TRAINING A MODEL TO DIFFERENTIATE BETWEEN DEFAULTERS AND NON DEFAULTERS

MATHS 231 ## Produced by: ##GROUP - 5 ##Muhammad Musa 23100004 ##Malik Muhammad Mussab 23110229 ## Suleman Mahmood 23100011

RESEARCH QUESTION:

Is there a relationship between a borrower being a defaulter (not returning the loan bank) ***and some attributes of the borrower available to the bank*** (family status, job, etc) ***, which could be used to make a model that could predict whther a borrower is going to return the loan or not?***

ABSTRACT:

There is a huge amount of people who take loans but do not return them to banks. This not only damages the banks but also makes them give out loans with a harder hand, possibly, bias. To reduce this problem, we used our dataset to train a binary choice model which could tell whether a borrower will potentially return the loan or not. For this, we used some variables from our dataset which are easily available to banks as well for them to run our model.

PROCESS:

1. Install the required libraries and the corresponding libraries.

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

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

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

```
library(caret)
```

```
## Loading required package: ggplot2
```

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

```
library(caTools)
library(ggplot2)
library(Amelia)
```

```
## Loading required package: Rcpp
```

```
## ##
## ## Amelia II: Multiple Imputation
## ## (Version 1.8.0, built: 2021-05-26)
## ## Copyright (C) 2005-2021 James Honaker, Gary King and Matthew Blackwell
## ## Refer to http://gking.harvard.edu/amelia/ for more information
## ##
```

```
library(mlbench)
library(MASS)
```

```
##
## Attaching package: 'MASS'
```

```
## The following object is masked from 'package:dplyr':
##
##     select
```

2. We will import the data and study it.

```
raw_data<-read.csv("D:/Musa/LUMS/Fall_21/Stats/application_data.csv")
```

Upon looking at the data, we can see that the column TARGET tells whether the the borrowe of loan paid it on time or not. It is a categorical variable where 1 means the client had payment difficulties.

