
100
150
200
250
300
350
df$training_hours

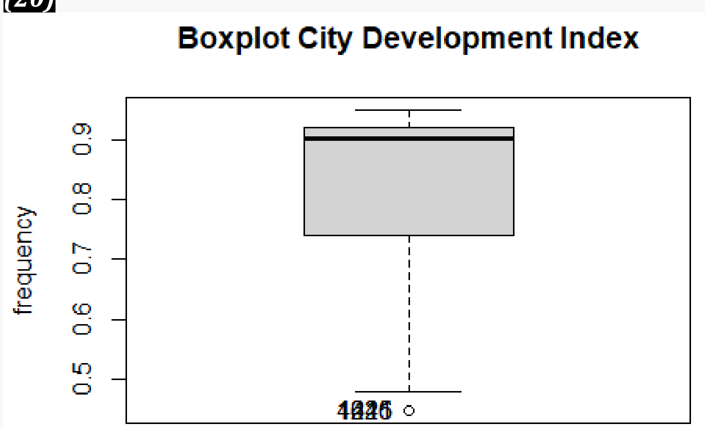
(18)


Target
0
1000
2000
3000
0
1

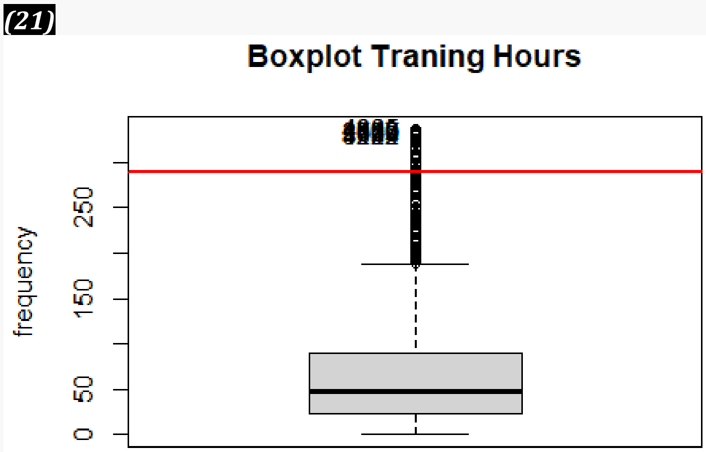
(19)


Boxplot Experience
frequency
0
5
10
15
20

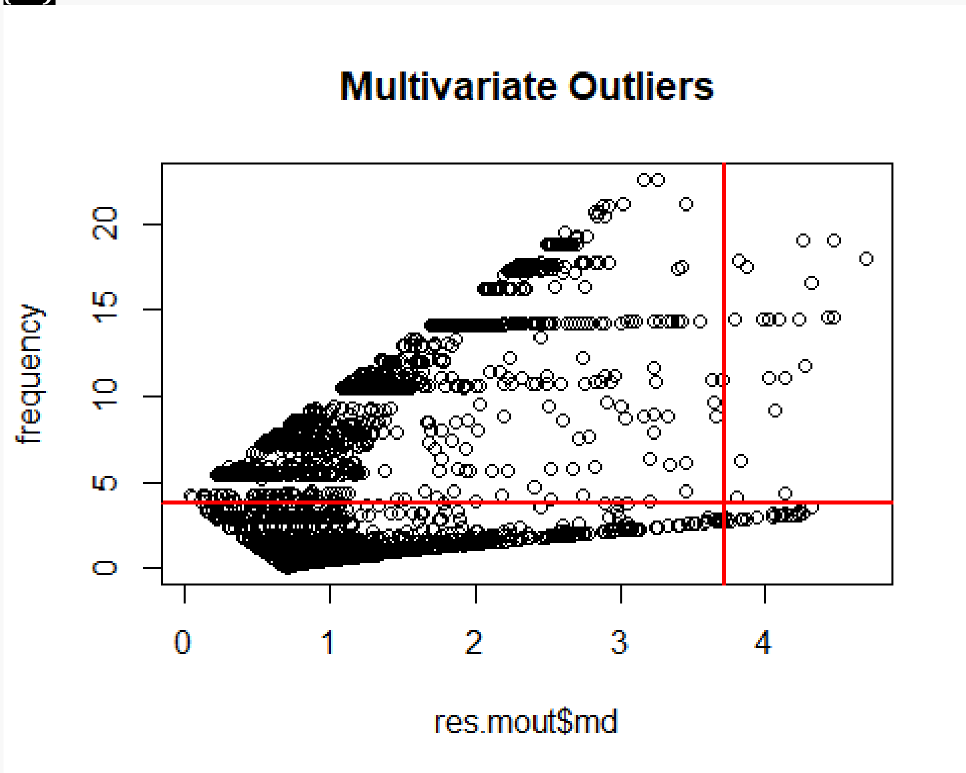
(20)


