Assignment 2

Diana Rocio Galindo Gonzalez

```
rm(list=ls())
set.seed(310883)

# Import initial dataset
csvf<-list.files(path=".",pattern=".csv")
df1<-data.frame(lapply(csvf, read.delim,stringsAsFactors = TRUE,header=T, sep=","))

# Reordering columns to facilitate my reading
df2<-df1[,c(1,14,3,13,2,4:12)]

# Initial data frame structure
data.frame(Variable = names(df2),
 Class = sapply(df2, class),
 Head = sapply(df2, function(x) paste0(head(x,n=4), collapse = ", ")),
 row.names = NULL) %>% kable(booktabs = T) %>%
  kable_styling(full_width = F, latex_options = c("striped", "scale_down")) %>%
  column_spec(1, width = "10em")
```

Data preparation

To structure the dataset to be used the data preparation includes fix structural errors associated to trailing blanks in labels and removing duplicate observations:

```
# Removing leading, trailing, multiple spaces on levels
for (i in c(1,6:14)){ df2[,i] <-gsub(" +$", " ", df2[,i]) }
for (i in c(1,6:14)){ df2[,i] <-trimws(df2[,i])}
```

Variable	Class	Head
enrollee_id	integer	8949, 29725, 11561, 33241
target	numeric	1, 0, 0, 1
city_development_inde:	numeric	0.92, 0.776, 0.624, 0.789
training_hours	integer	36, 47, 83, 52
city	factor	city_103, city_40, city_21, city_115
gender	factor	Male, Male, ,
relevent_experience	factor	Has relevent experience, No relevent experience, No relevent experience, No relevent experience
enrolled_university	factor	no_enrollment, no_enrollment, Full time course,
education_level	factor	Graduate, Graduate, Graduate, Graduate
major_discipline	factor	STEM, STEM, STEM, Business Degree
experience	factor	>20, 15, 5, <1
company_size	factor	, 50-99, ,
company_type	factor	, Pvt Ltd, , Pvt Ltd
last_new_job	factor	1, >4, never, never

```
colnames(df2)[3] <-'CDI'
names(df2)
```

```
##  [1] "enrollee_id"         "target"              "CDI"
##  [4] "training_hours"      "city"                "gender"
##  [7] "relevent_experience" "enrolled_university" "education_level"
## [10] "major_discipline"    "experience"          "company_size"
## [13] "company_type"        "last_new_job"
```

```
df3<-df2[,c("enrollee_id","target","city","CDI","training_hours","gender","education_level",
            "enrolled_university","major_discipline","experience",
            "relevent_experience","company_type","company_size","last_new_job")]
table(duplicated(df3))
```

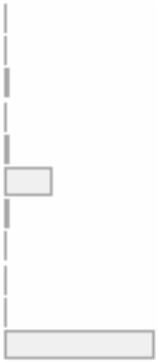
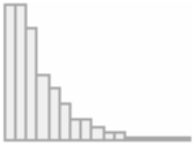
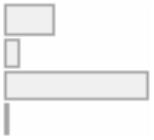
```
##
## FALSE
## 19158
```

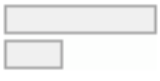
```
dfSummary(df3)
```

Initial data frame summary in PDF Format

df3
Dimensions: 19158 x 14
Duplicates: 0

No	Variable	Stats / Values	Freqs (% of Valid)	Graph	Missing
1	enrollee_id [character]	1. 1 2. 10 3. 10000 4. 10001 5. 10002 6. 10003 7. 10004 8. 10005 9. 10006 10. 10008 [19148 others]	1 (0.0%) 1 (0.0%) 1 (0.0%) 1 (0.0%) 1 (0.0%) 1 (0.0%) 1 (0.0%) 1 (0.0%) 1 (0.0%) 1 (0.0%) 19148 (99.9%)		0 (0.0%)
2	target [numeric]	Min : 0 Mean : 0.2 Max : 1	0 : 14381 (75.1%) 1 : 4777 (24.9%)		0 (0.0%)

No	Variable	Stats / Values	Freqs (% of Valid)	Graph	Missing
3	city [factor]	1. city_1 2. city_10 3. city_100 4. city_101 5. city_102 6. city_103 7. city_104 8. city_105 9. city_106 10. city_107 [113 others]	26 (0.1%) 86 (0.4%) 275 (1.4%) 75 (0.4%) 304 (1.6%) 4355 (22.7%) 301 (1.6%) 79 (0.4%) 9 (0.0%) 6 (0.0%) 13642 (71.2%)		0 (0.0%)
4	CDI [numeric]	Mean (sd) : 0.8 (0.1) min < med < max: 0.4 < 0.9 < 0.9 IQR (CV) : 0.2 (0.1)	93 distinct values		0 (0.0%)
5	training_hours [integer]	Mean (sd) : 65.4 (60.1) min < med < max: 1 < 47 < 336 IQR (CV) : 65 (0.9)	241 distinct values		0 (0.0%)
6	gender [character]	1. (Empty string) 2. Female 3. Male 4. Other	4508 (23.5%) 1238 (6.5%) 13221 (69.0%) 191 (1.0%)		0 (0.0%)
7	education_level [character]	1. (Empty string) 2. Graduate 3. High School 4. Masters 5. Phd 6. Primary School	460 (2.4%) 11598 (60.5%) 2017 (10.5%) 4361 (22.8%) 414 (2.2%) 308 (1.6%)		0 (0.0%)
8	enrolled_university [character]	1. (Empty string) 2. Full time course 3. no_enrollment 4. Part time course	386 (2.0%) 3757 (19.6%) 13817 (72.1%) 1198 (6.3%)		0 (0.0%)
9	major_discipline [character]	1. (Empty string) 2. Arts 3. Business Degree 4. Humanities 5. No Major 6. Other 7. STEM	2813 (14.7%) 253 (1.3%) 327 (1.7%) 669 (3.5%) 223 (1.2%) 381 (2.0%) 14492 (75.6%)		0 (0.0%)

No	Variable	Stats / Values	Freqs (% of Valid)	Graph	Missing
10	experience [character]	1. >20	3286 (17.2%)		0 (0.0%)
		2. 5	1430 (7.5%)		
		3. 4	1403 (7.3%)		
		4. 3	1354 (7.1%)		
		5. 6	1216 (6.3%)		
		6. 2	1127 (5.9%)		
		7. 7	1028 (5.4%)		
		8. 10	985 (5.1%)		
		9. 9	980 (5.1%)		
		10. 8	802 (4.2%)		
		[13 others]	5547 (29.0%)		
11	relevent_experience [character]	1. Has relevent experience	13792 (72.0%)		0 (0.0%)
		2. No relevent experience	5366 (28.0%)		
12	company_type [character]	1. (Empty string)	6140 (32.0%)		0 (0.0%)
		2. Early Stage Startup	603 (3.1%)		
		3. Funded Startup	1001 (5.2%)		
		4. NGO	521 (2.7%)		
		5. Other	121 (0.6%)		
		6. Public Sector	955 (5.0%)		
		7. Pvt Ltd	9817 (51.2%)		
13	company_size [character]	1. (Empty string)	5938 (31.0%)		0 (0.0%)
		2. <10	1308 (6.8%)		
		3. 10/49	1471 (7.7%)		
		4. 100-500	2571 (13.4%)		
		5. 1000-4999	1328 (6.9%)		
		6. 10000+	2019 (10.5%)		
		7. 50-99	3083 (16.1%)		
		8. 500-999	877 (4.6%)		
		9. 5000-9999	563 (2.9%)		
14	last_new_job [character]	1. (Empty string)	423 (2.2%)		0 (0.0%)
		2. >4	3290 (17.2%)		
		3. 1	8040 (42.0%)		
		4. 2	2900 (15.1%)		
		5. 3	1024 (5.3%)		
		6. 4	1029 (5.4%)		
		7. never	2452 (12.8%)		

The dataset has no duplicated data and there is no upper/lower cases to homogenize.

The second step includes check data types and levels of categorical data. For candidates dataset, counts with 11 variables and the identifier. Two variables are numeric (City Development Index and training hours). The rest of the variables are categorical variables. The dataset contains empty cells which are mostly associated to NA data.

For the purpose of the modeling the categorical variables are set as ordinal variables in the way that lower values of the ordinal variables are associated too have FALSE output as expected and counting with the assumption that individuals with higher education level, experience and working for bigger and consolidated companies have a better response.

To manage the empty data firstly and related with the consistency of the information the following assumptions were made:

- Candidates with primary school can not be enrolled in university hence if their enrolled university is empty changes to "no enrollment"
- Candidates with any Education major discipline has at least graduate level of education hence their if their education_level is empty changes to "graduate"
- Candidates with any Education major discipline has at least graduate level of education hence their if their education_level is empty changes to "graduate"

The rest of empty data were considered NA data, and it was imputed as it is explain in the next section.

```
str(df3)
table(df3$education_level)
table(df3$enrolled_university)
table(df3$major_discipline)

df3[df3$enrolled_university== "" & df3$education_level=="Primary School",
    c("enrolled_university")]<-"no_enrollment"
df3[df3$major_discipline != "" & df3$education_level=="",
    c("education_level")]<-"Graduate" # 0 records
```

As well, as part of the preprocessing and to perform the following rules to perform the predictive analysis:

- Consider: **City Development Index (CDI)**, **training hours**, **experience** and **last new job** as numerical variables due to its characteristics, the variables values are associated more to numbers than categories and are considered in the same way. The highest number of city development index, training hours and years of experience is expected to have associated higher probability of being hired. The last new job is interpreted as a variable to describe expectations to change of candidate.
- To make the string variables to numericFor the case of experience, candidates with less than one year of experience are relabelled as 0.5 and those with more than 20 years, 20.5 years to make the differences with the initial data and highlight they correspond to distinct numbers.

Under the same assumptions of higher levels of education and experience are more expected to be in the hired candidates, I re categorize the variables according to the observed distribution, including as well technical similarities and aiming for balanced the distribution of the levels as follows and based on data distribution (Column "graph" on summary description):

- **enrrolled_university** with two categories: NotEnrolled (2.no_enrollment) and Enrolled (3. Part time course and 1. Full time course)
- **education_level** with three categories: School (5. Primary School and 2. High School), Graduate (1. Graduate) and Specialized (3.Masters and 4. PhD)
- **major_discipline** with three categories: NoMajor(4. No Major), NoSTEM (1. Arts, 2. Business Degree, 3. Humanities and 5. Other) and STEM (6.)
- **company_size**: According to the quartile distribution of the data, Less than 100 employees, Between 50 and 500, between 100 and 5000, and More than 4.

```
## Numeric

df3$experience<-as.character(df3$experience)
df3$experience[df3$experience == '<1'] <- '0.9'
```

```
df3$experience[df3$experience == '>20'] <- '20.5'
df3$experience<-as.numeric(df3$experience)


## Levels of factor variables
df3$relevent_experience<-as.factor(df3$relevent_experience)
levels(df3$relevent_experience)<-list(NoRelevantExp="Has relevent experience",
                                      RelevantExp="No relevent experience")


df3$enrolled_university<-as.factor(df3$enrolled_university)
levels(df3$enrolled_university)<-list(NotEnrolled="no_enrollment", Enrolled=
                                        "Part time course", Enrolled="Full time course")
#str(df3$enrolled_university)


df3$education_level<-as.factor(df3$education_level)
levels(df3$education_level)<-list(School= "Primary School",School="High School",
                                  Grad="Graduate",Specialized="Masters",Specialized="Phd")
#str(df3$education_level)

df3$major_discipline<-as.factor(df3$major_discipline)
levels(df3$major_discipline)<-list(NoMajor="No Major",NoSTEM = c("Arts","Business Degree",
"Humanities","Other"), STEM="STEM")
#str(df3$major_discipline)

df3$company_size <- factor(df3$company_size, order = TRUE,
                           levels =c('<10','10/49','50-99','100-500','500-999',
                                     '1000-4999','5000-9999','10000+'))
#table(ntile(df3$company_size, 4), df3$company_size)
df3$compSizeCat<-ntile(df3$company_size, 4)
#table(ntile(df3$company_size, 4), df3$company_size)
#table(df3$compSizeCat, df3$company_size)
df3$compSizeCat<-as.factor(df3$compSizeCat)

levels(df3$compSizeCat)<-list(Less100='1', '50to500' = '2', '100to5000' ='3',
                              More1000='4')
summary(df3)
```

```
##  enrollee_id            target             city           CDI
##  Length:19158       Min.   :0.0000   city_103:4355   Min.   :0.4480
##  Class :character   1st Qu.:0.0000   city_21 :2702   1st Qu.:0.7400
##  Mode  :character   Median :0.0000   city_16 :1533   Median :0.9030
##                     Mean   :0.2493   city_114:1336   Mean   :0.8288
##                     3rd Qu.:0.0000   city_160: 845   3rd Qu.:0.9200
##                     Max.   :1.0000   city_136: 586   Max.   :0.9490
##                                      (Other) :7801
##  training_hours      gender           education_level   enrolled_university
##  Min.   :  1.00   Length:19158       School     : 2325   NotEnrolled:13826
##  1st Qu.: 23.00   Class :character   Grad       :11598   Enrolled   : 4955
##  Median : 47.00   Mode  :character   Specialized: 4775   NA's       :  377
##  Mean   : 65.37                      NA's       :  460
##  3rd Qu.: 88.00
```

```
##  Max.   :336.00
##
##  major_discipline   experience       relevent_experience company_type
##  NoMajor:  223    Min.   : 0.90   NoRelevantExp:13792    Length:19158
##  NoSTEM : 1630    1st Qu.: 4.00   RelevantExp  : 5366    Class :character
##  STEM   :14492    Median : 9.00                          Mode  :character
##  NA's   : 2813    Mean   :10.04
##                   3rd Qu.:16.00
##                   Max.   :20.50
##                   NA's   :65
##     company_size  last_new_job          compSizeCat
##  50-99    :3083   Length:19158       Less100  :3305
##  100-500  :2571   Class :character   50to500  :3305
##  10000+   :2019   Mode  :character   100to5000:3305
##  10/49    :1471                      More1000 :3305
##  1000-4999:1328                      NA's     :5938
##  (Other)  :2748
##  NA's     :5938
```

```
df4<-df3[,-13]
```