3. Data Cleaning

The data we have imported shows a lot of information and requires thorough going through. After we went through the information, the next important thing was to clean the data. This was done to make sure the model is accurate with little to no effect of any outliers false values. There were two things to be cleaned in the data. * **Variables** + We first eliminated some variables to make the dataset easier to work with. Since we had a lot of variables, the exact number being 122, we have to decrease it to only those variables that significantly effect the target variable. This will help us get a more accurate model along with making sure the model does not over fit for the given data. To decide which columns to keep and which columns not to keep, we will look at the data again.

```
summary(raw_data)
```

```
##    SK_ID_CURR         TARGET        NAME_CONTRACT_TYPE CODE_GENDER
##  Min.   :100002   Min.   :0.00000   Length:307511      Length:307511
##  1st Qu.:189146   1st Qu.:0.00000   Class :character   Class :character
##  Median :278202   Median :0.00000   Mode  :character   Mode  :character
##  Mean   :278181   Mean   :0.08073
##  3rd Qu.:367143   3rd Qu.:0.00000
##  Max.   :456255   Max.   :1.00000
##
##  FLAG_OWN_CAR       FLAG_OWN_REALTY     CNT_CHILDREN     AMT_INCOME_TOTAL
##  Length:307511      Length:307511      Min.   : 0.0000   Min.   :    25650
##  Class :character   Class :character   1st Qu.: 0.0000   1st Qu.:   112500
##  Mode  :character   Mode  :character   Median : 0.0000   Median :   147150
##                                        Mean   : 0.4171   Mean   :   168798
##                                        3rd Qu.: 1.0000   3rd Qu.:   202500
##                                        Max.   :19.0000   Max.   :117000000
##
##    AMT_CREDIT       AMT_ANNUITY     AMT_GOODS_PRICE   NAME_TYPE_SUITE
##  Min.   :  45000   Min.   :  1616   Min.   :  40500   Length:307511
##  1st Qu.: 270000   1st Qu.: 16524   1st Qu.: 238500   Class :character
##  Median : 513531   Median : 24903   Median : 450000   Mode  :character
##  Mean   : 599026   Mean   : 27109   Mean   : 538396
##  3rd Qu.: 808650   3rd Qu.: 34596   3rd Qu.: 679500
##  Max.   :4050000   Max.   :258026   Max.   :4050000
##                    NA's   :12       NA's   :278
##  NAME_INCOME_TYPE   NAME_EDUCATION_TYPE NAME_FAMILY_STATUS NAME_HOUSING_TYPE
##  Length:307511      Length:307511       Length:307511      Length:307511
##  Class :character   Class :character    Class :character   Class :character
##  Mode  :character   Mode  :character    Mode  :character   Mode  :character
##
##
##
##
##  REGION_POPULATION_RELATIVE   DAYS_BIRTH     DAYS_EMPLOYED    DAYS_REGISTRATION
##  Min.   :0.00029            Min.   :-25229   Min.   :-17912   Min.   :-24672
##  1st Qu.:0.01001            1st Qu.:-19682   1st Qu.: -2760   1st Qu.: -7480
##  Median :0.01885            Median :-15750   Median : -1213   Median : -4504
##  Mean   :0.02087            Mean   :-16037   Mean   : 63815   Mean   : -4986
##  3rd Qu.:0.02866            3rd Qu.:-12413   3rd Qu.:  -289   3rd Qu.: -2010
##  Max.   :0.07251            Max.   : -7489   Max.   :365243   Max.   :     0
##
##  DAYS_ID_PUBLISH  OWN_CAR_AGE       FLAG_MOBIL FLAG_EMP_PHONE
##  Min.   :-7197   Min.   : 0.00    Min.   :0    Min.   :0.0000
##  1st Qu.:-4299   1st Qu.: 5.00    1st Qu.:1    1st Qu.:1.0000
##  Median :-3254   Median : 9.00    Median :1    Median :1.0000
##  Mean   :-2994   Mean   :12.06    Mean   :1    Mean   :0.8199
##  3rd Qu.:-1720   3rd Qu.:15.00    3rd Qu.:1    3rd Qu.:1.0000
##  Max.   :    0   Max.   :91.00    Max.   :1    Max.   :1.0000
##                  NA's   :202929
##  FLAG_WORK_PHONE  FLAG_CONT_MOBILE   FLAG_PHONE       FLAG_EMAIL
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.00000
##  1st Qu.:0.0000   1st Qu.:1.0000   1st Qu.:0.0000   1st Qu.:0.00000
##  Median :0.0000   Median :1.0000   Median :0.0000   Median :0.00000
##  Mean   :0.1994   Mean   :0.9981   Mean   :0.2811   Mean   :0.05672
##  3rd Qu.:0.0000   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:0.00000
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.00000
##
##  OCCUPATION_TYPE    CNT_FAM_MEMBERS  REGION_RATING_CLIENT
##  Length:307511      Min.   : 1.000   Min.   :1.000
##  Class :character   1st Qu.: 2.000   1st Qu.:2.000
##  Mode  :character   Median : 2.000   Median :2.000
##                     Mean   : 2.153   Mean   :2.052
##                     3rd Qu.: 3.000   3rd Qu.:2.000
##                     Max.   :20.000   Max.   :3.000
##                     NA's   :2
##  REGION_RATING_CLIENT_W_CITY WEEKDAY_APPR_PROCESS_START HOUR_APPR_PROCESS_START
##  Min.   :1.000               Length:307511              Min.   : 0.00
##  1st Qu.:2.000               Class :character           1st Qu.:10.00
##  Median :2.000               Mode  :character           Median :12.00
##  Mean   :2.032                                          Mean   :12.06
##  3rd Qu.:2.000                                          3rd Qu.:14.00
##  Max.   :3.000                                          Max.   :23.00
##
##  REG_REGION_NOT_LIVE_REGION REG_REGION_NOT_WORK_REGION
##  Min.   :0.00000            Min.   :0.00000
##  1st Qu.:0.00000            1st Qu.:0.00000
##  Median :0.00000            Median :0.00000
##  Mean   :0.01514            Mean   :0.05077
##  3rd Qu.:0.00000            3rd Qu.:0.00000
##  Max.   :1.00000            Max.   :1.00000
##
##  LIVE_REGION_NOT_WORK_REGION REG_CITY_NOT_LIVE_CITY REG_CITY_NOT_WORK_CITY
##  Min.   :0.00000             Min.   :0.00000        Min.   :0.0000
##  1st Qu.:0.00000             1st Qu.:0.00000        1st Qu.:0.0000
```

```