Boxplot City Development Index
frequency
0.5
0.6
0.7
0.8
0.9

(21)

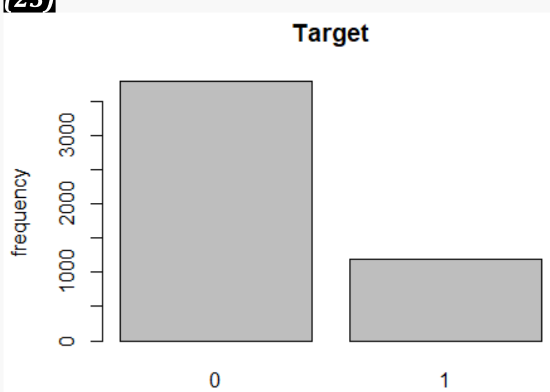

Boxplot Traning Hours
frequency
0
50
150
250

(22)



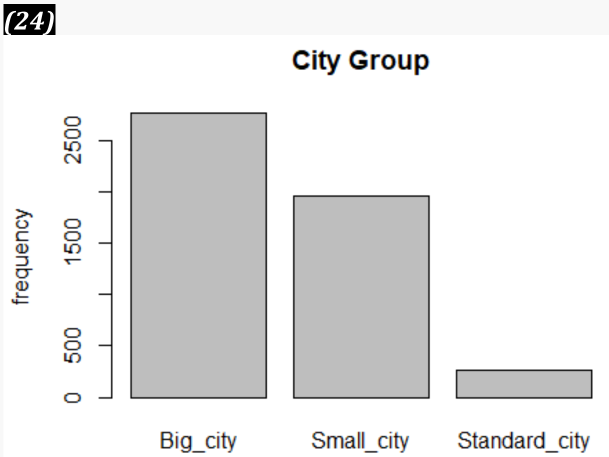
Multivariate Outliers
frequency
0
5
10
15
20
0
1
2
3
4
res.mout$md

(23)



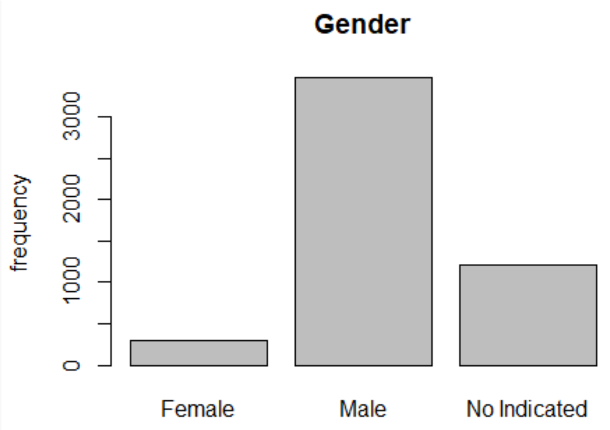
Target
frequency
0
1000
2000
3000
0
1

(24)

