Assignment2

Irene Fernández Rebollo i Àlex Martorell i Locascio

5/12/2021

Presentation

A company which is active in Big Data and Data Science wants to hire data scientists among people who successfully pass some courses which are conducted by the company. Many people signup for their training. Company wants to know which of these candidates really want to work for the company after training or looking for a new employment because it helps to reduce the cost and time as well as the quality of training or planning the courses and categorization of candidates. Information related to demographics, education, experience are in hands from candidates signup and enrollment.

This dataset designed to understand the factors that lead a person to leave current job for HR researches too. By model(s) that uses the current credentials, demographics, experience data you will predict the probability of a candidate to look for a new job or will work for the company, as well as interpreting affected factors on employee decision.

Data Preparation

```
df <- read.csv("aug_train.csv",header=T, sep=",", na.strings="NA")
summary(df)
```

```
##   enrollee_id        city           city_development_index    gender
##  Min.   :    1   Length:19158       Min.   :0.4480         Length:19158
##  1st Qu.: 8554   Class :character   1st Qu.:0.7400         Class :character
##  Median :16983   Mode  :character   Median :0.9030         Mode  :character
##  Mean   :16875                      Mean   :0.8288
##  3rd Qu.:25170                      3rd Qu.:0.9200
##  Max.   :33380                      Max.   :0.9490
##  relevent_experience enrolled_university education_level    major_discipline
##  Length:19158        Length:19158        Length:19158       Length:19158
##  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:19158       Length:19158       Length:19158       Length:19158
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
```

```
##
##
##  training_hours       target
##  Min.   :  1.00   Min.   :0.0000
##  1st Qu.: 23.00   1st Qu.:0.0000
##  Median : 47.00   Median :0.0000
##  Mean   : 65.37   Mean   :0.2493
##  3rd Qu.: 88.00   3rd Qu.:0.0000
##  Max.   :336.00   Max.   :1.0000
```

Removing Duplicates and Irrelavant Observations

```
df <- unique(df) #No duplicates
#nrow(df) == nrow(unique(df))

set.seed(130798) #Birthday of 1 member of the group as random seed:
samples <- as.vector(sort(sample(1:nrow(df),5000))) #Subset of 5000 observations
df <- df[samples,]
```

Fix structural errors

For coding purposes, blanks ("") in the dataframe are considered as NA's.

```
sum(is.na(df)) #0
df[df==""] <- NA
sum(is.na(df)) #5433
```

Some inconsistencies are also checked and corrected.

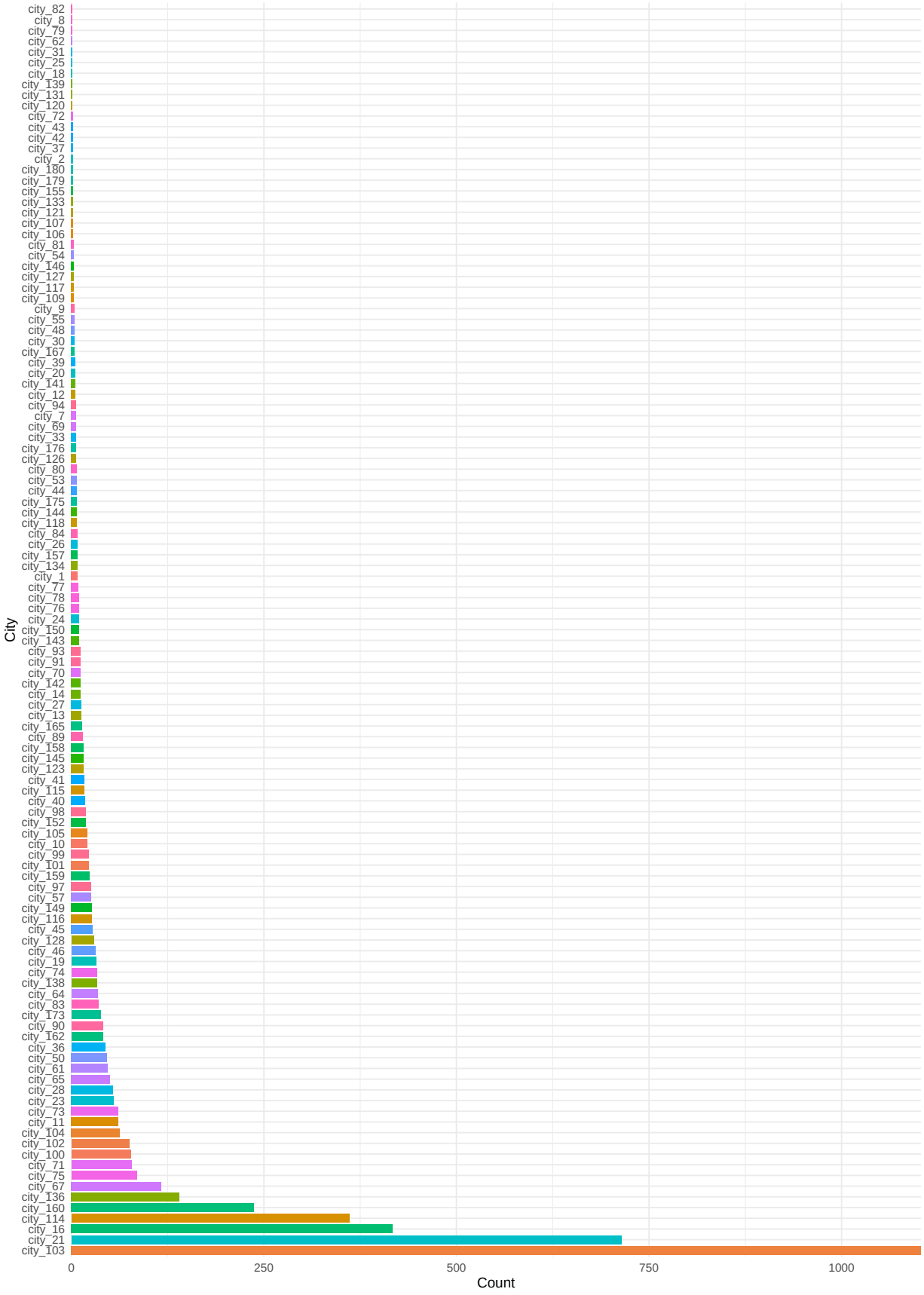
```
# Column name relevent_experience should be relevant_experience
colnames(df)[5] <- "relevant_experience"
# Also their values should be with the "relevant" word
df$relevant_experience <- gsub("relevent", "relevant", df$relevant_experience)

# Correct the format of the company size
df$company_size[df$company_size == "10/49"] <- "10-49"
```

Univariate Descriptive Analysis

Out of the 14 variables in the dataset, R detects 9 of them being character-type. They are transformed into factors, taking into consideration NA values. One of the more tricky aspects is the years of experience variable. Two possibilites are taken into consideration, a factor with levels based on quartiles and a numeric variable.

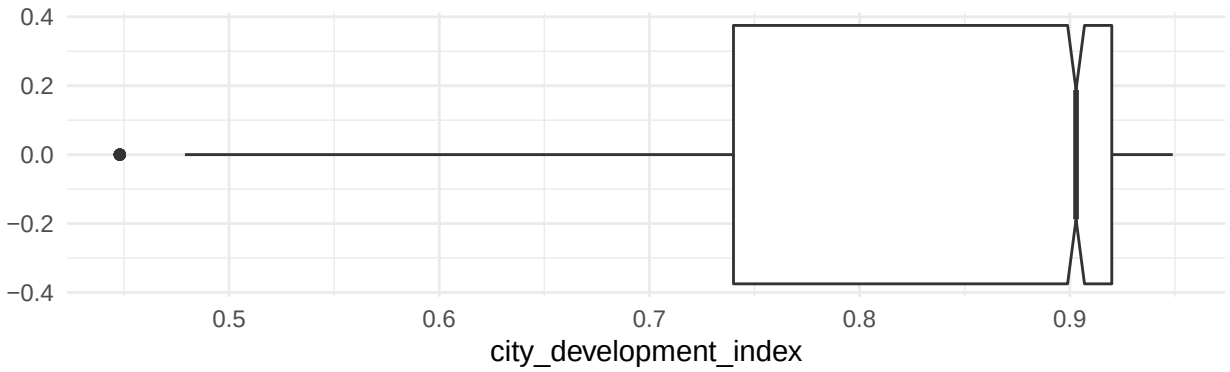
City code

```
class(df$city) #"character"
df$city <- factor(df$city) #Transform to "factor"
```


City
Count
0
250
500
750
1000
city_103
city_21
city_16
city_114
city_160
city_136
city_67
city_75
city_71
city_100
city_102
city_104
city_11
city_73
city_23
city_28
city_65
city_61
city_50
city_36
city_162
city_90
city_173
city_83
city_64
city_138
city_74
city_19
city_46
city_128
city_45
city_116
city_149
city_57
city_97
city_159
city_101
city_99
city_10
city_105
city_152
city_98
city_40
city_115
city_41
city_123
city_145
city_158
city_89
city_165
city_13
city_27
city_14
city_142
city_70
city_91
city_93
city_143
city_150
city_24
city_76
city_78
city_77
city_1
city_134
city_157
city_26
city_84
city_118
city_144
city_175
city_44
city_53
city_80
city_126
city_176
city_33
city_69
city_7
city_94
city_12
city_141
city_20
city_39
city_167
city_30
city_48
city_55
city_9
city_109
city_117
city_127
city_146
city_54
city_81
city_106
city_107
city_121
city_133
city_155
city_179
city_180
city_2
city_37
city_42
city_43
city_72
city_120
city_131
city_139
city_18
city_25
city_31
city_62
city_79
city_8
city_82

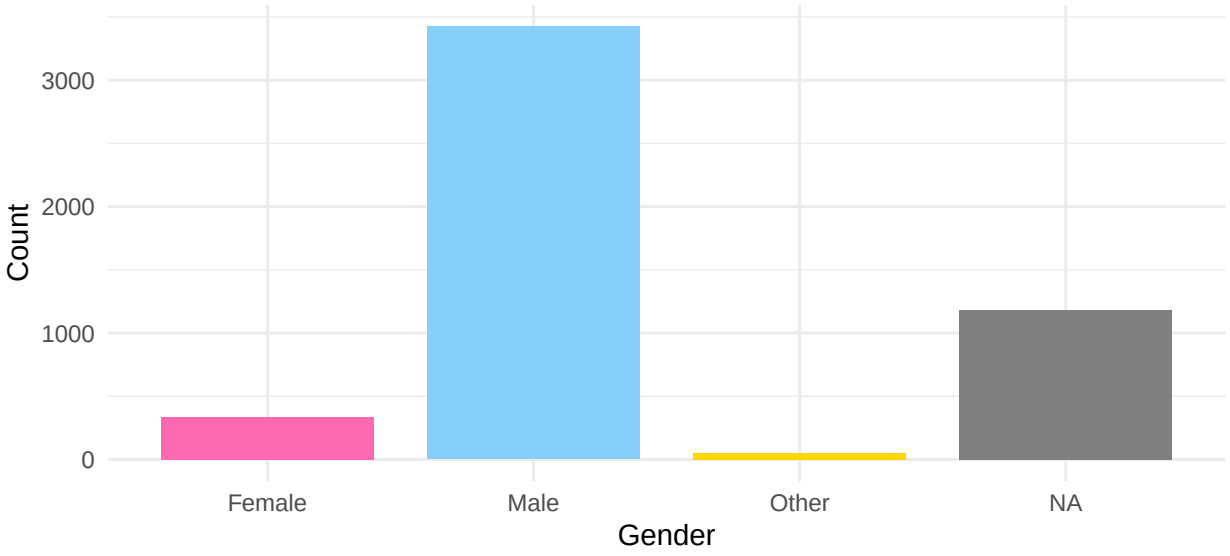
Developement index of the city

```
class(df$city_development_index) #"numeric"
# Kept as "numeric"
```



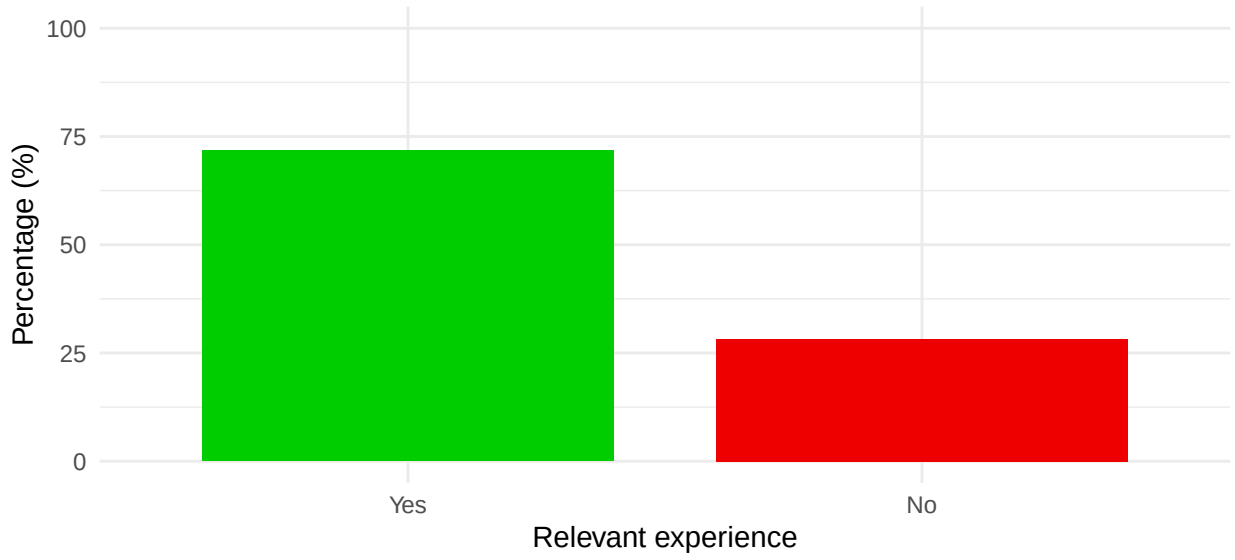
Gender of candidate

```
class(df$gender) #"character"
df$gender <- factor(df$gender) #Transform to "factor"
```



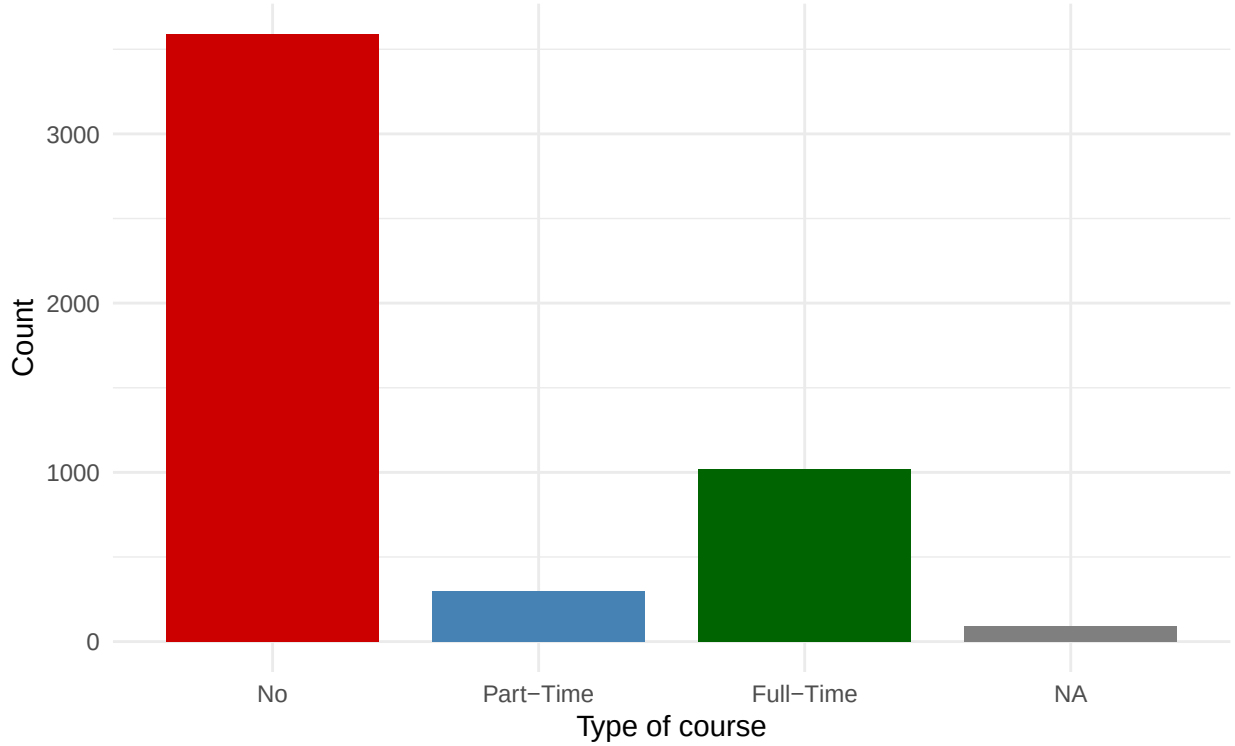
Relevant experience of candidate

```
class(df$relevant_experience) #"character"
df$relevant_experience <- factor(df$relevant_experience,
                                 levels=c("Has relevant experience", "No relevant experience"),
                                 labels = c("Yes", "No")) #Transform to "factor" and change to simpler
```



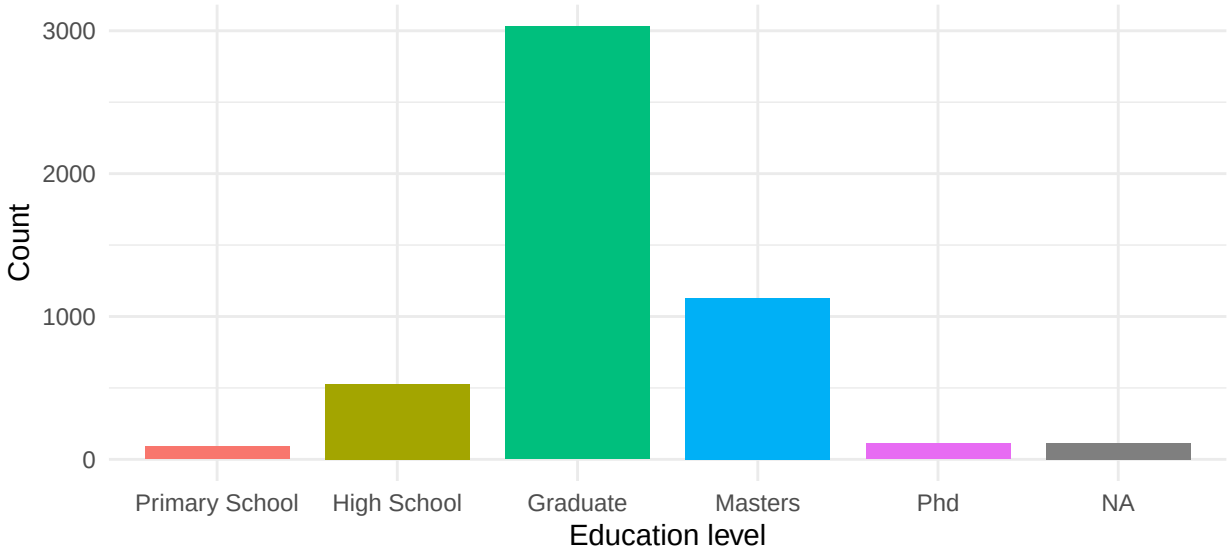
Type of University course enrolled

```
class(df$enrolled_university) #"character"
df$enrolled_university <- factor(df$enrolled_university,
                                 levels=c("no_enrollment", "Part time course", "Full time course"),
                                 labels= c("No", "Part-Time", "Full-Time")) #Transform to "factor" and c
```



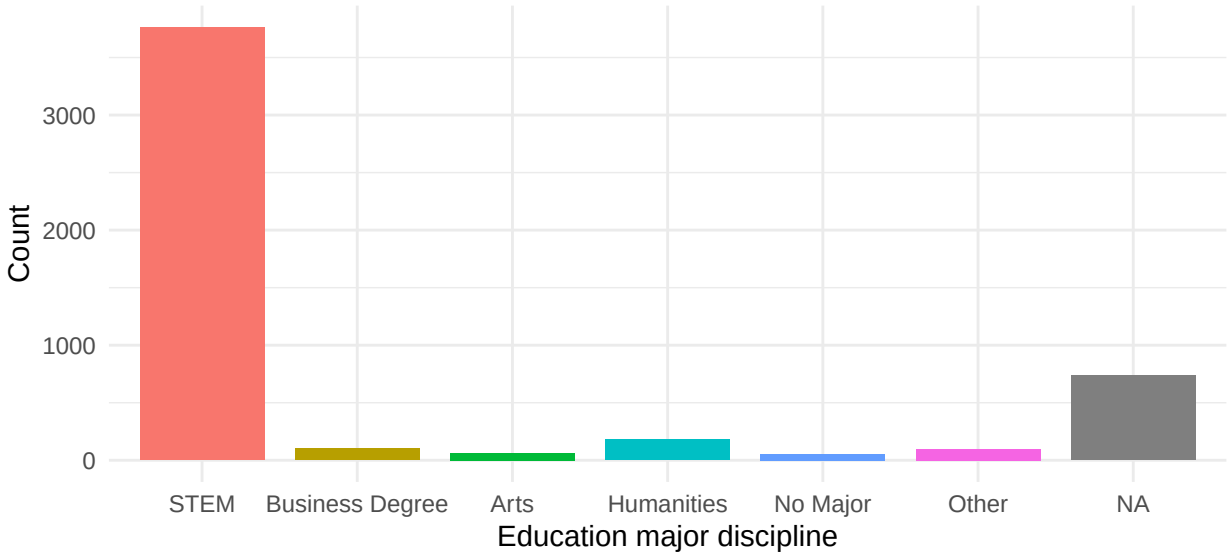
Education level of candidate

```
class(df$education_level) #"character"
df$education_level <- factor(df$education_level,
                             levels = c("Primary School", "High School",
                                        "Graduate", "Masters", "Phd")) #Transform to "factor"
```