##  Median :0.00000             Median :0.00000        Median :0.0000
##  Mean   :0.04066             Mean   :0.07817        Mean   :0.2305
##  3rd Qu.:0.00000             3rd Qu.:0.00000        3rd Qu.:0.0000
##  Max.   :1.00000             Max.   :1.00000        Max.   :1.0000
##
##  LIVE_CITY_NOT_WORK_CITY ORGANIZATION_TYPE   EXT_SOURCE_1     EXT_SOURCE_2
##  Min.   :0.0000          Length:307511      Min.   :0.01     Min.   :0.0000
##  1st Qu.:0.0000          Class :character   1st Qu.:0.33     1st Qu.:0.3925
##  Median :0.0000          Mode  :character   Median :0.51     Median :0.5660
##  Mean   :0.1796                             Mean   :0.50     Mean   :0.5144
##  3rd Qu.:0.0000                             3rd Qu.:0.68     3rd Qu.:0.6636
##  Max.   :1.0000                             Max.   :0.96     Max.   :0.8550
##                                             NA's   :173378   NA's   :660
##   EXT_SOURCE_3   APARTMENTS_AVG   BASEMENTAREA_AVG YEARS_BEGINEXPLUATATION_AVG
##  Min.   :0.00    Min.   :0.00     Min.   :0.00     Min.   :0.00
##  1st Qu.:0.37    1st Qu.:0.06     1st Qu.:0.04     1st Qu.:0.98
##  Median :0.54    Median :0.09     Median :0.08     Median :0.98
##  Mean   :0.51    Mean   :0.12     Mean   :0.09     Mean   :0.98
##  3rd Qu.:0.67    3rd Qu.:0.15     3rd Qu.:0.11     3rd Qu.:0.99
##  Max.   :0.90    Max.   :1.00     Max.   :1.00     Max.   :1.00
##  NA's   :60965   NA's   :156061   NA's   :179943   NA's   :150007
##  YEARS_BUILD_AVG  COMMONAREA_AVG   ELEVATORS_AVG    ENTRANCES_AVG
##  Min.   :0.00     Min.   :0.00     Min.   :0.00     Min.   :0.00
##  1st Qu.:0.69     1st Qu.:0.01     1st Qu.:0.00     1st Qu.:0.07
##  Median :0.76     Median :0.02     Median :0.00     Median :0.14
##  Mean   :0.75     Mean   :0.04     Mean   :0.08     Mean   :0.15
##  3rd Qu.:0.82     3rd Qu.:0.05     3rd Qu.:0.12     3rd Qu.:0.21
##  Max.   :1.00     Max.   :1.00     Max.   :1.00     Max.   :1.00
##  NA's   :204488   NA's   :214865   NA's   :163891   NA's   :154828
##  FLOORSMAX_AVG    FLOORSMIN_AVG     LANDAREA_AVG    LIVINGAPARTMENTS_AVG
##  Min.   :0.00     Min.   :0.00     Min.   :0.00     Min.   :0.00
##  1st Qu.:0.17     1st Qu.:0.08     1st Qu.:0.02     1st Qu.:0.05
##  Median :0.17     Median :0.21     Median :0.05     Median :0.08
##  Mean   :0.23     Mean   :0.23     Mean   :0.07     Mean   :0.10
##  3rd Qu.:0.33     3rd Qu.:0.38     3rd Qu.:0.09     3rd Qu.:0.12
##  Max.   :1.00     Max.   :1.00     Max.   :1.00     Max.   :1.00
##  NA's   :153020   NA's   :208642   NA's   :182590   NA's   :210199
##  LIVINGAREA_AVG   NONLIVINGAPARTMENTS_AVG NONLIVINGAREA_AVG APARTMENTS_MODE
##  Min.   :0.00     Min.   :0.00            Min.   :0.00      Min.   :0.00
##  1st Qu.:0.05     1st Qu.:0.00            1st Qu.:0.00      1st Qu.:0.05
##  Median :0.07     Median :0.00            Median :0.00      Median :0.08
##  Mean   :0.11     Mean   :0.01            Mean   :0.03      Mean   :0.11
##  3rd Qu.:0.13     3rd Qu.:0.00            3rd Qu.:0.03      3rd Qu.:0.14
##  Max.   :1.00     Max.   :1.00            Max.   :1.00      Max.   :1.00
##  NA's   :154350   NA's   :213514          NA's   :169682    NA's   :156061
##  BASEMENTAREA_MODE YEARS_BEGINEXPLUATATION_MODE YEARS_BUILD_MODE
##  Min.   :0.00      Min.   :0.00                 Min.   :0.00
##  1st Qu.:0.04      1st Qu.:0.98                 1st Qu.:0.70
##  Median :0.07      Median :0.98                 Median :0.76
##  Mean   :0.09      Mean   :0.98                 Mean   :0.76
##  3rd Qu.:0.11      3rd Qu.:0.99                 3rd Qu.:0.82
##  Max.   :1.00      Max.   :1.00                 Max.   :1.00
##  NA's   :179943    NA's   :150007               NA's   :204488
##  COMMONAREA_MODE  ELEVATORS_MODE   ENTRANCES_MODE   FLOORSMAX_MODE
##  Min.   :0.00     Min.   :0.00     Min.   :0.00     Min.   :0.00
##  1st Qu.:0.01     1st Qu.:0.00     1st Qu.:0.07     1st Qu.:0.17
##  Median :0.02     Median :0.00     Median :0.14     Median :0.17
##  Mean   :0.04     Mean   :0.07     Mean   :0.15     Mean   :0.22
##  3rd Qu.:0.05     3rd Qu.:0.12     3rd Qu.:0.21     3rd Qu.:0.33
##  Max.   :1.00     Max.   :1.00     Max.   :1.00     Max.   :1.00
##  NA's   :214865   NA's   :163891   NA's   :154828   NA's   :153020
##  FLOORSMIN_MODE   LANDAREA_MODE    LIVINGAPARTMENTS_MODE LIVINGAREA_MODE
##  Min.   :0.00     Min.   :0.00     Min.   :0.00          Min.   :0.00
##  1st Qu.:0.08     1st Qu.:0.02     1st Qu.:0.05          1st Qu.:0.04
##  Median :0.21     Median :0.05     Median :0.08          Median :0.07
##  Mean   :0.23     Mean   :0.06     Mean   :0.11          Mean   :0.11
##  3rd Qu.:0.38     3rd Qu.:0.08     3rd Qu.:0.13          3rd Qu.:0.13
##  Max.   :1.00     Max.   :1.00     Max.   :1.00          Max.   :1.00
##  NA's   :208642   NA's   :182590   NA's   :210199        NA's   :154350
##  NONLIVINGAPARTMENTS_MODE NONLIVINGAREA_MODE APARTMENTS_MEDI  BASEMENTAREA_MEDI
##  Min.   :0.00             Min.   :0.00       Min.   :0.00     Min.   :0.00
##  1st Qu.:0.00             1st Qu.:0.00       1st Qu.:0.06     1st Qu.:0.04
##  Median :0.00             Median :0.00       Median :0.09     Median :0.08
##  Mean   :0.01             Mean   :0.03       Mean   :0.12     Mean   :0.09
##  3rd Qu.:0.00             3rd Qu.:0.02       3rd Qu.:0.15     3rd Qu.:0.11
##  Max.   :1.00             Max.   :1.00       Max.   :1.00     Max.   :1.00
##  NA's   :213514           NA's   :169682     NA's   :156061   NA's   :179943
##  YEARS_BEGINEXPLUATATION_MEDI YEARS_BUILD_MEDI COMMONAREA_MEDI
##  Min.   :0.00                 Min.   :0.00     Min.   :0.00
##  1st Qu.:0.98                 1st Qu.:0.69     1st Qu.:0.01
##  Median :0.98                 Median :0.76     Median :0.02
##  Mean   :0.98                 Mean   :0.76     Mean   :0.04
##  3rd Qu.:0.99                 3rd Qu.:0.83     3rd Qu.:0.05
```

```