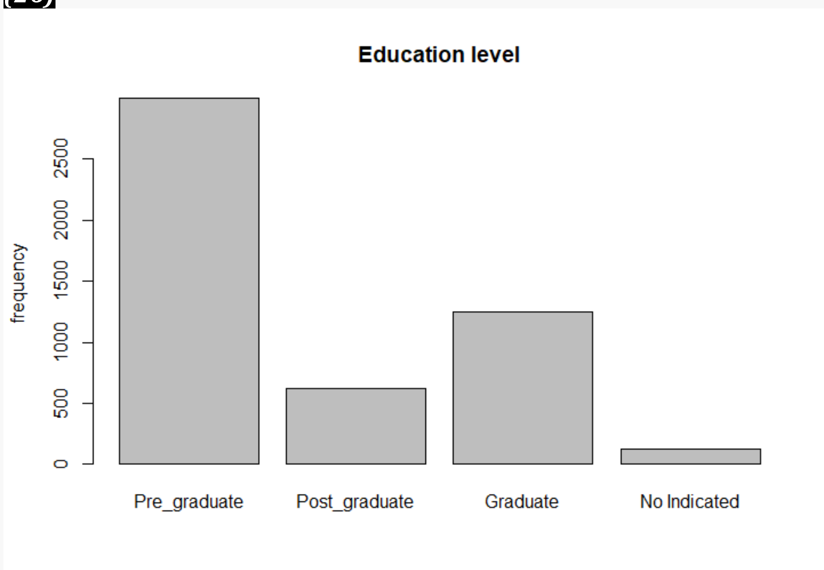


City Group
frequency
0
500
1500
2500
Big_city
Small_city
Standard_city

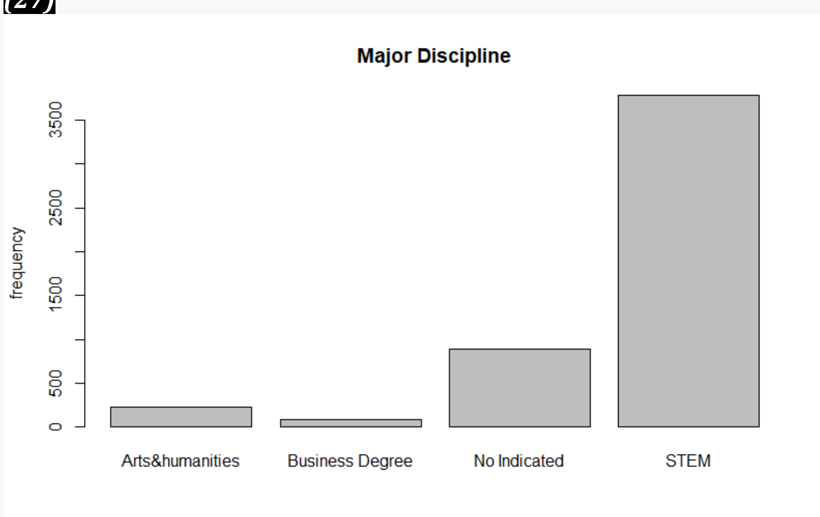
(25)


Gender
frequency
0
1000
2000
3000
Female
Male
No Indicated

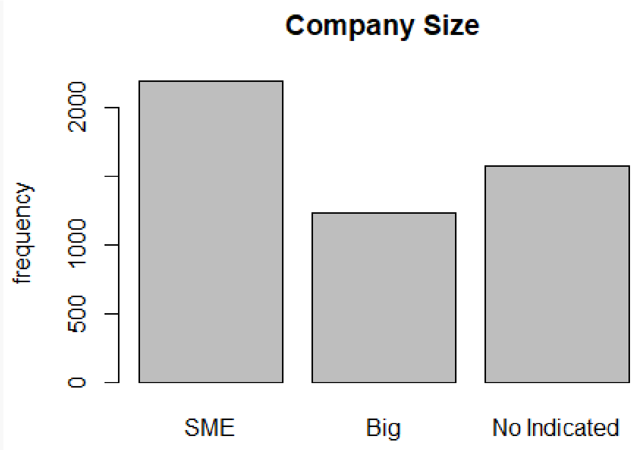
(26)


Education level
frequency
0
500
1000
1500
2000
2500
Pre_graduate
Post_graduate
Graduate
No Indicated