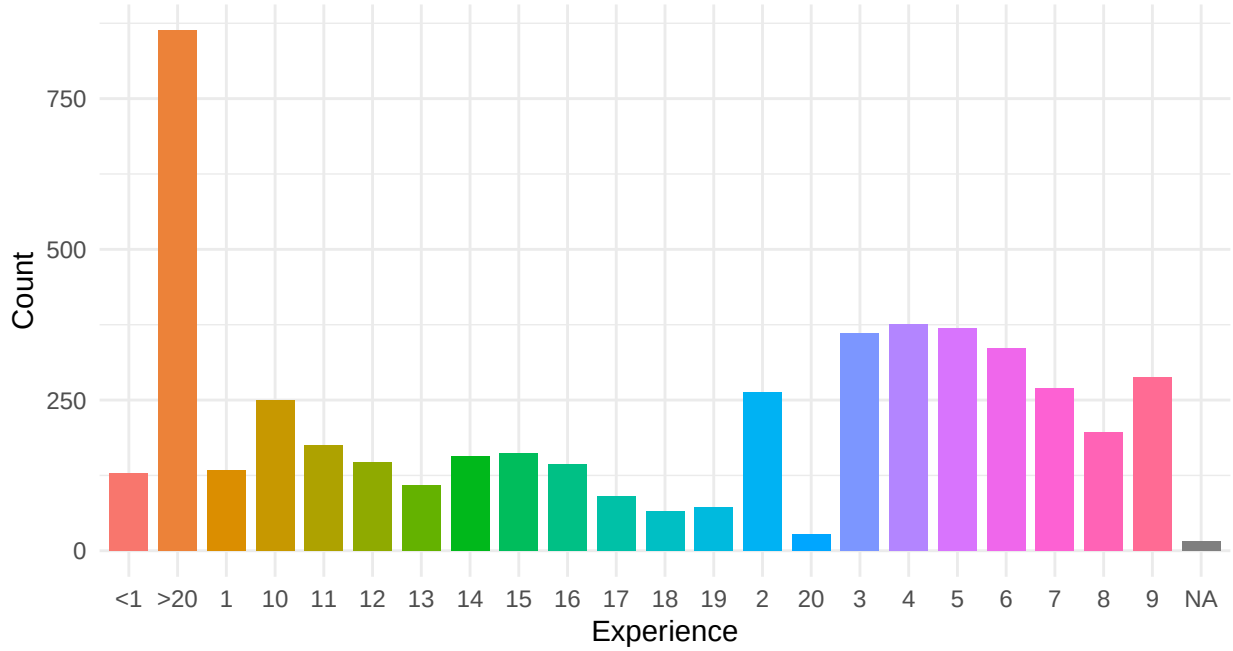
Education major discipline of candidate

```
class(df$major_discipline) #"character"
df$major_discipline <-factor(df$major_discipline,
                             levels=c("STEM", "Business Degree", "Arts",
                                        "Humanities", "No Major", "Other")) #Transform to "factor"
```



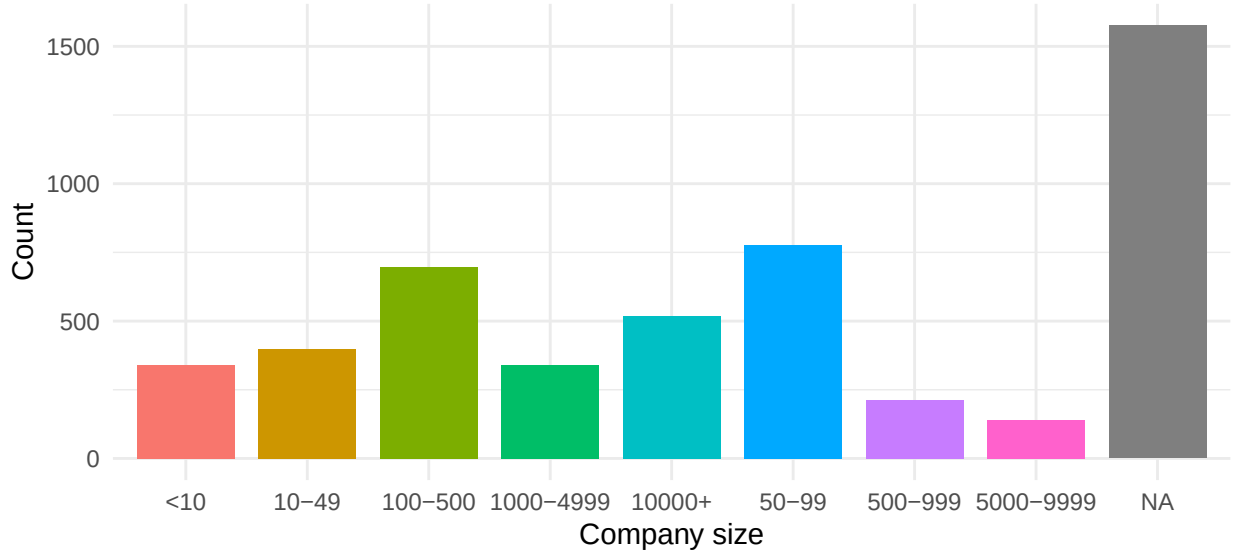
Candidate total experience in years

```
class(df$experience) #"character"
df$experience <- factor(df$experience) #Transform to "factor"
```



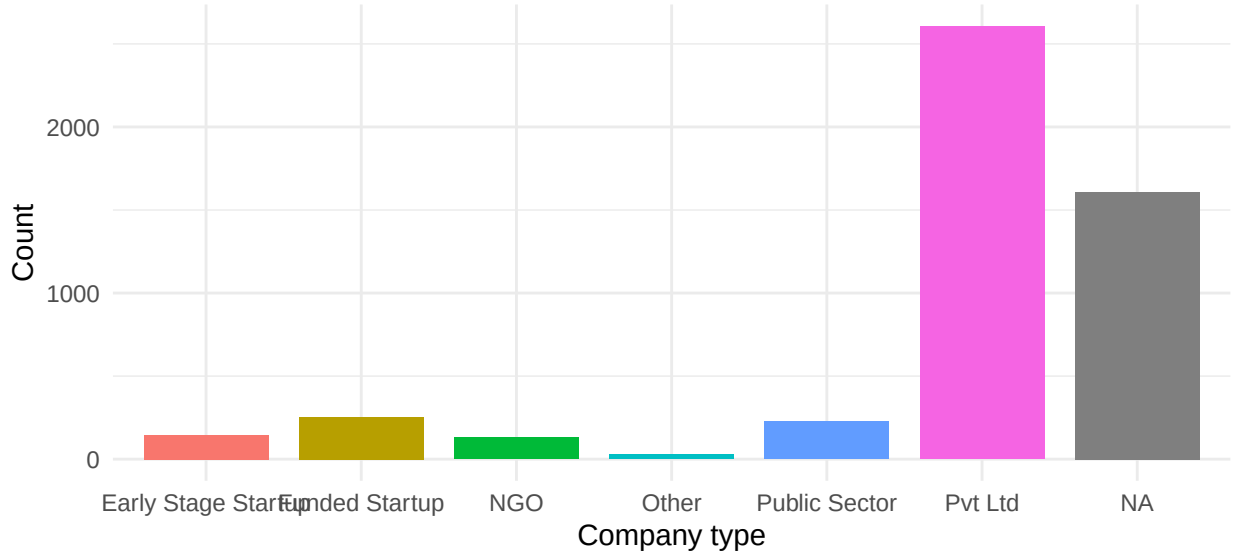
Number of employees in current employer's company

```
class(df$company_size) #"character"
df$company_size <- factor(df$company_size) #Transform to "factor"
```



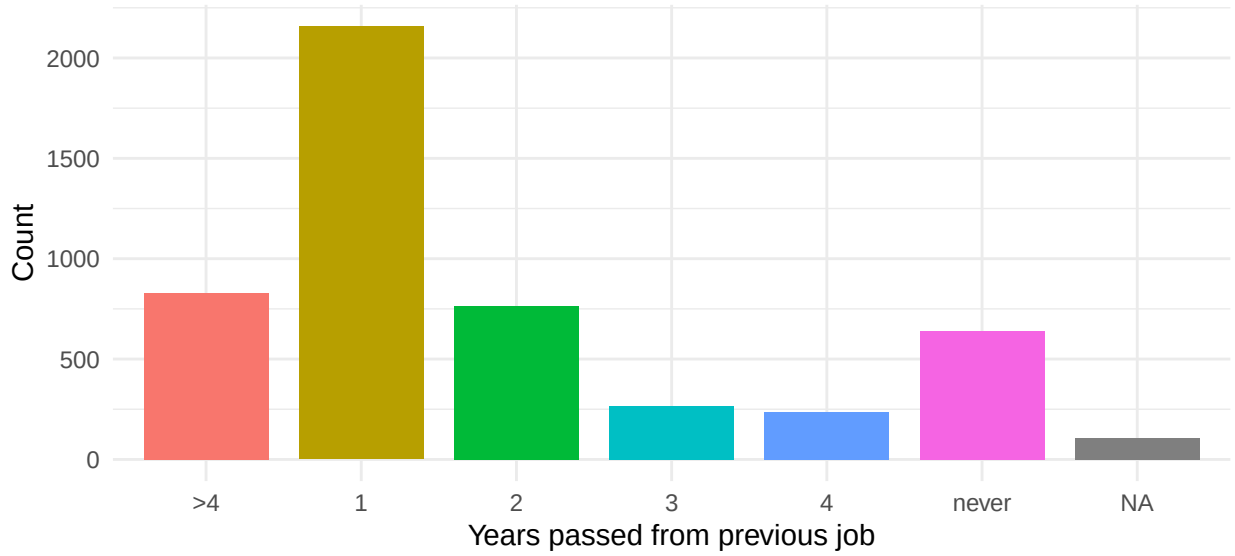
Type of current employer

```
class(df$company_type) #"character"
df$company_type <- factor(df$company_type) #Transform to "factor"
```



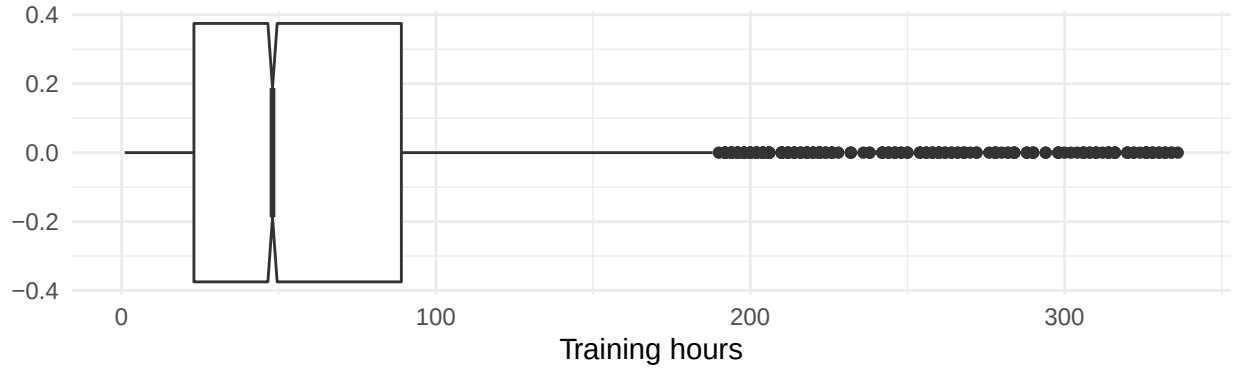
Difference in years between previous job and current job

```
class(df$last_new_job) #"character"
df$last_new_job <-factor(df$last_new_job) #Transform to "factor"
```



Training hours completed

```
class(df$training_hours) #"integer"
#Keep as "integer"
```



Missing values

The percentage of missing values per column with respect to their total varies significantly along different variables. `company_type` and `company_size` have over 30% of missing values. The numeric variables, the target, `relevant_experience`, and `city` return no missing values.

```
missing <- unlist(lapply(df, function(x) sum(is.na(x))))/nrow(df)
sort(missing[missing >= 0], decreasing = TRUE)
```

```
##           company_type           company_size                 gender
##                 0.3216                 0.3152                 0.2368
##       major_discipline        education_level           last_new_job
##                 0.1470                 0.0230                 0.0212
##    enrolled_university             experience            enrollee_id
##                 0.0186                 0.0032                 0.0000
##                   city city_development_index    relevant_experience
##                 0.0000                 0.0000                 0.0000
##         training_hours                 target
##                 0.0000                 0.0000
```

```
count_na<-colSums(is.na(df[, 1:14]))
count_na
```

```
##            enrollee_id                   city city_development_index
##                      0                      0                      0
##                 gender    relevant_experience    enrolled_university
##                   1184                      0                     93
##        education_level       major_discipline             experience
##                    115                    735                     16
##           company_size           company_type           last_new_job
##                   1576                   1608                    106
##         training_hours                 target
##                      0                      0
```

```
sum(is.na(df)) #5433
```

Outliers

Univariate Outlier detection

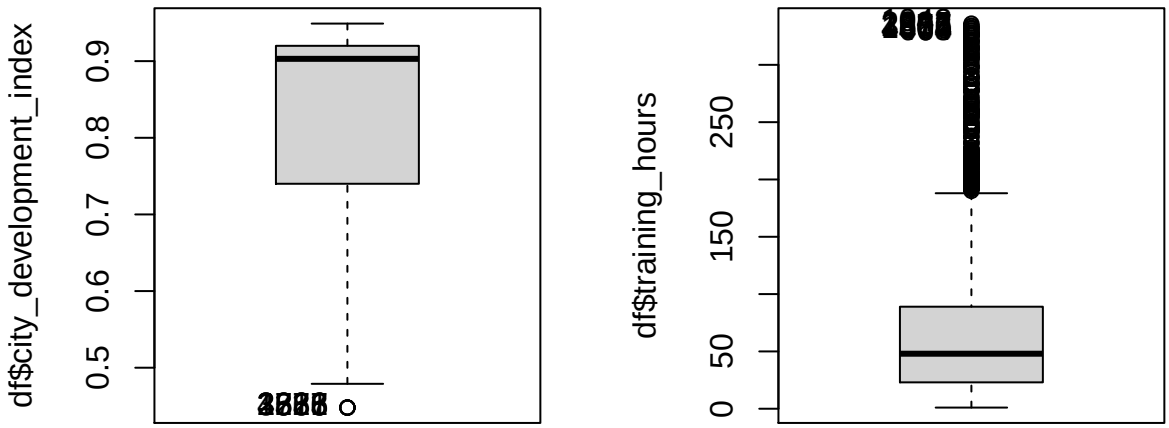
Outliers for the three numerical variables are considered. Training hours and city_development_index are fairly straightforward to analyze, but with regards to the experience variable there are no outliers due to the manner in which the variable was built. The characters "<1" and ">20" do not allow enough precision in order to identify for possible outliers, as they had to be converted to numbers 0 and 21 respectively during the conversion to a numeric variable (noting a significant bias).

The outliers found are registered in two different ways. A separate table is built for the Data Quality Report named `dqind` (stands for Data Quality per Individual). This table stores the rows where mild and extreme outliers for training hours and mild outliers for city development index occur. A new variable called `quality` is added to the main data frame also for the Data Quality Report.

```
par(mfrow = c(1,2))
Boxplot(df$city_development_index)
```

```
## [1] 1267 2376 2733 3518 3886 4621
```

```
Boxplot(df$training_hours)
```



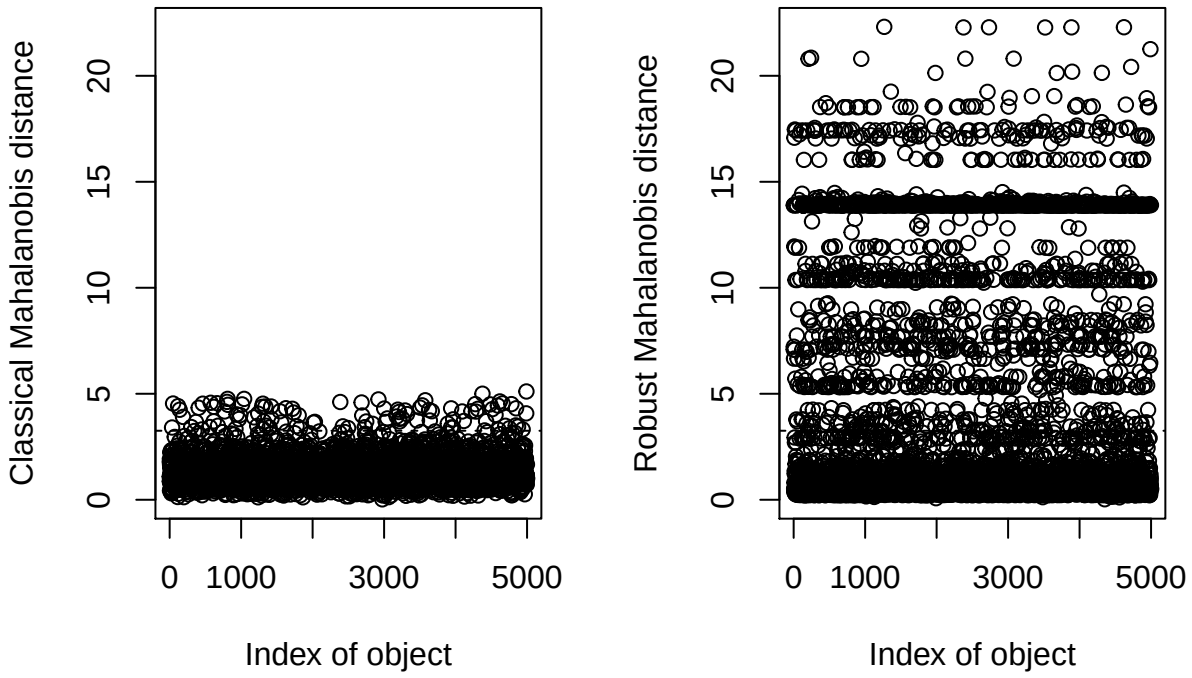

```
##  [1] 1042  811 2685  803 2387 3576 1304 4376  463  664
```

```
mout_city = quantile(df$city_development_index)[[2]]-1.5*IQR(df$city_development_index)
sum(df$city_development_index < mout_city) #6
mout_th = quantile(df$training_hours)[[4]]+1.5*IQR(df$training_hours)
eout_th = quantile(df$training_hours)[[4]]+3*IQR(df$training_hours)
sum(df$training_hours > mout_th) #227
sum(df$training_hours > eout_th) #58
```

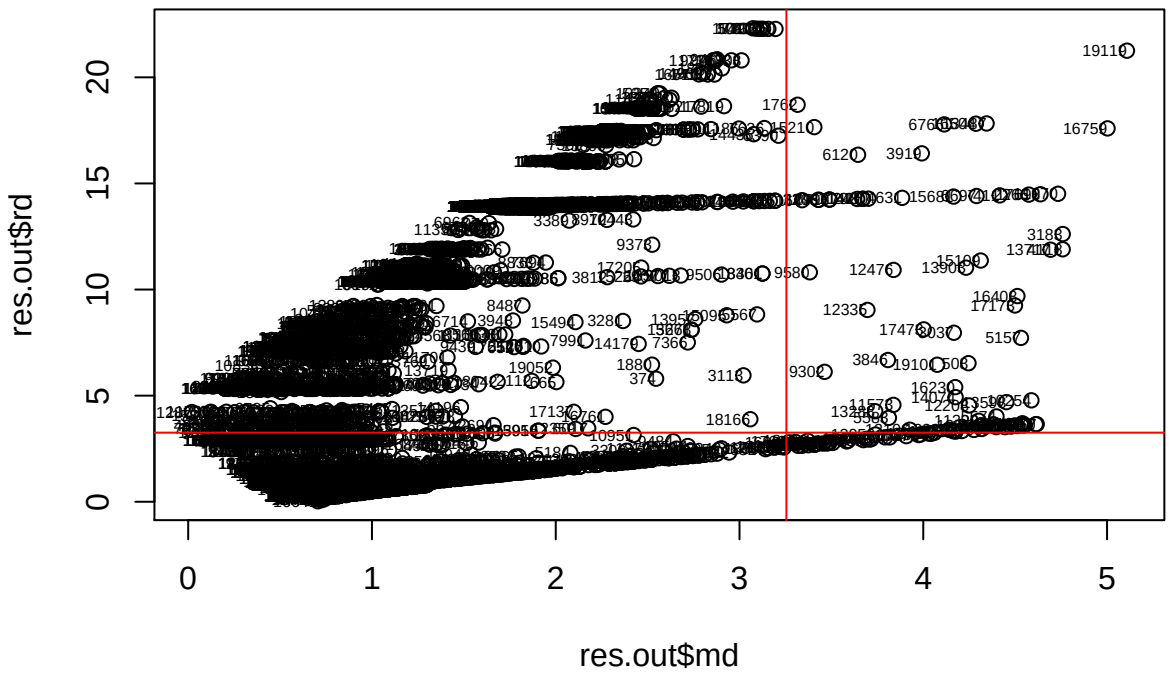
Multivariate outlier detection

With the two numerical variables, we apply the Moutlier function at 99.5%. 105 multivariate outliers are returned.

```
res.out<-Moutlier(df[,c(3,13)],quantile=0.995)
```




```
#quantile(res.out$md,seq(0,1,0.005))
multout<-which(res.out$md > res.out$cutoff)
length(multout) #105
```



Data Quality Report

To summarize all the its and buts of the Data Processing in a readable manner, two dataframes are created, one refers to variables and the other to individuals. This is the *Data Quality Report*. For obvious reasons, it is possible to print out the whole data quality report per variable but not per individual (5000 in total). Also, a data quality variable is created.

```
df$quality <- 0
```

Per variable

```
dqvar <- data.frame(colnames(df[, 1:14]))
dqvar$outliers <-0
dqvar$missing <-0
dqvar$errors <-0

# Outliers
dqvar[13, "outliers"] <- sum(df$training_hours > mout_th) + sum(df$training_hours > eout_th)
dqvar[3, "outliers"] <- sum(df$city_development_index < mout_city)

#Missing
#We add missing values to the column per variable
```

```
dqvar$missing <- (colSums(is.na(df[, 1:14])))
dqvar
```

```
##      colnames.df...1.14.. outliers missing errors
## 1             enrollee_id        0       0      0
## 2                    city        0       0      0
## 3  city_development_index        6       0      0
## 4                  gender        0    1184      0
## 5     relevant_experience        0       0      0
## 6     enrolled_university        0      93      0
## 7         education_level        0     115      0
## 8        major_discipline        0     735      0
## 9              experience        0      16      0
## 10           company_size        0    1576      0
## 11           company_type        0    1608      0
## 12           last_new_job        0     106      0
## 13         training_hours      285       0      0
## 14                 target        0       0      0
```

Per individual

```
dqind <- data.frame(df$enrollee_id)
colnames(dqind)[1] <- "enrollee_id"
dqind$missing <-0
dqind$outliers <-0
dqind$errors <-0

#Outliers
regmout_th <- subset(df$enrollee_id, df$training_hours > mout_th)
regeout_th <-subset(df$enrollee_id, df$training_hours > eout_th)
regmout_city <-subset(df$enrollee_id, df$city_development_index < mout_city)

for(i in 1:length(regmout_th)) {
  s<-which(df$enrollee_id == regmout_th[i])
  w<-which(dqind$enrollee_id == regmout_th[i])
  df[s, "quality"] <- df[s , "quality"] + 1
  dqind[w, "outliers"] <- dqind[w, "outliers"] + 1}

for(i in 1:length(regeout_th)) {   # for-loop over rows
  s<-which(df$enrollee_id == regeout_th[i])
  w<-which(dqind$enrollee_id == regeout_th[i])
  df[s, "quality"] <- df[s , "quality"] + 1
  dqind[w, "outliers"] <- dqind[w, "outliers"] + 1}

for(i in 1:length(regmout_city)) {   # for-loop over rows
  s<-which(df$enrollee_id == regmout_city[i])
  w<-which(dqind$enrollee_id == regmout_city[i])
  df[s , "quality"] <- df[s , "quality"] + 1
  dqind[w, "outliers"] <- dqind[w, "outliers"] + 1}

df[multout, "quality"] <- df[multout, "quality"] + 1
```

```
dqind[multout, "outliers"] <- dqind[multout, "outliers"] + 1

dqind$missing <- rowSums(is.na(df))
df$quality <- df$quality + rowSums(is.na(df))
head(dqind)
```

```
##   enrollee_id missing outliers errors
## 1         666       0        0      0
## 2       21651       3        0      0
## 3        8722       4        0      0
## 4        5764       1        0      0
## 5        7041       0        0      0
## 6       10408       0        0      0
```