##  Max.   :1.00                Max.   :1.00     Max.   :1.00
##  NA's   :150007              NA's   :204488   NA's   :214865
##  ELEVATORS_MEDI   ENTRANCES_MEDI   FLOORSMAX_MEDI   FLOORSMIN_MEDI
##  Min.   :0.00     Min.   :0.00     Min.   :0.00     Min.   :0.00
##  1st Qu.:0.00     1st Qu.:0.07     1st Qu.:0.17     1st Qu.:0.08
##  Median :0.00     Median :0.14     Median :0.17     Median :0.21
##  Mean   :0.08     Mean   :0.15     Mean   :0.23     Mean   :0.23
##  3rd Qu.:0.12     3rd Qu.:0.21     3rd Qu.:0.33     3rd Qu.:0.38
##  Max.   :1.00     Max.   :1.00     Max.   :1.00     Max.   :1.00
##  NA's   :163891   NA's   :154828   NA's   :153020   NA's   :208642
##  LANDAREA_MEDI    LIVINGAPARTMENTS_MEDI LIVINGAREA_MEDI
##  Min.   :0.00     Min.   :0.00          Min.   :0.00
##  1st Qu.:0.02     1st Qu.:0.05          1st Qu.:0.05
##  Median :0.05     Median :0.08          Median :0.07
##  Mean   :0.07     Mean   :0.10          Mean   :0.11
##  3rd Qu.:0.09     3rd Qu.:0.12          3rd Qu.:0.13
##  Max.   :1.00     Max.   :1.00          Max.   :1.00
##  NA's   :182590   NA's   :210199        NA's   :154350
##  NONLIVINGAPARTMENTS_MEDI NONLIVINGAREA_MEDI FONDKAPREMONT_MODE
##  Min.   :0.00             Min.   :0.00       Length:307511
##  1st Qu.:0.00             1st Qu.:0.00       Class :character
##  Median :0.00             Median :0.00       Mode  :character
##  Mean   :0.01             Mean   :0.03
##  3rd Qu.:0.00             3rd Qu.:0.03
##  Max.   :1.00             Max.   :1.00
##  NA's   :213514           NA's   :169682
##  HOUSETYPE_MODE     TOTALAREA_MODE   WALLSMATERIAL_MODE EMERGENCYSTATE_MODE
##  Length:307511      Min.   :0.00     Length:307511      Length:307511
##  Class :character   1st Qu.:0.04     Class :character   Class :character
##  Mode  :character   Median :0.07     Mode  :character   Mode  :character
##                     Mean   :0.10
##                     3rd Qu.:0.13
##                     Max.   :1.00
##                     NA's   :148431
##  OBS_30_CNT_SOCIAL_CIRCLE DEF_30_CNT_SOCIAL_CIRCLE OBS_60_CNT_SOCIAL_CIRCLE
##  Min.   :  0.000          Min.   : 0.0000          Min.   :  0.000
##  1st Qu.:  0.000          1st Qu.: 0.0000          1st Qu.:  0.000
##  Median :  0.000          Median : 0.0000          Median :  0.000
##  Mean   :  1.422          Mean   : 0.1434          Mean   :  1.405
##  3rd Qu.:  2.000          3rd Qu.: 0.0000          3rd Qu.:  2.000
##  Max.   :348.000          Max.   :34.0000          Max.   :344.000
##  NA's   :1021             NA's   :1021             NA's   :1021
##  DEF_60_CNT_SOCIAL_CIRCLE DAYS_LAST_PHONE_CHANGE FLAG_DOCUMENT_2
##  Min.   : 0.0             Min.   :-4292.0        Min.   :0.00e+00
##  1st Qu.: 0.0             1st Qu.:-1570.0        1st Qu.:0.00e+00
##  Median : 0.0             Median : -757.0        Median :0.00e+00
##  Mean   : 0.1             Mean   : -962.9        Mean   :4.23e-05
##  3rd Qu.: 0.0             3rd Qu.: -274.0        3rd Qu.:0.00e+00
##  Max.   :24.0             Max.   :    0.0        Max.   :1.00e+00
##  NA's   :1021             NA's   :1
##  FLAG_DOCUMENT_3 FLAG_DOCUMENT_4    FLAG_DOCUMENT_5   FLAG_DOCUMENT_6
##  Min.   :0.00    Min.   :0.00e+00   Min.   :0.00000   Min.   :0.00000
##  1st Qu.:0.00    1st Qu.:0.00e+00   1st Qu.:0.00000   1st Qu.:0.00000
##  Median :1.00    Median :0.00e+00   Median :0.00000   Median :0.00000
##  Mean   :0.71    Mean   :8.13e-05   Mean   :0.01511   Mean   :0.08806
##  3rd Qu.:1.00    3rd Qu.:0.00e+00   3rd Qu.:0.00000   3rd Qu.:0.00000
##  Max.   :1.00    Max.   :1.00e+00   Max.   :1.00000   Max.   :1.00000
##
##  FLAG_DOCUMENT_7     FLAG_DOCUMENT_8   FLAG_DOCUMENT_9    FLAG_DOCUMENT_10
##  Min.   :0.0000000   Min.   :0.00000   Min.   :0.000000   Min.   :0.00e+00
##  1st Qu.:0.0000000   1st Qu.:0.00000   1st Qu.:0.000000   1st Qu.:0.00e+00
##  Median :0.0000000   Median :0.00000   Median :0.000000   Median :0.00e+00
##  Mean   :0.0001919   Mean   :0.08138   Mean   :0.003896   Mean   :2.28e-05
##  3rd Qu.:0.0000000   3rd Qu.:0.00000   3rd Qu.:0.000000   3rd Qu.:0.00e+00
##  Max.   :1.0000000   Max.   :1.00000   Max.   :1.000000   Max.   :1.00e+00
##
##  FLAG_DOCUMENT_11   FLAG_DOCUMENT_12  FLAG_DOCUMENT_13   FLAG_DOCUMENT_14
##  Min.   :0.000000   Min.   :0.0e+00   Min.   :0.000000   Min.   :0.000000
##  1st Qu.:0.000000   1st Qu.:0.0e+00   1st Qu.:0.000000   1st Qu.:0.000000
##  Median :0.000000   Median :0.0e+00   Median :0.000000   Median :0.000000
##  Mean   :0.003912   Mean   :6.5e-06   Mean   :0.003525   Mean   :0.002936
##  3rd Qu.:0.000000   3rd Qu.:0.0e+00   3rd Qu.:0.000000   3rd Qu.:0.000000
##  Max.   :1.000000   Max.   :1.0e+00   Max.   :1.000000   Max.   :1.000000
##
##  FLAG_DOCUMENT_15  FLAG_DOCUMENT_16   FLAG_DOCUMENT_17    FLAG_DOCUMENT_18
##  Min.   :0.00000   Min.   :0.000000   Min.   :0.0000000   Min.   :0.00000
##  1st Qu.:0.00000   1st Qu.:0.000000   1st Qu.:0.0000000   1st Qu.:0.00000
##  Median :0.00000   Median :0.000000   Median :0.0000000   Median :0.00000
##  Mean   :0.00121   Mean   :0.009928   Mean   :0.0002667   Mean   :0.00813
##  3rd Qu.:0.00000   3rd Qu.:0.000000   3rd Qu.:0.0000000   3rd Qu.:0.00000
##  Max.   :1.00000   Max.   :1.000000   Max.   :1.0000000   Max.   :1.00000
##
##  FLAG_DOCUMENT_19    FLAG_DOCUMENT_20    FLAG_DOCUMENT_21
```

```
##  Min.   :0.0000000   Min.   :0.0000000   Min.   :0.0000000
##  1st Qu.:0.0000000   1st Qu.:0.0000000   1st Qu.:0.0000000
##  Median :0.0000000   Median :0.0000000   Median :0.0000000
##  Mean   :0.0005951   Mean   :0.0005073   Mean   :0.0003349
##  3rd Qu.:0.0000000   3rd Qu.:0.0000000   3rd Qu.:0.0000000
##  Max.   :1.0000000   Max.   :1.0000000   Max.   :1.0000000
##
##  AMT_REQ_CREDIT_BUREAU_HOUR AMT_REQ_CREDIT_BUREAU_DAY
##  Min.   :0.00               Min.   :0.00
##  1st Qu.:0.00               1st Qu.:0.00
##  Median :0.00               Median :0.00
##  Mean   :0.01               Mean   :0.01
##  3rd Qu.:0.00               3rd Qu.:0.00
##  Max.   :4.00               Max.   :9.00
##  NA's   :41519              NA's   :41519
##  AMT_REQ_CREDIT_BUREAU_WEEK AMT_REQ_CREDIT_BUREAU_MON AMT_REQ_CREDIT_BUREAU_QRT
##  Min.   :0.00               Min.   : 0.00             Min.   :  0.00
##  1st Qu.:0.00               1st Qu.: 0.00             1st Qu.:  0.00
##  Median :0.00               Median : 0.00             Median :  0.00
##  Mean   :0.03               Mean   : 0.27             Mean   :  0.27
##  3rd Qu.:0.00               3rd Qu.: 0.00             3rd Qu.:  0.00
##  Max.   :8.00               Max.   :27.00             Max.   :261.00
##  NA's   :41519              NA's   :41519             NA's   :41519
##  AMT_REQ_CREDIT_BUREAU_YEAR
##  Min.   : 0.0
##  1st Qu.: 0.0
##  Median : 1.0
##  Mean   : 1.9
##  3rd Qu.: 3.0
##  Max.   :25.0
##  NA's   :41519
```