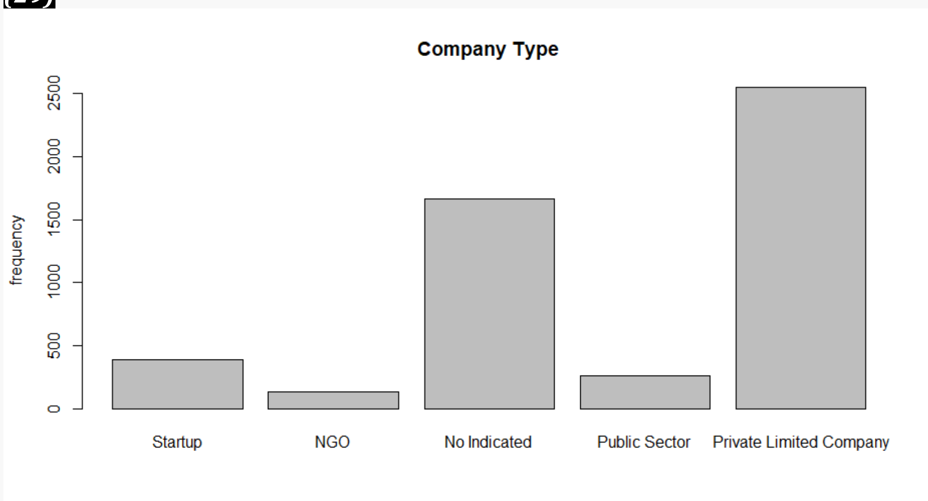
(27)


Major Discipline
frequency
0
500
1500
2500
3500
Arts&humanities
Business Degree
No Indicated
STEM

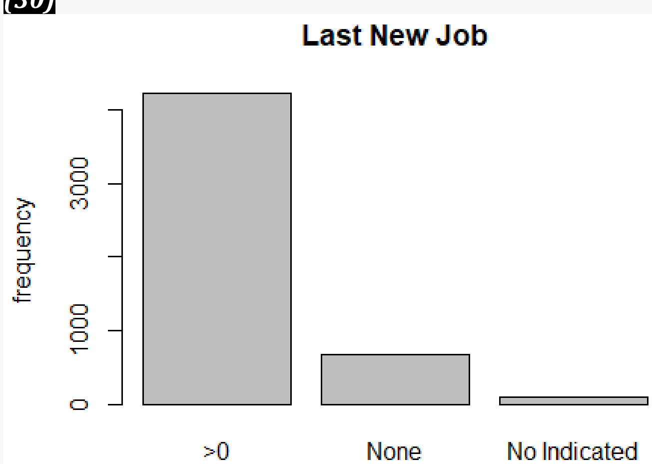
(28)


Company Size
frequency
0
500
1000
2000
SME
Big
No Indicated

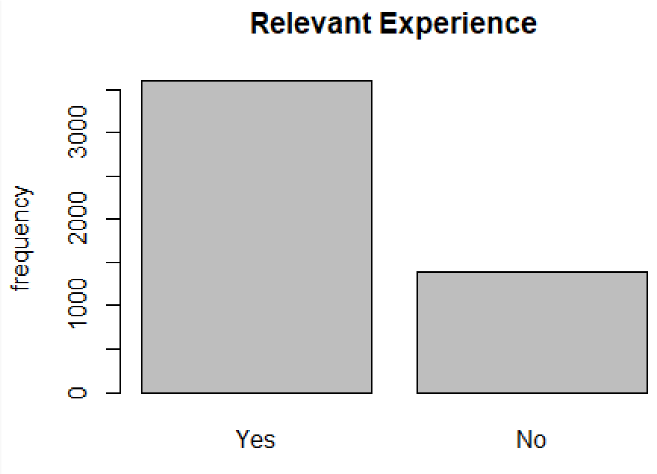
(29)


Company Type
frequency
0
500
1000
1500
2000
2500
Startup
NGO
No Indicated
Public Sector
Private Limited Company

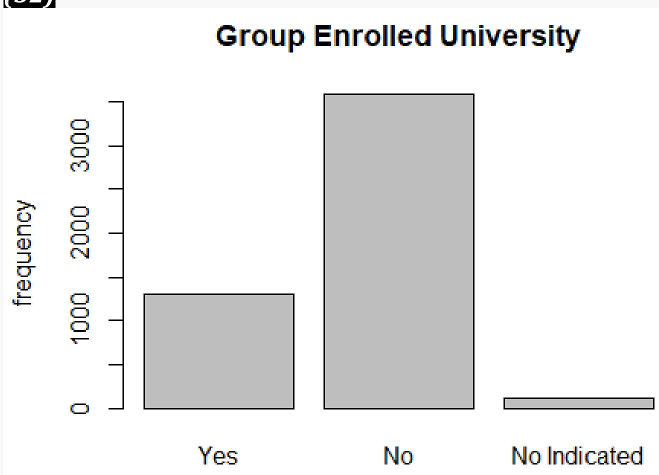
(30)