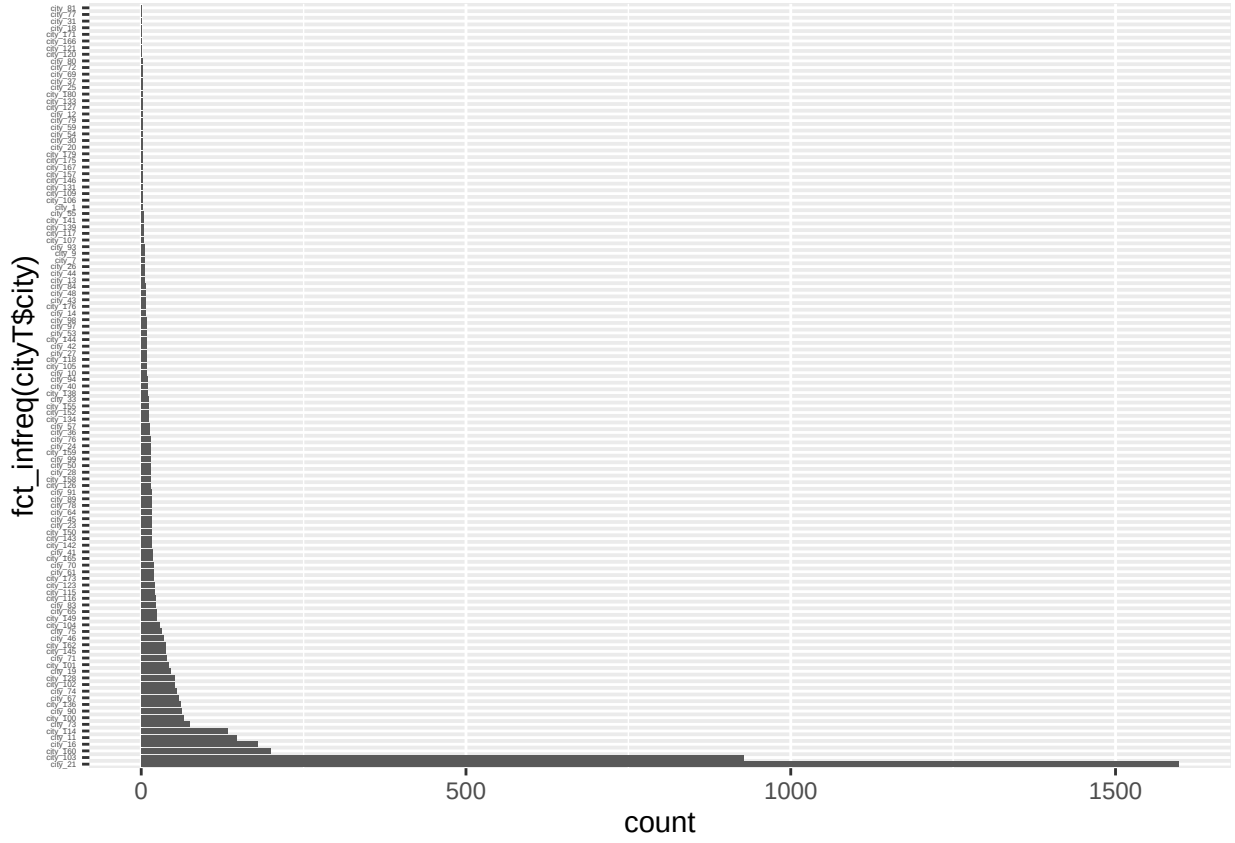
For the remaining variables:**city**, **company__type** and **last__new__job** it is necessary to review the distribution of the data.

```
print(str(df4))
```

```
## 'data.frame':    19158 obs. of  14 variables:
##  $ enrollee_id        : chr  "8949" "29725" "11561" "33241" ...
##  $ target             : num  1 0 0 1 0 1 0 1 1 0 ...
##  $ city               : Factor w/ 123 levels "city_1","city_10",..: 6 78 65 15 51 58 50 84 6 6 ...
##  $ CDI                : num  0.92 0.776 0.624 0.789 0.767 0.764 0.92 0.762 0.92 0.92 ...
##  $ training_hours     : int  36 47 83 52 8 24 24 18 46 123 ...
##  $ gender             : chr  "Male" "Male" "" "" ...
##  $ education_level    : Factor w/ 3 levels "School","Grad",..: 2 2 2 2 3 2 1 2 2 2 ...
##  $ enrolled_university: Factor w/ 2 levels "NotEnrolled",..: 1 1 2 NA 1 2 1 1 1 1 ...
##  $ major_discipline   : Factor w/ 3 levels "NoMajor","NoSTEM",..: 3 3 3 2 3 3 NA 3 3 3 ...
##  $ experience         : num  20.5 15 5 0.9 20.5 11 5 13 7 17 ...
##  $ relevent_experience: Factor w/ 2 levels "NoRelevantExp",..: 1 2 2 2 1 1 1 1 1 1 ...
##  $ company_type       : chr  "" "Pvt Ltd" "" "Pvt Ltd" ...
##  $ last_new_job       : chr  "1" ">4" "never" "never" ...
##  $ compSizeCat        : Factor w/ 4 levels "Less100","50to500",..: NA 1 NA NA 1 NA 1 1 1 4 ...
## NULL
```

```
resp_cat<-df4[,c("target","city","company_type","last_new_job")]
#str(resp_cat)

resp_cat$target<-as.logical(resp_cat$target)
#as.factor(resp_cat$target)
resp_cat[sapply(resp_cat, is.character)] <- lapply(resp_cat[sapply(resp_cat,
                                                                  is.character)],
                                                   as.factor)
#str(resp_cat)
```

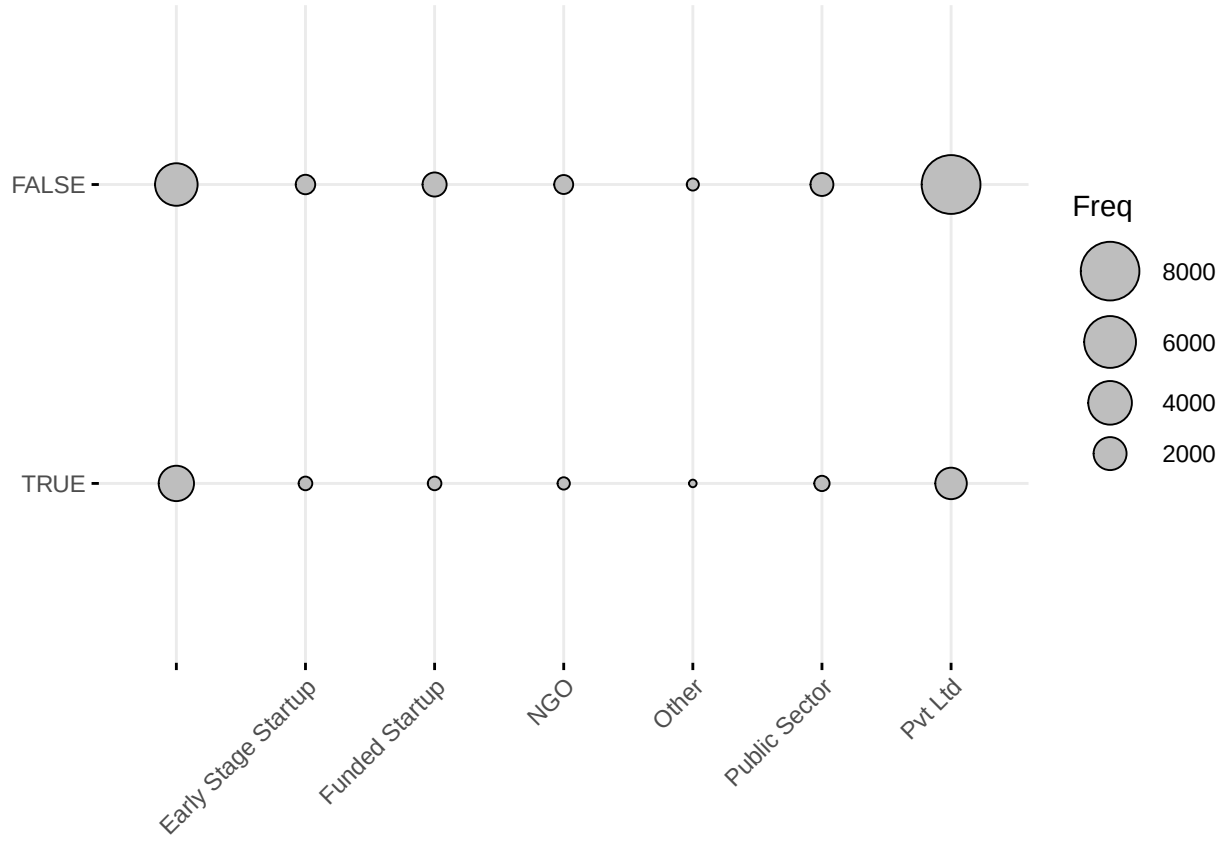
```
#City
cityT<-df4[df4$target==TRUE,c("target","city")]
tcityt<-as.data.frame(table(cityT))
#head(tcityt)

#tcityt$perc<-tcityt$Freq/sum(tcityt$Freq)*100
#head(tcityt)

#tcityt<-tcityt[,c(2,4)]
#head(tcityt)

#tcityt<-tcityt[order(-tcityt$perc),]
#head(tcityt)

ggplot(cityT, aes(y=fct_infreq(cityT$city)), fill = cityT$Freq) + geom_bar()+
  theme(axis.text.y = element_text(size =3))
```



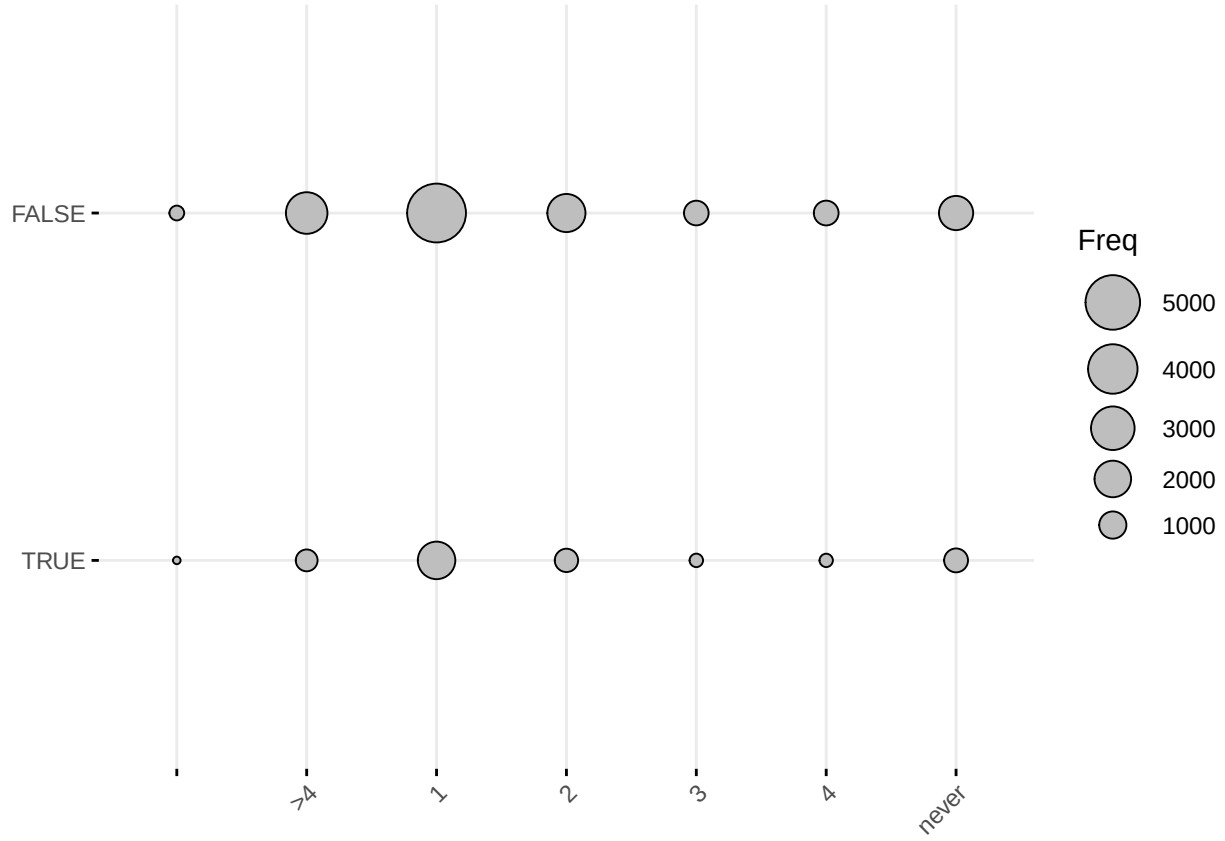

```
df4$CityQ<-ntile(df4$city, 4)
df4$CityQ<-as.factor(df4$CityQ)
levels(df4$CityQ)<-list(CityGroup1=1, CityGroup2=2,
                                  CityGroup3=3, CityGroup4=4)

#ggballoonplot(d,fill = "#0073C2FF",size.range = c(0.01, 1), show.label = TRUE,
#font.label = c(2, "plain")) + theme(axis.text.y = element_blank())
```

The city variable is grouped by the distribution in percentiles given the high number of levels.

```
# Company Type
b<-table(df4$company_type,resp_cat$target)
c<-as.data.frame(b)

ggballoonplot(c)
```




```
df4$company_type<-as.factor(df4$company_type)
levels(df4$company_type)<-list(Private= "Pvt Ltd",Startup="Funded Startup",
                               Startup="Early Stage Startup", Other = "NGO",
                               Other ="Public Sector", Other = "Other")
#str(df4$company_type)

# Last New Job
d<-table(df4$last_new_job,resp_cat$target)
e<-as.data.frame(d)

ggballoonplot(e)
```




```
df4$last_new_job<-as.character(df4$last_new_job)
df4$last_new_job[df4$last_new_job == 'never'] <- '0'
df4$last_new_job[df4$last_new_job == '>4'] <- '4.5'
df4$last_new_job<-as.numeric(df4$last_new_job)

df4$target<-as.logical(df4$target)
df4$gender<-as.factor(df4$gender)

#ordering the columns
df5<-df4[,c("enrollee_id","target","CDI","training_hours",
            "experience", "last_new_job", "gender","education_level",
            "enrolled_university", "major_discipline", "relevent_experience",
            "company_type","compSizeCat","CityQ")]

df5$training_hours<-as.double(df5$training_hours)
df5$last_new_job<-as.double(df5$last_new_job)
df5$target<-as.logical(df5$target)
str(df5)
```

```
## 'data.frame':    19158 obs. of  14 variables:
##  $ enrollee_id        : chr  "8949" "29725" "11561" "33241" ...
##  $ target             : logi  TRUE FALSE FALSE TRUE FALSE TRUE ...
##  $ CDI                : num  0.92 0.776 0.624 0.789 0.767 0.764 0.92 0.762 0.92 0.92 ...
##  $ training_hours     : num  36 47 83 52 8 24 24 18 46 123 ...
##  $ experience         : num  20.5 15 5 0.9 20.5 11 5 13 7 17 ...
##  $ last_new_job       : num  1 4.5 0 0 4 1 1 4.5 1 4.5 ...
```

```
##  $ gender             : Factor w/ 4 levels "","Female","Male",..: 3 3 1 1 3 1 3 3 3 1 ...
##  $ education_level    : Factor w/ 3 levels "School","Grad",..: 2 2 2 2 3 2 1 2 2 2 ...
##  $ enrolled_university: Factor w/ 2 levels "NotEnrolled",..: 1 1 2 NA 1 2 1 1 1 1 ...
##  $ major_discipline   : Factor w/ 3 levels "NoMajor","NoSTEM",..: 3 3 3 2 3 3 NA 3 3 3 ...
##  $ relevent_experience: Factor w/ 2 levels "NoRelevantExp",..: 1 2 2 2 1 1 1 1 1 1 ...
##  $ company_type       : Factor w/ 3 levels "Private","Startup",..: NA 1 NA 1 2 NA 2 1 1 1 ...
##  $ compSizeCat        : Factor w/ 4 levels "Less100","50to500",..: NA 1 NA NA 1 NA 1 1 1 4 ...
##  $ CityQ              : Factor w/ 4 levels "CityGroup1","CityGroup2",..: 1 4 3 2 3 3 3 4 1 1 ...
```

```
#b1 <- getBins(df, "target",
# c("CDI","training_hours","experience","last_new_job","gender","education_level",
#"enrolled_university","major_discipline","relevent_experience","company_type",
#"compSizeCat"), minCr = 0.6, nCores = 2)
#b1
```

According to the distribution of the data, the **company_type** variable is recategorized with three levels: Private (Pvt Ltd), Startup (Funded Startup and Early Stage Startup) and Other (NGO,Public Sector and Other). The variable **last_new_job** is converted to numeric.