- We see that the data can be divided into sections:
- *Information about where the client lives:* There are a lot of columns that show normalized information about the different statistics (MEAN, MEDIAN, MODE) about measurements regarding where the client lives. Not only are these columns extra and will have negligible effect on our data, but most of the values in these columns are NAs as it is. As an example, lets look at two columns from this list:

```
summary(raw_data$NONLIVINGAREA_MODE)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    0.00    0.00    0.00    0.03    0.02    1.00  169682
```

```
summary(raw_data$BASEMENTAREA_MEDI)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    0.00    0.04    0.08    0.09    0.11    1.00  179943
```

- As we can see, the values already stay close to 0 and the number of values of NA are 169682, which is more than 50% of our data. Hence, we will remove these columns from our data.

```
clean<-subset(raw_data,select = -c( NONLIVINGAREA_MODE,OWN_CAR_AGE,EXT_SOURCE_1,APARTMENTS_AVG,BASEMENTAREA_AVG,YEARS_BEGINE
XPLUATATION_AVG,YEARS_BUILD_AVG,COMMONAREA_AVG,ELEVATORS_AVG,ENTRANCES_AVG,FLOORSMAX_AVG,FLOORSMIN_AVG,LANDAREA_AVG,LIVINGAP
ARTMENTS_AVG,LIVINGAREA_AVG,NONLIVINGAPARTMENTS_AVG,NONLIVINGAREA_AVG,APARTMENTS_MODE,BASEMENTAREA_MODE,YEARS_BEGINEXPLUATAT
ION_MODE,YEARS_BUILD_MODE,COMMONAREA_MODE,ELEVATORS_MODE,ENTRANCES_MODE,FLOORSMAX_MODE,FLOORSMIN_MODE,LANDAREA_MODE,LIVINGAP
ARTMENTS_MODE,LIVINGAREA_MODE,NONLIVINGAPARTMENTS_MODE,NONLIVINGAPARTMENTS_MODE,APARTMENTS_MEDI,BASEMENTAREA_MEDI,YEARS_BEGI
NEXPLUATATION_MEDI,YEARS_BUILD_MEDI,COMMONAREA_MEDI,ELEVATORS_MEDI,ENTRANCES_MEDI,FLOORSMAX_MEDI,FLOORSMIN_MEDI,LANDAREA_MED
I,LIVINGAPARTMENTS_MEDI,LIVINGAREA_MEDI,NONLIVINGAPARTMENTS_MEDI,NONLIVINGAREA_MEDI,FONDKAPREMONT_MODE,HOUSETYPE_MODE,TOTALA
REA_MODE,WALLSMATERIAL_MODE,EMERGENCYSTATE_MODE))
```

- *Variables with no variation:* There are some FLAG_DOCUMENTS that were supposed to be provided by clients. These flag variables represent whether particular documents have been submitted or not by the clients. However, we can see that the variation in these is very less. For example, in the column 'FLAG_DOCUMENT_10', the number of 0s are 307504 , while the total number of rows are 307511. This trend is followed throughout these columns and hence we will remove them from our data as well.

```
clean<-subset(clean,select=-c(FLAG_DOCUMENT_2, FLAG_DOCUMENT_3,FLAG_DOCUMENT_4, FLAG_DOCUMENT_5, FLAG_DOCUMENT_6,FLAG_DOCUME
NT_7, FLAG_DOCUMENT_8, FLAG_DOCUMENT_9,FLAG_DOCUMENT_10, FLAG_DOCUMENT_11, FLAG_DOCUMENT_12,FLAG_DOCUMENT_13, FLAG_DOCUMENT_
14, FLAG_DOCUMENT_15,FLAG_DOCUMENT_16, FLAG_DOCUMENT_17, FLAG_DOCUMENT_18,FLAG_DOCUMENT_19, FLAG_DOCUMENT_20, FLAG_DOCUMENT_
21))
```