Note that the quality variable is the sum of missing values and outliers. Using the `condes()` package we obtain the correlation with the quantitative and qualitative variables. Note that company_size and company_type as factors seem to influence the most on the quality variable (R2 around 0.6). That is because of the large missing value number that these variables have.

```
# Correlation
table(df$quality)
```

```
##
##    0    1    2    3    4    5    6    7    8
## 2171 1022 1000  545  178   54   23    4    3
```

```
cor(df[, c(3, 13:15)])
```

```
##                        city_development_index training_hours      target
## city_development_index            1.000000000   -0.006019761 -0.32971376
## training_hours                   -0.006019761    1.000000000 -0.02210974
## target                           -0.329713763   -0.022109738  1.00000000
## quality                          -0.105990752    0.201059724  0.15757980
##                            quality
## city_development_index -0.1059908
## training_hours          0.2010597
## target                  0.1575798
## quality                 1.0000000
```

```
res.con <- condes(df, 15)
res.con$quali
```

```
##                             R2       p.value
## education_level     0.29160790  0.000000e+00
## company_size        0.57934677  0.000000e+00
## company_type        0.57105470  0.000000e+00
## major_discipline    0.26162324 4.940656e-324
## last_new_job        0.19613276 2.362457e-232
## relevant_experience 0.17783188 8.198615e-215
## gender              0.16471079 1.284160e-194
## enrolled_university 0.13233214 2.088264e-153
## experience          0.09964441  9.313658e-97
## city                0.05151068  3.334249e-13
```

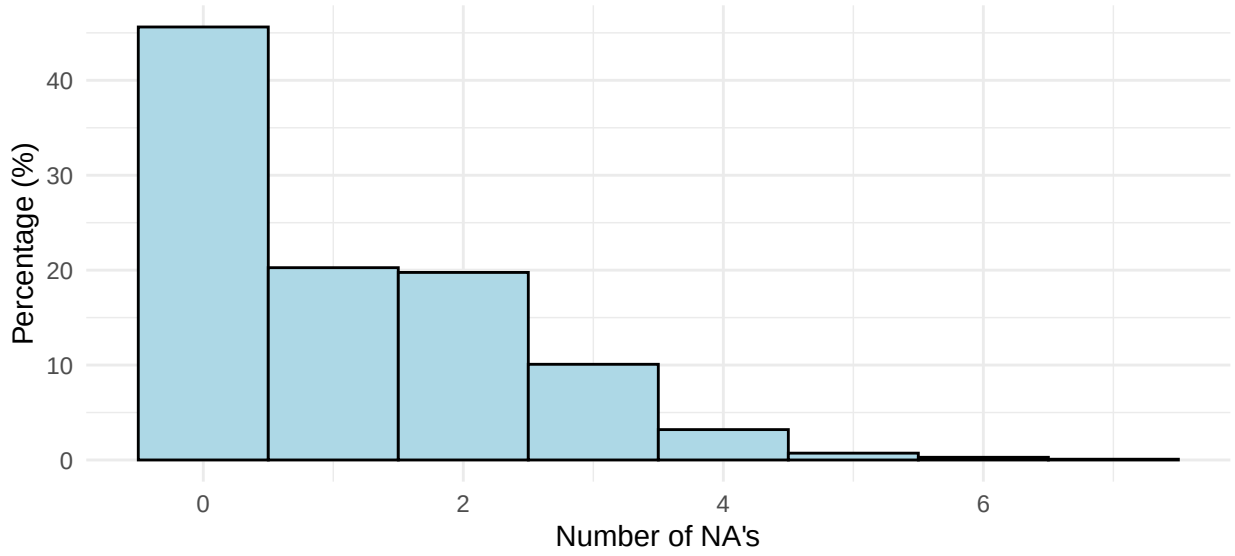
```
res.con$quanti
```

```
##                          correlation      p.value
## training_hours             0.2010597 9.170131e-47
## target                     0.1575798 3.646471e-29
## city_development_index    -0.1059908 5.735764e-14
```

Imputation of values

There are 14 variables and 5000 individuals; for each individual we can have from 0 to 7 NA's along the variables. We can obtain an histogram were we can observe that there is a tail for higher number of missing values, so we determine a first rule of thumb that is that any row containing 5 missing values or more will be removed because of the cost that results when imputing so many values and only represents aprox. 1% of the samples.

```
#Create column with na counts for every observation
df$na_count <- rowSums(is.na(df))

# Actualitzem aquí el Data Quality per Missing's
for(i in 1:nrow(df)) {    # for-loop over rows
  df[i , "quality"] <- df[i , "quality"] + df[i, "na_count"]
  #w <- which(dqind$enrollee_id == df[i, "enrollee_id"])
  dqind[i, "missing"] <- df$na_count[i]}
```




```
df <- df[df$na_count < 5,]
nrow(df) #4946
100 - nrow(df)/5000*100 #only removed a 1.08% of the observations
```

The variables with some NA value are: gender (1184), enrolled_university (93), education_level (115), major_discipline (735), experience (16), company_size (1576), company_type (1608), last_new_job (106). Each one is analysed separately to decide the required imputation method.

- Gender: In this case, there are a lot of NA's. We decide to create a new category "Missing" because probably there are individuals that do not want to share this information.
- Enrolled university: In this case, we cannot assign a new category for NA's since the variable stores already the three possible categories, so we decide to do imputation applying MCA.
- Education level: For this variable we also decide to impute NA's with MCA because we cannot find any similarity of these values with the values of one category.
- Major discipline: Here the "STEM" category is more represented that the rest, so we try to collapse the rest in "Others" since the probabilities in the target are similar between them compared with the "STEM" value; following this approach, the NA's were assigned into "Others"
- Experience: In experience, since we only have 10 missing values, we will use MCA imputation.
- Company size: There are a lot of NA's in this variable, we create another category "Unknown" because is possible that the individual does not know this value.
- Company type: In this case, the number of missing values is very high. We tried to find some pattern in the probabilities of the categories in the target but the NA's are very different from the rest. We decide to assign these NA's values into the "Other" category because the number of individuals in the category are very low and does not make sense to create a "Missing" or "Unknown" category for this variable.
- Last new job: For this variable, analyzing the probabilities we saw some similarity with the category "4", but as we are not sure of this assignation and the number of missing values not exceeds 100, we applied MCA imputation

```
summary(df)
df$target <- factor(df$target, levels=c(1, 0), labels=c("Target.Yes", "Target.No"))
table(df$target)

#Gender
levels(df$gender)[length(levels(df$gender)) + 1] <- "Missing"
df[is.na(df$gender), "gender"] <- "Missing"

#Enrolled university
#levels(df$enrolled_university)[length(levels(df$enrolled_university)) + 1] <- "Missing"
#df[is.na(df$enrolled_university), "enrolled_university"] <- "Missing"
#prop.table(table(df$enrolled_university, df$target)) #missMDA
levels(df$enrolled_university)
nrow(df[is.na(df$enrolled_university),]) #66

#Education level
#levels(df$education_level)[length(levels(df$education_level)) + 1] <- "Missing"
#df[is.na(df$education_level), "education_level"] <- "Missing"
#prop.table(table(df$education_level, df$target)) #missMDA
levels(df$education_level)
nrow(df[is.na(df$education_level),]) #75

#Major discipline
levels(df$major_discipline)[length(levels(df$major_discipline)) + 1] <- "Missing"
df[is.na(df$major_discipline), "major_discipline"] <- "Missing"
prop.table(table(df$major_discipline, df$target)) #STEM & other only?
df[df$major_discipline %in% c("Business Degree", "Arts", "Humanities", "No Major"), "major_discipline"]
table(df$major_discipline)
prop.table(table(df$major_discipline, df$target)) #STEM & other only?
df[df$major_discipline == "Missing", "major_discipline"] <- "Other"
table(df$major_discipline)
prop.table(table(df$major_discipline, df$target)) #STEM & other only?
df$major_discipline <- factor(df$major_discipline)
```

```
levels(df$major_discipline)

#Experience
#levels(df$experience)[length(levels(df$experience)) + 1] <- "Missing"
#df[is.na(df$experience), "experience"] <- "Missing"
#prop.table(table(df$experience, df$target)) #missMDA
levels(df$experience)
nrow(df[is.na(df$experience),]) #10

#Company size
levels(df$company_size)[length(levels(df$company_size)) + 1] <- "Unknown"
df[is.na(df$company_size), "company_size"] <- "Unknown"
prop.table(table(df$company_size, df$target)) #unknown

#Company type
#levels(df$company_type)[length(levels(df$company_type)) + 1] <- "Missing"
#df[is.na(df$company_type), "company_type"] <- "Missing"
#prop.table(table(df$company_type, df$target)) #assign to other
df[is.na(df$company_type), "company_type"] <- "Other"

#Last new job
#levels(df$last_new_job)[length(levels(df$last_new_job)) + 1] <- "Missing"
#df[is.na(df$last_new_job), "last_new_job"] <- "Missing"
#prop.table(table(df$last_new_job, df$target)) #missMDA
levels(df$last_new_job)
nrow(df[is.na(df$last_new_job),]) #91
```

For some variables with NA's we use missDNA library to impute them. As it is a process that take some time, the code appear commented to avoid its execution; the resulting dataframe without NA's is stored and we load it in the next step.

```
#library(missMDA)
#dfimp <- df
#colnames(dfimp)
#res <- MCA(dfimp[, c(4:12)])
#res <- MCA(dfimp[, c(6, 7, 9, 12)])
#vars_dis <- names(dfimp)[c(6, 7, 9, 12)]
#summary(dfimp[,vars_dis])
#nb <- estim_ncpMCA(dfimp[,vars_dis],ncp.max=25)
#res.input<-imputeMCA(dfimp[,vars_dis],ncp=10)

#Result of Imputation
#summary(res.input$completeObs)
#summary(df)
#df$enrolled_university <- res.input$completeObs$enrolled_university
#df$education_level <- res.input$completeObs$education_level
#df$experience <- res.input$completeObs$experience
#df$last_new_job <- res.input$completeObs$last_new_job
#summary(df)
#write.csv(df, "df_imputation.csv", row.names = FALSE)
```

Data Profiling

```
# Check if either training hours or city development index are normally distributed
shapiro.test(df$city_development_index)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  df$city_development_index
## W = 0.76156, p-value < 2.2e-16
```

```
shapiro.test(df$training_hours)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  df$training_hours
## W = 0.83127, p-value < 2.2e-16
```

```
# B ~ X where B is the response factor variable and X is the explanatory cont. variable
str(df)
```

```
## 'data.frame':    4946 obs. of  16 variables:
##  $ enrollee_id          : int  666 21651 8722 5764 7041 10408 4324 26966 13643 5568 ...
##  $ city                 : Factor w/ 117 levels "city_1","city_10",..: 48 53 60 60 73 60 6 47 88 4 .
##  $ city_development_index: num  0.767 0.764 0.624 0.624 0.776 0.624 0.92 0.92 0.666 0.558 ...
##  $ gender               : Factor w/ 4 levels "Female","Male",..: 2 4 4 4 2 2 1 1 4 2 ...
##  $ relevant_experience  : Factor w/ 2 levels "Yes","No": 1 1 2 1 1 1 2 1 2 2 ...
##  $ enrolled_university  : Factor w/ 3 levels "No","Part-Time",..: 1 2 3 1 1 1 3 1 1 1 ...
##  $ education_level      : Factor w/ 5 levels "Primary School",..: 4 3 2 3 3 3 3 3 3 3 ...
##  $ major_discipline     : Factor w/ 2 levels "STEM","Other": 1 1 2 1 2 1 1 1 2 1 ...
##  $ experience           : Factor w/ 22 levels "<1",">20","1",..: 2 5 18 14 1 12 18 2 22 17 ...
##  $ company_size         : Factor w/ 9 levels "<10","10-49",..: 6 9 9 8 4 6 9 3 6 9 ...
##  $ company_type         : Factor w/ 6 levels "Early Stage Startup",..: 2 4 4 6 6 2 4 6 6 4 ...
##  $ last_new_job         : Factor w/ 6 levels ">4","1","2","3",..: 5 2 6 3 2 3 2 1 6 2 ...
##  $ training_hours       : int  8 24 26 7 65 68 24 82 108 92 ...
##  $ target               : Factor w/ 2 levels "Target.Yes","Target.No": 2 1 2 2 2 1 2 2 2 1 ...
##  $ quality              : num  0 6 8 2 0 0 4 0 2 4 ...
##  $ na_count             : num  0 3 4 1 0 0 2 0 1 2 ...
```

```
# Test on means
oneway.test(df$city_development_index ~ df$target)
```

```
##
##  One-way analysis of means (not assuming equal variances)
##
## data:  df$city_development_index and df$target
## F = 451.62, num df = 1.0, denom df = 1628.2, p-value < 2.2e-16
```

```
kruskal.test(df$city_development_index ~ df$target)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  df$city_development_index by df$target
## Kruskal-Wallis chi-squared = 359.53, df = 1, p-value < 2.2e-16
```

```
oneway.test(df$training_hours ~ df$target)
```

```
##
##  One-way analysis of means (not assuming equal variances)
##
## data:  df$training_hours and df$target
## F = 3.0753, num df = 1.0, denom df = 2156.4, p-value = 0.07963
```

```
kruskal.test(df$training_hours ~ df$target)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  df$training_hours by df$target
## Kruskal-Wallis chi-squared = 1.1232, df = 1, p-value = 0.2892
```

```
#Test on variances
fligner.test(df$city_development_index ~ df$target)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  df$city_development_index by df$target
## Fligner-Killeen:med chi-squared = 403.39, df = 1, p-value < 2.2e-16
```

```
fligner.test(df$training_hours ~ df$target)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  df$training_hours by df$target
## Fligner-Killeen:med chi-squared = 2.6715, df = 1, p-value = 0.1022
```

```
#Catdes
str(df)
```

```
## 'data.frame':    4946 obs. of  16 variables:
##  $ enrollee_id           : int  666 21651 8722 5764 7041 10408 4324 26966 13643 5568 ...
##  $ city                  : Factor w/ 117 levels "city_1","city_10",..: 48 53 60 60 73 60 6 47 88 4 .
##  $ city_development_index: num  0.767 0.764 0.624 0.624 0.776 0.624 0.92 0.92 0.666 0.558 ...
##  $ gender                : Factor w/ 4 levels "Female","Male",..: 2 4 4 4 2 2 1 1 4 2 ...
```

```
##  $ relevant_experience   : Factor w/ 2 levels "Yes","No": 1 1 2 1 1 1 2 1 2 2 ...
##  $ enrolled_university   : Factor w/ 3 levels "No","Part-Time",..: 1 2 3 1 1 1 3 1 1 1 ...
##  $ education_level       : Factor w/ 5 levels "Primary School",..: 4 3 2 3 3 3 3 3 3 3 ...
##  $ major_discipline      : Factor w/ 2 levels "STEM","Other": 1 1 2 1 2 1 1 1 2 1 ...
##  $ experience            : Factor w/ 22 levels "<1",">20","1",..: 2 5 18 14 1 12 18 2 22 17 ...
##  $ company_size          : Factor w/ 9 levels "<10","10-49",..: 6 9 9 8 4 6 9 3 6 9 ...
##  $ company_type          : Factor w/ 6 levels "Early Stage Startup",..: 2 4 4 6 6 2 4 6 6 4 ...
##  $ last_new_job          : Factor w/ 6 levels ">4","1","2","3",..: 5 2 6 3 2 3 2 1 6 2 ...
##  $ training_hours        : int  8 24 26 7 65 68 24 82 108 92 ...
##  $ target                : Factor w/ 2 levels "Target.Yes","Target.No": 2 1 2 2 2 1 2 2 2 1 ...
##  $ quality               : num  0 6 8 2 0 0 4 0 2 4 ...
##  $ na_count              : num  0 3 4 1 0 0 2 0 1 2 ...
```

```
res.cat <- catdes(df, 14)
res.cat$test.chi2
```

```
##                           p.value  df
## city                 1.766878e-101 116
## company_size          6.098994e-55   8
## company_type          6.317547e-44   5
## experience            4.590485e-28  22
## enrolled_university   4.928251e-28   3
## relevant_experience   3.393692e-20   1
## education_level       5.009888e-09   5
## last_new_job          8.042028e-07   6
## gender                2.415835e-06   3
## major_discipline      6.442908e-06   1
```

```
res.cat$category
```

```
## $Target.Yes
##                                    Cla/Mod     Mod/Cla     Global      p.value
## city=city_21                     57.142857 33.94957983 14.2943793 8.992131e-95
## company_size=Unknown             38.845144 49.74789916 30.8127780 1.984988e-56
## company_type=Other               36.782334 48.99159664 32.0460979 5.992081e-45
## enrolled_university=Full-Time    37.179487 31.68067227 20.5014153 3.724207e-26
## relevant_experience=No           33.090379 38.15126050 27.7395875 2.015804e-19
## education_level=Graduate         26.930693 68.57142857 61.2616256 2.062562e-09
## city=city_11                     58.333333  2.94117647  1.2131015 1.447377e-08
## gender=Missing                   29.991205 28.65546218 22.9882734 1.603446e-07
## experience=<1                    44.800000  4.70588235  2.5272948 2.949942e-07
## experience=3                     35.112360 10.50420168  7.1977355 1.130762e-06
## major_discipline=STEM            25.598086 80.92436975 76.0614638 4.535028e-06
## experience=2                     35.177866  7.47899160  5.1152446 4.555783e-05
## city=city_128                    55.172414  1.34453782  0.5863324 3.677732e-04
## experience=1                     37.692308  4.11764706  2.6283866 4.559624e-04
## experience=4                     31.720430  9.91596639  7.5212293 4.757892e-04
## city=city_74                     50.000000  1.34453782  0.6469875 1.605583e-03
## last_new_job=NA                  38.461538  2.94117647  1.8398706 2.092943e-03
## experience=5                     30.494505  9.32773109  7.3594824 3.570292e-03
## city=city_101                    52.173913  1.00840336  0.4650222 4.057290e-03
## last_new_job=1                   25.872499 46.72268908 43.4492519 9.082625e-03
## enrolled_university=NA           36.363636  2.01680672  1.3344116 2.455487e-02
```

```
## last_new_job=never            27.463651 14.28571429 12.5151638 3.620120e-02
## city=city_19                  40.625000  1.09243697  0.6469875 3.878893e-02
## city=city_102                 14.666667  0.92436975  1.5163769 4.833198e-02
## city=city_65                  12.244898  0.50420168  0.9906996 4.358249e-02
## city=city_98                   5.263158  0.08403361  0.3841488 4.264933e-02
## city=city_104                 13.114754  0.67226891  1.2333199 3.687820e-02
## city=city_50                  10.869565  0.42016807  0.9300445 2.751642e-02
## experience=15                 16.875000  2.26890756  3.2349373 2.660526e-02
## last_new_job=4                18.067227  3.61344538  4.8119693 2.354601e-02
## experience=18                 12.121212  0.67226891  1.3344116 1.666229e-02
## city=city_99                   4.347826  0.08403361  0.4650222 1.635534e-02
## city=city_159                  4.166667  0.08403361  0.4852406 1.282823e-02
## company_size=<10              18.584071  5.29411765  6.8540235 1.263706e-02
## education_level=Masters       21.276596 20.16806723 22.8063081 1.214591e-02
## city=city_138                  6.060606  0.16806723  0.6672058 8.452341e-03
## city=city_97                   3.846154  0.08403361  0.5256773 7.865192e-03
## city=city_28                   9.433962  0.42016807  1.0715730 7.521024e-03
## city=city_36                   6.976744  0.25210084  0.8693894 4.235659e-03
## city=city_173                  5.263158  0.16806723  0.7682976 2.694176e-03
## city=city_61                   6.521739  0.25210084  0.9300445 2.206352e-03
## city=city_75                  10.714286  0.75630252  1.6983421 2.083906e-03
## city=city_103                 20.437956 18.82352941 22.1593207 1.302616e-03
## experience=17                 10.000000  0.75630252  1.8196522 7.127141e-04
## education_level=Phd           10.810811  1.00840336  2.2442378 4.074426e-04
## company_size=50-99            19.151671 12.52100840 15.7298827 3.805208e-04
## city=city_23                   5.555556  0.25210084  1.0917913 3.689071e-04
## experience=16                 11.805556  1.42857143  2.9114436 2.019153e-04
## company_size=500-999          13.615023  2.43697479  4.3065103 1.243606e-04
## city=city_136                 10.948905  1.26050420  2.7699151 9.538838e-05
## city=city_67                   9.401709  0.92436975  2.3655479 4.644195e-05
## education_level=High School   16.926070  7.31092437 10.3922362 3.732626e-05
## company_size=10000+           16.763006  7.31092437 10.4933279 2.189535e-05
## last_new_job=>4               18.325243 12.68907563 16.6599272 1.606116e-05
## company_type=Funded Startup   13.043478  2.77310924  5.1152446 8.493319e-06
## major_discipline=Other        19.172297 19.07563025 23.9385362 4.535028e-06
## gender=Male                   22.170225 63.69747899 69.1265669 4.102767e-06
## company_size=1000-4999        14.076246  4.03361345  6.8944602 2.644380e-06
## city=city_16                  13.734940  4.78991597  8.3906187 6.134178e-08
## company_size=100-500          15.395683  8.99159664 14.0517590 2.034976e-09
## experience=>20                14.836449 10.67226891 17.3069147 4.717188e-13
## city=city_114                  9.269663  2.77310924  7.1977355 1.365236e-13
## relevant_experience=Yes       20.593173 61.84873950 72.2604125 2.015804e-19
## enrolled_university=No        20.212766 60.67226891 72.2199757 1.967427e-23
## company_type=Pvt Ltd          18.073676 39.57983193 52.6890416 2.440778e-25
##                                  v.test
## city=city_21                   20.653968
## company_size=Unknown           15.828304
## company_type=Other             14.067780
## enrolled_university=Full-Time  10.579125
## relevant_experience=No          9.012408
## education_level=Graduate        5.992802
## city=city_11                    5.667687
## gender=Missing                  5.240267
## experience=<1                   5.126617
```

```
## experience=3                     4.867400
## major_discipline=STEM            4.585223
## experience=2                     4.077320
## city=city_128                    3.562191
## experience=1                     3.505376
## experience=4                     3.494029
## city=city_74                     3.154891
## last_new_job=NA                  3.076716
## experience=5                     2.913826
## city=city_101                    2.873673
## last_new_job=1                   2.608928
## enrolled_university=NA           2.248335
## last_new_job=never               2.094660
## city=city_19                     2.066419
## city=city_102                   -1.974438
## city=city_65                    -2.018084
## city=city_98                    -2.027128
## city=city_104                   -2.087110
## city=city_50                    -2.204113
## experience=15                   -2.217261
## last_new_job=4                  -2.264457
## experience=18                   -2.394076
## city=city_99                    -2.400888
## city=city_159                   -2.488502
## company_size=<10                -2.493837
## education_level=Masters         -2.507876
## city=city_138                   -2.633445
## city=city_97                    -2.657803
## city=city_28                    -2.672848
## city=city_36                    -2.860056
## city=city_173                   -3.000635
## city=city_61                    -3.060951
## city=city_75                    -3.078006
## city=city_103                   -3.215403
## experience=17                   -3.384641
## education_level=Phd             -3.535216
## company_size=50-99              -3.553237
## city=city_23                    -3.561383
## experience=16                   -3.716608
## company_size=500-999            -3.837367
## city=city_136                   -3.902032
## city=city_67                    -4.072847
## education_level=High School     -4.123436
## company_size=10000+             -4.244631
## last_new_job=>4                 -4.313608
## company_type=Funded Startup     -4.452356
## major_discipline=Other          -4.585223
## gender=Male                     -4.606107
## company_size=1000-4999          -4.696669
## city=city_16                    -5.414845
## company_size=100-500            -5.994990
## experience=>20                  -7.233207
## city=city_114                   -7.399670
## relevant_experience=Yes         -9.012408
```

```
## enrolled_university=No         -9.974676
## company_type=Pvt Ltd          -10.401493
##
## $Target.No
##                                 Cla/Mod    Mod/Cla     Global      p.value
## company_type=Pvt Ltd           81.92632 56.8423855 52.6890416 2.440778e-25
## enrolled_university=No         79.78723 75.8785942 72.2199757 1.967427e-23
## relevant_experience=Yes        79.40683 75.5591054 72.2604125 2.015804e-19
## city=city_114                  90.73034  8.5995740  7.1977355 1.365236e-13
## experience=>20                 85.16355 19.4089457 17.3069147 4.717188e-13
## company_size=100-500           84.60432 15.6549521 14.0517590 2.034976e-09
## city=city_16                   86.26506  9.5314164  8.3906187 6.134178e-08
## company_size=1000-4999         85.92375  7.8008520  6.8944602 2.644380e-06
## gender=Male                    77.82977 70.8466454 69.1265669 4.102767e-06
## major_discipline=Other         80.82770 25.4792332 23.9385362 4.535028e-06
## company_type=Funded Startup    86.95652  5.8572950  5.1152446 8.493319e-06
## last_new_job=>4                81.67476 17.9179979 16.6599272 1.606116e-05
## company_size=10000+            83.23699 11.5015974 10.4933279 2.189535e-05
## education_level=High School    83.07393 11.3684771 10.3922362 3.732626e-05
## city=city_67                   90.59829  2.8221512  2.3655479 4.644195e-05
## city=city_136                  89.05109  3.2481363  2.7699151 9.538838e-05
## company_size=500-999           86.38498  4.8988285  4.3065103 1.243606e-04
## experience=16                  88.19444  3.3812567  2.9114436 2.019153e-04
## city=city_23                   94.44444  1.3578275  1.0917913 3.689071e-04
## company_size=50-99             80.84833 16.7465389 15.7298827 3.805208e-04
## education_level=Phd            89.18919  2.6357827  2.2442378 4.074426e-04
## experience=17                  90.00000  2.1565495  1.8196522 7.127141e-04
## city=city_103                  79.56204 23.2161874 22.1593207 1.302616e-03
## city=city_75                   89.28571  1.9968051  1.6983421 2.083906e-03
## city=city_61                   93.47826  1.1448349  0.9300445 2.206352e-03
## city=city_173                  94.73684  0.9584665  0.7682976 2.694176e-03
## city=city_36                   93.02326  1.0649627  0.8693894 4.235659e-03
## city=city_28                   90.56604  1.2779553  1.0715730 7.521024e-03
## city=city_97                   96.15385  0.6656017  0.5256773 7.865192e-03
## city=city_138                  93.93939  0.8253461  0.6672058 8.452341e-03
## education_level=Masters        78.72340 23.6421725 22.8063081 1.214591e-02
## company_size=<10               81.41593  7.3482428  6.8540235 1.263706e-02
## city=city_159                  95.83333  0.6123536  0.4852406 1.282823e-02
## city=city_99                   95.65217  0.5857295  0.4650222 1.635534e-02
## experience=18                  87.87879  1.5441960  1.3344116 1.666229e-02
## last_new_job=4                 81.93277  5.1916933  4.8119693 2.354601e-02
## experience=15                  83.12500  3.5410011  3.2349373 2.660526e-02
## city=city_50                   89.13043  1.0915868  0.9300445 2.751642e-02
## city=city_104                  86.88525  1.4110756  1.2333199 3.687820e-02
## city=city_98                   94.73684  0.4792332  0.3841488 4.264933e-02
## city=city_65                   87.75510  1.1448349  0.9906996 4.358249e-02
## city=city_102                  85.33333  1.7039404  1.5163769 4.833198e-02
## city=city_19                   59.37500  0.5058573  0.6469875 3.878893e-02
## last_new_job=never             72.53635 11.9542066 12.5151638 3.620120e-02
## enrolled_university=NA         63.63636  1.1182109  1.3344116 2.455487e-02
## last_new_job=1                 74.12750 42.4121406 43.4492519 9.082625e-03
## city=city_101                  47.82609  0.2928647  0.4650222 4.057290e-03
## experience=5                   69.50549  6.7358892  7.3594824 3.570292e-03
## last_new_job=NA                61.53846  1.4909478  1.8398706 2.092943e-03
```

```
## city=city_74                   50.00000  0.4259851  0.6469875 1.605583e-03
## experience=4                   68.27957  6.7625133  7.5212293 4.757892e-04
## experience=1                   62.30769  2.1565495  2.6283866 4.559624e-04
## city=city_128                  44.82759  0.3461129  0.5863324 3.677732e-04
## experience=2                   64.82213  4.3663472  5.1152446 4.555783e-05
## major_discipline=STEM          74.40191 74.5207668 76.0614638 4.535028e-06
## experience=3                   64.88764  6.1501597  7.1977355 1.130762e-06
## experience=<1                  55.20000  1.8370607  2.5272948 2.949942e-07
## gender=Missing                 70.00880 21.1927583 22.9882734 1.603446e-07
## city=city_11                   41.66667  0.6656017  1.2131015 1.447377e-08
## education_level=Graduate       73.06931 58.9456869 61.2616256 2.062562e-09
## relevant_experience=No         66.90962 24.4408946 27.7395875 2.015804e-19
## enrolled_university=Full-Time 62.82051 16.9595314 20.5014153 3.724207e-26
## company_type=Other             63.21767 26.6773163 32.0460979 5.992081e-45
## company_size=Unknown           61.15486 24.8136315 30.8127780 1.984988e-56
## city=city_21                   42.85714  8.0670927 14.2943793 8.992131e-95
##                                    v.test
## company_type=Pvt Ltd           10.401493
## enrolled_university=No          9.974676
## relevant_experience=Yes         9.012408
## city=city_114                   7.399670
## experience=>20                  7.233207
## company_size=100-500            5.994990
## city=city_16                    5.414845
## company_size=1000-4999          4.696669
## gender=Male                     4.606107
## major_discipline=Other          4.585223
## company_type=Funded Startup     4.452356
## last_new_job=>4                 4.313608
## company_size=10000+             4.244631
## education_level=High School     4.123436
## city=city_67                    4.072847
## city=city_136                   3.902032
## company_size=500-999            3.837367
## experience=16                   3.716608
## city=city_23                    3.561383
## company_size=50-99              3.553237
## education_level=Phd             3.535216
## experience=17                   3.384641
## city=city_103                   3.215403
## city=city_75                    3.078006
## city=city_61                    3.060951
## city=city_173                   3.000635
## city=city_36                    2.860056
## city=city_28                    2.672848
## city=city_97                    2.657803
## city=city_138                   2.633445
## education_level=Masters         2.507876
## company_size=<10                2.493837
## city=city_159                   2.488502
## city=city_99                    2.400888
## experience=18                   2.394076
## last_new_job=4                  2.264457
## experience=15                   2.217261
```

```
## city=city_50                    2.204113
## city=city_104                   2.087110
## city=city_98                    2.027128
## city=city_65                    2.018084
## city=city_102                   1.974438
## city=city_19                   -2.066419
## last_new_job=never             -2.094660
## enrolled_university=NA         -2.248335
## last_new_job=1                 -2.608928
## city=city_101                  -2.873673
## experience=5                   -2.913826
## last_new_job=NA                -3.076716
## city=city_74                   -3.154891
## experience=4                   -3.494029
## experience=1                   -3.505376
## city=city_128                  -3.562191
## experience=2                   -4.077320
## major_discipline=STEM          -4.585223
## experience=3                   -4.867400
## experience=<1                  -5.126617
## gender=Missing                 -5.240267
## city=city_11                   -5.667687
## education_level=Graduate       -5.992802
## relevant_experience=No         -9.012408
## enrolled_university=Full-Time -10.579125
## company_type=Other            -14.067780
## company_size=Unknown          -15.828304
## city=city_21                  -20.653968
```

```
res.cat$quanti.var
```

```
##                              Eta2       P-value
## city_development_index 0.109671247 6.648278e-127
## na_count               0.038763117  2.066963e-44
## quality                0.036106693  1.965044e-41
## enrollee_id            0.003306029  5.209160e-05
```

```
res.cat$quanti
```

```
## $Target.Yes
##                            v.test Mean in category Overall mean sd in category
## na_count                13.844985     1.447059e+00 1.039426e+00      1.1980954
## quality                 13.362170     2.954622e+00 2.157501e+00      2.4012514
## enrollee_id              4.043305     1.785119e+04 1.687536e+04   9164.6365607
## city_development_index -23.287858     7.560092e-01 8.288837e-01      0.1435887
##                          Overall sd       p.value
## na_count                  1.1653884  1.364089e-43
## quality                   2.3612447  1.005879e-40
## enrollee_id            9552.7825825  5.270305e-05
## city_development_index    0.1238625 5.885515e-120
##
## $Target.No
##                            v.test Mean in category Overall mean sd in category
```

```
## city_development_index  23.287858     8.519723e-01 8.288837e-01      0.1070269
## enrollee_id             -4.043305     1.656619e+04 1.687536e+04   9651.9505598
## quality                -13.362170     1.904952e+00 2.157501e+00      2.2912918
## na_count               -13.844985     9.102769e-01 1.039426e+00      1.1244172
##                           Overall sd       p.value
## city_development_index    0.1238625 5.885515e-120
## enrollee_id            9552.7825825  5.270305e-05
## quality                   2.3612447  1.005879e-40
## na_count                  1.1653884  1.364089e-43
```