With the previous preprocessed data, the empty values are considered null.

```
df5[df5 == ""] <- NA
dim(df5)
```

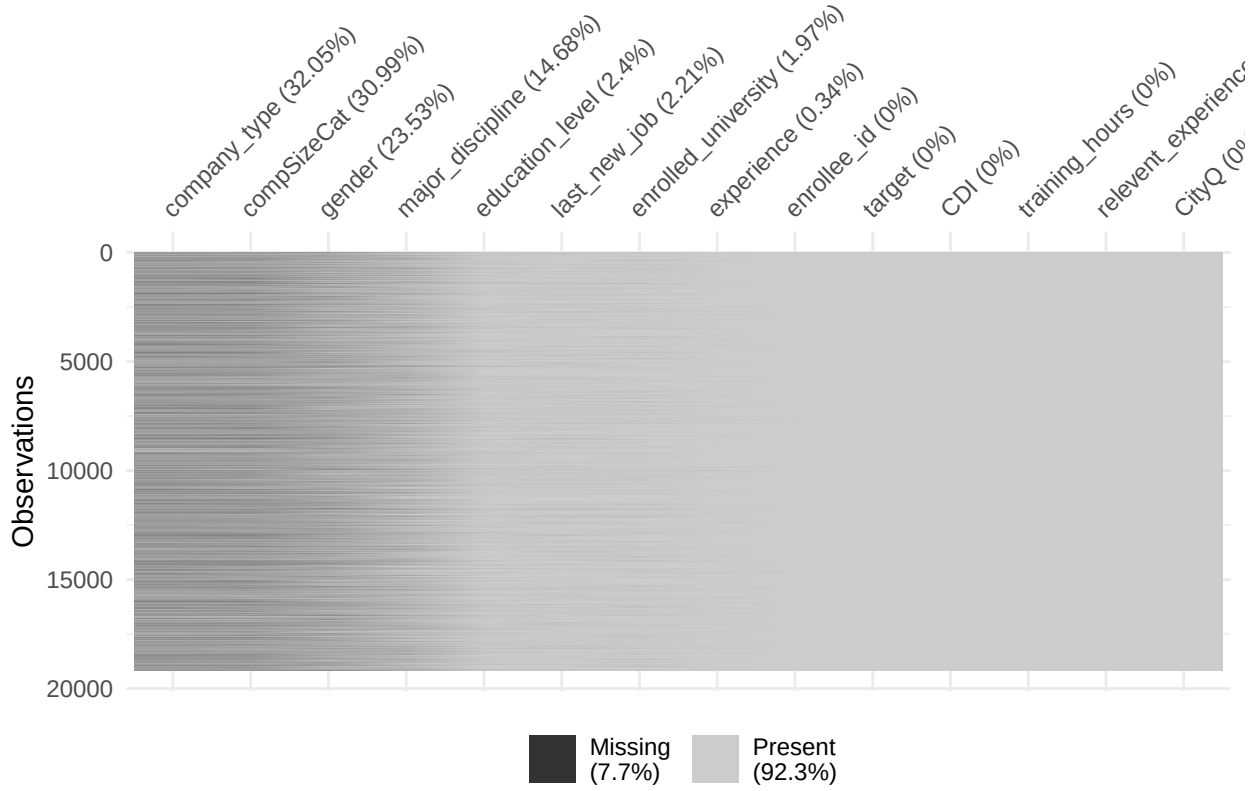
```
## [1] 19158    14
```

```
dim(df5[complete.cases(df5),])
```

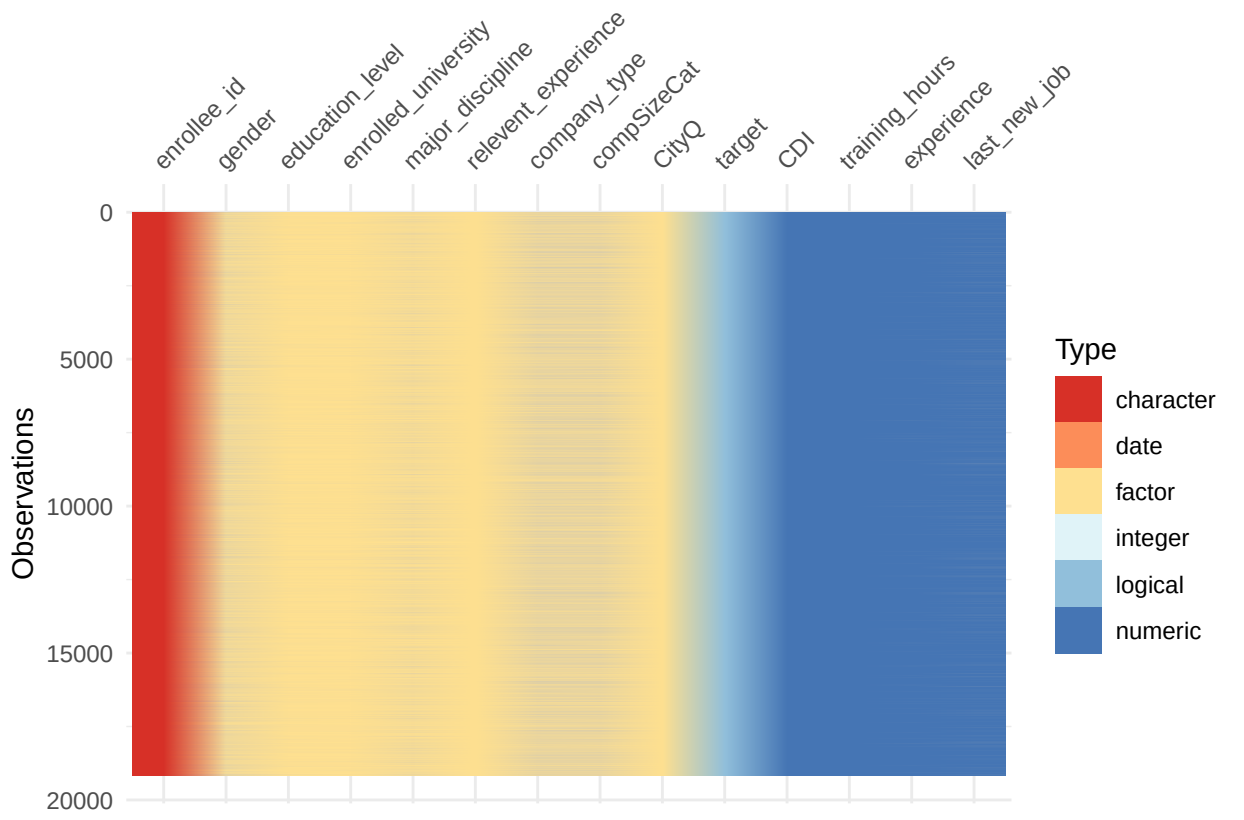
```
## [1] 8955   14
```

The dataset contains 19.158 rows and the complete cases are 8.955 corresponding to the 46.74% of the cases. 7.7% of the data is missing in the data. The following graphs depicts the distribution of missing data.

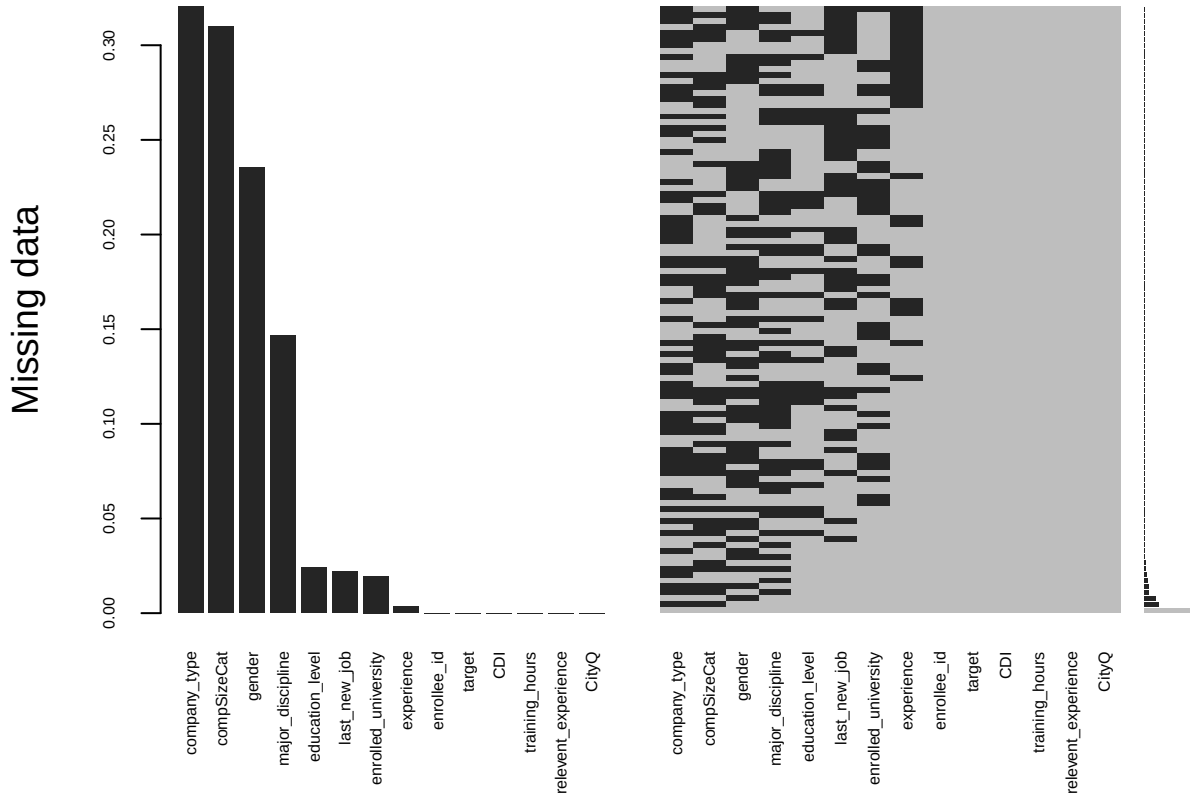
```
# Missing Data Graphs
vis_miss(df5,sort_miss = TRUE, show_perc = TRUE,
  show_perc_col = TRUE)
```




```
vis_dat(df5, sort_type = TRUE, palette = "cb_safe")
```




```
aggr(df5, col=c('grey','#252525'), numbers=TRUE, sortVars=TRUE, labels=names(df5),
     cex.axis=.5, gap=1, ylab=c("Missing data"," "),border=NA, prop=TRUE)
```




```
##
##  Variables sorted by number of missings:
##              Variable       Count
##          company_type 0.320492745
##           compSizeCat 0.309948846
##                gender 0.235306399
##     major_discipline 0.146831611
##       education_level 0.024010857
##          last_new_job 0.022079549
##  enrolled_university 0.019678463
##            experience 0.003392839
##           enrollee_id 0.000000000
##                target 0.000000000
##                   CDI 0.000000000
##        training_hours 0.000000000
##  relevent_experience 0.000000000
##                 CityQ 0.000000000
```

company_type is the variable with highest proportion of missing values (32.05%) of the data, followed by the number of employees in current employer's company (30.99%) and gender (23.53%).

Regarding the outliers, univariate outliers are calculated and set as NA. In accordance with the summary table, there is presence of univariate outlier values per numeric variable.The threshold established corresponds to be > 1.5*Interquartile Range from the borders of the box.

```
df6<-df5
nm<-names(df5[,3:6])
a<-as.list(nm)
funbox<-function(i){out<-boxplot.stats(df5[,i])$out
                  label=ifelse(df5[,i] %in% out,df5$enrollee_id,"")
                  print(ggplot(df5,aes(x=df5[,2],y=df5[,i])) +
                  geom_boxplot(outlier.colour="red",
                               outlier.size=0.8) +
                    geom_text(aes(label=label),hjust=3, size=0.1) +
                    labs(title=nm[i-2], x="target", y=NULL) +
                    theme(plot.title = element_text(size = rel(0.7),face ="bold",
                                                    hjust = 0.5),
                       axis.title.y = element_text(size = rel(0.6)),
                       axis.text = element_text(size = rel(0.6))))
                  print(paste("Enrolled_id of outliers in ",title=nm[i-2]))
                  print(df5$enrollee_id[df5[,i] %in% out])}
boxplots<-lapply(3:6,funbox)
```

```
## [1] "Enrolled_id of outliers in  CDI"
##  [1] "8238"  "30985" "27970" "31194" "598"   "18564" "31179" "26838" "4858"
## [10] "19463" "28317" "3891"  "4776"  "30131" "10486" "24256" "16548"
```

```
## [1] "Enrolled_id of outliers in  training_hours"
##   [1] "4866"  "4830"  "32401" "4789"  "14199" "32776" "13333" "4064"  "3024"
##  [10] "7672"  "5278"  "399"   "12454" "16994" "25418" "12117" "20322" "22194"
##  [19] "13788" "27435" "10481" "26388" "21466" "4626"  "15796" "6283"  "33367"
##  [28] "22954" "19555" "22420" "19702" "33181" "12349" "6995"  "24280" "24937"
##  [37] "21442" "26126" "6502"  "12217" "3708"  "7951"  "8852"  "25834" "19560"
##  [46] "23295" "31368" "13273" "21667" "29151" "27539" "20048" "21691" "27153"
##  [55] "32082" "7726"  "6511"  "11045" "15264" "6838"  "10653" "15942" "17679"
##  [64] "21192" "4029"  "7174"  "19544" "28992" "2691"  "24034" "22113" "8270"
##  [73] "26477" "2602"  "8251"  "19448" "3674"  "24874" "32502" "29705" "241"
##  [82] "25317" "5490"  "861"   "21960" "26170" "13713" "21784" "24420" "13378"
##  [91] "6608"  "7580"  "18654" "30499" "24915" "29698" "6410"  "11160" "23754"
## [100] "12615" "24351" "29249" "19581" "12556" "26177" "28345" "18993" "19175"
## [109] "23841" "13231" "33130" "2756"  "294"   "31322" "29837" "28097" "17219"
## [118] "6863"  "1384"  "17776" "15655" "30238" "18056" "2909"  "12962" "32262"
## [127] "20368" "26140" "1495"  "5682"  "11396" "8355"  "7251"  "3288"  "25976"
## [136] "15444" "30412" "6475"  "663"   "10452" "13799" "262"   "29893" "29770"
## [145] "24574" "5316"  "28986" "4753"  "18828" "28678" "23107" "16313" "451"
## [154] "5920"  "20915" "24733" "21329" "19497" "8675"  "4022"  "31256" "29273"
## [163] "8888"  "16982" "24537" "8697"  "14352" "3222"  "12441" "14146" "28682"
## [172] "16445" "13310" "9447"  "28906" "7662"  "18714" "12783" "30610" "7344"
## [181] "16778" "8775"  "9293"  "24520" "24686" "10926" "25500" "23285" "8841"
## [190] "14087" "26641" "11203" "5499"  "26396" "29932" "4711"  "315"   "23504"
## [199] "562"   "28186" "5262"  "13989" "27391" "26239" "9171"  "18562" "9538"
## [208] "28414" "17053" "30127" "9346"  "8945"  "24624" "14332" "17752" "31933"
## [217] "12260" "31390" "6228"  "7975"  "27651" "5679"  "27795" "14212" "31735"
## [226] "5656"  "5602"  "24019" "19910" "28793" "31780" "33282" "9477"  "21846"
## [235] "29017" "3820"  "5013"  "7271"  "28975" "391"   "33370" "748"   "2890"
## [244] "30034" "9553"  "9146"  "1340"  "27313" "3886"  "7099"  "12383" "20446"
## [253] "32407" "30230" "2206"  "22350" "17089" "7556"  "33278" "31134" "8094"
## [262] "30931" "25080" "10983" "3992"  "4442"  "19589" "6176"  "24242" "31127"
```