- *removing vaiables based on insight and research question:* We were able to remove some further variables absed upon logical thinking. These variables would clearly not have a significant effect on the TARGET value and removing these still leaves us with enough predictors to make a good model.Furthermore, some of these variables were used to remove possible bias. For example, 'REGION_RATING_CLIENT' was a variable which gave a rating for the region where the client lives. This rating was determined by the bank itself and hence could cause a bias.

```
clean <-subset(clean,select=-c(FLAG_EMP_PHONE, FLAG_WORK_PHONE, FLAG_CONT_MOBILE, FLAG_PHONE, REGION_RATING_CLIENT, REGION_R
ATING_CLIENT_W_CITY,OBS_30_CNT_SOCIAL_CIRCLE, DEF_30_CNT_SOCIAL_CIRCLE, OBS_60_CNT_SOCIAL_CIRCLE, DEF_60_CNT_SOCIAL_CIRCLE,A
MT_REQ_CREDIT_BUREAU_HOUR, AMT_REQ_CREDIT_BUREAU_DAY, AMT_REQ_CREDIT_BUREAU_WEEK, AMT_REQ_CREDIT_BUREAU_MON, AMT_REQ_CREDIT_
BUREAU_QRT, AMT_REQ_CREDIT_BUREAU_YEAR,NAME_TYPE_SUITE,REGION_POPULATION_RELATIVE, WEEKDAY_APPR_PROCESS_START, HOUR_APPR_PRO
CESS_START, REG_REGION_NOT_LIVE_REGION, REG_REGION_NOT_WORK_REGION, LIVE_REGION_NOT_WORK_REGION, REG_CITY_NOT_LIVE_CITY, REG
_CITY_NOT_WORK_CITY, LIVE_CITY_NOT_WORK_CITY, SK_ID_CURR))
summary(clean)
```

```
##      TARGET        NAME_CONTRACT_TYPE CODE_GENDER        FLAG_OWN_CAR
##  Min.   :0.00000   Length:307511      Length:307511      Length:307511
##  1st Qu.:0.00000   Class :character   Class :character   Class :character
##  Median :0.00000   Mode  :character   Mode  :character   Mode  :character
##  Mean   :0.08073
##  3rd Qu.:0.00000
##  Max.   :1.00000
##
##  FLAG_OWN_REALTY     CNT_CHILDREN     AMT_INCOME_TOTAL      AMT_CREDIT
##  Length:307511      Min.   : 0.0000   Min.   :    25650   Min.   :  45000
##  Class :character   1st Qu.: 0.0000   1st Qu.:   112500   1st Qu.: 270000
##  Mode  :character   Median : 0.0000   Median :   147150   Median : 513531
##                     Mean   : 0.4171   Mean   :   168798   Mean   : 599026
##                     3rd Qu.: 1.0000   3rd Qu.:   202500   3rd Qu.: 808650
##                     Max.   :19.0000   Max.   :117000000   Max.   :4050000
##
##   AMT_ANNUITY     AMT_GOODS_PRICE   NAME_INCOME_TYPE   NAME_EDUCATION_TYPE
##  Min.   :  1616   Min.   :  40500   Length:307511      Length:307511
##  1st Qu.: 16524   1st Qu.: 238500   Class :character   Class :character
##  Median : 24903   Median : 450000   Mode  :character   Mode  :character
##  Mean   : 27109   Mean   : 538396
##  3rd Qu.: 34596   3rd Qu.: 679500
##  Max.   :258026   Max.   :4050000
##  NA's   :12       NA's   :278
##  NAME_FAMILY_STATUS NAME_HOUSING_TYPE    DAYS_BIRTH     DAYS_EMPLOYED
##  Length:307511      Length:307511      Min.   :-25229   Min.   :-17912
##  Class :character   Class :character   1st Qu.:-19682   1st Qu.: -2760
##  Mode  :character   Mode  :character   Median :-15750   Median : -1213
##                                        Mean   :-16037   Mean   : 63815
##                                        3rd Qu.:-12413   3rd Qu.:  -289
##                                        Max.   : -7489   Max.   :365243
##
##  DAYS_REGISTRATION DAYS_ID_PUBLISH   FLAG_MOBIL   FLAG_EMAIL
##  Min.   :-24672    Min.   :-7197   Min.   :0    Min.   :0.00000
##  1st Qu.: -7480    1st Qu.:-4299   1st Qu.:1    1st Qu.:0.00000
##  Median : -4504    Median :-3254   Median :1    Median :0.00000
##  Mean   : -4986    Mean   :-2994   Mean   :1    Mean   :0.05672
##  3rd Qu.: -2010    3rd Qu.:-1720   3rd Qu.:1    3rd Qu.:0.00000
##  Max.   :     0    Max.   :    0   Max.   :1    Max.   :1.00000
##
##  OCCUPATION_TYPE    CNT_FAM_MEMBERS  ORGANIZATION_TYPE   EXT_SOURCE_2
##  Length:307511      Min.   : 1.000   Length:307511      Min.   :0.0000
##  Class :character   1st Qu.: 2.000   Class :character   1st Qu.:0.3925
##  Mode  :character   Median : 2.000   Mode  :character   Median :0.5660
##                     Mean   : 2.153                      Mean   :0.5144
##                     3rd Qu.: 3.000                      3rd Qu.:0.6636
##                     Max.   :20.000                      Max.   :0.8550
##                     NA's   :2                           NA's   :660
##   EXT_SOURCE_3   DAYS_LAST_PHONE_CHANGE
##  Min.   :0.00    Min.   :-4292.0
##  1st Qu.:0.37    1st Qu.:-1570.0
##  Median :0.54    Median : -757.0
##  Mean   :0.51    Mean   : -962.9
##  3rd Qu.:0.67    3rd Qu.: -274.0
##  Max.   :0.90    Max.   :    0.0
##  NA's   :60965   NA's   :1
```