The following conclusions are reached after the Data Profiling: *It is observed using the Shapiro-Wilk Test that neither* ***city_development_index*** *nor* ***training_hours*** *follow a normal distribution.* It is tested if the Target Factor is impacted by the numerical variables. In other words, does a certain value in these continuous variables affect the outcome.? Observe that even though both available tests are done, the valid results are the ones from Kruskal-Wallis test, since it does not assume normality on the explanatory variable. The output shows a 0 p-value for `city_development_index` and around 0.29 for `training_hours`. In the first case it is clear, there is influence on the target depending on the city where the employee resides. For the number of training hours the null hypothesis is not rejected, so the mean is spread equally around the tearget variable *It is also important to check if dispersion of the numerical variables affects the target, and we get the similar answer as in the means.* Using the `catdes()` package from `FactoMinR` global association between the target factor and categorical variables is assessed. The chi-squared test gives p-values a lot lower than 0.05 for all the factors. *The `quanti.var` output shows if the quantitative variables influence the target. Since all p-values are very low, the is answer Yes. Similarly, the `quanti` output shows if the mean in category varies with the overall mean.

Interpretation of all the results before modelling

Separation between Train and Test

```
# We upload the new dataframe with imputation results
df <- read.csv("df_imputation.csv",header=T, sep=",", na.strings="NA")
for(i in 1:ncol(df)){
  if(is.character(df[, i])){
    df[, i] <-factor(df[, i])
  }
}
#summary(df)
df$quality <-NULL
df$na_count <-NULL
```

```
set.seed(130798)
s <- sample(1:nrow(df),round(0.75*nrow(df),0))
dfall<-df
df <- df[s,]
dftest <-dfall[-s,]
```

Modelling the Target

Observation. To simplify and differentiate kinds of models, the following table states the notation

m0	Null model
m1, m11, m12...	Variations with numerical explanatory variables
m2, m21, m22...	Variations with factor variables
m3, m31, m32...	Variations with numerical and factor variables

The first model to try is the null model. A quick verifaction that the intercept in fact corresponds with the logit link function is performed.

```
# First model
m0 <- glm(target ~ 1, family="binomial", data=df)
ptt <- prop.table(table(df$target))
summary(m0)
```

```
##
## Call:
## glm(formula = target ~ 1, family = "binomial", data = df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.7411  -0.7411  -0.7411  -0.7411   1.6890
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.15179    0.03843  -29.97   <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: 4090.4  on 3709  degrees of freedom
## Residual deviance: 4090.4  on 3709  degrees of freedom
## AIC: 4092.4
##
## Number of Fisher Scoring iterations: 4
```

```
oddm0 <- ptt[2] / (1- ptt[2])
log(oddm0)
```

```
## Target.Yes
##  -1.151793
```

Modelling the target using numeric variables

The numerical variables that we originally have are training hours and city development index. A significant improvement with regards to the AIC and the Deviance is observed with respect to the null model. The step function suggests to remove the training hours variable, as the AIC slightly decreases. We have labeled the models in this family as m1. Note that a quick overview of the marginal plots of `m11` suggests a variable transformation on city development index. Both a polynomial of degree 2 (`m12`) and a logarithmic transformation are assessed (`m13`), getting a significantly better fit with the latter option, as the marginal plots show. However, the AIC worsens slightly in comparison to the polynomial model, although it is still better than the first model. We choose to keep model `m13`. Let us check also for influential observations: The Influence plot shows several observations with a large Cook's distance (based on the area of the circle). To

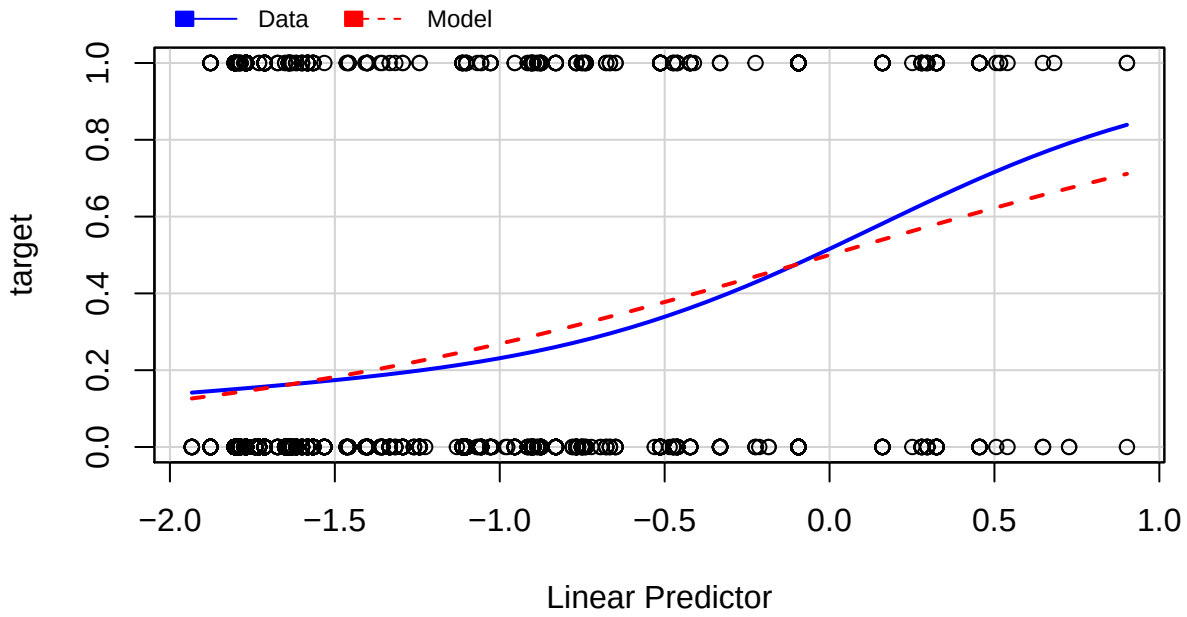
filter the important observations, we use a BoxPlot for the Cook's distance which will yield the 10 instances with the largest Cook's D.It is prudent to delete these observations, as it is a general way of proceeding with a posteriori influential data. The model is reestimated, and although the AIC decreases, this is not relevant because the data frame has 2 fewer observations. The Residual Deviance also shows a significant drop.

With regards to the final numerical model, a rule of thumb is applied to check for goodness of fit since the data is not aggregated. Since the residual deviance and the degrees of freedom are of the same order, the model can be labeled a good fit.

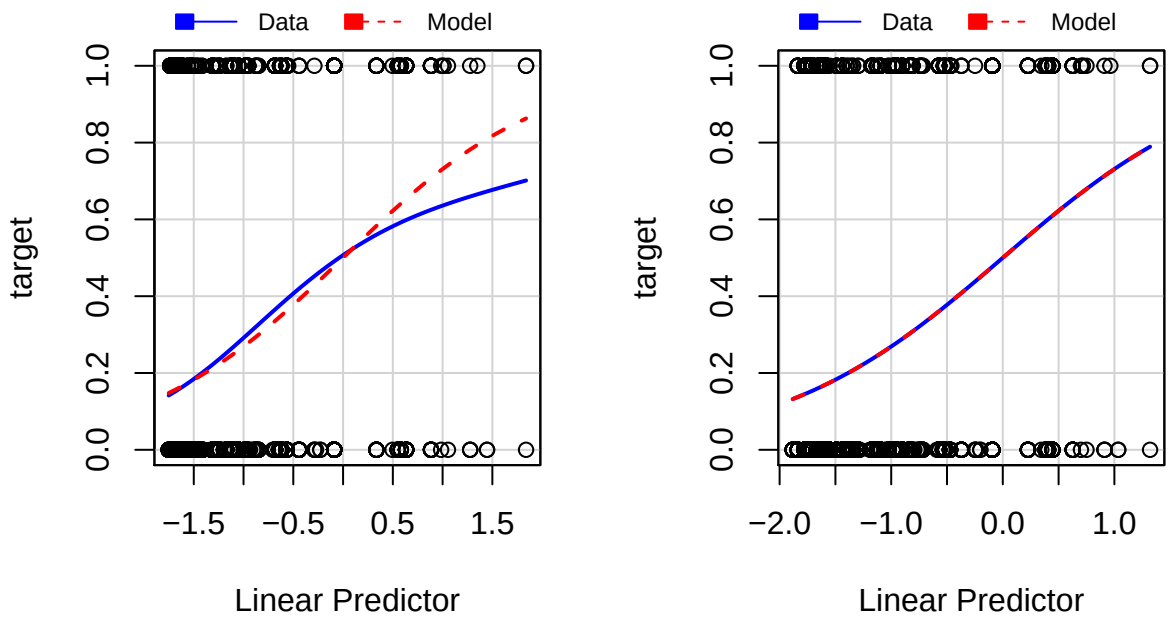
Finally, with function `allEffects()` we see the logarithmic effect of the explanatory variable.

```
m1 <- glm(target ~ . , family="binomial", data=df[, c(3, 13, 14)])
m11<- step(m1) #city_development_index + training_hours
m11 <- step(m1, k=log(nrow(df))) #BIC case   city_development_index
```

```
marginalModelPlot(m11) #some transformation needed
```



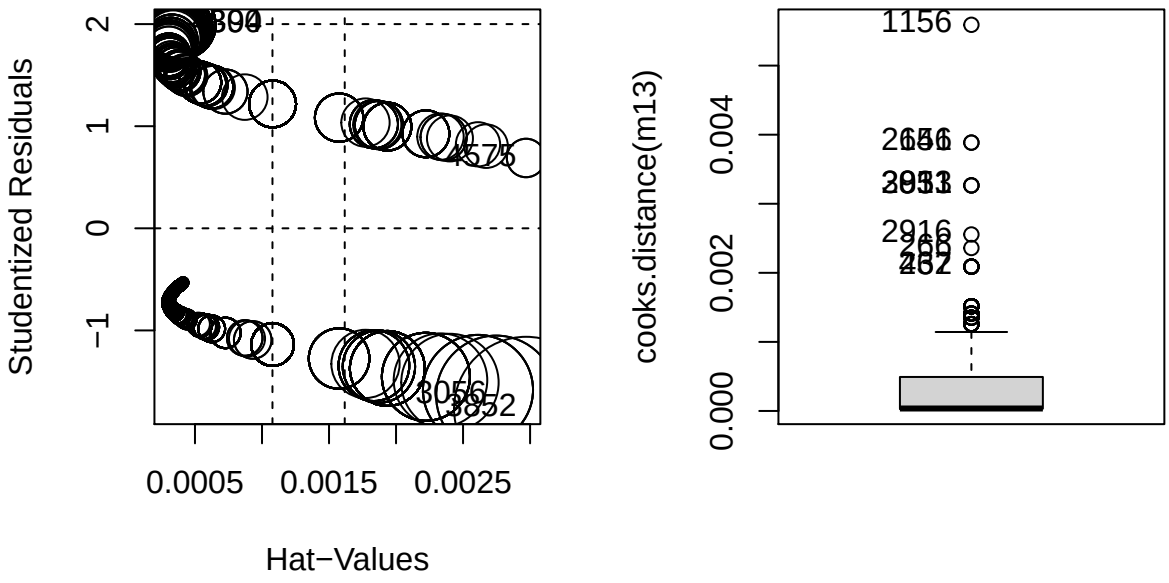

```
m12 <- glm(target ~ poly(city_development_index,2), family="binomial", data=df)
m13 <- glm(target ~ log(city_development_index), family="binomial", data=df)
#AIC(m11, m12, m13)

par(mfrow = c(1, 2))
marginalModelPlot(m12)
marginalModelPlot(m13) #better fit
```




```
influencePlot(m13)
```

```
##         StudRes          Hat        CookD
## 4800  1.9931671 0.0004000836 0.0012572246
## 4575  0.6887775 0.0029700418 0.0003991642
## 4394  1.9931671 0.0004000836 0.0012572246
## 3056 -1.6384995 0.0027469216 0.0038851200
## 3852 -1.7675654 0.0029700418 0.0055908864
```

```
cook <- Boxplot(cooks.distance(m13))
```




```
cookd <- sort(cooks.distance(m13)[cook], decreasing=TRUE)
cookd #10 influent observations
```