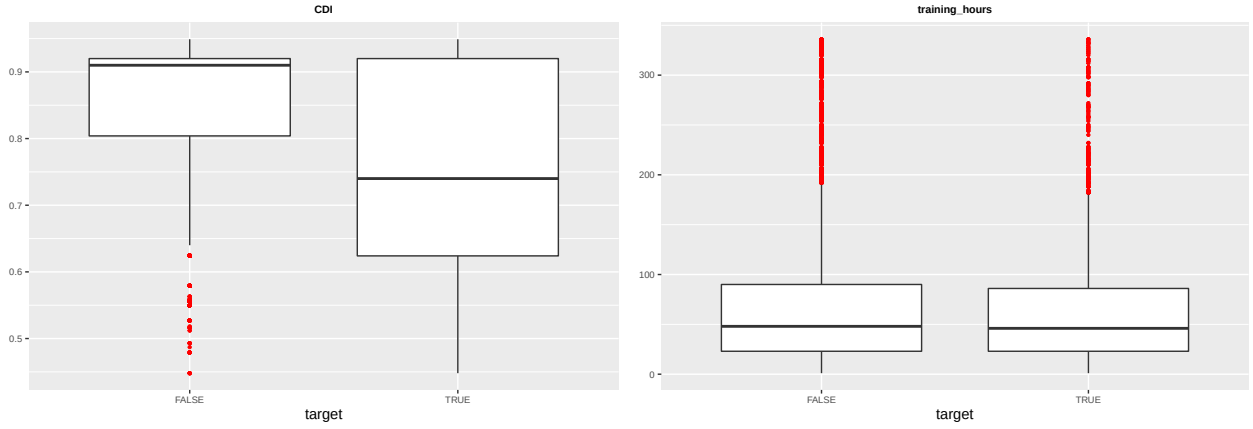
```
## [271] "27231" "26194" "25412" "2363"  "14486" "19879" "30806" "8572"  "14461"
## [280] "12720" "3157"  "25032" "24640" "7919"  "15053" "12715" "32999" "26889"
## [289] "9931"  "6248"  "11428" "12084" "2711"  "10751" "2341"  "4396"  "13618"
## [298] "6476"  "20054" "6417"  "23722" "17811" "16795" "31117" "7648"  "21363"
## [307] "30691" "15037" "23827" "25631" "2977"  "8158"  "1278"  "22466" "16610"
## [316] "22411" "26409" "19623" "32554" "16063" "4275"  "10043" "7757"  "17784"
## [325] "27197" "9220"  "1954"  "4537"  "15186" "23531" "5948"  "458"   "24047"
## [334] "23916" "31227" "15740" "31792" "15855" "13243" "8527"  "10755" "64"
## [343] "11321" "26102" "17765" "15058" "6489"  "9065"  "17262" "23252" "13841"
## [352] "10910" "20726" "19280" "16930" "14560" "15464" "745"   "26749" "10041"
## [361] "14912" "31105" "30804" "29239" "14753" "8774"  "8591"  "28863" "31462"
## [370] "7669"  "27691" "4576"  "868"   "28003" "29148" "15687" "15248" "3942"
## [379] "17186" "1508"  "13415" "8015"  "5469"  "11030" "3367"  "13617" "22243"
## [388] "11763" "18317" "2778"  "25183" "17895" "32462" "28823" "15360" "5876"
## [397] "14432" "17705" "32218" "3984"  "32588" "32217" "28135" "12544" "11873"
## [406] "27680" "2831"  "1369"  "22771" "31041" "29179" "662"   "32702" "12942"
## [415] "27656" "5054"  "5513"  "18704" "13120" "28052" "30754" "19720" "26678"
## [424] "649"   "16926" "26657" "24276" "15475" "5760"  "5121"  "15301" "33317"
## [433] "20920" "13679" "29419" "7433"  "2354"  "2260"  "21741" "13805" "28908"
## [442] "16590" "22486" "15401" "10135" "15610" "17678" "19870" "16006" "18430"
## [451] "28720" "29470" "31866" "1623"  "4342"  "11551" "5848"  "17955" "5474"
## [460] "21569" "17382" "13351" "29834" "6348"  "28382" "21051" "32135" "18392"
## [469] "6275"  "25887" "4193"  "11194" "28152" "8611"  "574"   "31346" "28184"
## [478] "24291" "8246"  "16308" "21736" "19370" "22514" "20539" "28813" "3174"
## [487] "21427" "7941"  "15308" "17951" "1766"  "18624" "1359"  "11487" "23946"
## [496] "1919"  "33220" "7735"  "4235"  "6856"  "9454"  "12959" "24792" "10502"
## [505] "25356" "23945" "1084"  "600"   "6198"  "9615"  "19733" "13077" "22179"
## [514] "19236" "5859"  "27831" "32899" "10534" "21866" "27110" "31373" "15584"
## [523] "12370" "5041"  "11453" "14924" "20502" "27375" "1502"  "32161" "13404"
## [532] "29677" "29436" "23925" "1756"  "20272" "21473" "3388"  "26560" "20641"
## [541] "15439" "5486"  "22484" "3364"  "21269" "23237" "23112" "7796"  "23074"
## [550] "30435" "26959" "7004"  "16227" "2499"  "6071"  "24066" "23180" "29251"
## [559] "27253" "4940"  "10557" "14418" "19038" "17409" "24696" "30022" "30707"
## [568] "3064"  "24765" "7734"  "9659"  "27492" "24972" "22901" "10163" "22220"
## [577] "2844"  "21609" "8137"  "5473"  "22438" "7977"  "16576" "7345"  "30852"
## [586] "14913" "20808" "10507" "24491" "7287"  "23086" "15906" "15741" "14640"
## [595] "8150"  "16925" "22783" "22072" "27665" "32239" "28687" "15765" "3715"
## [604] "11206" "27558" "1315"  "21088" "26644" "17315" "10396" "24799" "15819"
## [613] "3152"  "24044" "19656" "5706"  "6145"  "13054" "8985"  "15356" "26634"
## [622] "6855"  "10220" "26968" "30932" "19246" "11660" "710"   "18294" "6707"
## [631] "1092"  "7783"  "25153" "32984" "33167" "25985" "14320" "8348"  "1059"
## [640] "32044" "3454"  "21267" "18213" "7208"  "7616"  "20702" "7540"  "31857"
## [649] "16028" "17099" "13136" "31015" "32924" "11663" "20815" "23385" "26520"
## [658] "31979" "3066"  "5634"  "4572"  "21348" "9131"  "18288" "20188" "15952"
## [667] "24024" "3232"  "30993" "1753"  "23349" "8810"  "4360"  "2634"  "17176"
## [676] "16510" "2637"  "8763"  "22818" "31577" "14712" "25584" "8866"  "5119"
## [685] "23680" "31271" "22601" "5881"  "30329" "12641" "2775"  "239"   "30589"
## [694] "26822" "4983"  "22751" "2408"  "31331" "5725"  "19187" "23892" "29016"
## [703] "13744" "29120" "28963" "16064" "14868" "24122" "29163" "25386" "13247"
## [712] "13366" "20804" "11156" "6046"  "13059" "21753" "18815" "23227" "13821"
## [721] "25841" "30710" "7006"  "22921" "12337" "23111" "26871" "30835" "17311"
## [730] "32091" "32712" "11238" "31030" "17735" "30508" "23291" "7445"  "14096"
## [739] "16861" "24917" "19253" "25581" "17539" "12353" "23089" "15075" "14752"
## [748] "14961" "3915"  "25355" "19410" "20276" "6302"  "25796" "948"   "6080"
```

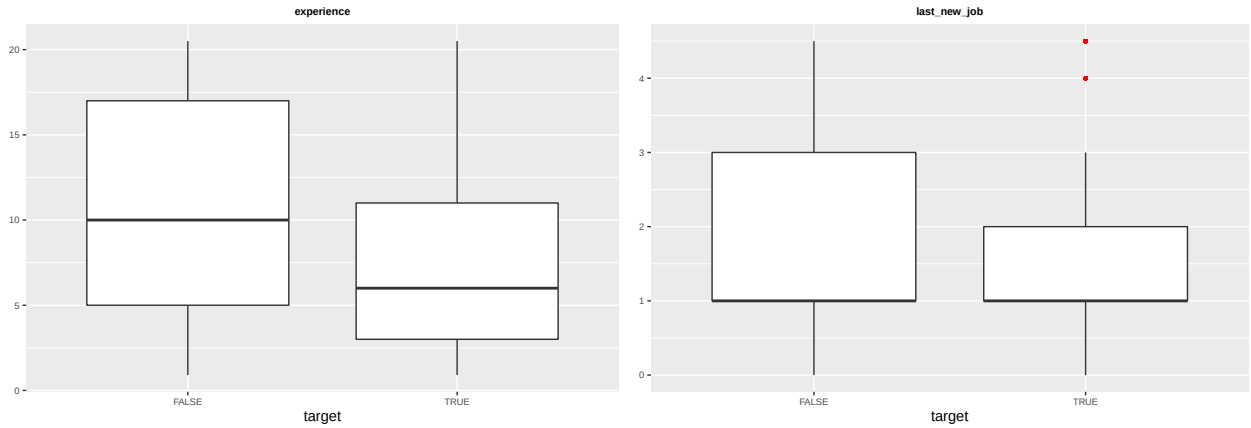
```
## [757] "767"   "22455" "5854"  "12089" "15080" "29331" "13028" "32342" "18206"
## [766] "20556" "20882" "4218"  "26394" "17864" "12644" "19231" "26839" "13723"
## [775] "11383" "25118" "1013"  "1215"  "14769" "6564"  "26624" "17016" "10106"
## [784] "17636" "18773" "10600" "28903" "31392" "18899" "31454" "10223" "13574"
## [793] "2771"  "25597" "6252"  "3753"  "21048" "32935" "502"   "11246" "18029"
## [802] "32582" "9744"  "4454"  "8288"  "29593" "8381"  "8627"  "13052" "6667"
## [811] "18394" "22306" "11112" "22681" "28887" "13198" "26588" "31968" "3473"
## [820] "23127" "16179" "22864" "19915" "7635"  "5933"  "26257" "5063"  "11954"
## [829] "22807" "14245" "25718" "33301" "2304"  "4821"  "31956" "27200" "6436"
## [838] "17534" "33147" "23182" "10695" "29954" "13911" "10265" "20600" "6251"
## [847] "2007"  "14328" "5183"  "6485"  "2867"  "14100" "28795" "21924" "12787"
## [856] "3112"  "5949"  "21635" "23769" "2468"  "29437" "26858" "844"   "1797"
## [865] "28266" "22521" "3726"  "18196" "28071" "31205" "15738" "7829"  "11213"
## [874] "6215"  "10156" "6937"  "29912" "2186"  "2235"  "12559" "18499" "30424"
## [883] "18315" "15830" "12662" "7355"  "7686"  "1345"  "612"   "28439" "5421"
## [892] "13045" "14103" "32566" "31009" "10189" "32507" "1430"  "33005" "21436"
## [901] "7598"  "21589" "28625" "19297" "12690" "10005" "3102"  "7640"  "24529"
## [910] "24390" "21014" "17693" "14612" "3641"  "19044" "8097"  "17399" "26531"
## [919] "10063" "5652"  "6901"  "24095" "24135" "10764" "31044" "25856" "2651"
## [928] "28895" "9357"  "12159" "24672" "29208" "11562" "17220" "32301" "26084"
## [937] "21467" "29994" "635"   "4319"  "28711" "13941" "32476" "17004" "27229"
## [946] "6443"  "15471" "1073"  "9219"  "19062" "26869" "32716" "1809"  "22721"
## [955] "5324"  "23888" "9905"  "20672" "15667" "28966" "20406" "9520"  "30612"
## [964] "29365" "15954" "25065" "21456" "4092"  "3662"  "2208"  "6230"  "16416"
## [973] "32072" "27000" "2242"  "6877"  "29181" "6023"  "22375" "16368" "15133"
## [982] "3458"  "12211" "155"
```

```
## [1] "Enrolled_id of outliers in  experience"
## character(0)
```

```
## [1] "Enrolled_id of outliers in  last_new_job"
## character(0)
```

```
outCDI<-boxplot.stats(df6$CDI)$out
df6[(df6$CDI %in% outCDI),c("CDI")] <- NA

outTH<-boxplot.stats(df6$training_hours)$out
df6[(df6$training_hours %in% outTH),c("training_hours")] <- NA
```






```
summary(df6)
```

```
##  enrollee_id          target             CDI         training_hours
##  Length:19158       Mode :logical   Min.   :0.4790   Min.   :  1.0
##  Class :character   FALSE:14381     1st Qu.:0.7400   1st Qu.: 22.0
##  Mode  :character   TRUE :4777      Median :0.9030   Median : 45.0
##                                     Mean   :0.8292   Mean   : 55.5
##                                     3rd Qu.:0.9200   3rd Qu.: 80.0
##                                     Max.   :0.9490   Max.   :184.0
##                                     NA's   :17       NA's   :984
##    experience     last_new_job      gender         education_level
##  Min.   : 0.90   Min.   :0.000         :    0   School     : 2325
##  1st Qu.: 4.00   1st Qu.:1.000   Female: 1238   Grad       :11598
##  Median : 9.00   Median :1.000   Male  :13221   Specialized: 4775
##  Mean   :10.04   Mean   :1.913   Other :  191   NA's       :  460
##  3rd Qu.:16.00   3rd Qu.:3.000   NA's  : 4508
##  Max.   :20.50   Max.   :4.500
##  NA's   :65      NA's   :423
##   enrolled_university major_discipline    relevent_experience  company_type
##  NotEnrolled:13826    NoMajor:  223    NoRelevantExp:13792    Private:9817
##  Enrolled   : 4955    NoSTEM : 1630    RelevantExp  : 5366    Startup:1604
##  NA's       :  377    STEM   :14492                           Other  :1597
##                       NA's   : 2813                           NA's   :6140
##
##
##
##     compSizeCat          CityQ
##  Less100  :3305   CityGroup1:4790
##  50to500  :3305   CityGroup2:4790
##  100to5000:3305   CityGroup3:4789
##  More1000 :3305   CityGroup4:4789
##  NA's     :5938
##
##
```

```
str(df6)
```

```
## 'data.frame':    19158 obs. of  14 variables:
```

```
##  $ enrollee_id        : chr  "8949" "29725" "11561" "33241" ...
##  $ target             : logi  TRUE FALSE FALSE TRUE FALSE TRUE ...
##  $ CDI                : num  0.92 0.776 0.624 0.789 0.767 0.764 0.92 0.762 0.92 0.92 ...
##  $ training_hours     : num  36 47 83 52 8 24 24 18 46 123 ...
##  $ experience         : num  20.5 15 5 0.9 20.5 11 5 13 7 17 ...
##  $ last_new_job       : num  1 4.5 0 0 4 1 1 4.5 1 4.5 ...
##  $ gender             : Factor w/ 4 levels "","Female","Male",..: 3 3 NA NA 3 NA 3 3 3 NA ...
##  $ education_level    : Factor w/ 3 levels "School","Grad",..: 2 2 2 2 3 2 1 2 2 2 ...
##  $ enrolled_university: Factor w/ 2 levels "NotEnrolled",..: 1 1 2 NA 1 2 1 1 1 1 ...
##  $ major_discipline   : Factor w/ 3 levels "NoMajor","NoSTEM",..: 3 3 3 2 3 3 NA 3 3 3 ...
##  $ relevent_experience: Factor w/ 2 levels "NoRelevantExp",..: 1 2 2 2 1 1 1 1 1 1 ...
##  $ company_type       : Factor w/ 3 levels "Private","Startup",..: NA 1 NA 1 2 NA 2 1 1 1 ...
##  $ compSizeCat        : Factor w/ 4 levels "Less100","50to500",..: NA 1 NA NA 1 NA 1 1 1 4 ...
##  $ CityQ              : Factor w/ 4 levels "CityGroup1","CityGroup2",..: 1 4 3 2 3 3 3 4 1 1 ...
```

Imputing the corresponding values:

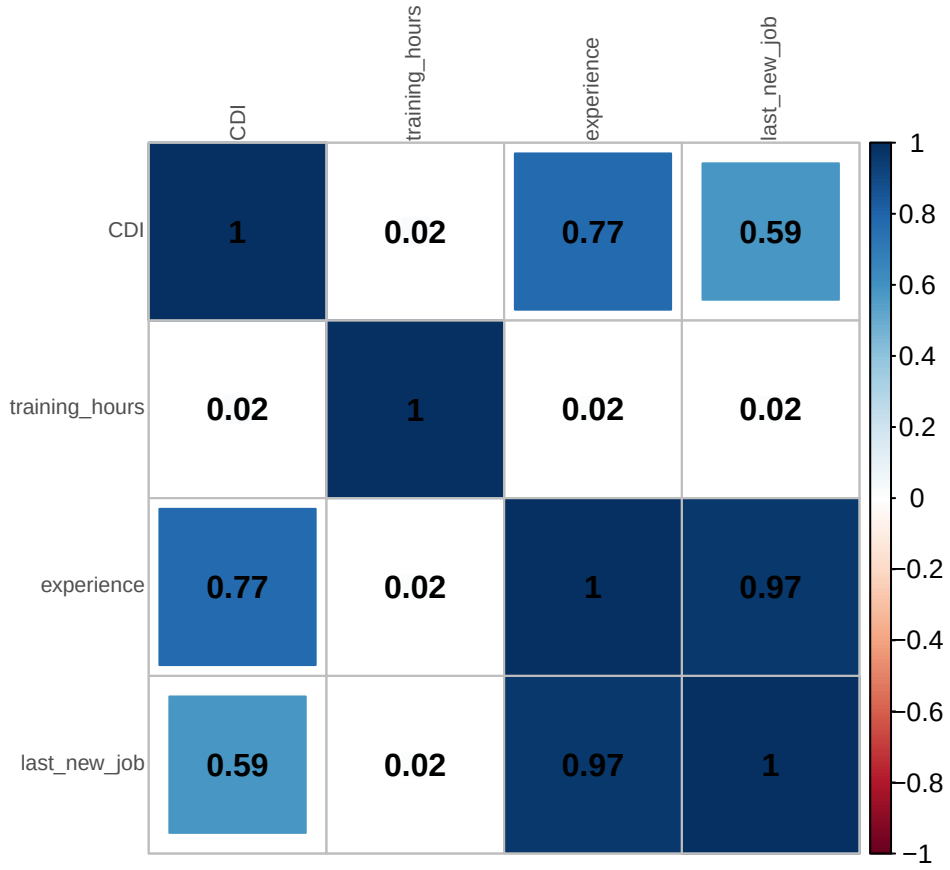
```
numv<-df6[,3:6]
nbdimn0 <- estim_ncpPCA(numv)
imputedNum<-imputePCA(numv, ncp = 3)

catv<-df6[,7:14]
#nbdimn <- estim_ncpMCA(catv)
imputedCat<-imputeMCA(catv, ncp = 6)

names(df6)[1:6]
```

```
## [1] "enrollee_id"    "target"         "CDI"            "training_hours"
## [5] "experience"     "last_new_job"
```

```
df7<-as.data.frame(cbind(df6$enrollee_id,df6$target,imputedNum$fittedX,imputedCat$completeObs))
names(df7)[1:6]<-names(df6)[1:6]
```