- Hence, we now have a data set with statistically significant columns to serve as predictors for our model. We have removed labels which would not have helped our model and would have only added noise. While doing so, we made sure not to cause any **omitted variable bias**. This bias could happen if we remove variables which have a strong effect on the dependent variable, and could cause an overstatement of the effect of the predictors we chose to keep.
- **Rows**
- We have a large amount of data available to us and we now cleaned it to have an effective data set. Let us visualise what the rows contain:

```
str(raw_data)
```

```
## 'data.frame':    307511 obs. of  122 variables:
##  $ SK_ID_CURR                  : int  100002 100003 100004 100006 100007 100008 100009 100010 100011 100012 ...
##  $ TARGET                      : int  1 0 0 0 0 0 0 0 0 0 ...
##  $ NAME_CONTRACT_TYPE          : chr  "Cash loans" "Cash loans" "Revolving loans" "Cash loans" ...
##  $ CODE_GENDER                 : chr  "M" "F" "M" "F" ...
##  $ FLAG_OWN_CAR                : chr  "N" "N" "Y" "N" ...
##  $ FLAG_OWN_REALTY             : chr  "Y" "N" "Y" "Y" ...
##  $ CNT_CHILDREN                : int  0 0 0 0 0 0 1 0 0 0 ...
##  $ AMT_INCOME_TOTAL            : num  202500 270000 67500 135000 121500 ...
##  $ AMT_CREDIT                  : num  406598 1293503 135000 312683 513000 ...
##  $ AMT_ANNUITY                 : num  24701 35699 6750 29687 21866 ...
##  $ AMT_GOODS_PRICE             : num  351000 1129500 135000 297000 513000 ...
##  $ NAME_TYPE_SUITE             : chr  "Unaccompanied" "Family" "Unaccompanied" "Unaccompanied" ...
##  $ NAME_INCOME_TYPE            : chr  "Working" "State servant" "Working" "Working" ...
##  $ NAME_EDUCATION_TYPE         : chr  "Secondary / secondary special" "Higher education" "Secondary / secondary special"
"Secondary / secondary special" ...
##  $ NAME_FAMILY_STATUS          : chr  "Single / not married" "Married" "Single / not married" "Civil marriage" ...
##  $ NAME_HOUSING_TYPE           : chr  "House / apartment" "House / apartment" "House / apartment" "House / apartment" ...
##  $ REGION_POPULATION_RELATIVE  : num  0.0188 0.00354 0.01003 0.00802 0.02866 ...
##  $ DAYS_BIRTH                  : int  -9461 -16765 -19046 -19005 -19932 -16941 -13778 -18850 -20099 -14469 ...
##  $ DAYS_EMPLOYED               : int  -637 -1188 -225 -3039 -3038 -1588 -3130 -449 365243 -2019 ...
##  $ DAYS_REGISTRATION           : num  -3648 -1186 -4260 -9833 -4311 ...
##  $ DAYS_ID_PUBLISH             : int  -2120 -291 -2531 -2437 -3458 -477 -619 -2379 -3514 -3992 ...
##  $ OWN_CAR_AGE                 : num  NA NA 26 NA NA NA 17 8 NA NA ...
##  $ FLAG_MOBIL                  : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ FLAG_EMP_PHONE              : int  1 1 1 1 1 1 1 1 0 1 ...
##  $ FLAG_WORK_PHONE             : int  0 0 1 0 0 1 0 1 0 0 ...
##  $ FLAG_CONT_MOBILE            : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ FLAG_PHONE                  : int  1 1 1 0 0 1 1 0 0 0 ...
##  $ FLAG_EMAIL                  : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ OCCUPATION_TYPE             : chr  "Laborers" "Core staff" "Laborers" "Laborers" ...
##  $ CNT_FAM_MEMBERS             : num  1 2 1 2 1 2 3 2 2 1 ...
##  $ REGION_RATING_CLIENT        : int  2 1 2 2 2 2 2 3 2 2 ...
##  $ REGION_RATING_CLIENT_W_CITY : int  2 1 2 2 2 2 2 3 2 2 ...
##  $ WEEKDAY_APPR_PROCESS_START  : chr  "WEDNESDAY" "MONDAY" "MONDAY" "WEDNESDAY" ...
##  $ HOUR_APPR_PROCESS_START     : int  10 11 9 17 11 16 16 16 14 8 ...
##  $ REG_REGION_NOT_LIVE_REGION  : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ REG_REGION_NOT_WORK_REGION  : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ LIVE_REGION_NOT_WORK_REGION : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ REG_CITY_NOT_LIVE_CITY      : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ REG_CITY_NOT_WORK_CITY      : int  0 0 0 0 1 0 0 1 0 0 ...
##  $ LIVE_CITY_NOT_WORK_CITY     : int  0 0 0 0 1 0 0 1 0 0 ...
##  $ ORGANIZATION_TYPE           : chr  "Business Entity Type 3" "School" "Government" "Business Entity Type 3" ...
##  $ EXT_SOURCE_1                : num  0.083 0.311 NA NA NA ...
##  $ EXT_SOURCE_2                : num  0.263 0.622 0.556 0.65 0.323 ...
##  $ EXT_SOURCE_3                : num  0.139 NA 0.73 NA NA ...
##  $ APARTMENTS_AVG              : num  0.0247 0.0959 NA NA NA NA NA NA NA NA ...
##  $ BASEMENTAREA_AVG            : num  0.0369 0.0529 NA NA NA NA NA NA NA NA ...
##  $ YEARS_BEGINEXPLUATATION_AVG : num  0.972 0.985 NA NA NA ...
##  $ YEARS_BUILD_AVG             : num  0.619 0.796 NA NA NA ...
##  $ COMMONAREA_AVG              : num  0.0143 0.0605 NA NA NA NA NA NA NA NA ...