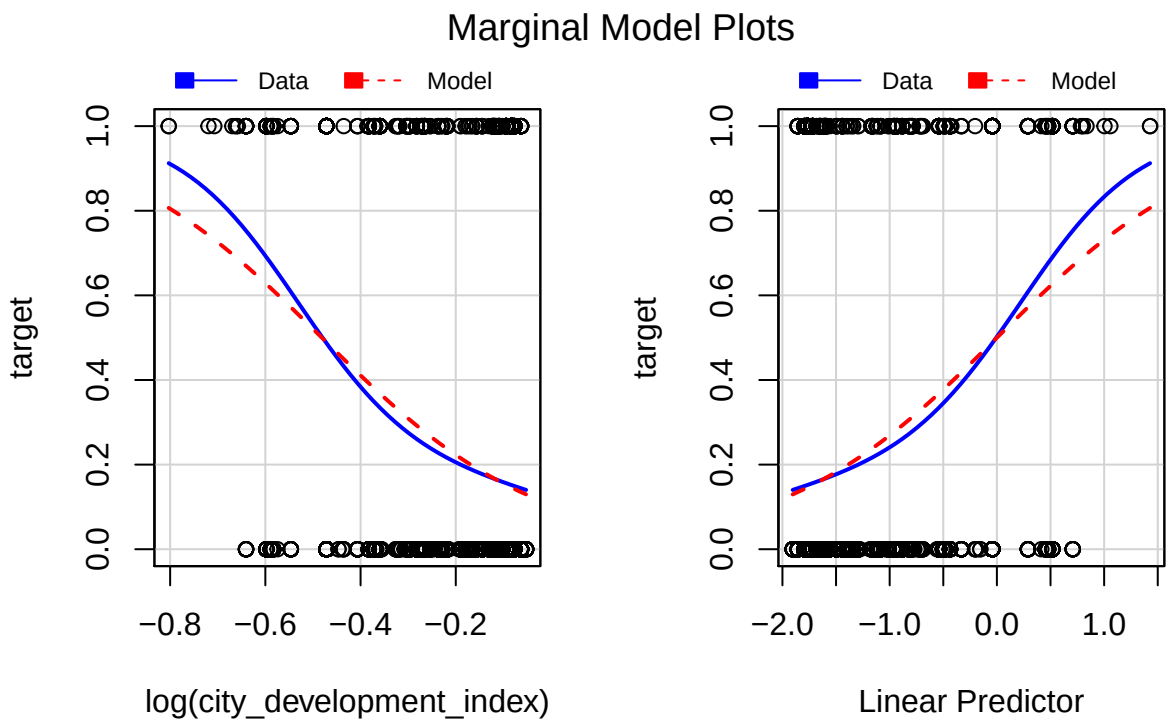
```
##        3852        3056         204        4271        3648        1969
## 0.005590886 0.003885120 0.003885120 0.003265413 0.003265413 0.003265413
##        1350        2998        1561         487
## 0.002555228 0.002359506 0.002089487 0.002089487
```

```
df <- df[!(rownames(df) %in% names(cookd)),]

m131 <- glm(target ~ log(city_development_index), family="binomial", data=df)
AIC(m13, m131)
```

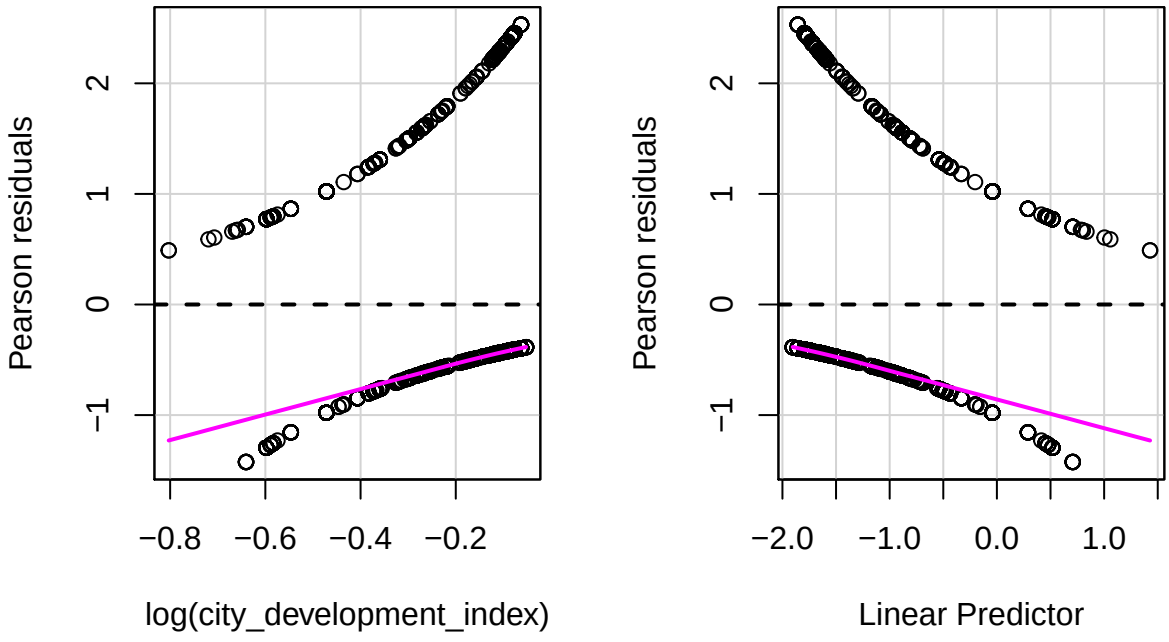
```
## Warning in AIC.default(m13, m131): models are not all fitted to the same number
## of observations
```

```
##      df      AIC
## m13   2 3731.910
## m131  2 3706.603
```

```
#summary(m13)
#summary(m131)
marginalModelPlots(m131)
```




```
residualPlots(m131) #some values of the tail in m13 are removed
```




```
##                 Test stat Pr(>|Test stat|)
```

```
## log(city_development_index)     8.417         0.003717 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#plot(allEffects((m131)))

# Temporarily remove outliers: compare models
#dfcheck <-df
#ll <- which(df$enrollee_id %in% regmout_th)
#dfcheck <- dfcheck[-ll,]
#ll <- which(df$enrollee_id %in% regeout_th)
#df <- df[-ll, ]
#ll <- which(df$enrollee_id %in% regmout_city)
#dfcheck <- dfcheck[-ll, ]

#m14 <- glm(target ~ log(city_development_index) + training_hours, family="binomial", data=dfcheck)
#AIC(m14, m13)
```

Introducing factor variables

The first important thing to note here is the difference between the Akaike Information Criterion (AIC from now on) and the Bayesian Information Criterion(BIC). The AIC has many different formulas that describe it, but the main idea is that takes into account the number of parameters in the model and the total of number observations. It applies a penalty based on the number of parameters. The BIC, however, applies a larger penalty. So in search of the best model, different results will be achieved according to the criteria used.

First, a general model m2 containing all . A significant drop in comparison with respect to the deviance from the null model is noted, however, there is an egregious amount of parameters and that is not good news. It will be shown that reducing the amount of factors (especially the ones with many factor levels)

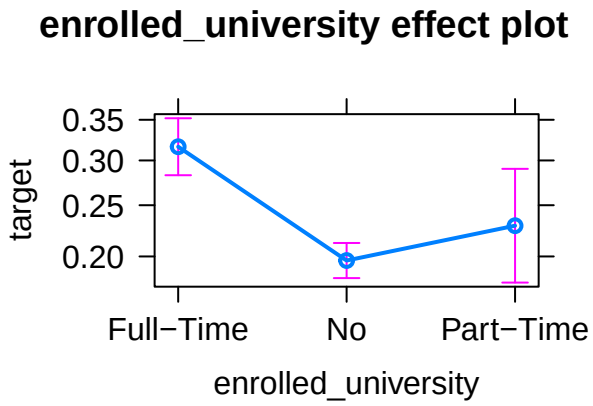
In this specific m2 model with the results printed in Appendix C, many p-values close to 1 are observed for some factor levels. This is clearly noticeable with the city factor. The step functions are now applied, for AIC and BIC criteria respectively. Both eliminate several variables: * Using AIC criteria we obtain a model that eliminates factors `experience`, `company_type` and `gender` * Using BIC criteria we obtain a model that only considers 3 regressors: `enrolled_university` , `major_discipline` and `company_size`. We choose the latter model because it will be much easier to check for factor interactions in the enxt section. The `Anova()` function gives low p-values for all regresssors, which suggests a good fit. gives got results with the latter model, showing that all regressors are useful. Let us follow with a bit of interpretation:

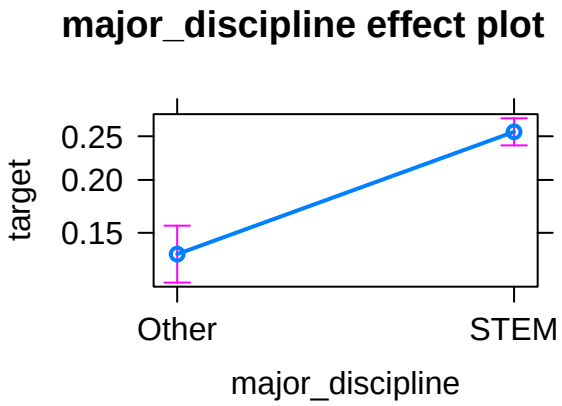
```
m00 <-glm(target ~ 1, family="binomial", data=df)
summary(m00)
m2 <- glm(target ~ ., family="binomial", data=df[, c(2, 4:12, 14)])
AIC(m2);BIC(m2)

# AIC option
#m21<-step(m2, trace=0)
#m21$anova
#summary(m21)

#BIC option
m211<-step(m2, k=log(nrow(df)), trace=0)
m211$anova
```

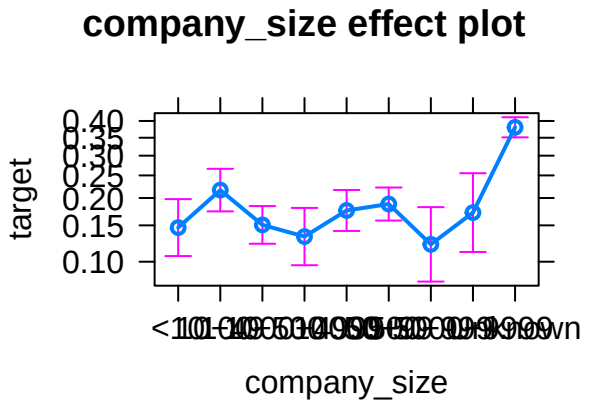
```r
summary(m211)
```

```
##
## Call:
## glm(formula = target ~ enrolled_university + major_discipline +
##     company_size, family = "binomial", data = df[, c(2, 4:12,
##     14)])
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.2626  -0.7110  -0.5838  -0.3718   2.3266
##
## Coefficients:
##                               Estimate Std. Error z value Pr(>|z|)
## (Intercept)                   -1.88409    0.21639  -8.707  < 2e-16 ***
## enrolled_universityNo         -0.63887    0.09468  -6.747 1.50e-11 ***
## enrolled_universityPart-Time -0.44335    0.18114  -2.448   0.0144 *
## major_disciplineSTEM           0.80557    0.10447   7.711 1.25e-14 ***
## company_size10-49              0.47738    0.23042   2.072   0.0383 *
## company_size100-500            0.03440    0.22126   0.155   0.8765
## company_size1000-4999         -0.11436    0.26185  -0.437   0.6623
## company_size10000+             0.21821    0.22710   0.961   0.3366
## company_size50-99              0.30009    0.21328   1.407   0.1594
## company_size500-999           -0.21338    0.30400  -0.702   0.4827
## company_size5000-9999          0.19096    0.31521   0.606   0.5446
## company_sizeUnknown            1.27656    0.19651   6.496 8.24e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4084.9  on 3699  degrees of freedom
## Residual deviance: 3769.0  on 3688  degrees of freedom
## AIC: 3793
##
## Number of Fisher Scoring iterations: 4
```

```r
#vif(m211)
#Anova(m21, test="LR")
#Anova(m211, test="LR")

#AIC(m00, m211)
plot(allEffects(m211), axes=list(y=list(lab="target")))
```







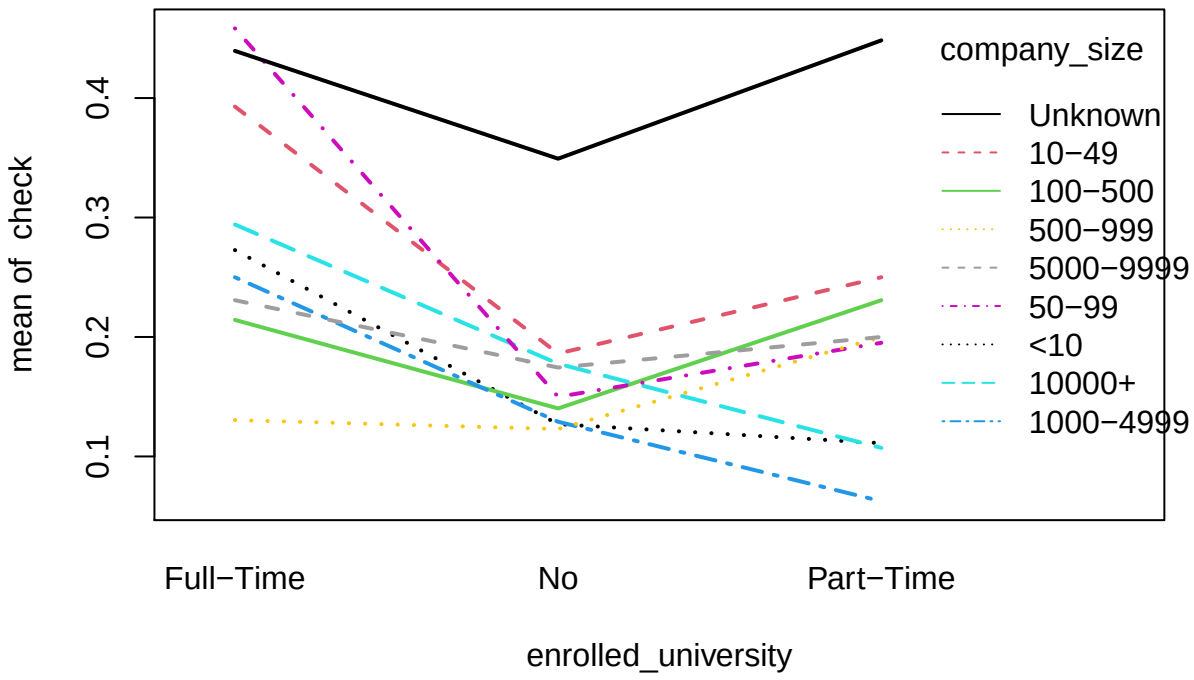
Factor interactions

We go ahead and study some more properties for this model. Note that all the factors are additive, and we want to check interaction. We have four factors, and we calculate the different factor interactions:

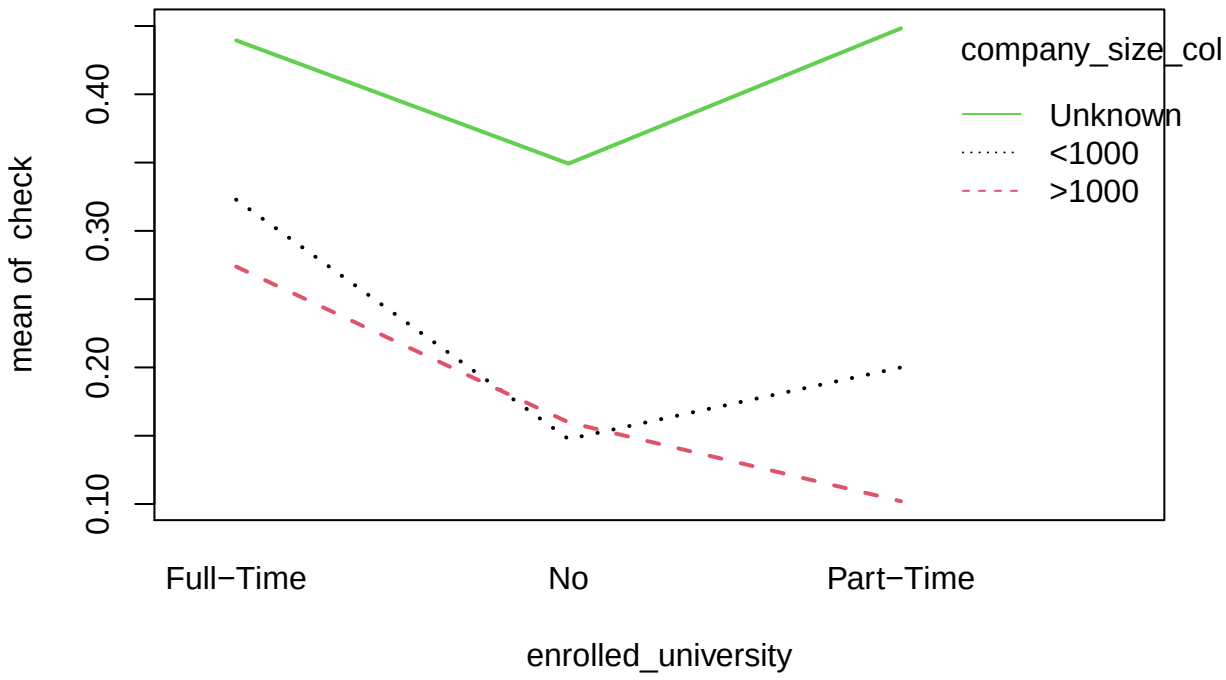
$$\binom{3}{2} = 3$$

Firstly, two by two factor interactions show that gender and en

```
check <- as.numeric(df$target) - 1

# enrolled_university - major_discipline
#with(df, interaction.plot(enrolled_university, major_discipline, check, lwd = 2, col = 1:2))
m22 <- glm(target ~ enrolled_university + major_discipline, family="binomial", data=df)
m23 <- glm(target ~ enrolled_university * major_discipline, family="binomial", data=df)
Anova(m23, test="LR") #No interaction

# enrolled_university - company_size
with(df, interaction.plot(enrolled_university, company_size, check, lwd = 2, col = 1:9))
```




```
m22 <- glm(target ~ enrolled_university + company_size, family="binomial", data=df)
m23 <- glm(target ~ enrolled_university * company_size, family="binomial", data=df)
Anova(m23, test="LR") #No interaction, not clear...

prop.table(table(df$enrolled_university, df$company_size), 2)
table(df$company_size)
df$company_size_col <- fct_collapse(df$company_size, "<1000"= c("<10", "10-49", "50-99", "100-500", "500
table(df$company_size_col)
with(df, interaction.plot(enrolled_university, company_size_col, check, lwd = 2, col = 1:9))
```




```
# We try enrolled_university *company_size_col
m24 <- glm(target ~ enrolled_university * company_size_col, family="binomial", data=df)
Anova(m24, test="LR")
AIC(m23, m24)
summary(m23)
summary(m24)
BIC(m23, m24)
BIC(m2)

# major_discipline - company_size
#with(df, interaction.plot(major_discipline, company_size, check, lwd = 2, col = 1:2))
m22 <- glm(target ~ major_discipline + company_size, family="binomial", data=df)
m23 <- glm(target ~ major_discipline * company_size, family="binomial", data=df)
Anova(m23, test="LR") #No interaction
```

The following table summarizes the different possibilities factor interactions:

Models	Deviance	n-p	AIC
`enrolled_university + major_discipline`	3946.5	3696	3954.5
`enrolled_university * major_discipline`	3942.8	3694	3954.8
`enrolled_university + company_size`	3834	3689	3856
`enrolled_university * company_size`	3809	3673	3863
`major_discipline + company_size`	3813.5	3690	3833.5
`major_discipline * company_size`	3804.3	3682	3840.3

Best model

We end up with the best three models so far. The table below summarizes these models according to their Residual Deviance and their BIC. We end up chooosing model `m13` because it has the lowest BIC.

```
m3 <- glm(target ~ log(city_development_index) + enrolled_university + major_discipline + company_size,
summary(m3)
```

```
##
## Call:
## glm(formula = target ~ log(city_development_index) + enrolled_university +
##     major_discipline + company_size, family = "binomial", data = df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8760  -0.6761  -0.4917  -0.3127   2.4920
##
## Coefficients:
##                               Estimate Std. Error z value Pr(>|z|)
## (Intercept)                   -2.91921    0.23699 -12.318  < 2e-16 ***
## log(city_development_index)   -4.20814    0.24445 -17.215  < 2e-16 ***
## enrolled_universityNo         -0.40489    0.10038  -4.033 5.50e-05 ***
## enrolled_universityPart-Time -0.31651    0.18995  -1.666   0.0957 .
## major_disciplineSTEM           0.65921    0.11036   5.973 2.33e-09 ***
## company_size10-49              0.37147    0.24212   1.534   0.1250
## company_size100-500            0.08381    0.23095   0.363   0.7167
## company_size1000-4999         -0.02022    0.27240  -0.074   0.9408
## company_size10000+             0.28032    0.23731   1.181   0.2375
## company_size50-99              0.27582    0.22339   1.235   0.2169
## company_size500-999           -0.15387    0.31716  -0.485   0.6276
## company_size5000-9999          0.28832    0.32755   0.880   0.3787
## company_sizeUnknown            1.35312    0.20566   6.579 4.73e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4084.9  on 3699  degrees of freedom
## Residual deviance: 3462.6  on 3687  degrees of freedom
## AIC: 3488.6
##
## Number of Fisher Scoring iterations: 4
```

```
m31 <- glm(target ~ log(city_development_index) + enrolled_university + major_discipline + company_size
summary(m31)
```

```
##
## Call:
## glm(formula = target ~ log(city_development_index) + enrolled_university +
##     major_discipline + company_size_col, family = "binomial",
##     data = df)
##
## Deviance Residuals:
```

```
##     Min       1Q   Median       3Q      Max
## -1.8820  -0.6743  -0.4738  -0.3380   2.4277
##
## Coefficients:
##                              Estimate Std. Error z value Pr(>|z|)
## (Intercept)                  -2.75301    0.15577 -17.674  < 2e-16 ***
## log(city_development_index)  -4.24099    0.24403 -17.379  < 2e-16 ***
## enrolled_universityNo        -0.40674    0.10031  -4.055 5.02e-05 ***
## enrolled_universityPart-Time -0.31484    0.18996  -1.657   0.0974 .
## major_disciplineSTEM          0.66068    0.11027   5.991 2.08e-09 ***
## company_size_col>1000         0.01484    0.12254   0.121   0.9036
## company_size_colUnknown       1.17960    0.09848  11.979  < 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: 4084.9  on 3699  degrees of freedom
## Residual deviance: 3470.1  on 3693  degrees of freedom
## AIC: 3484.1
##
## Number of Fisher Scoring iterations: 4
```

```
m32 <- glm(target ~ log(city_development_index) + enrolled_university * company_size_col + major_discipl
summary(m32)
```

```
##
## Call:
## glm(formula = target ~ log(city_development_index) + enrolled_university *
##     company_size_col + major_discipline, family = "binomial",
##     data = df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8426  -0.6697  -0.4669  -0.3261   2.4562
##
## Coefficients:
##                                                    Estimate Std. Error
## (Intercept)                                       -2.489717   0.186039
## log(city_development_index)                       -4.219902   0.244814
## enrolled_universityNo                             -0.742248   0.165637
## enrolled_universityPart-Time                      -0.551527   0.287386
## company_size_col>1000                             -0.007614   0.298179
## company_size_colUnknown                            0.757026   0.179496
## major_disciplineSTEM                               0.665158   0.110119
## enrolled_universityNo:company_size_col>1000        0.088418   0.329220
## enrolled_universityPart-Time:company_size_col>1000 -0.551788   0.619952
## enrolled_universityNo:company_size_colUnknown      0.596906   0.213501
## enrolled_universityPart-Time:company_size_colUnknown  0.591548   0.414765
##                                                    z value Pr(>|z|)
## (Intercept)                                       -13.383  < 2e-16 ***
## log(city_development_index)                       -17.237  < 2e-16 ***
## enrolled_universityNo                              -4.481 7.42e-06 ***
## enrolled_universityPart-Time                       -1.919  0.05497 .
```

```
## company_size_col>1000                                    -0.026  0.97963
## company_size_colUnknown                                   4.218 2.47e-05 ***
## major_disciplineSTEM                                      6.040 1.54e-09 ***
## enrolled_universityNo:company_size_col>1000               0.269  0.78826
## enrolled_universityPart-Time:company_size_col>1000       -0.890  0.37344
## enrolled_universityNo:company_size_colUnknown             2.796  0.00518 **
## enrolled_universityPart-Time:company_size_colUnknown      1.426  0.15380
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4084.9  on 3699  degrees of freedom
## Residual deviance: 3459.5  on 3689  degrees of freedom
## AIC: 3481.5
##
## Number of Fisher Scoring iterations: 4
```