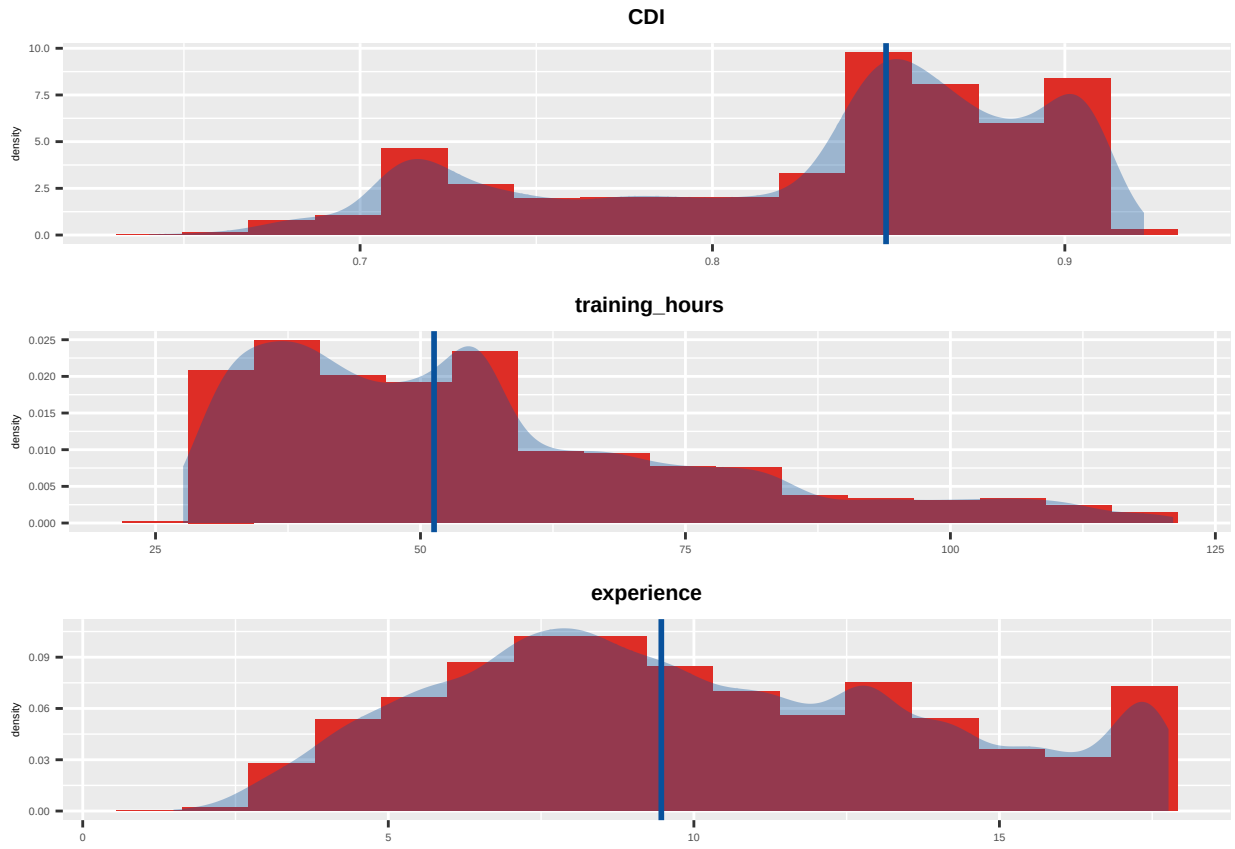
Correlation matrix and histograms are calculated for the numeric variables. The last_new_job variable it is removed from the data because its high correlation with experience. Histograms depict the distribution for the numeric variables.

```
# Correlation between numeric variables

corrplot(cor(df7[3:6]),cex.main=0.7,method = c("square"),
         number.cex = 1,tl.cex=0.7,tl.col="gray31",
         cl.align="c",tl.offset = 0.1,addCoef.col=TRUE)
```



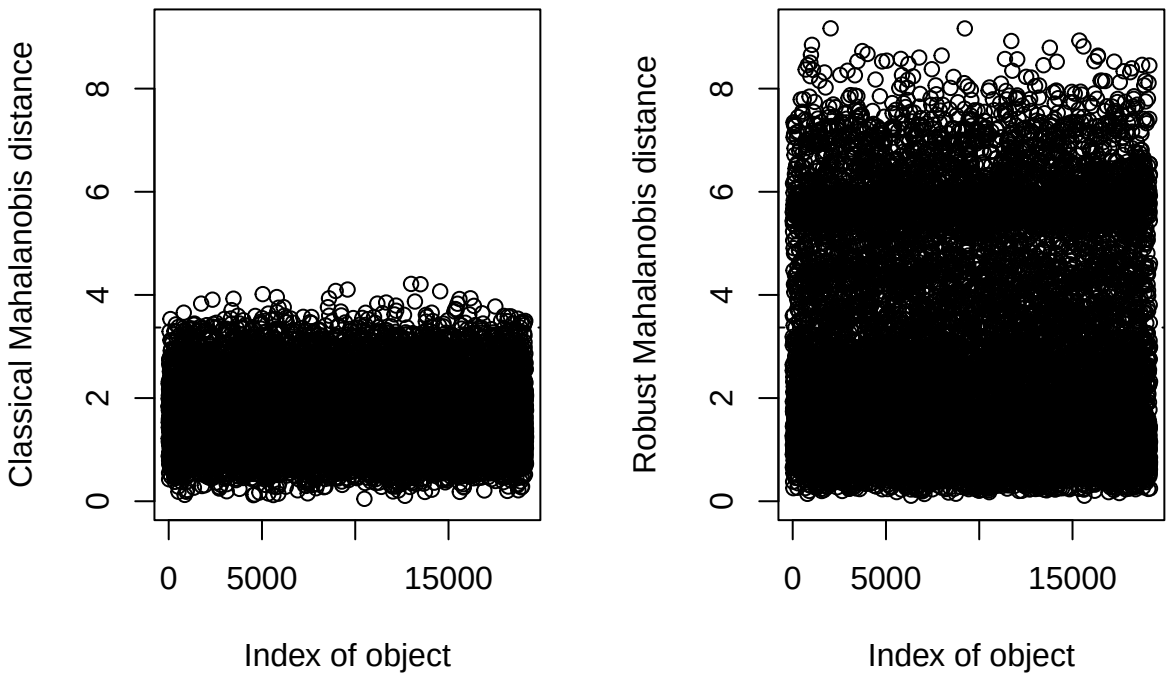

```
# Removing last_new_job variable for high correlation with experience
df8<-df7[,c("target","CDI","training_hours","experience","gender",
            "education_level","enrolled_university","major_discipline"
            ,"relevent_experience","company_type","compSizeCat" )]


numvars<-df8[2:4]
a<-names(numvars)
a<-as.list(a)
fun02<-function(i){
                  bw <- nclass.Sturges(numvars[,i]) # Freedman-Diaconis
                  nm=a[i]
                  assign(paste("g",i,sep=""),
                  ggplot(numvars, aes(numvars[,i])) +
                  geom_histogram(bins = bw,aes(y=..density..), fill="#de2d26") +
                  geom_density(alpha=.35, fill="#08519c",color = NA)  +
                  geom_vline (aes(xintercept=median(numvars[,i])),
           color="#08519c", size=1) +
                  labs(title=nm, x=NULL)) +
                  theme(plot.title = element_text(size = rel(0.7),face ="bold",
                                                  hjust = 0.5),
                       axis.title.y = element_text(size = rel(0.4)),
                       axis.text = element_text(size = rel(0.4)))
                  }
Histos<-lapply(1:length(a),fun02)
#Histos<-lapply(1,fun02)
```

```
do.call(grid.arrange, Histos)
```

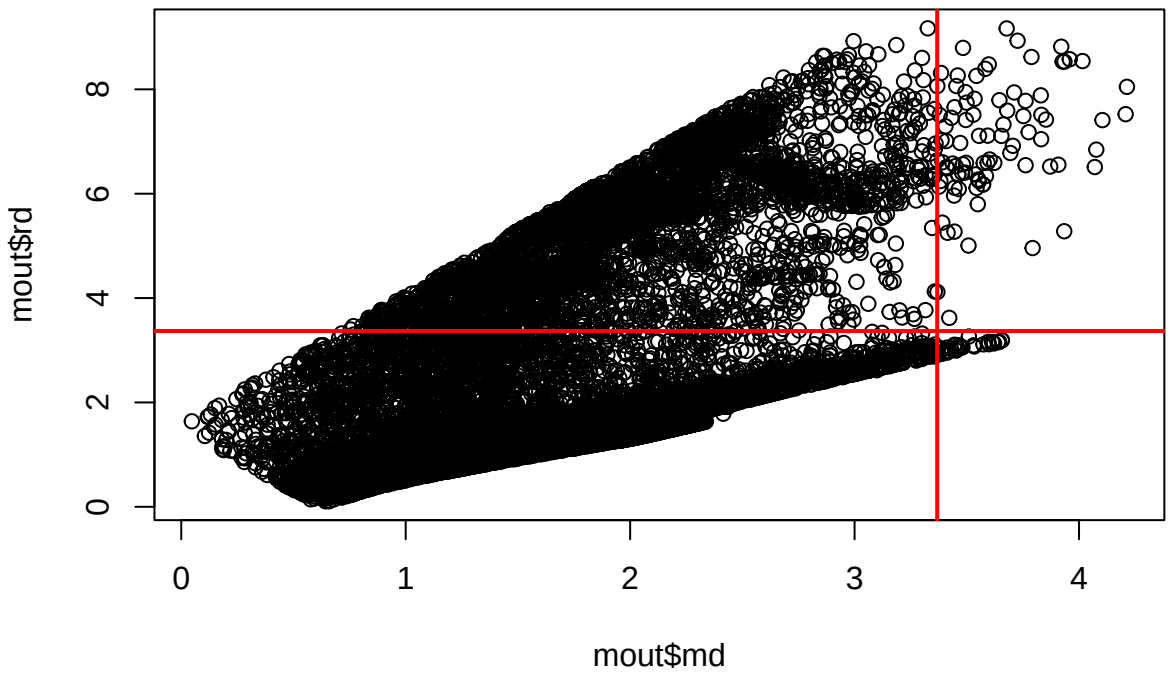


Multivariate outliers are calculated and identified.

```
mout <- Moutlier(numvars, quantile = 0.99, plot=T )
```




```
par(mfrow=c(1,1))
plot(mout$md, mout$rd)
abline( h=mout$cutoff, lwd=2, col="red")
abline( v=mout$cutoff, lwd=2, col="red")
```




```
llmout <- which((mout$md>mout$cutoff) & (mout$rd > mout$cutoff) )
llmout
```

```
##  [1]   523  1704  1739  2042  2330  2347  2395  2549  2994  3334  3370  3373
## [13]  3473  3812  4639  4898  4915  5012  5041  5180  5551  5795  5810  5891
## [25]  6167  6204  6254  6281  6533  6662  7209  7654  7664  8537  8595  8763
## [37]  8952  9106  9226  9253  9590  9890 10174 10233 10386 10555 11008 11055
## [49] 11186 11244 11470 11658 11930 12187 12195 12772 12994 13201 13499 13785
## [61] 14092 14158 14497 14553 14725 14741 14958 15320 15369 15432 15486 15587
## [73] 15699 16116 16177 16362 16428 16851 17509 17819 17965 18144 18181 18338
## [85] 18450 18715 18849 19087 19115
```

```
numvars[llmout,]
```

```
##             CDI training_hours experience
## 523   0.7257271       34.84057  12.353099
## 1704  0.6961259       35.80252  10.308560
## 1739  0.7407783      115.57142  11.231267
## 2042  0.7525298      105.40946  11.792571
## 2330  0.6856053      108.08976   2.429707
## 2347  0.7665717      112.56449  13.671178
## 2395  0.7080475      115.30718   4.194401
## 2549  0.7043389       45.45045  11.428671
## 2994  0.7220625       40.40686  12.750278
## 3334  0.7093944       65.24487  12.006254
```

```
## 3370  0.7190453      117.29089   4.101990
## 3373  0.7439332      111.51677  11.535383
## 3473  0.7121897       49.50529  13.372978
## 3812  0.7805404      120.64204  10.867395
## 4639  0.7567440      105.44654  12.754255
## 4898  0.7289298       33.32301  12.657817
## 4915  0.7631017       99.87433  13.363089
## 5012  0.7819490      108.54555  13.756474
## 5041  0.7119510       36.81982  13.369968
## 5180  0.7378296      115.47428   9.261453
## 5551  0.7330681       81.53247  12.974333
## 5795  0.7125238       67.26494  13.377191
## 5810  0.6911862      117.28956   4.006415
## 5891  0.7242105      110.85622  10.401343
## 6167  0.7240359       99.28193  11.832771
## 6204  0.6970901       87.05180  10.320720
## 6254  0.7299964      109.94335   9.658100
## 6281  0.7572177      115.50247  11.091928
## 6533  0.7223339      120.34010   4.407791
## 6662  0.7210301      118.37049   5.415440
## 7209  0.7113552      119.37123   4.500443
## 7654  0.7191025      120.33540   4.102712
## 7664  0.7090799       37.34356  11.529238
## 8537  0.7632735      109.00786  13.365256
## 8595  0.8109172      118.74943  15.429901
## 8763  0.7077852      116.31275   4.006637
## 8952  0.7570113      119.65426  12.757626
## 9106  0.7267364      108.41639   9.352660
## 9226  0.6781747       44.88920  10.168563
## 9253  0.7081239      119.36653   4.195364
## 9590  0.7433206      112.51146  12.946808
## 9890  0.7126017      118.29633   3.492072
## 10174 0.7757940      119.46368   4.815409
## 10233 0.8579601      112.77426  16.416120
## 10386 0.7309602      108.96089  10.314464
## 10555 0.7042608      119.34169   5.301468
## 11008 0.7191025      120.33540   4.102712
## 11055 0.7210301      118.37049   5.415440
## 11186 0.7127130      119.40348   8.174943
## 11244 0.7726525       92.27714  14.276521
## 11470 0.7296546       55.63620  13.737808
## 11658 0.7445782       92.20471  14.044639
## 11930 0.7335798      113.28648   8.765305
## 12187 0.8157595      119.78368  14.937654
## 12195 0.7503673      110.00391  12.145180
## 12772 0.7564201       35.91239  14.301486
## 12994 0.7239564      112.47092  12.066439
## 13201 0.7624338      116.58676  12.710457
## 13499 0.7470890      107.43263  14.094838
## 13785 0.6782114       30.69773   8.984546
## 14092 0.7092749      113.21745   3.185789
## 14158 0.7081345       95.01106  10.100790
## 14497 0.6875996      109.67678   3.743277
## 14553 0.7730916      115.61840  14.282059
```

```
## 14725 0.7711349      119.45813  10.144408
## 14741 0.7436876      120.43304   7.551476
## 14958 0.8106880      106.57139  15.427012
## 15320 0.6855666      116.18596   4.113181
## 15369 0.6498233      105.51997   3.452863
## 15432 0.7162195       98.25448  11.194419
## 15486 0.6726604      117.18201   2.893105
## 15587 0.6447383      111.53945   1.834378
## 15699 0.7112788      115.31188   4.499480
## 16116 0.7411183      118.52349   9.567254
## 16177 0.7121515       47.47561  13.372496
## 16362 0.6805476      107.30282   7.412093
## 16428 0.7961852      108.57992  14.054637
## 16851 0.6814648      108.64066   4.113493
## 17509 0.7503617       92.21867  14.165767
## 17819 0.6724885      108.04847   2.890938
## 17965 0.7456475       30.81243  14.076660
## 18144 0.7265558      119.42064   6.433059
## 18181 0.6612802      106.05805   3.453744
## 18338 0.7093704      118.29163   3.186993
## 18450 0.7112979      116.32672   4.499721
## 18715 0.7061008      116.25726   2.881432
## 18849 0.6678771      102.01481   3.592544
## 19087 0.7372250      120.42365   6.941317
## 19115 0.7144708      108.42951   8.758440
```

```
mout$md[llmout]
```

```
##  [1] 3.379191 3.385237 3.831259 3.414353 3.387619 3.908005 3.437029 3.459558
##  [9] 3.534696 3.543208 3.445494 3.705280 3.927034 3.445245 3.592582 3.431444
## [17] 3.479344 3.549000 4.015742 3.496521 3.529220 3.958804 3.663471 3.752236
## [25] 3.761932 3.597857 3.470573 3.554412 3.543589 3.461998 3.575274 3.572467
## [33] 3.453444 3.762024 3.934058 3.482679 4.076931 3.408484 3.677701 3.602003
## [41] 4.104368 3.551341 3.369052 3.421939 3.554547 3.654209 3.572467 3.461998
## [49] 3.832434 3.375910 3.645348 3.852705 3.391934 3.793282 3.694088 3.368905
## [57] 4.213600 3.871622 4.207769 3.483187 3.382168 3.494356 3.404911 4.070259
## [65] 3.391037 3.443257 3.507668 3.680109 3.726291 3.710977 3.829607 3.920782
## [73] 3.409064 3.622606 3.935079 3.788173 3.414466 3.445600 3.775602 3.498990
## [81] 3.593086 3.492450 3.582708 3.585387 3.450343 3.539750 3.368601 3.464282
## [89] 3.495680
```

```
numvars$mout <- 0
numvars$mout[ llmout ] <- 1
numvars$mout <- factor( numvars$mout, labels = c("MvOut.No","MvOut.Yes"))
```