Last New Job
frequency
0
1000
3000
>0
None
No Indicated

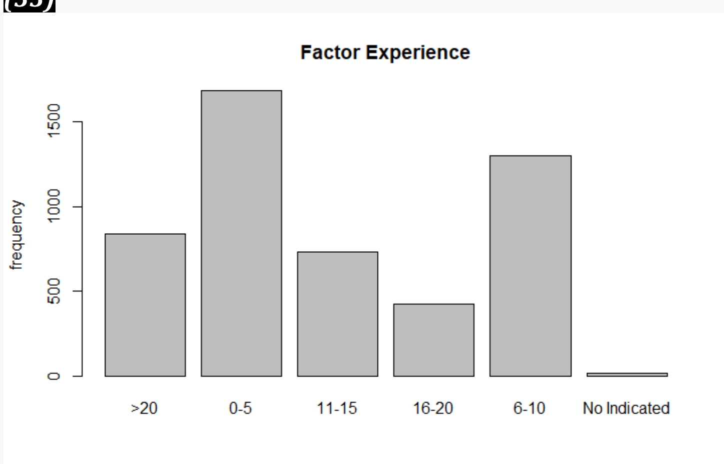
(31)


Relevant Experience
frequency
0
1000
2000
3000
Yes
No

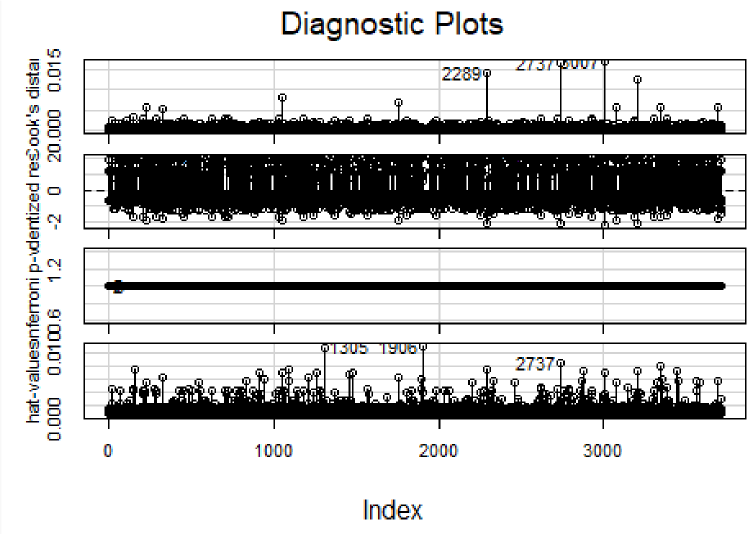
(32)


Group Enrolled University
frequency
0
1000
2000
3000
Yes
No
No Indicated

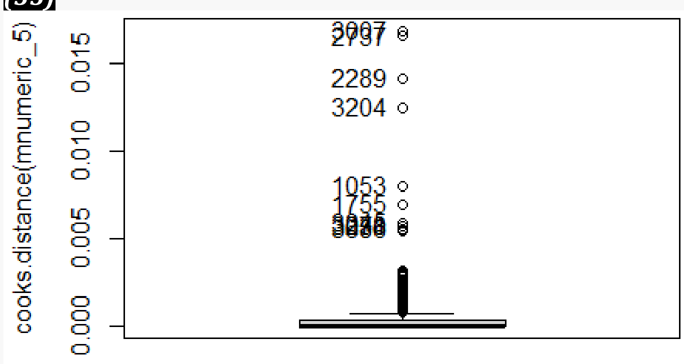
(33)


Factor Experience
frequency
0
500
1000
1500
>20
0-5
11-15
16-20
6-10
No Indicated