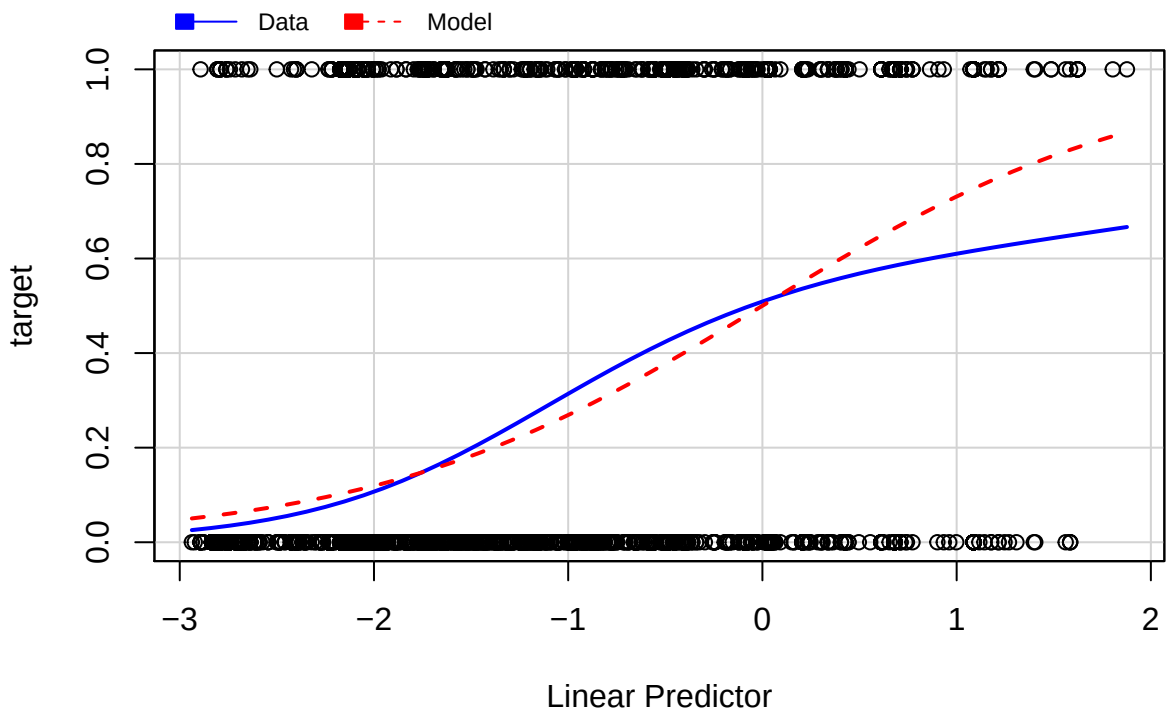
```
anova(m31, m32, test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: target ~ log(city_development_index) + enrolled_university +
##     major_discipline + company_size_col
## Model 2: target ~ log(city_development_index) + enrolled_university *
##     company_size_col + major_discipline
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1      3693     3470.1
## 2      3689     3459.5  4   10.597  0.03149 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

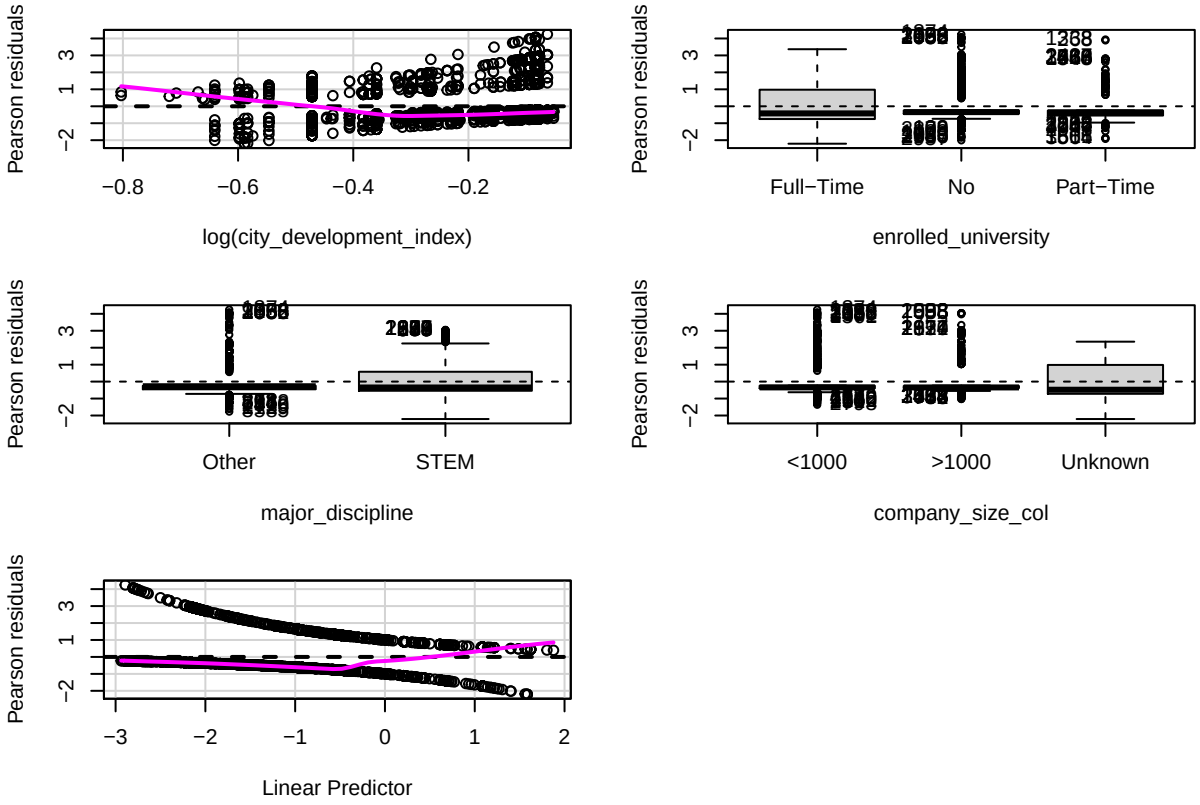
```
BIC(m3, m31, m32)
```

```
##     df      BIC
## m3  13 3569.370
## m31  7 3527.602
## m32 11 3549.870
```

```
marginalModelPlot(m31)
```

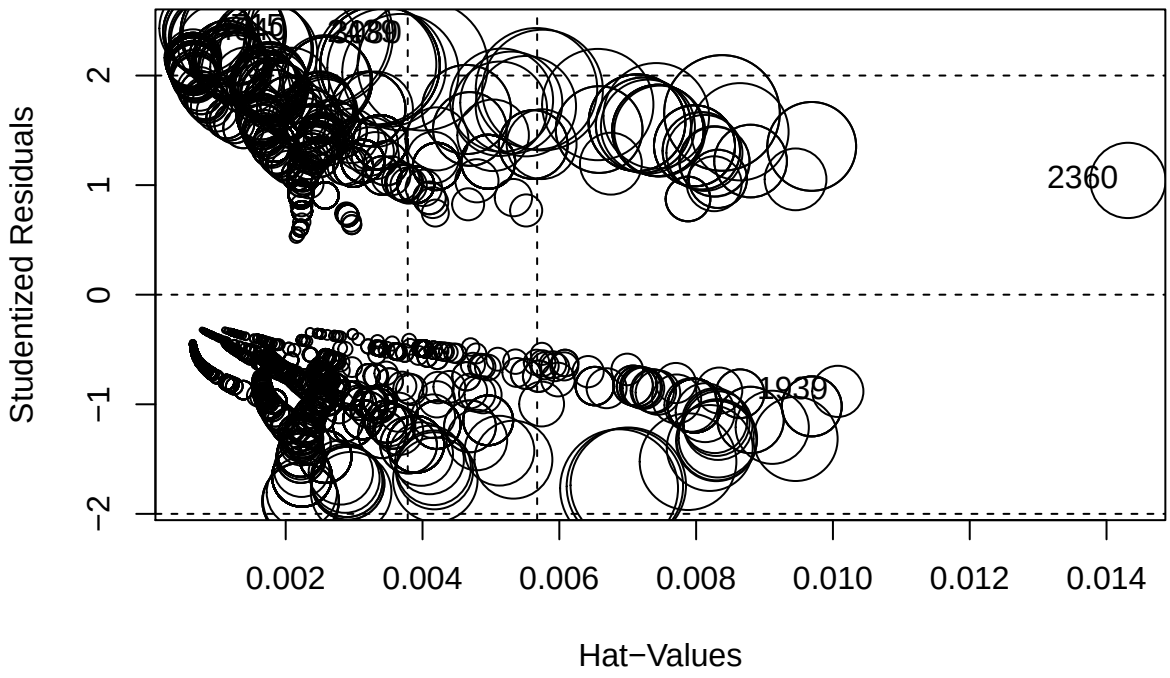



```
residualPlots(m31)
```



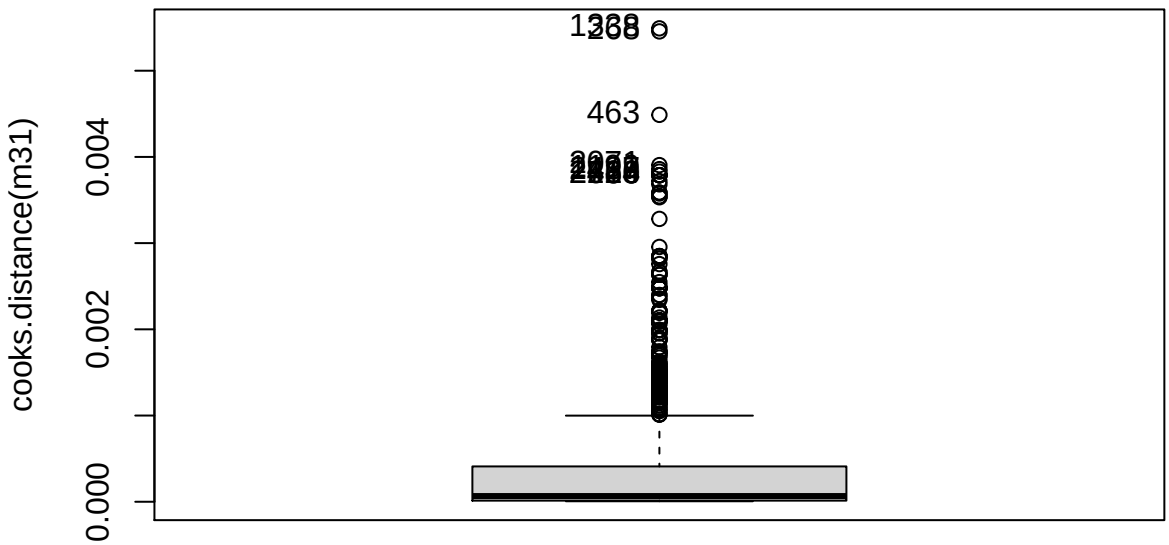

```
##                               Test stat Pr(>|Test stat|)
## log(city_development_index)    16.618        4.572e-05 ***
## enrolled_university
## major_discipline
## company_size_col
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
influencePlot(m31)
```




```
##        StudRes          Hat        CookD
## 4040  2.396571 0.0008409606 0.0019911097
## 3080  2.365347 0.0025176483 0.0054558582
## 2439  2.376315 0.0024654093 0.0054914187
## 4715  2.430624 0.0007980146 0.0020603622
## 2360  1.042831 0.0143131740 0.0015017039
## 1939 -0.882225 0.0100717567 0.0006933083
```

```
cook <- Boxplot(cooks.distance(m31))
```




```
cookd <- sort(cooks.distance(m31)[cook], decreasing=TRUE)
cookd
```

```
##        2439        3080        1930        3769        2942        4177
## 0.005491419 0.005455858 0.004488095 0.003904845 0.003858279 0.003829276
##        2369        2720        2812         561
## 0.003788255 0.003788255 0.003788255 0.003788255
```

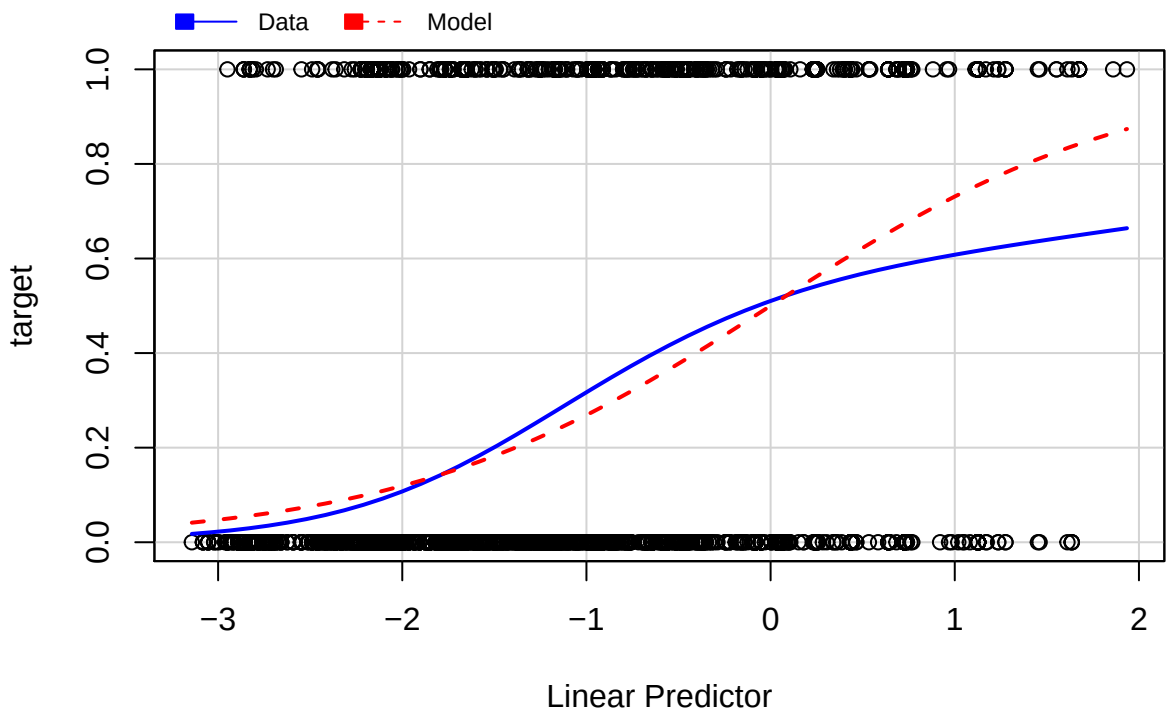
```
df<-df[!(rownames(df) %in% names(cookd)),]

m4 <- glm(target ~ log(city_development_index) + enrolled_university + major_discipline + company_size_
```

Model	Deviance	BIC
m3	3462.6	3569.37
m31	3470.1	3527.602
m32	3459.5	3459.870

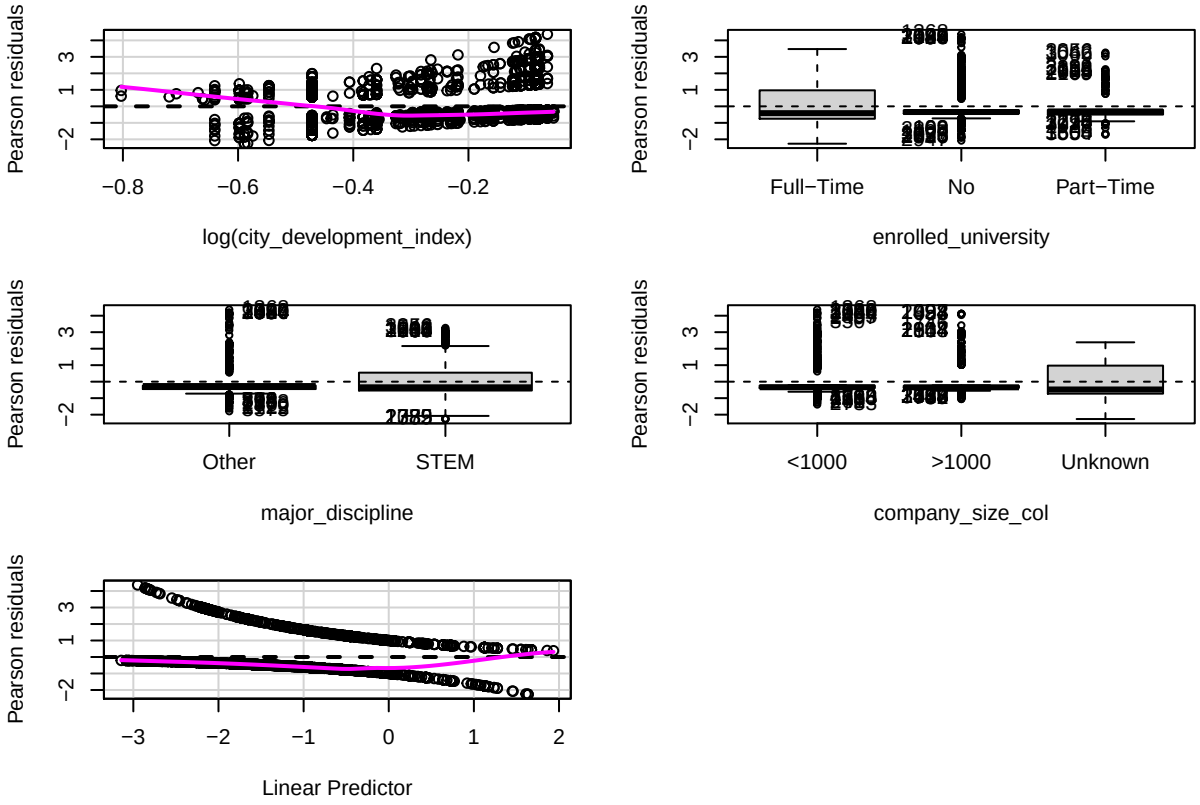
Analysis, Goodness of Fit for our Final model.

First, we output all the important analysis and results for a final model.