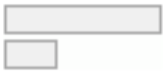
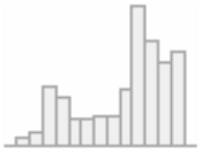
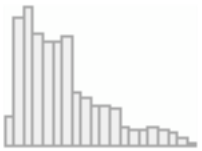
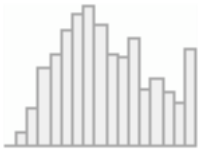
Setting the sample and splitting the dataset

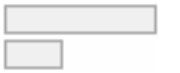
```
df9<-df8
#str(df9)
df9$target<-as.logical(df9$target)
```

```
rownames(df9)<-df8$enrollee_id


# Random selection of x registers:
sam<-as.vector(sort(sample(1:nrow(df9),5000)))
df<-df9[sam,] # Subset of rows _ It will be my sample

llwork <- sample(1:nrow(df),round(0.75*nrow(df),0))

dfTR<-df[llwork,]
dfTS<-df[-llwork,]

dfSummary(df)
```

Initial data frame summary in PDF Format

df
Dimensions: 5000 x 11
Duplicates: 13

No	Variable	Stats / Values	Freqs (% of Valid)	Graph	Missing
1	target [logical]	1. FALSE 2. TRUE	3741 (74.8%) 1259 (25.2%)		0 (0.0%)
2	CDI [numeric]	Mean (sd) : 0.8 (0.1) min < med < max: 0.6 < 0.8 < 0.9 IQR (CV) : 0.1 (0.1)	4679 distinct values		0 (0.0%)
3	training_hours [numeric]	Mean (sd) : 55.2 (20.9) min < med < max: 27.9 < 50.7 < 120.9 IQR (CV) : 27.1 (0.4)	4679 distinct values		0 (0.0%)
4	experience [numeric]	Mean (sd) : 10 (4) min < med < max: 1.9 < 9.5 < 17.8 IQR (CV) : 5.9 (0.4)	4679 distinct values		0 (0.0%)
5	gender [factor]	1. Female 2. Male 3. Other	328 (6.6%) 4623 (92.5%) 49 (1.0%)		0 (0.0%)
6	education_level [factor]	1. School 2. Grad 3. Specialized	602 (12.0%) 3146 (62.9%) 1252 (25.0%)		0 (0.0%)
7	enrolled_university [factor]	1. NotEnrolled 2. Enrolled	3666 (73.3%) 1334 (26.7%)		0 (0.0%)

No	Variable	Stats / Values	Freqs (% of Valid)	Graph	Missing
8	major_discipline [factor]	1. NoMajor 2. NoSTEM 3. STEM	44 (0.9%) 452 (9.0%) 4504 (90.1%)		0 (0.0%)
9	relevent_experience [factor]	1. NoRelevantExp 2. RelevantExp	3598 (72.0%) 1402 (28.0%)		0 (0.0%)
10	company_type [factor]	1. Private 2. Startup 3. Other	4160 (83.2%) 427 (8.5%) 413 (8.3%)		0 (0.0%)
11	compSizeCat [factor]	1. Less100 2. 50to500 3. 100to5000 4. More1000	1422 (28.4%) 1097 (21.9%) 978 (19.6%) 1503 (30.1%)		0 (0.0%)

Profiling and feature selection

the target is profiled:

```
profile<-catdes(dfTR, num.var=1, prob = 0.01)
profile
```

```
##
## Link between the cluster variable and the categorical variables (chi-square test)
## ================================================================================
##                            p.value df
## enrolled_university 6.708606e-21  1
## relevent_experience 4.480654e-15  1
## compSizeCat         1.863996e-12  3
## education_level     8.873017e-08  2
## company_type        9.028976e-07  2
##
## Description of each cluster by the categories
## =============================================
## $`FALSE`
##                                      Cla/Mod   Mod/Cla    Global      p.value
## enrolled_university=NotEnrolled   78.44137 77.204301 73.226667 5.119072e-20
## relevent_experience=NoRelevantExp 77.91000 75.089606 71.706667 1.414419e-14
## compSizeCat=100to5000             81.09244 20.752688 19.040000 3.160139e-06
## company_type=Startup              84.13174 10.071685  8.906667 8.578431e-06
## education_level=Specialized       79.70244 26.881720 25.093333 1.227194e-05
## compSizeCat=50to500               78.06604 23.727599 22.613333 4.996516e-03
## education_level=School            79.60526 13.010753 12.160000 5.717425e-03
## company_type=Other                80.69620  9.139785  8.426667 6.171421e-03
## company_type=Private              72.70968 80.788530 82.666667 9.390459e-08
## education_level=Grad              71.27072 60.107527 62.746667 8.567288e-09
## compSizeCat=More1000              66.75532 26.989247 30.080000 4.281917e-12
## relevent_experience=RelevantExp   65.50424 24.910394 28.293333 1.414419e-14
## enrolled_university=Enrolled      63.34661 22.795699 26.773333 5.119072e-20
```

```
##                                      v.test
## enrolled_university=NotEnrolled    9.161479
## relevent_experience=NoRelevantExp  7.695047
## compSizeCat=100to5000              4.660127
## company_type=Startup               4.450215
## education_level=Specialized        4.372700
## compSizeCat=50to500                2.807258
## education_level=School             2.763562
## company_type=Other                 2.738532
## company_type=Private              -5.338141
## education_level=Grad              -5.756899
## compSizeCat=More1000              -6.927550
## relevent_experience=RelevantExp   -7.695047
## enrolled_university=Enrolled      -9.161479
##
## $`TRUE`
##                                       Cla/Mod   Mod/Cla    Global      p.value
## enrolled_university=Enrolled      36.65339 38.333333 26.773333 5.119072e-20
## relevent_experience=RelevantExp   34.49576 38.125000 28.293333 1.414419e-14
## compSizeCat=More1000              33.24468 39.062500 30.080000 4.281917e-12
## education_level=Grad              28.72928 70.416667 62.746667 8.567288e-09
## company_type=Private              27.29032 88.125000 82.666667 9.390459e-08
## company_type=Other                19.30380  6.354167  8.426667 6.171421e-03
## education_level=School            20.39474  9.687500 12.160000 5.717425e-03
## compSizeCat=50to500               21.93396 19.375000 22.613333 4.996516e-03
## education_level=Specialized       20.29756 19.895833 25.093333 1.227194e-05
## company_type=Startup              15.86826  5.520833  8.906667 8.578431e-06
## compSizeCat=100to5000             18.90756 14.062500 19.040000 3.160139e-06
## relevent_experience=NoRelevantExp 22.09000 61.875000 71.706667 1.414419e-14
## enrolled_university=NotEnrolled   21.55863 61.666667 73.226667 5.119072e-20
##                                      v.test
## enrolled_university=Enrolled       9.161479
## relevent_experience=RelevantExp    7.695047
## compSizeCat=More1000               6.927550
## education_level=Grad               5.756899
## company_type=Private               5.338141
## company_type=Other                -2.738532
## education_level=School            -2.763562
## compSizeCat=50to500               -2.807258
## education_level=Specialized       -4.372700
## company_type=Startup              -4.450215
## compSizeCat=100to5000             -4.660127
## relevent_experience=NoRelevantExp -7.695047
## enrolled_university=NotEnrolled   -9.161479
##
##
## Link between the cluster variable and the quantitative variables
## ================================================================
##                   Eta2       P-value
## CDI        0.12846941 4.461523e-114
## experience 0.06042481  9.915811e-53
##
## Description of each cluster by quantitative variables
## =====================================================
```

```
## $`FALSE`
##               v.test Mean in category Overall mean sd in category Overall sd
## CDI         21.94611        0.8422598    0.8281231     0.05803206 0.06723793
## experience 15.05100       10.6146897   10.0413007     3.81479256 3.97656589
##                    p.value
## CDI         9.433246e-107
## experience   3.400530e-51
##
## $`TRUE`
##                 v.test Mean in category Overall mean sd in category Overall sd
## experience -15.05100        8.3748890   10.0413007     3.96784357 3.97656589
## CDI        -21.94611        0.7870383    0.8281231     0.07485801 0.06723793
##                    p.value
## experience   3.400530e-51
## CDI         9.433246e-107
```

From this analysis it can be concluded that the link between each variable and the response variable are significant. Are highlighted the variables: CDI, experience and enrolled university.

The characteristics of the average candidate in the TRUE target condition, looking for a new job or will work for the company is coming from a city with a city development index of 0.79, with 8.37 years. Those candidates inf FALSE target condition, coming from a city with a city development index of 0.84, with 10.61 years of experience and a difference of 1.98 years between previous and current job.

The category associated to the enrolled in university program is over represented among the true target and not Enrolled is under represented. On the opposite, false cases have over represented Not enrolled to a university and Enrolled under represented.

Modeling using numeric variables using transformations if needed

```
# Y Target No variables
m0 <- glm( target ~ 1, family="binomial", data = dfTR)
ptt<-prop.table(table(dfTR$target));ptt
```

```
##
## FALSE  TRUE
## 0.744 0.256
```

```
summary(m0)
```

```
##
## Call:
## glm(formula = target ~ 1, family = "binomial", data = dfTR)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -0.769  -0.769  -0.769   1.651   1.651
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.06686    0.03742  -28.51   <2e-16 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4266.2  on 3749  degrees of freedom
## Residual deviance: 4266.2  on 3749  degrees of freedom
## AIC: 4268.2
##
## Number of Fisher Scoring iterations: 4
```

```
oddm0<-ptt[2]/ptt[1];oddm0
```

```
##     TRUE
## 0.344086
```

```
logoddm0 <- log( oddm0 ); logoddm0
```

```
##      TRUE
## -1.066864
```

Candidates looking for a new job or with a probability to be hired is positive in 25.6% of the cases. The odds of the positive case is 0.34 and its marginal probability is -1.07.

```
# Y Target numeric variables
m1 <- glm( target ~ CDI + training_hours + experience, family="binomial", data = dfTR)
summary(m1)
```

```
##
## Call:
## glm(formula = target ~ CDI + training_hours + experience, family = "binomial",
##     data = dfTR)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6157  -0.6605  -0.5617   0.9234   2.1865
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)      9.592090   0.624185  15.367   <2e-16 ***
## CDI            -13.197694   0.890449 -14.821   <2e-16 ***
## training_hours  -0.002859   0.001918  -1.491   0.1360
## experience       0.028765   0.016120   1.784   0.0743 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4266.2  on 3749  degrees of freedom
## Residual deviance: 3793.9  on 3746  degrees of freedom
## AIC: 3801.9
##
## Number of Fisher Scoring iterations: 4
```

```
#m1a <- glm( target ~ logCDI + logTH + logexp, family="binomial", data = dfTR)
#summary(m1a)

ptt1<-prop.table(table(dfTR$CDI,dfTR$target))
oddm1<-ptt1[,2]/ptt1[,1]
logoddm1 <- log( oddm1 )
```

For this initial model the logit transformation for the probability of the candidates looking for a job or to be hired for true cases is 9.59. As well, the logit probability for TRUE cases decreases in 13.19 units if City Development Index (CDI) increases, decreases in 0.002 units per one additional hour of training and increases 0.02 units for each increment of one year of experience.

The residual deviance is 3801.9.

```
anova(m0, m1, test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: target ~ 1
## Model 2: target ~ CDI + training_hours + experience
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1      3749     4266.2
## 2      3746     3793.9  3   472.31 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Performing the comparison of the deviance of the models using a Chi-squared test, it can be concluded that the model using the CDI variable and the model using just the intercept with a significance or 0.01 are not equivalent.

```
vif(m1)
```

```
##            CDI training_hours     experience
##       2.383142       1.000037       2.383183
```

```
m2<-step(m1, k=log(nrow(dfTR)))
```

```
## Start:  AIC=3826.84
## target ~ CDI + training_hours + experience
##
##                  Df Deviance    AIC
## - training_hours  1   3796.2 3820.9
## - experience      1   3797.1 3821.8
## <none>                3793.9 3826.8
## - CDI             1   4023.3 4048.0
##
## Step:  AIC=3820.86
## target ~ CDI + experience
##
##              Df Deviance    AIC
## - experience  1   3799.4 3815.8
```

```
## <none>            3796.2 3820.9
## - CDI         1   4026.1 4042.5
##
## Step:  AIC=3815.83
## target ~ CDI
##
##        Df Deviance    AIC
## <none>      3799.4 3815.8
## - CDI   1   4266.2 4274.5
```

```
summary(m2)
```

```
##
## Call:
## glm(formula = target ~ CDI, family = "binomial", data = dfTR)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5851  -0.6794  -0.5636   0.9293   2.1781
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   8.7333     0.4690   18.62   <2e-16 ***
## CDI         -12.0051     0.5795  -20.72   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4266.2  on 3749  degrees of freedom
## Residual deviance: 3799.4  on 3748  degrees of freedom
## AIC: 3803.4
##
## Number of Fisher Scoring iterations: 4
```

The stepwise method is performed to establish the best numeric model. In this case, after the comparison of AIC criteria, the model just includes the CDI variable. Some

```
m3 <- glm( target ~ poly(CDI,2), family="binomial", data = dfTR)
summary(m3)
```

```
##
## Call:
## glm(formula = target ~ poly(CDI, 2), family = "binomial", data = dfTR)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9731  -0.6295  -0.5722   0.7466   1.9847
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -1.18843    0.04176 -28.457  < 2e-16 ***
```

```
## poly(CDI, 2)1 -47.31111    2.37896 -19.887  < 2e-16 ***
## poly(CDI, 2)2  12.29448    2.51035   4.898 9.71e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4266.2  on 3749  degrees of freedom
## Residual deviance: 3775.2  on 3747  degrees of freedom
## AIC: 3781.2
##
## Number of Fisher Scoring iterations: 4
```

The polytomic variable of order two is significant, explaining that CDI in quadratic form is better than linear approach. However due to the high number of categorical variables and the complexity of the interpretation of this parameter it is not considered for the main effect model.

Residual analysis: unusual and influent data filtering

Residuals of both models (not constant) considered, are performed in order to check the effect of the transformation of CDI index. To observe the residual data are used the plots of the model using the CDI variable, the marginal plots and the influence plot. Cook's Distance is considered for m2 only.