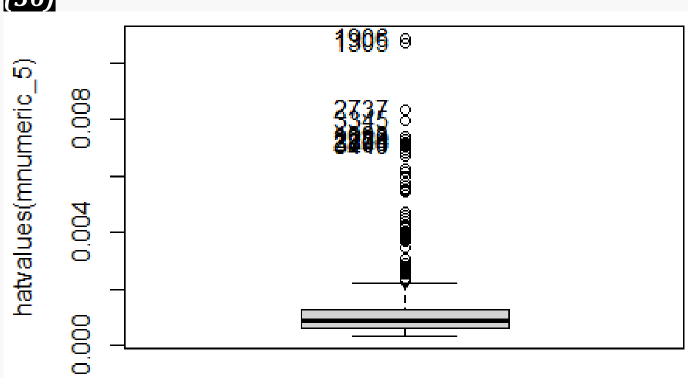
(34)


Diagnostic Plots
2289
2737
1305
1906
2737
0
1000
2000
3000
Index

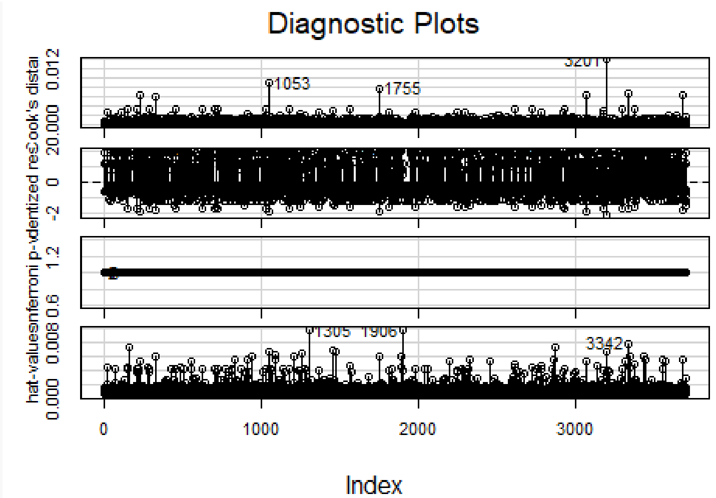
(35)


cooks.distance(mnumeric_5)
0.000
0.005
0.010
0.015
2289
3204
1053
1755

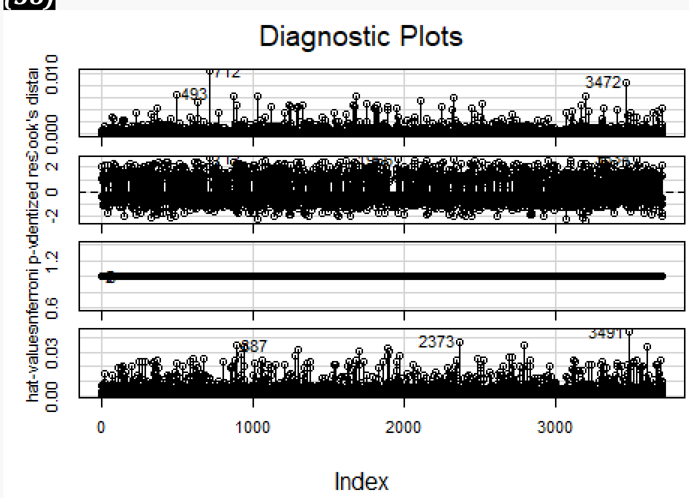
(36)


hatvalues(mnumeric_5)
0.000
0.004
0.008
2737

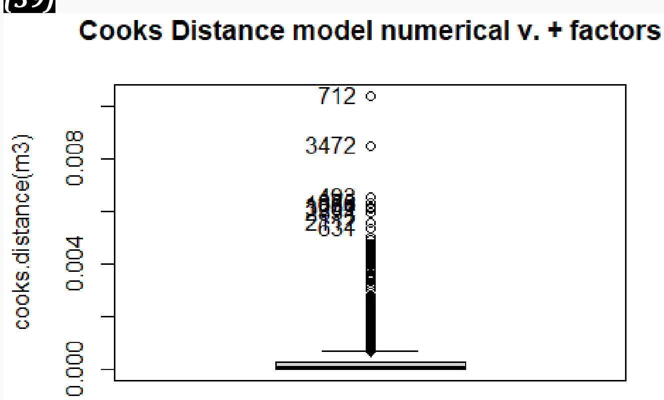
(37)