```
marginalModelPlot(m4)
```

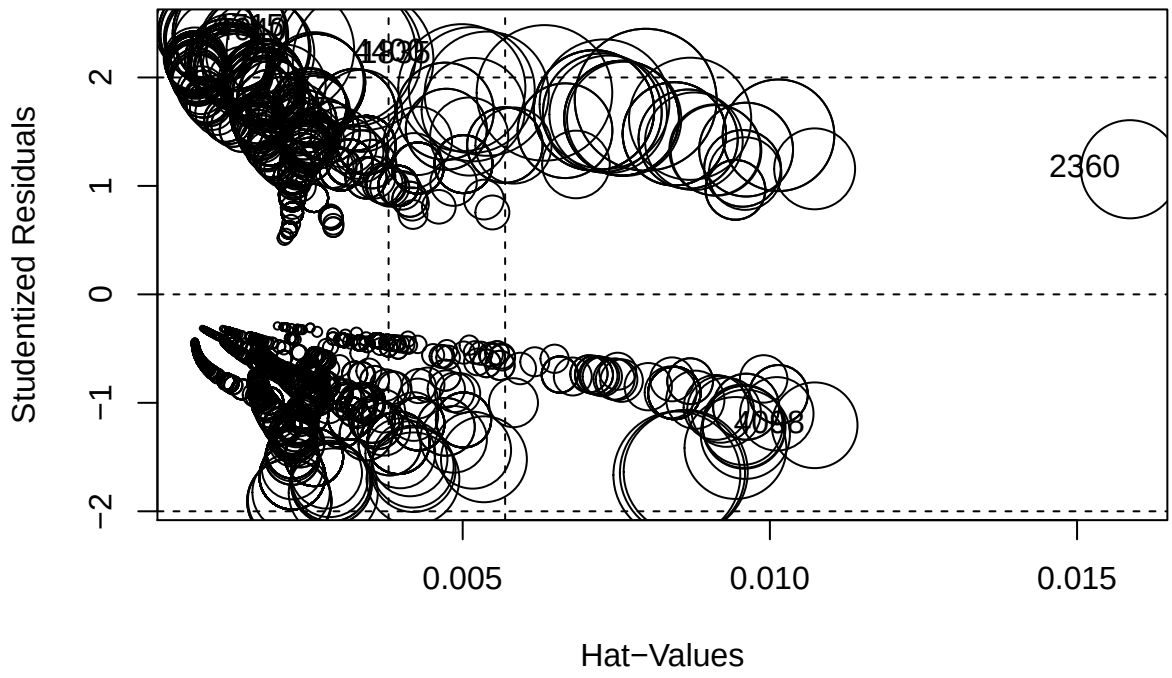



```
residualPlots(m4)
```



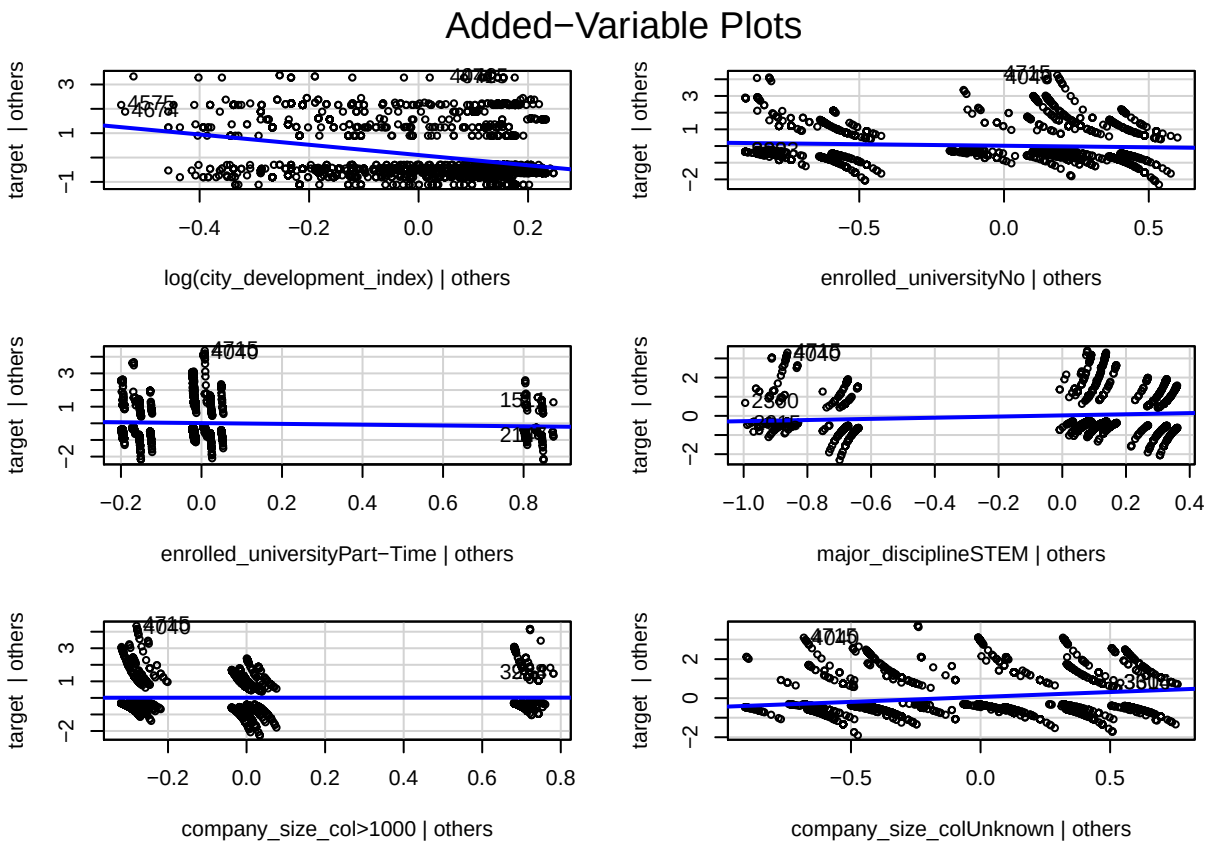

```
##                               Test stat Pr(>|Test stat|)
## log(city_development_index)     14.78        0.0001208 ***
## enrolled_university
## major_discipline
## company_size_col
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
influencePlot(m4)
```




```
##        StudRes          Hat       CookD
## 4040  2.418009 0.0008267887 0.002066861
## 4098 -1.206881 0.0107284969 0.001659478
## 4715  2.452682 0.0007828851 0.002138661
## 1835  2.202387 0.0031677991 0.004611018
## 2360  1.153485 0.0158606369 0.002176775
## 4400  2.222074 0.0030685641 0.004682812
```

```
avPlots(m4)
```




```
summary(m4)
```

```
##
## Call:
## glm(formula = target ~ log(city_development_index) + enrolled_university +
##     major_discipline + company_size_col, family = "binomial",
##     data = df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9044  -0.6719  -0.4724  -0.3291   2.4496
##
## Coefficients:
##                               Estimate Std. Error z value Pr(>|z|)
## (Intercept)                   -2.82334    0.15782 -17.890  < 2e-16 ***
## log(city_development_index)   -4.34405    0.24585 -17.669  < 2e-16 ***
## enrolled_universityNo         -0.39937    0.10072  -3.965 7.34e-05 ***
## enrolled_universityPart-Time -0.59271    0.20209  -2.933  0.00336 **
## major_disciplineSTEM           0.69138    0.11172   6.188 6.08e-10 ***
## company_size_col>1000          0.03253    0.12350   0.263  0.79224
## company_size_colUnknown        1.20947    0.09941  12.166  < 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: 4056.3  on 3689  degrees of freedom
## Residual deviance: 3423.1  on 3683  degrees of freedom
## AIC: 3437.1
##
## Number of Fisher Scoring iterations: 4
```

```
df[c("4040"),]
```

```
##      enrollee_id     city city_development_index gender relevant_experience
## 4040       28010 city_103                   0.92 Female                  No
##      enrolled_university education_level major_discipline experience
## 4040                  No        Graduate            Other          2
##      company_size  company_type last_new_job training_hours     target
## 4040      100-500 Public Sector            2             35 Target.Yes
##      company_size_col
## 4040            <1000
```

```
df[c("4715"),]
```

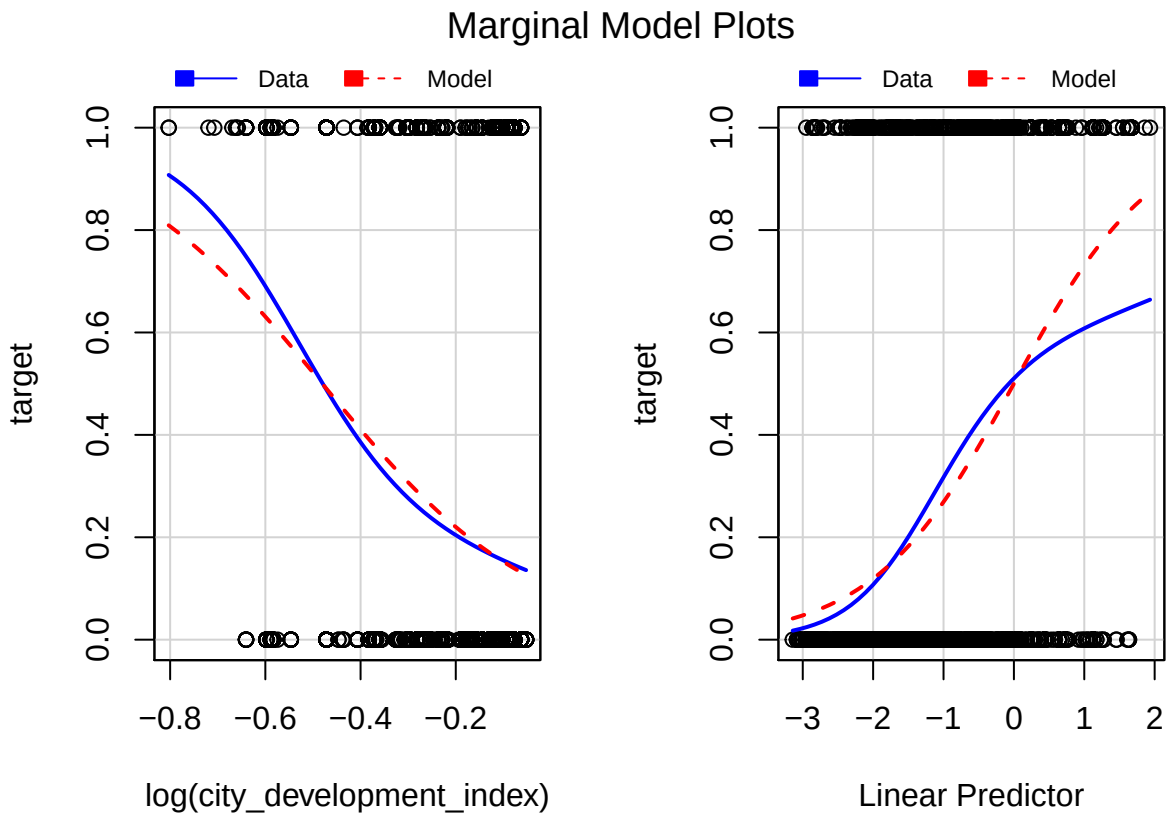
```
##      enrollee_id    city city_development_index gender relevant_experience
## 4715       27293 city_75                   0.939   Male                 Yes
##      enrolled_university education_level major_discipline experience
## 4715                  No  Primary School            Other         13
##      company_size   company_type last_new_job training_hours     target
## 4715          <10 Funded Startup            4             32 Target.Yes
##      company_size_col
## 4715            <1000
```

```
sum( resid( m4, "pearson") ^2 )
```

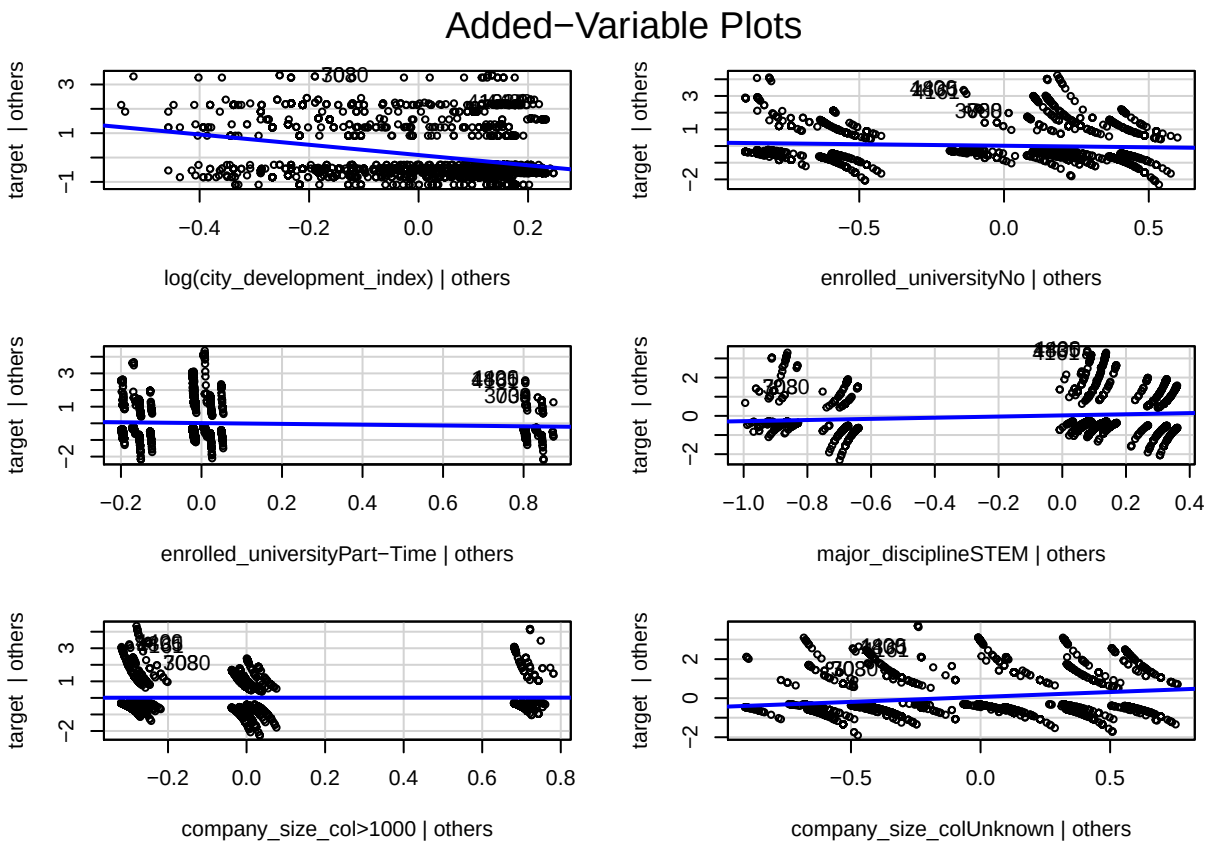
```
## [1] 3559.038
```

```
marginalModelPlots(m4,id=list(labels=row.names(df),method=abs(cooks.distance(m4)), n=5) )
```

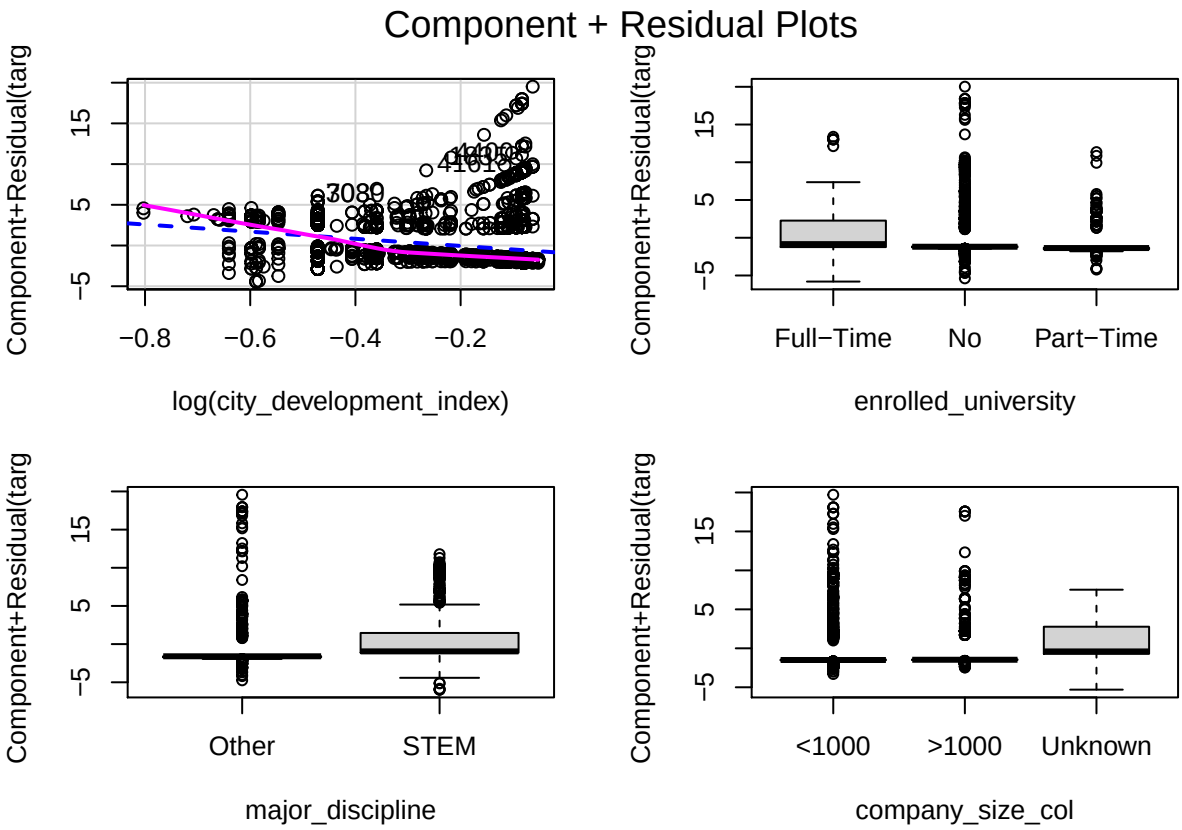
```
## Warning in mmps(...): Interactions and/or factors skipped
```



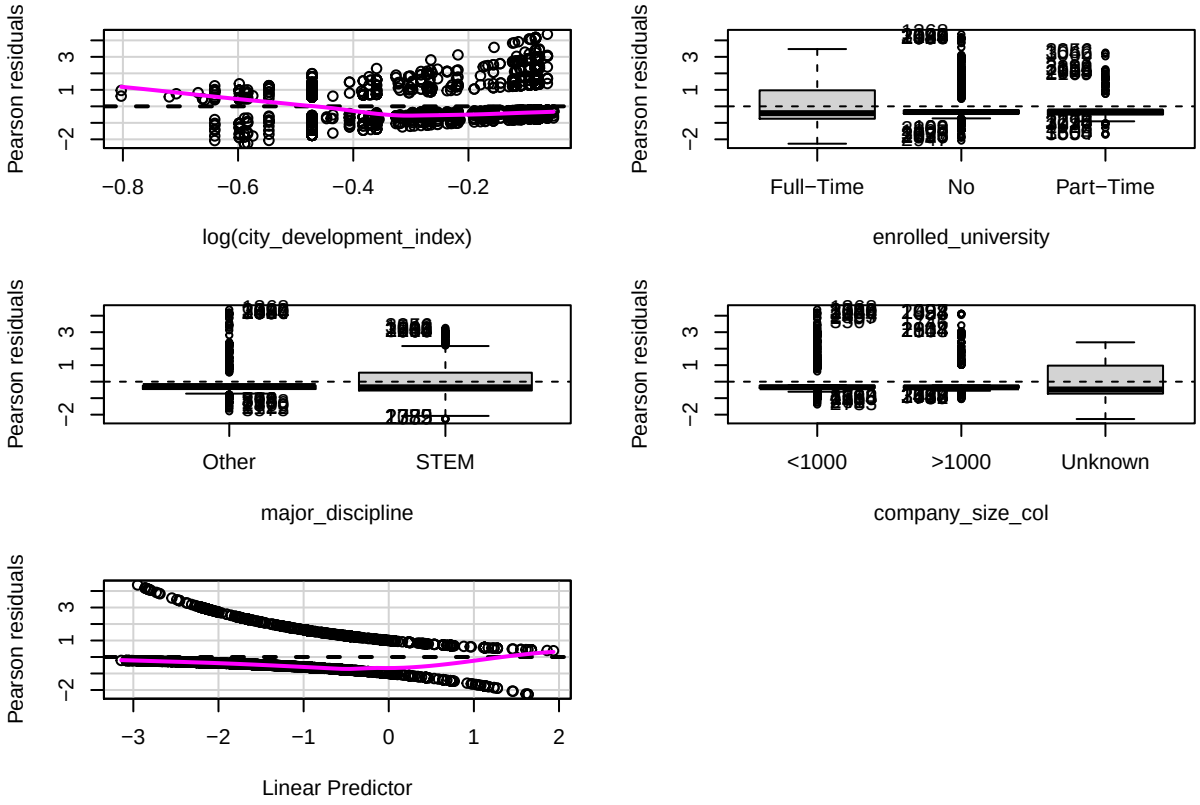

```
avPlots(m4,id=list(labels=row.names(df),method=abs(cooks.distance(m4)), n=5) )
```




```
crPlots(m4,id=list(labels=row.names(df),method=abs(cooks.distance(m4)), n=5) )
```




```
residualPlots(m4, layout=c(3, 2))
```

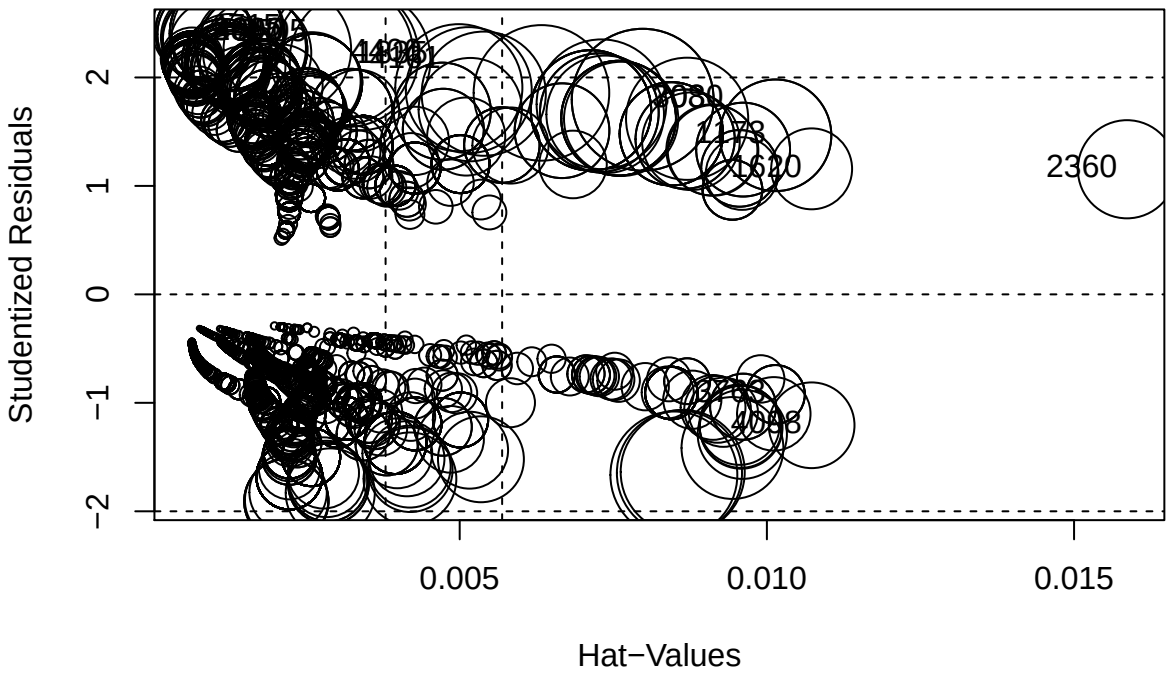



```
##                               Test stat Pr(>|Test stat|)
## log(city_development_index)     14.78        0.0001208 ***
## enrolled_university
## major_discipline
## company_size_col
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
outlierTest(m4)
```

```
## No Studentized residuals with Bonferroni p < 0.05
## Largest |rstudent|:
##      rstudent unadjusted p-value Bonferroni p
## 4715 2.452682           0.01418           NA
```

```
par(mfrow=c(1,1))
influencePlot(m4,id=list(n=5) )
```




```
##         StudRes          Hat        CookD
## 2505  2.4065056 0.0012075197 0.0029242437
## 4040  2.4180088 0.0008267887 0.0020668610
## 3763 -0.9219496 0.0101404856 0.0007768966
## 1173  1.4666355 0.0101404856 0.0028136102
## 4098 -1.2068811 0.0107284969 0.0016594781
## 1620  1.1573378 0.0107284969 0.0014779106
## 4715  2.4526817 0.0007828851 0.0021386614
## 1835  2.2023867 0.0031677991 0.0046110184
## 2360  1.1534845 0.0158606369 0.0021767747
## 3030  1.7944989 0.0079789667 0.0045462022
## 26    2.4180088 0.0008267887 0.0020668610
## 708   1.7944989 0.0079789667 0.0045462022
## 4400  2.2220736 0.0030685641 0.0046828120
## 4161  2.1620738 0.0033790765 0.0044678156
## 3395  2.4180088 0.0008267887 0.0020668610
```

```
model.final <- lrm(target ~ log(city_development_index) + enrolled_university + major_discipline + compa
model.final
```