```
marginalModelPlots(m2)
marginalModelPlots(m3)
influencePlot(m2)
```



```
##          StudRes          Hat        CookD
## 3733   0.7832468 0.0020759141 0.0003736344
## 19072  2.1791593 0.0004829220 0.0023490328
## 12036  0.8099366 0.0020094759 0.0003910522
## 2607   2.1777238 0.0004824738 0.0023387956
## 944   -1.5809645 0.0019738583 0.0024580991
## 14156 -1.5866537 0.0019854752 0.0025038654
```

```
influencePlot(m3)
```



```
##          StudRes         Hat        CookD
## 3733   0.4967104 0.005619886 0.0002479447
## 19072  1.9871812 0.001578257 0.0032549490
## 12036  0.5405115 0.005365584 0.0002834664
## 2607   1.9871570 0.001561210 0.0032196977
## 944   -1.9685510 0.005207079 0.0102372384
## 14156 -1.9811428 0.005260171 0.0106402962
```

```
dist <- cooks.distance(m3)
influential <- Boxplot(dist)
length(influential)
```

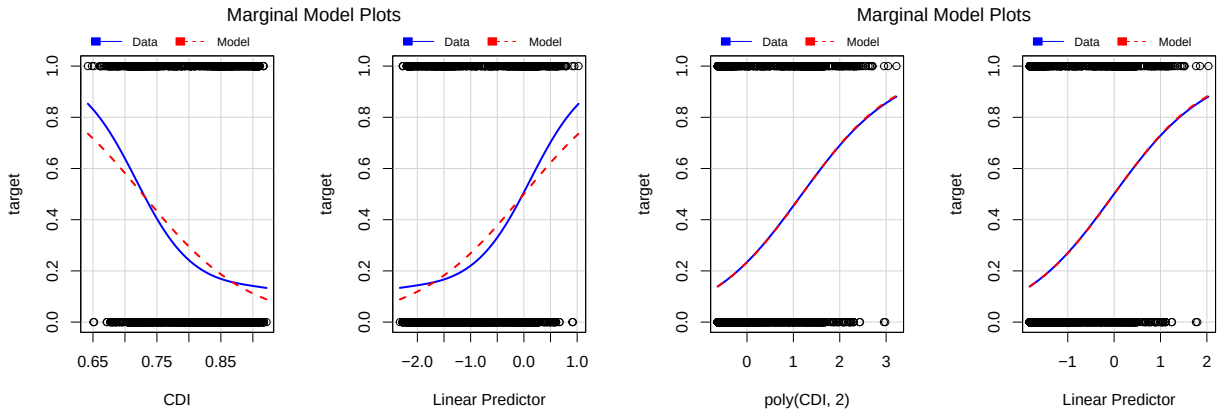
```
## [1] 10
```

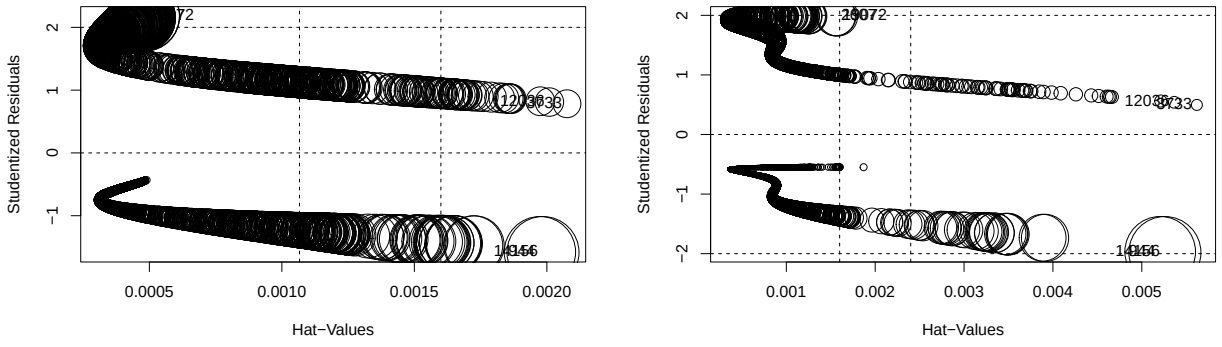
```r
Anova(m2, test="LR")
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: target
##     LR Chisq Df Pr(>Chisq)
## CDI   466.86  1  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

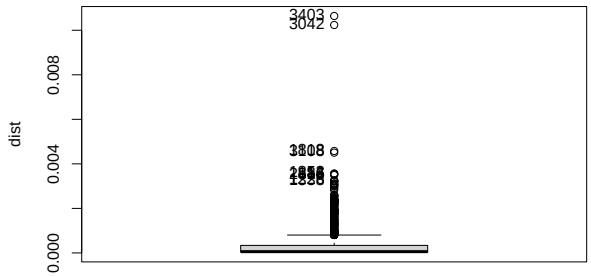
```r
Anova(m3, test="LR")
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: target
##               LR Chisq Df Pr(>Chisq)
## poly(CDI, 2)   491.05  2  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```





The marginal model plots shows difference between



the data and the model, for the lineal model data has a trend to follow a curve instead straight line. For the quadratic model the line fits better over the model one but eventually overfitted. Cook's Distance from model 2 identify 10 influential data which are removed from the dataset.

Adding factor main effects to the best model containing numeric variables

The model with the factor variables and the CDI index is calculated.

```
#Obtaining a model with input CDI variable and categorical variables trough stepwise
#regression
#dfTR[, c(1,2,5:10)]
dfTR<-dfTR[-influential,c(1,2,5:10)]

m4<-glm(target ~ ., family="binomial", data = dfTR)
m5<-step(m4, k=log(nrow(dfTR)), direction="both", data=dfTR)
```

```
## Start:  AIC=3777.23
## target ~ CDI + gender + education_level + enrolled_university +
##     major_discipline + relevent_experience + company_type
##
##                         Df Deviance    AIC
## - gender                 2   3678.7 3760.9
## - major_discipline       2   3679.7 3762.0
## <none>                       3678.5 3777.2
## - enrolled_university    1   3690.7 3781.2
## - company_type           2   3701.8 3784.0
## - relevent_experience    1   3698.7 3789.2
## - education_level        2   3718.7 3801.0
## - CDI                    1   4068.1 4158.6
##
## Step:  AIC=3760.94
## target ~ CDI + education_level + enrolled_university + major_discipline +
##     relevent_experience + company_type
##
##                         Df Deviance    AIC
## - major_discipline       2   3679.9 3745.7
## <none>                       3678.7 3760.9
```

```
## - enrolled_university  1   3690.8 3764.9
## - company_type         2   3701.8 3767.6
## - relevent_experience  1   3699.0 3773.0
## + gender               2   3678.5 3777.2
## - education_level      2   3719.1 3784.9
## - CDI                  1   4068.7 4142.8
##
## Step:  AIC=3745.74
## target ~ CDI + education_level + enrolled_university + relevent_experience +
##     company_type
##
##                       Df Deviance    AIC
## <none>                     3679.9 3745.7
## - enrolled_university  1   3691.5 3749.1
## - company_type         2   3702.9 3752.2
## - relevent_experience  1   3701.6 3759.2
## + major_discipline     2   3678.7 3760.9
## + gender               2   3679.7 3762.0
## - education_level      2   3722.4 3771.8
## - CDI                  1   4071.0 4128.6
```

```
summary(m5)
```

```
##
## Call:
## glm(formula = target ~ CDI + education_level + enrolled_university +
##     relevent_experience + company_type, family = "binomial",
##     data = dfTR)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7473  -0.6863  -0.5219   0.8002   2.6654
##
## Coefficients:
##                                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)                      7.56207    0.51250  14.755  < 2e-16 ***
## CDI                            -11.66186    0.60880 -19.155  < 2e-16 ***
## education_levelGrad              0.88198    0.14215   6.205 5.48e-10 ***
## education_levelSpecialized       0.69817    0.16169   4.318 1.58e-05 ***
## enrolled_universityEnrolled      0.32035    0.09369   3.419 0.000628 ***
## relevent_experienceRelevantExp   0.44817    0.09573   4.681 2.85e-06 ***
## company_typeStartup             -0.68866    0.16630  -4.141 3.46e-05 ***
## company_typeOther               -0.36510    0.15899  -2.296 0.021654 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4258.2  on 3739  degrees of freedom
## Residual deviance: 3679.9  on 3732  degrees of freedom
## AIC: 3695.9
##
## Number of Fisher Scoring iterations: 4
```

The optimal model with a AIC of 3695.9 considers the CDI, education level, enrolled university, relevant experience and company types to model the probability of a candidate to be hired. The levels of reference of the variables are: School (education level), Not Enrolled (Enrolled University), No Relevant Experience (Relevant experience) and private company.

Residual analysis: unusual and influent data filtering.

```
marginalModelPlots(m5)
influencePlot(m5)
```

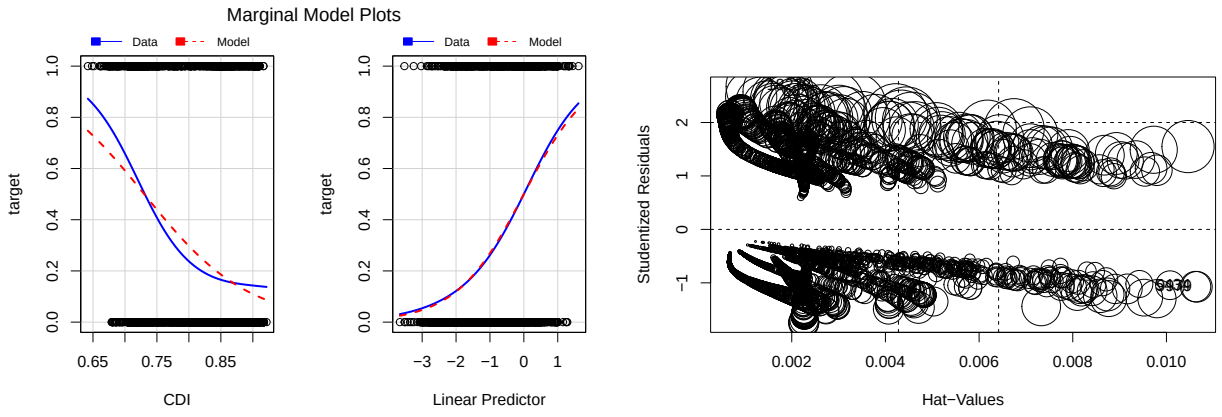
```
##         StudRes         Hat       CookD
## 18233  2.674113 0.001366506 0.005805026
## 5434  -1.070468 0.010621333 0.001039236
## 6856   2.572825 0.001708499 0.005524490
## 9979  -1.074490 0.010649857 0.001052346
```

```
dist <- cooks.distance(m5)
inf2 <- Boxplot(dist)
length(inf2)
```

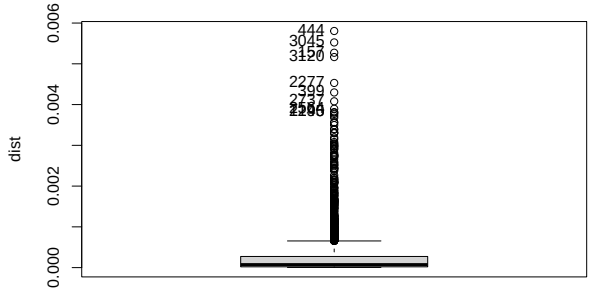
```
## [1] 10
```

```
Anova(m5, test="LR")
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: target
##                       LR Chisq Df Pr(>Chisq)
## CDI                     391.10  1  < 2.2e-16 ***
## education_level          42.51  2  5.870e-10 ***
## enrolled_university      11.57  1   0.000669 ***
## relevent_experience      21.69  1  3.205e-06 ***
## company_type             22.96  2  1.036e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```



For this model,



Adding factor main effects and interactions (limit your statement to order 2) to the best

To characterize the main effects and interactions are evaluated with Anova method evaluating interactions between CDI and each of the selected categorical values. As well the corresponding chi-squared test and the AIC.

```
#Main effects

summary(m5)
```

```
##
## Call:
## glm(formula = target ~ CDI + education_level + enrolled_university +
##     relevent_experience + company_type, family = "binomial",
##     data = dfTR)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7473  -0.6863  -0.5219   0.8002   2.6654
##
## Coefficients:
##                                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)                      7.56207    0.51250  14.755  < 2e-16 ***
## CDI                            -11.66186    0.60880 -19.155  < 2e-16 ***
## education_levelGrad              0.88198    0.14215   6.205 5.48e-10 ***
## education_levelSpecialized       0.69817    0.16169   4.318 1.58e-05 ***
## enrolled_universityEnrolled      0.32035    0.09369   3.419 0.000628 ***
## relevent_experienceRelevantExp   0.44817    0.09573   4.681 2.85e-06 ***
## company_typeStartup             -0.68866    0.16630  -4.141 3.46e-05 ***
## company_typeOther               -0.36510    0.15899  -2.296 0.021654 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4258.2  on 3739  degrees of freedom
```

```
## Residual deviance: 3679.9  on 3732  degrees of freedom
## AIC: 3695.9
##
## Number of Fisher Scoring iterations: 4
```

```
m6<-glm(target ~ (CDI + enrolled_university + relevent_experience +
                    company_type) * education_level, family="binomial", data= dfTR)
m7<-glm(target ~ (CDI + education_level +  relevent_experience +
                    company_type) * enrolled_university , family="binomial", data= dfTR)
m8<-glm(target ~ (CDI + education_level + enrolled_university  +
                    company_type) * relevent_experience , family="binomial", data= dfTR)
m9<-glm(target ~ (CDI + education_level + enrolled_university +
                    relevent_experience) * company_type , family="binomial", data= dfTR)

Anova(m6, test="LR")
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: target
##                                      LR Chisq Df Pr(>Chisq)
## CDI                                    387.59  1  < 2.2e-16 ***
## enrolled_university                      8.91  1   0.002829 **
## relevent_experience                     21.71  1  3.165e-06 ***
## company_type                            21.98  2  1.688e-05 ***
## education_level                         42.51  2  5.870e-10 ***
## CDI:education_level                      3.15  2   0.207124
## enrolled_university:education_level      5.83  2   0.054104 .
## relevent_experience:education_level      4.83  2   0.089510 .
## company_type:education_level            10.79  4   0.029013 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Anova(m7, test="LR")
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: target
##                                         LR Chisq Df Pr(>Chisq)
## CDI                                       388.44  1  < 2.2e-16 ***
## education_level                            40.25  2  1.820e-09 ***
## relevent_experience                        20.14  1  7.196e-06 ***
## company_type                               23.66  2  7.268e-06 ***
## enrolled_university                        11.57  1   0.000669 ***
## CDI:enrolled_university                     5.56  1   0.018361 *
## education_level:enrolled_university         4.99  2   0.082477 .
## relevent_experience:enrolled_university     2.60  1   0.106946
## company_type:enrolled_university            1.46  2   0.481156
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Anova(m8, test="LR")
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: target
##                                         LR Chisq Df Pr(>Chisq)
## CDI                                       392.23  1  < 2.2e-16 ***
## education_level                            40.51  2  1.598e-09 ***
## enrolled_university                         9.93  1   0.001622 **
## company_type                               22.30  2  1.440e-05 ***
## relevent_experience                        21.69  1  3.205e-06 ***
## CDI:relevent_experience                     6.05  1   0.013872 *
## education_level:relevent_experience         4.08  2   0.130051
## enrolled_university:relevent_experience     2.17  1   0.140605
## company_type:relevent_experience            0.15  2   0.925819
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Anova(m9, test="LR")
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: target
##                                    LR Chisq Df Pr(>Chisq)
## CDI                                  392.25  1  < 2.2e-16 ***
## education_level                       41.43  2  1.009e-09 ***
## enrolled_university                   10.17  1  0.0014287 **
## relevent_experience                   20.84  1  4.984e-06 ***
## company_type                          22.96  2  1.036e-05 ***
## CDI:company_type                      14.45  2  0.0007264 ***
## education_level:company_type          16.00  4  0.0030256 **
## enrolled_university:company_type       4.55  2  0.1026195
## relevent_experience:company_type       2.40  2  0.3012295
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(m5,m6, test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: target ~ CDI + education_level + enrolled_university + relevent_experience +
##     company_type
## Model 2: target ~ (CDI + enrolled_university + relevent_experience + company_type) *
##     education_level
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1      3732     3679.9
## 2      3722     3657.8 10   22.128  0.01446 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(m5,m7, test="Chisq")
```

```
## Analysis of Deviance Table
##
```

```
## Model 1: target ~ CDI + education_level + enrolled_university + relevent_experience +
##     company_type
## Model 2: target ~ (CDI + education_level + relevent_experience + company_type) *
##     enrolled_university
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1      3732     3679.9
## 2      3726     3666.4  6   13.503   0.0357 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(m5,m8, test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: target ~ CDI + education_level + enrolled_university + relevent_experience +
##     company_type
## Model 2: target ~ (CDI + education_level + enrolled_university + company_type) *
##     relevent_experience
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1      3732     3679.9
## 2      3726     3668.3  6   11.634  0.07065 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(m5,m9, test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: target ~ CDI + education_level + enrolled_university + relevent_experience +
##     company_type
## Model 2: target ~ (CDI + education_level + enrolled_university + relevent_experience) *
##     company_type
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1      3732     3679.9
## 2      3722     3651.4 10   28.548 0.001474 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
waldtest(m5,m6,test="Chisq")
```

```
## Wald test
##
## Model 1: target ~ CDI + education_level + enrolled_university + relevent_experience +
##     company_type
## Model 2: target ~ (CDI + enrolled_university + relevent_experience + company_type) *
##     education_level
##   Res.Df Df  Chisq Pr(>Chisq)
## 1   3732
## 2   3722 10 22.477    0.01285 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
waldtest(m5,m7,test="Chisq")
```

```
## Wald test
##
## Model 1: target ~ CDI + education_level + enrolled_university + relevent_experience +
##     company_type
## Model 2: target ~ (CDI + education_level + relevent_experience + company_type) *
##     enrolled_university
##   Res.Df Df  Chisq Pr(>Chisq)
## 1   3732
## 2   3726  6 13.581    0.03468 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
waldtest(m5,m8,test="Chisq")
```

```
## Wald test
##
## Model 1: target ~ CDI + education_level + enrolled_university + relevent_experience +
##     company_type
## Model 2: target ~ (CDI + education_level + enrolled_university + company_type) *
##     relevent_experience
##   Res.Df Df  Chisq Pr(>Chisq)
## 1   3732
## 2   3726  6 11.631    0.07072 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
waldtest(m5,m9,test="Chisq")
```

```
## Wald test
##
## Model 1: target ~ CDI + education_level + enrolled_university + relevent_experience +
##     company_type
## Model 2: target ~ (CDI + education_level + enrolled_university + relevent_experience) *
##     company_type
##   Res.Df Df  Chisq Pr(>Chisq)
## 1   3732
## 2   3722 10 25.114   0.005133 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
AIC(m5,m6,m7,m8,m9)
```

```
##    df      AIC
## m5  8 3695.924
## m6 18 3693.795
## m7 14 3694.420
## m8 14 3696.290
## m9 18 3687.375
```