Diagnostic Plots
Cook's distance
Studentized residuals
Bonferroni p-value
hat-values
1053
1755
3201
1305
1906
3342
0
1000
2000
3000
Index

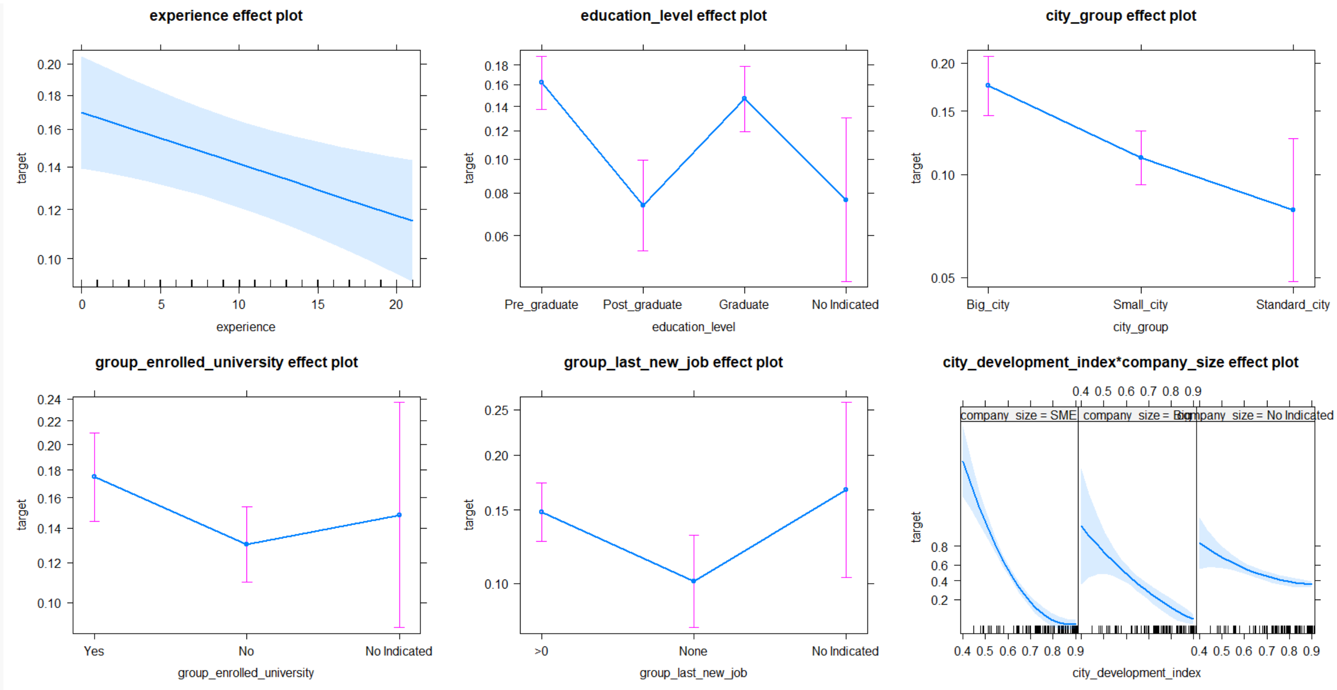
(38)


Diagnostic Plots
Cook's distance
Studentized residuals
Bonferroni p-value
hat-values
712
493
3472
887
2373
3491
0
1000
2000
3000
Index