```
## Logistic Regression Model
##
##  lrm(formula = target ~ log(city_development_index) + enrolled_university +
##      major_discipline + company_size_col, data = df)
##
```

```
##                         Model Likelihood    Discrimination    Rank Discrim.
##                               Ratio Test           Indexes          Indexes
##  Obs          3690    LR chi2     633.23    R2       0.236    C       0.775
##   Target.No   2809    d.f.             6    g        1.124    Dxy     0.549
##   Target.Yes   881    Pr(> chi2) <0.0001    gr       3.078    gamma   0.556
##  max |deriv| 8e-14                         gp       0.190    tau-a   0.200
##                                            Brier    0.151
##
##                                Coef    S.E.   Wald Z Pr(>|Z|)
##  Intercept                     -2.8233 0.1578 -17.89 <0.0001
##  city_development_index        -4.3441 0.2459 -17.67 <0.0001
##  enrolled_university=No        -0.3994 0.1007  -3.97 <0.0001
##  enrolled_university=Part-Time -0.5927 0.2021  -2.93 0.0034
##  major_discipline=STEM          0.6914 0.1117   6.19 <0.0001
##  company_size_col=>1000         0.0325 0.1235   0.26 0.7922
##  company_size_col=Unknown       1.2095 0.0994  12.17 <0.0001
##
```

```
m0 <- glm(target ~ 1, family="binomial", data=df)
NagelkerkeR2(m4)
```

```
## $N
## [1] 3690
##
## $R2
## [1] 0.236457
```

```
summary(m4)
```

```
##
## Call:
## glm(formula = target ~ log(city_development_index) + enrolled_university +
##     major_discipline + company_size_col, family = "binomial",
##     data = df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9044  -0.6719  -0.4724  -0.3291   2.4496
##
## Coefficients:
##                               Estimate Std. Error z value Pr(>|z|)
## (Intercept)                   -2.82334    0.15782 -17.890  < 2e-16 ***
## log(city_development_index)   -4.34405    0.24585 -17.669  < 2e-16 ***
## enrolled_universityNo         -0.39937    0.10072  -3.965 7.34e-05 ***
## enrolled_universityPart-Time -0.59271    0.20209  -2.933  0.00336 **
## major_disciplineSTEM           0.69138    0.11172   6.188 6.08e-10 ***
## company_size_col>1000          0.03253    0.12350   0.263  0.79224
## company_size_colUnknown        1.20947    0.09941  12.166  < 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: 4056.3  on 3689  degrees of freedom
## Residual deviance: 3423.1  on 3683  degrees of freedom
## AIC: 3437.1
##
## Number of Fisher Scoring iterations: 4
```

```
100*(1-m4$dev/m4$null.dev)
```

```
## [1] 15.61095
```

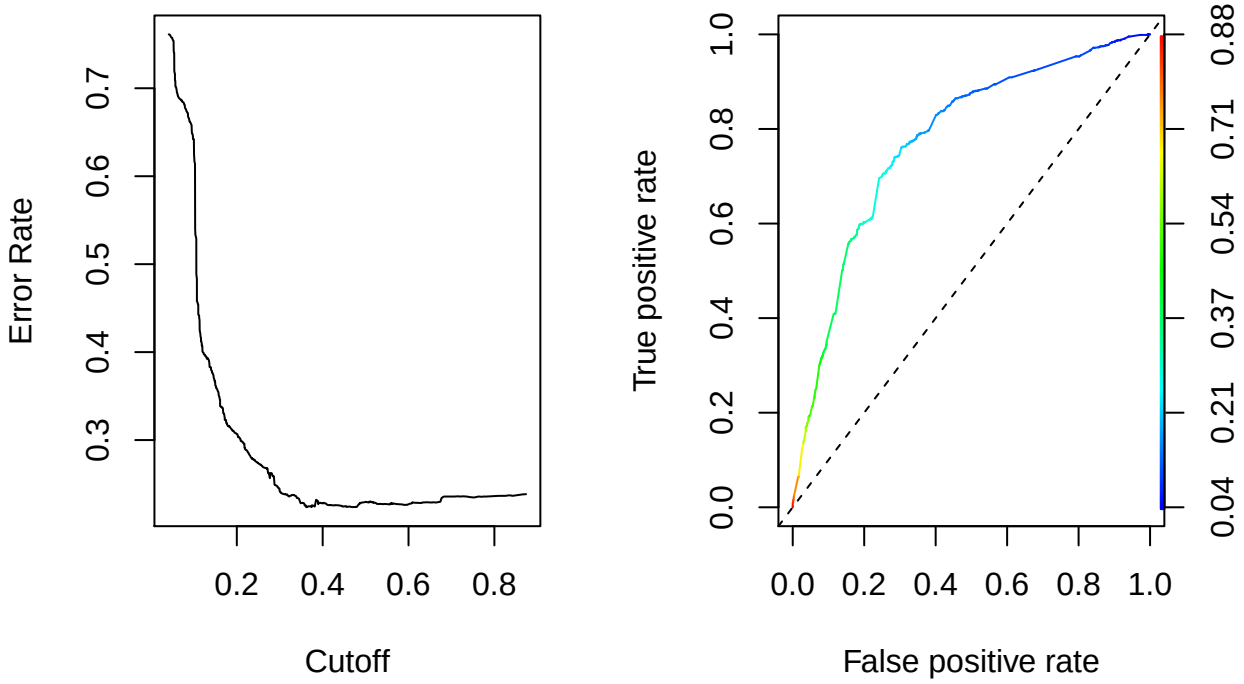
```
m4$dev
```

```
## [1] 3423.101
```

```
m4$null.dev
```

```
## [1] 4056.334
```

```
dadesroc<-prediction(predict(m4,type="response"),df$target)
par(mfrow=c(1,2))
plot(performance(dadesroc,"err"))
plot(performance(dadesroc,"tpr","fpr"), colorize=TRUE)
abline(0,1,lty=2)
```



```
library(cvAUC)
```

```
## Loading required package: data.table
```

```
##
## Attaching package: 'data.table'
```

```
## The following object is masked from 'package:purrr':
##
##     transpose
```

```
## The following objects are masked from 'package:dplyr':
##
##     between, first, last
```

```
##
```

```
## cvAUC version: 1.1.0
```

```
## Notice to cvAUC users: Major speed improvements in version 1.1.0
```

```
##
```

```
AUC(predict(m4,type="response"),df$target)
```

```
## [1] 0.7747458
```

1. The summary for the final model shows good p-values for most intercepts, except for the company size bigger than 1000.
2. With regards to the ROC curve, we seem to get a decent fit. The AUC (Area under curve) is 0.77, which is a labeled as a good fit by statisticians.
3. The Nagel-Kerke test returns a pseudo-coefficient of determination of 0.24. Not knowing if this a good fit, we compare it with Lab 6 Election 92 Diagnostics, that presented a final model that returned an R2 of 0.25, so our fit seems adequate.
4. In the previous section when we considered the best three models so far, we deleted a posteriori influential data based on the Cook's distance value. In the end, our test data frame has ended up with 3690 observations (20 less than in the beginning)
5. In the Marginal Model Plots we see that up until the linear predictor is 0, the model seems quite a good fit. The problem is that the Model expects that with higher values in the linear predictor, the target should be 1. However, our data has values with high linear predictor that have are Target 0. This introduced a bias in our model and that is why the data curve deviates from the model curve.
6. The Residual Plots for the log(city development index) variable show very high Pearson Residuals as the city development index gets bigger. This because our model is not predicting well the Target variable if the city development index increases.
7. If we look at the Influential Plot, we see many a priori influential values, but find less a posteriori influential values. That is because they may have high leverage, but the Cook's Distance is not high enough for them to be removed.
8. The Added Variable Plots consistently show observations 4040 and 4715 with very high Pearson residuals in most of the plots. We observe that the city development index is very high (0.92 and 0.93) respectively. But there is no other similarity between these instances that allows us to draw any conclusion as to why they systematically appear as outliers.

9. The component residual plots do not add any more important information. Again, it is shown that problems appear with high logarithm city development index values, as the residuals are far away from the pink line which is the optimal fit.
10. The final conclusion is that although these aforementioned values negatively impact our model, since we cannot label them as influential a posteriori data, we cannot directly remove them. This is the best model obtained considering all the improvements.

Forecasting capability of the final model

```
df.fin <- m4
dftest$company_size_col <- fct_collapse(dftest$company_size, "<1000"= c("<10", "10-49", "50-99", "100-5
job.vot <- predict(df.fin, newdata=dftest, type="response")
job.est <- ifelse(job.vot < 0.5, 0, 1)
table(job.est, dftest$target)
```

```
##
## job.est Target.No Target.Yes
##       0       888        234
##       1        49         65
```

```
sum(diag(table(job.est,dftest$target)))/dim(dftest)[1]
```

```
## [1] 0.7710356
```

```
# Null model
m0<-glm(target ~ 1, family=binomial, data=dftest)
job.vot0 <- m0$fit
job.est0 <- ifelse(job.vot0<0.5,0,1)
table(job.est0,dftest$target)
```

```
##
## job.est0 Target.No Target.Yes
##        0       937        299
```

```
table(job.est0,dftest$target)[1,2]/dim(dftest)[1]
```

```
## [1] 0.2419094
```

```
library(ResourceSelection)
```

```
## ResourceSelection 0.3-5   2019-07-22
```

```
hoslem.test(dftest$target, job.vot)
```

```
## Warning in Ops.factor(1, y): '-' not meaningful for factors
```

```
##
##  Hosmer and Lemeshow goodness of fit (GOF) test
##
## data:  dftest$target, job.vot
## X-squared = 1236, df = 8, p-value < 2.2e-16
```

Appendix

This last section includes extra information or plots that are not part of the main report in an attempt to avoid overloading the plot ## Appendix A: More on Data Preparation

Appendix B: More on the Target Modelling using Covariates

Appendix C: More on the Target Moddeling using Factors

```
summary(m2)
```

```
##
## Call:
## glm(formula = target ~ ., family = "binomial", data = df[, c(2,
##     4:12, 14)])
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.3981  -0.6128  -0.4180  -0.0003   2.9047
##
## Coefficients:
##                               Estimate Std. Error z value Pr(>|z|)
## (Intercept)                 -1.610e+01  1.063e+03  -0.015  0.98791
## citycity_10                  1.429e+01  1.063e+03   0.013  0.98927
## citycity_100                 1.545e+01  1.063e+03   0.015  0.98840
## citycity_101                 1.606e+01  1.063e+03   0.015  0.98795
## citycity_102                 1.455e+01  1.063e+03   0.014  0.98908
## citycity_103                 1.525e+01  1.063e+03   0.014  0.98855
## citycity_104                 1.465e+01  1.063e+03   0.014  0.98901
## citycity_105                 1.365e+01  1.063e+03   0.013  0.98975
## citycity_106                -1.241e+00  2.624e+03   0.000  0.99962
## citycity_107                 3.156e+01  2.624e+03   0.012  0.99041
## citycity_109                 3.676e-01  2.624e+03   0.000  0.99989
## citycity_11                  1.658e+01  1.063e+03   0.016  0.98755
## citycity_114                 1.439e+01  1.063e+03   0.014  0.98920
## citycity_115                 1.586e+01  1.063e+03   0.015  0.98809
## citycity_116                 1.549e+01  1.063e+03   0.015  0.98837
## citycity_117                 1.690e+01  1.063e+03   0.016  0.98732
## citycity_118                 1.559e+01  1.063e+03   0.015  0.98830
## citycity_12                  7.325e-01  1.957e+03   0.000  0.99970
## citycity_120                 4.291e-01  2.624e+03   0.000  0.99987
## citycity_121                -5.503e-01  2.624e+03   0.000  0.99983
## citycity_123                 1.708e+01  1.063e+03   0.016  0.98718
## citycity_127                -1.411e-01  1.729e+03   0.000  0.99993
## citycity_128                 1.709e+01  1.063e+03   0.016  0.98717
## citycity_13                  1.451e+01  1.063e+03   0.014  0.98910
## citycity_131                 3.204e+01  2.624e+03   0.012  0.99026
## citycity_133                 1.911e-01  2.624e+03   0.000  0.99994
## citycity_134                -4.249e-02  1.986e+03   0.000  0.99998
## citycity_136                 1.413e+01  1.063e+03   0.013  0.98939
## citycity_138                 1.438e+01  1.063e+03   0.014  0.98920
```

```
## citycity_139                3.326e+01  2.624e+03   0.013  0.98989
## citycity_14                 1.581e+01  1.063e+03   0.015  0.98813
## citycity_141                1.618e+01  1.063e+03   0.015  0.98786
## citycity_142                1.541e+01  1.063e+03   0.014  0.98843
## citycity_143                1.638e+01  1.063e+03   0.015  0.98770
## citycity_144                1.491e+01  1.063e+03   0.014  0.98881
## citycity_145                1.496e+01  1.063e+03   0.014  0.98877
## citycity_146                1.244e-01  2.624e+03   0.000  0.99996
## citycity_149                1.501e+01  1.063e+03   0.014  0.98873
## citycity_150               -2.299e-01  1.377e+03   0.000  0.99987
## citycity_152                1.498e+01  1.063e+03   0.014  0.98875
## citycity_155                3.313e+01  2.624e+03   0.013  0.98993
## citycity_157                1.539e+01  1.063e+03   0.014  0.98845
## citycity_158                1.541e+01  1.063e+03   0.014  0.98843
## citycity_159                1.365e+01  1.063e+03   0.013  0.98975
## citycity_16                 1.475e+01  1.063e+03   0.014  0.98893
## citycity_160                1.509e+01  1.063e+03   0.014  0.98867
## citycity_162                1.548e+01  1.063e+03   0.015  0.98838
## citycity_165                1.536e+01  1.063e+03   0.014  0.98847
## citycity_167                1.468e+01  1.063e+03   0.014  0.98898
## citycity_173                1.305e+01  1.063e+03   0.012  0.99021
## citycity_175                1.455e+01  1.063e+03   0.014  0.98908
## citycity_176                1.637e+01  1.063e+03   0.015  0.98771
## citycity_179                3.350e+01  2.624e+03   0.013  0.98982
## citycity_18                 1.647e-01  2.624e+03   0.000  0.99995
## citycity_180               -2.123e+00  2.624e+03  -0.001  0.99935
## citycity_19                 1.553e+01  1.063e+03   0.015  0.98835
## citycity_2                  4.655e-01  2.624e+03   0.000  0.99986
## citycity_20                 1.516e+01  1.063e+03   0.014  0.98862
## citycity_21                 1.687e+01  1.063e+03   0.016  0.98733
## citycity_23                 1.377e+01  1.063e+03   0.013  0.98967
## citycity_24                 1.472e+01  1.063e+03   0.014  0.98895
## citycity_25                 3.183e+01  2.624e+03   0.012  0.99032
## citycity_26                 1.749e+01  1.063e+03   0.016  0.98687
## citycity_27                 1.468e+01  1.063e+03   0.014  0.98898
## citycity_28                 1.410e+01  1.063e+03   0.013  0.98941
## citycity_30                 1.089e+00  1.736e+03   0.001  0.99950
## citycity_31                -9.395e-01  2.624e+03   0.000  0.99971
## citycity_33                 3.330e+01  1.962e+03   0.017  0.98646
## citycity_36                 1.416e+01  1.063e+03   0.013  0.98937
## citycity_37                 1.572e+01  1.063e+03   0.015  0.98820
## citycity_39                 3.988e-01  1.509e+03   0.000  0.99979
## citycity_40                 1.413e+01  1.063e+03   0.013  0.98939
## citycity_41                 1.336e+01  1.063e+03   0.013  0.98997
## citycity_42                 1.613e+01  1.063e+03   0.015  0.98789
## citycity_43                 3.183e+01  1.995e+03   0.016  0.98727
## citycity_44                 1.607e+01  1.063e+03   0.015  0.98793
## citycity_45                 1.611e+01  1.063e+03   0.015  0.98791
## citycity_46                 1.567e+01  1.063e+03   0.015  0.98824
## citycity_48                 3.345e+01  2.624e+03   0.013  0.98983
## citycity_50                 1.463e+01  1.063e+03   0.014  0.98902
## citycity_53                -2.239e-01  1.397e+03   0.000  0.99987
## citycity_54                 6.850e-01  2.001e+03   0.000  0.99973
## citycity_55                 1.552e+01  1.063e+03   0.015  0.98835
```

```
## citycity_57                    1.458e+01  1.063e+03   0.014  0.98905
## citycity_61                    1.365e+01  1.063e+03   0.013  0.98976
## citycity_62                    6.525e-01  2.624e+03   0.000  0.99980
## citycity_64                    1.460e+01  1.063e+03   0.014  0.98904
## citycity_65                    1.483e+01  1.063e+03   0.014  0.98887
## citycity_67                    1.388e+01  1.063e+03   0.013  0.98958
## citycity_69                    1.588e+01  1.063e+03   0.015  0.98808
## citycity_7                     1.474e+01  1.063e+03   0.014  0.98894
## citycity_70                    1.578e+01  1.063e+03   0.015  0.98816
## citycity_71                    1.490e+01  1.063e+03   0.014  0.98881
## citycity_72                    1.074e+00  2.624e+03   0.000  0.99967
## citycity_73                    1.515e+01  1.063e+03   0.014  0.98862
## citycity_74                    1.696e+01  1.063e+03   0.016  0.98727
## citycity_75                    1.482e+01  1.063e+03   0.014  0.98888
## citycity_76                    1.647e+01  1.063e+03   0.015  0.98764
## citycity_77                    5.594e-01  1.489e+03   0.000  0.99970
## citycity_78                    1.603e+01  1.063e+03   0.015  0.98797
## citycity_8                    -1.158e-03  2.624e+03   0.000  1.00000
## citycity_80                    1.495e+01  1.063e+03   0.014  0.98877
## citycity_81                   -2.535e-01  1.956e+03   0.000  0.99990
## citycity_82                    2.443e-01  2.624e+03   0.000  0.99993
## citycity_83                    1.503e+01  1.063e+03   0.014  0.98872
## citycity_84                   -5.385e-01  1.649e+03   0.000  0.99974
## citycity_89                    1.544e+01  1.063e+03   0.015  0.98841
## citycity_9                     1.657e+01  1.063e+03   0.016  0.98757
## citycity_90                    1.480e+01  1.063e+03   0.014  0.98889
## citycity_91                    1.527e+01  1.063e+03   0.014  0.98854
## citycity_93                    1.659e+01  1.063e+03   0.016  0.98754
## citycity_94                   -2.100e-02  1.694e+03   0.000  0.99999
## citycity_97                    1.346e+01  1.063e+03   0.013  0.98990
## citycity_98                    1.907e-01  1.241e+03   0.000  0.99988
## citycity_99                    5.112e-02  1.188e+03   0.000  0.99997
## genderMale                     6.565e-02  1.826e-01   0.360  0.71912
## genderMissing                  1.821e-01  1.964e-01   0.927  0.35388
## genderOther                    1.366e-01  5.045e-01   0.271  0.78664
## relevant_experienceYes        -2.263e-01  1.262e-01  -1.792  0.07308 .
## enrolled_universityNo         -3.403e-01  1.233e-01  -2.761  0.00577 **
## enrolled_universityPart-Time  -2.529e-01  2.097e-01  -1.206  0.22785
## education_levelHigh School    -6.740e-01  2.188e-01  -3.081  0.00207 **
## education_levelMasters        -1.395e-01  1.193e-01  -1.170  0.24217
## education_levelPhd            -2.202e-01  3.836e-01  -0.574  0.56584
## education_levelPrimary School -9.613e-01  4.309e-01  -2.231  0.02567 *
## major_disciplineSTEM           3.320e-01  1.527e-01   2.174  0.02971 *
## experience>20                 -7.105e-01  3.047e-01  -2.332  0.01969 *
## experience1                   -8.226e-02  3.515e-01  -0.234  0.81500
## experience10                  -2.622e-01  3.286e-01  -0.798  0.42485
## experience11                  -4.982e-01  3.536e-01  -1.409  0.15883
## experience12                  -4.259e-01  3.851e-01  -1.106  0.26872
## experience13                  -5.322e-01  4.164e-01  -1.278  0.20121
## experience14                  -3.982e-01  3.749e-01  -1.062  0.28813
## experience15                  -7.311e-01  3.846e-01  -1.901  0.05727 .
## experience16                  -1.148e+00  4.330e-01  -2.651  0.00803 **
## experience17                  -1.375e+00  5.156e-01  -2.667  0.00766 **
## experience18                  -6.482e-01  5.451e-01  -1.189  0.23440
```

```
## experience19                  5.682e-03  4.646e-01   0.012  0.99024
## experience2                  -3.157e-01  3.101e-01  -1.018  0.30867
## experience20                 -7.279e-01  9.005e-01  -0.808  0.41891
## experience3                  -2.001e-01  2.944e-01  -0.680  0.49671
## experience4                  -5.171e-01  2.926e-01  -1.767  0.07715 .
## experience5                  -4.538e-01  3.001e-01  -1.512  0.13054
## experience6                  -5.912e-01  3.099e-01  -1.908  0.05641 .
## experience7                  -2.968e-01  3.151e-01  -0.942  0.34635
## experience8                  -3.715e-01  3.343e-01  -1.111  0.26652
## experience9                  -3.823e-01  3.233e-01  -1.182  0.23707
## company_size10-49             3.596e-01  2.650e-01   1.357  0.17474
## company_size100-500          -1.300e-02  2.591e-01  -0.050  0.95997
## company_size1000-4999        -1.877e-01  3.016e-01  -0.622  0.53366
## company_size10000+            1.966e-02  2.677e-01   0.073  0.94146
## company_size50-99             2.091e-01  2.503e-01   0.836  0.40330
## company_size500-999          -2.826e-01  3.447e-01  -0.820  0.41225
## company_size5000-9999         1.179e-01  3.598e-01   0.328  0.74321
## company_sizeUnknown           1.280e+00  2.810e-01   4.555 5.24e-06 ***
## company_typeFunded Startup   -4.269e-01  3.703e-01  -1.153  0.24901
## company_typeNGO              -3.929e-01  4.382e-01  -0.897  0.36986
## company_typeOther             1.454e-01  3.319e-01   0.438  0.66124
## company_typePublic Sector     3.647e-01  3.636e-01   1.003  0.31588
## company_typePvt Ltd          -1.417e-01  2.946e-01  -0.481  0.63062
## last_new_job1                -1.950e-01  1.532e-01  -1.273  0.20302
## last_new_job2                -3.124e-01  1.780e-01  -1.755  0.07923 .
## last_new_job3                -1.117e-01  2.352e-01  -0.475  0.63474
## last_new_job4                -1.841e-01  2.513e-01  -0.733  0.46378
## last_new_jobnever            -1.018e+00  2.068e-01  -4.923 8.53e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4084.9  on 3699  degrees of freedom
## Residual deviance: 3142.0  on 3535  degrees of freedom
## AIC: 3472
##
## Number of Fisher Scoring iterations: 15
```