According to the Likelihood Ratio Test, adding each of the individual interactions of the selected variables to the optimal model so far: The education level interaction at 0.05 of significance is relevant for education level but not relevant for the rest of the analysis. For enrolled university and company size the interactions are significant with the CDI, as well company size is significant as well with education level. This interactions allow to identify the numeric variable is relate to the qualitative variables.

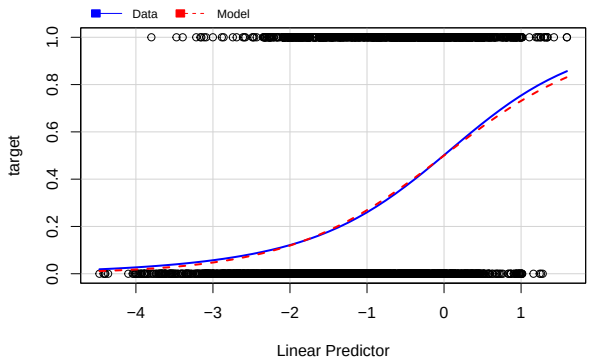
Evaluating if the models containing these interactions vs the model including just the main effect, it is considered the chi-squared test through the method anova, at a 0.05 of significance there is difference adding the interactions: education level, enrrollment university and company type. The wald test shows significance at 0.05 for the same interaction.

The chosen model includes: CDI, education level, enrolled university, relevant experience, company type and interactions of CDI, education level with company and relevent experience with company. It is the one with the lowest AIC: 3687.375 and includes

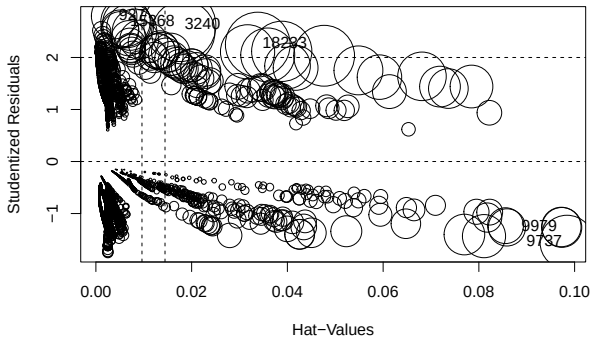
Final Residual analysis: unusual and influent data filtering

The final residual plots are calculated:

```
marginalModelPlot(m9)
```

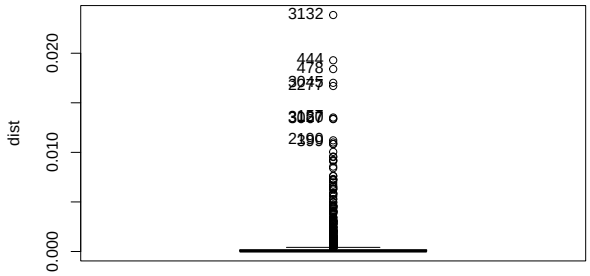



```
influencePlot(m9)
```



```
##          StudRes         Hat       CookD
## 18233  2.248119 0.033817481 0.019287539
## 15368  2.684871 0.006181693 0.011212822
## 927    2.795333 0.003811170 0.009531030
## 9737  -1.550487 0.098405038 0.013351380
## 9979  -1.266725 0.097331268 0.007313084
## 3240   2.613325 0.017541237 0.023853684
```

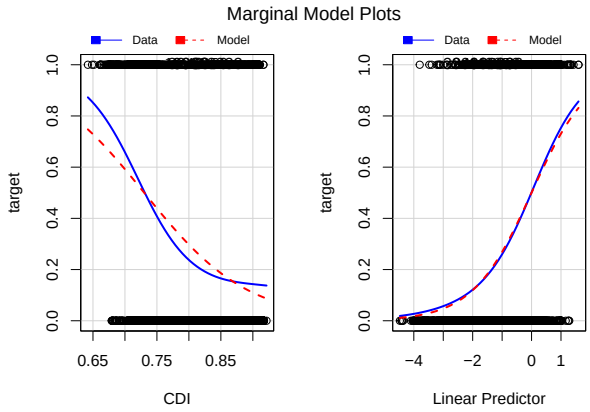
```
dist <- cooks.distance(m9)
influential <- Boxplot(dist)
```



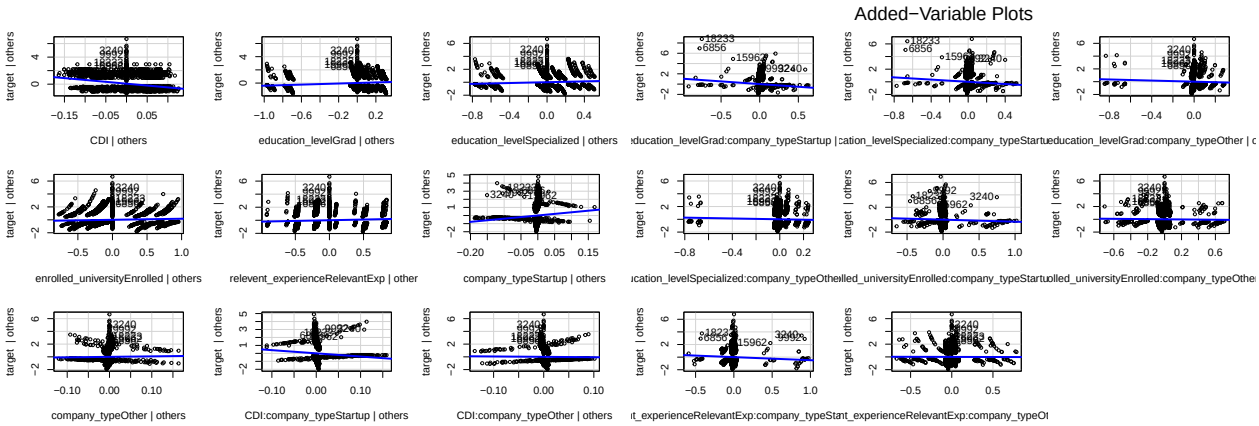

```
length(influential)
```

```
## [1] 10
```

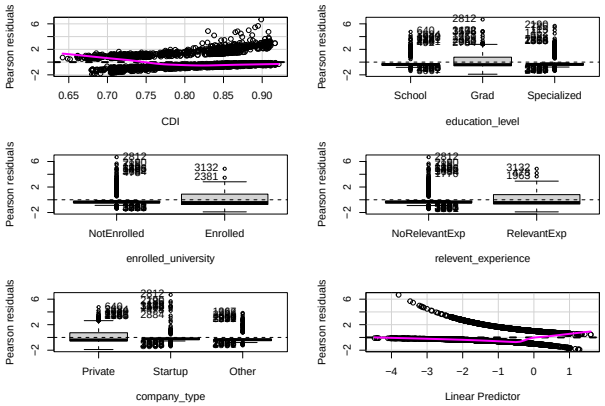
```
marginalModelPlots(m9,id=list(labels=row.names(dfTR),method=abs(cooks.distance(m9)), n=5) )
```




```
avPlots(m9,id=list(labels=row.names(dfTR),method=abs(cooks.distance(m9)), n=5) )
```




```
residualPlots(m9, layout=c(3, 2))
```




```
##                      Test stat Pr(>|Test stat|)
## CDI                     39.129        3.966e-10 ***
## education_level
## enrolled_university
## relevent_experience
## company_type
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
outlierTest(m9)
```

```
## No Studentized residuals with Bonferroni p < 0.05
## Largest |rstudent|:
##     rstudent unadjusted p-value Bonferroni p
## 927 2.795333          0.0051846           NA
```

```
Anova(m9, test="LR")
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: target
```

```
##                                   LR Chisq Df Pr(>Chisq)
## CDI                                 392.25  1  < 2.2e-16 ***
## education_level                      41.43  2  1.009e-09 ***
## enrolled_university                  10.17  1  0.0014287 **
## relevent_experience                  20.84  1  4.984e-06 ***
## company_type                         22.96  2  1.036e-05 ***
## CDI:company_type                     14.45  2  0.0007264 ***
## education_level:company_type         16.00  4  0.0030256 **
## enrolled_university:company_type      4.55  2  0.1026195
## relevent_experience:company_type      2.40  2  0.3012295
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
m10<-glm(target ~ CDI + education_level + enrolled_university + relevent_experience +
           company_type + CDI*company_type + education_level*company_type,
         family="binomial", data= dfTR)
```

The final selected model presents a smoother and closer to fit between the model and the data. This model do not present influential data.

Goodness of fit and Model Interpretation.

The figures associated with the goodness of fit are presented:

```
final<- m10
Anova(m10)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: target
##                               LR Chisq Df Pr(>Chisq)
## CDI                             392.96  1  < 2.2e-16 ***
## education_level                  42.33  2  6.443e-10 ***
## enrolled_university              10.67  1   0.001088 **
## relevent_experience              22.21  1  2.448e-06 ***
## company_type                     22.96  2  1.036e-05 ***
## CDI:company_type                 10.96  2   0.004164 **
## education_level:company_type     11.00  4   0.026540 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
probsel<- predict(final, newdata=dfTS, type = "response")
selTest<-ifelse(probsel<0.5,0,1)
CM<-table(selTest,dfTS$target)
CM
```

```
##
## selTest FALSE TRUE
##       0   886  209
##       1    65   90
```

```
NagelkerkeR2(final)
```

```
## $N
## [1] 3740
##
## $R2
## [1] 0.2178473
```

```
summary(final)
```

```
##
## Call:
## glm(formula = target ~ CDI + education_level + enrolled_university +
##     relevent_experience + company_type + CDI * company_type +
##     education_level * company_type, family = "binomial", data = dfTR)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7263  -0.6908  -0.5238   0.8102   2.8081
##
## Coefficients:
##                                                Estimate Std. Error z value
## (Intercept)                                     7.00967    0.54496  12.863
## CDI                                           -11.10208    0.65029 -17.073
## education_levelGrad                             1.01176    0.15069   6.714
## education_levelSpecialized                      0.76329    0.17439   4.377
## enrolled_universityEnrolled                     0.30875    0.09407   3.282
## relevent_experienceRelevantExp                  0.45473    0.09603   4.735
## company_typeStartup                             7.46553    2.29385   3.255
## company_typeOther                               1.73553    2.08780   0.831
## CDI:company_typeStartup                        -8.60889    2.79493  -3.080
## CDI:company_typeOther                          -1.59894    2.46696  -0.648
## education_levelGrad:company_typeStartup        -1.69394    0.58983  -2.872
## education_levelSpecialized:company_typeStartup -1.11973    0.64622  -1.733
## education_levelGrad:company_typeOther          -0.99308    0.54336  -1.828
## education_levelSpecialized:company_typeOther   -0.68100    0.56239  -1.211
##                                                Pr(>|z|)
## (Intercept)                                     < 2e-16 ***
## CDI                                             < 2e-16 ***
## education_levelGrad                            1.89e-11 ***
## education_levelSpecialized                     1.20e-05 ***
## enrolled_universityEnrolled                     0.00103 **
## relevent_experienceRelevantExp                 2.19e-06 ***
## company_typeStartup                             0.00114 **
## company_typeOther                               0.40582
## CDI:company_typeStartup                         0.00207 **
## CDI:company_typeOther                           0.51689
## education_levelGrad:company_typeStartup         0.00408 **
## education_levelSpecialized:company_typeStartup  0.08314 .
## education_levelGrad:company_typeOther           0.06760 .
## education_levelSpecialized:company_typeOther    0.22593
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4258.2  on 3739  degrees of freedom
## Residual deviance: 3658.8  on 3726  degrees of freedom
## AIC: 3686.8
##
## Number of Fisher Scoring iterations: 5
```

```
100*(1-m10$dev/m10$null.dev)
```

```
## [1] 14.07546
```

```
100*(1-(logLik(m10)/m10$df.residual)/(logLik(m0)/(m0$df.residual)))
```

```
## 'log Lik.' 13.70822 (df=14)
```

```
PseudoR2(final, which='all') # Not working for grouped data
```

```
##        McFadden     McFaddenAdj        CoxSnell      Nagelkerke   AldrichNelson
##       0.1407546       0.1341790       0.1480746       0.2178473       0.1381215
## VeallZimmermann           Efron McKelveyZavoina            Tjur             AIC
##       0.2594349       0.1675023       0.2349443       0.1650157    3686.8246819
##             BIC          logLik         logLik0              G2
##    3774.0004543   -1829.4123409   -2129.0917487     599.3588155
```

```
# Sheather
1 - (final$deviance / final$null.deviance)
```

```
## [1] 0.1407546
```

```
# McFadden
1-(as.numeric(logLik(final))/as.numeric(logLik(m0)))
```

```
## [1] 0.1423762
```

```
# Hosmer-Lemershow
seque<-quantile(fitted(m10),probs=seq(0,1,by=0.1))
fitgrup<-cut(fitted(m10),breaks=seque)

AUC(predict(m10,type="response"), dfTR$target)
```

```
## [1] 0.3416424
```

The logit transformation of the probability of being hired in the reference groups is 7.006 and the corresponding logit probabilities for the true outcome is increased:

- One unit per graduated education level, 0.76 per specialized level with respect of high school.

- 0.30 units per being enrolled in a university program respect not being.
- 0.45 units per have relevant experience respect not having it.
- 7.46 units per coming from an startup company and 1.73 units per being in other types of company, respect to being in a private company.

and is decreasing in 11.10 units per unit increasing in the city development index of the candidate.

The godness of fit of the model using the Hosmer-Lemeshow statistics and McFadden is above 14%. From the test dataset this model predicts TRUE cases in the 9.2% of the cases. The area under the curve is only. 0.34