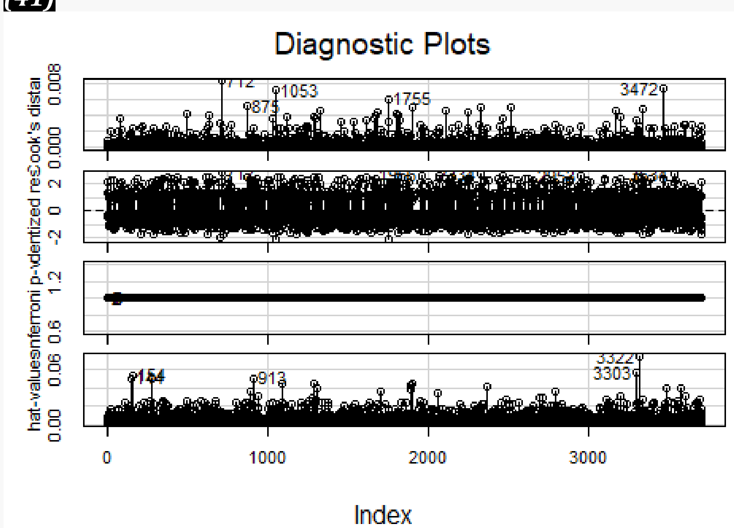
(39)


Cooks Distance model numerical v. + factors
cooks.distance(m3)
0.000
0.004
0.008
712
3472

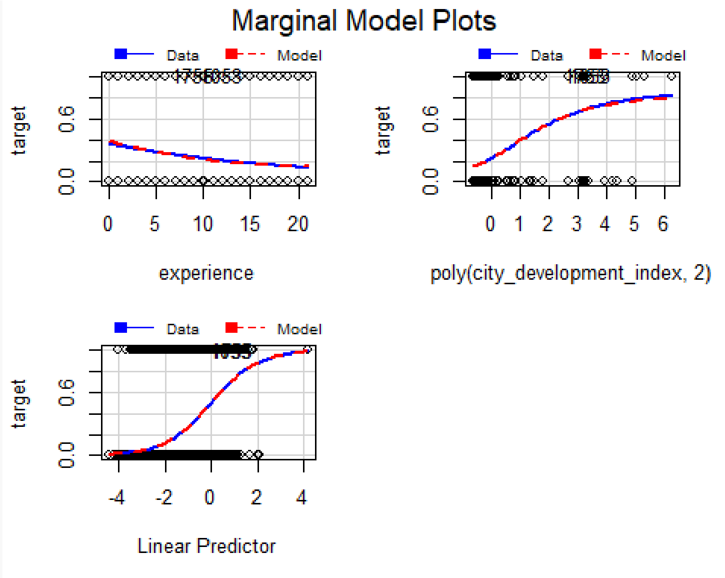
(40)


experience effect plot
education_level effect plot
city_group effect plot
group_enrolled_university effect plot
group_last_new_job effect plot
city_development_index*company_size effect plot
target
experience
Pre_graduate
Post_graduate
Graduate
No Indicated
education_level
Big_city
Small_city
Standard_city
city_group
Yes
No
No Indicated
group_enrolled_university
>0
None
No Indicated
group_last_new_job
company_size = SME
city_development_index

(41)

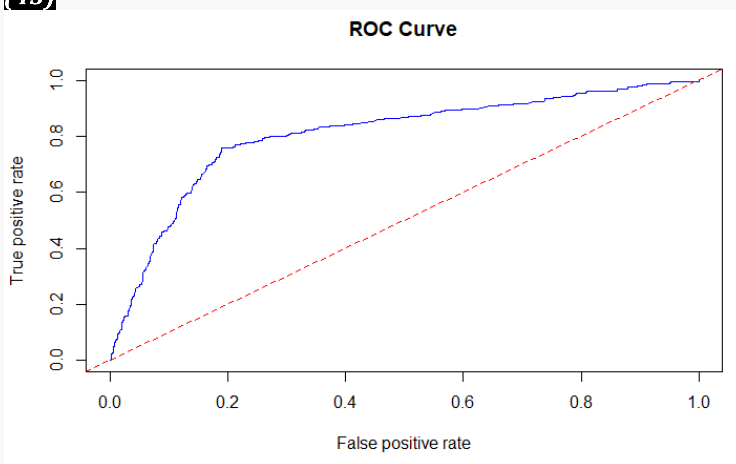

Diagnostic Plots
Index

(42)



Marginal Model Plots
Data
Model
target
0.0
0.6
0
5
10
15
20
experience
poly(city_development_index, 2)
0
1
2
3
4
5
6
Linear Predictor
-4
-2
0
2
4

(43)



ROC Curve
True positive rate
False positive rate
0.0
0.2
0.4
0.6
0.8
1.0