##  $ ELEVATORS_AVG               : num  0 0.08 NA NA NA NA NA NA NA NA ...
##  $ ENTRANCES_AVG               : num  0.069 0.0345 NA NA NA NA NA NA NA NA ...
##  $ FLOORSMAX_AVG               : num  0.0833 0.2917 NA NA NA ...
##  $ FLOORSMIN_AVG               : num  0.125 0.333 NA NA NA ...
##  $ LANDAREA_AVG                : num  0.0369 0.013 NA NA NA NA NA NA NA NA ...
##  $ LIVINGAPARTMENTS_AVG        : num  0.0202 0.0773 NA NA NA NA NA NA NA NA ...
##  $ LIVINGAREA_AVG              : num  0.019 0.0549 NA NA NA NA NA NA NA NA ...
##  $ NONLIVINGAPARTMENTS_AVG     : num  0 0.0039 NA NA NA NA NA NA NA NA ...
##  $ NONLIVINGAREA_AVG           : num  0 0.0098 NA NA NA NA NA NA NA NA ...
##  $ APARTMENTS_MODE             : num  0.0252 0.0924 NA NA NA NA NA NA NA NA ...
##  $ BASEMENTAREA_MODE           : num  0.0383 0.0538 NA NA NA NA NA NA NA NA ...
##  $ YEARS_BEGINEXPLUATATION_MODE: num  0.972 0.985 NA NA NA ...
##  $ YEARS_BUILD_MODE            : num  0.634 0.804 NA NA NA ...
##  $ COMMONAREA_MODE             : num  0.0144 0.0497 NA NA NA NA NA NA NA NA ...
##  $ ELEVATORS_MODE              : num  0 0.0806 NA NA NA NA NA NA NA NA ...
##  $ ENTRANCES_MODE              : num  0.069 0.0345 NA NA NA NA NA NA NA NA ...
##  $ FLOORSMAX_MODE              : num  0.0833 0.2917 NA NA NA ...
##  $ FLOORSMIN_MODE              : num  0.125 0.333 NA NA NA ...
##  $ LANDAREA_MODE               : num  0.0377 0.0128 NA NA NA NA NA NA NA NA ...
##  $ LIVINGAPARTMENTS_MODE       : num  0.022 0.079 NA NA NA NA NA NA NA NA ...
##  $ LIVINGAREA_MODE             : num  0.0198 0.0554 NA NA NA NA NA NA NA NA ...
##  $ NONLIVINGAPARTMENTS_MODE    : num  0 0 NA NA NA NA NA NA NA NA ...
##  $ NONLIVINGAREA_MODE          : num  0 0 NA NA NA NA NA NA NA NA ...
##  $ APARTMENTS_MEDI             : num  0.025 0.0968 NA NA NA NA NA NA NA NA ...
##  $ BASEMENTAREA_MEDI           : num  0.0369 0.0529 NA NA NA NA NA NA NA NA ...
##  $ YEARS_BEGINEXPLUATATION_MEDI: num  0.972 0.985 NA NA NA ...
##  $ YEARS_BUILD_MEDI            : num  0.624 0.799 NA NA NA ...
##  $ COMMONAREA_MEDI             : num  0.0144 0.0608 NA NA NA NA NA NA NA NA ...
##  $ ELEVATORS_MEDI              : num  0 0.08 NA NA NA NA NA NA NA NA ...
##  $ ENTRANCES_MEDI              : num  0.069 0.0345 NA NA NA NA NA NA NA NA ...
##  $ FLOORSMAX_MEDI              : num  0.0833 0.2917 NA NA NA ...
##  $ FLOORSMIN_MEDI              : num  0.125 0.333 NA NA NA ...
```

```
##  $ LANDAREA_MEDI               : num  0.0375 0.0132 NA NA NA NA NA NA NA NA ...
##  $ LIVINGAPARTMENTS_MEDI       : num  0.0205 0.0787 NA NA NA NA NA NA NA NA ...
##  $ LIVINGAREA_MEDI             : num  0.0193 0.0558 NA NA NA NA NA NA NA NA ...
##  $ NONLIVINGAPARTMENTS_MEDI    : num  0 0.0039 NA NA NA NA NA NA NA NA ...
##  $ NONLIVINGAREA_MEDI          : num  0 0.01 NA NA NA NA NA NA NA NA ...
##  $ FONDKAPREMONT_MODE          : chr  "reg oper account" "reg oper account" "" "" ...
##  $ HOUSETYPE_MODE              : chr  "block of flats" "block of flats" "" "" ...
##  $ TOTALAREA_MODE              : num  0.0149 0.0714 NA NA NA NA NA NA NA NA ...
##  $ WALLSMATERIAL_MODE          : chr  "Stone, brick" "Block" "" "" ...
##  $ EMERGENCYSTATE_MODE         : chr  "No" "No" "" "" ...
##  $ OBS_30_CNT_SOCIAL_CIRCLE    : num  2 1 0 2 0 0 1 2 1 2 ...
##  $ DEF_30_CNT_SOCIAL_CIRCLE    : num  2 0 0 0 0 0 0 0 0 0 ...
##  $ OBS_60_CNT_SOCIAL_CIRCLE    : num  2 1 0 2 0 0 1 2 1 2 ...
##  $ DEF_60_CNT_SOCIAL_CIRCLE    : num  2 0 0 0 0 0 0 0 0 0 ...
##  $ DAYS_LAST_PHONE_CHANGE      : num  -1134 -828 -815 -617 -1106 ...
##  $ FLAG_DOCUMENT_2             : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ FLAG_DOCUMENT_3             : int  1 1 0 1 0 1 0 1 1 0 ...
##  $ FLAG_DOCUMENT_4             : int  0 0 0 0 0 0 0 0 0 0 ...
##   [list output truncated]
```

- Here we noticed a couple of problems:

1. The variables for which factors or categorical values were used were being cast to integer or char data types.

- For this we converted these values to factor by using the `as.factor` function in R. However, we realized that our model while training the model automatically treats for this so we ended up letting the model do it and kept this values in the already assigned data types.

2. The variables that measured days had values in negative.

- A few simple lines were able to help us solve this problem.

```
clean$DAYS_BIRTH<-abs(clean$DAYS_BIRTH)
clean$DAYS_EMPLOYED<-abs(clean$DAYS_EMPLOYED)
clean$DAYS_REGISTRATION<-abs(clean$DAYS_REGISTRATION)
clean$DAYS_ID_PUBLISH<-abs(clean$DAYS_ID_PUBLISH)
clean$DAYS_LAST_PHONE_CHANGE<-abs(clean$DAYS_LAST_PHONE_CHANGE)
```

3. There were quite a lot of NULL values.

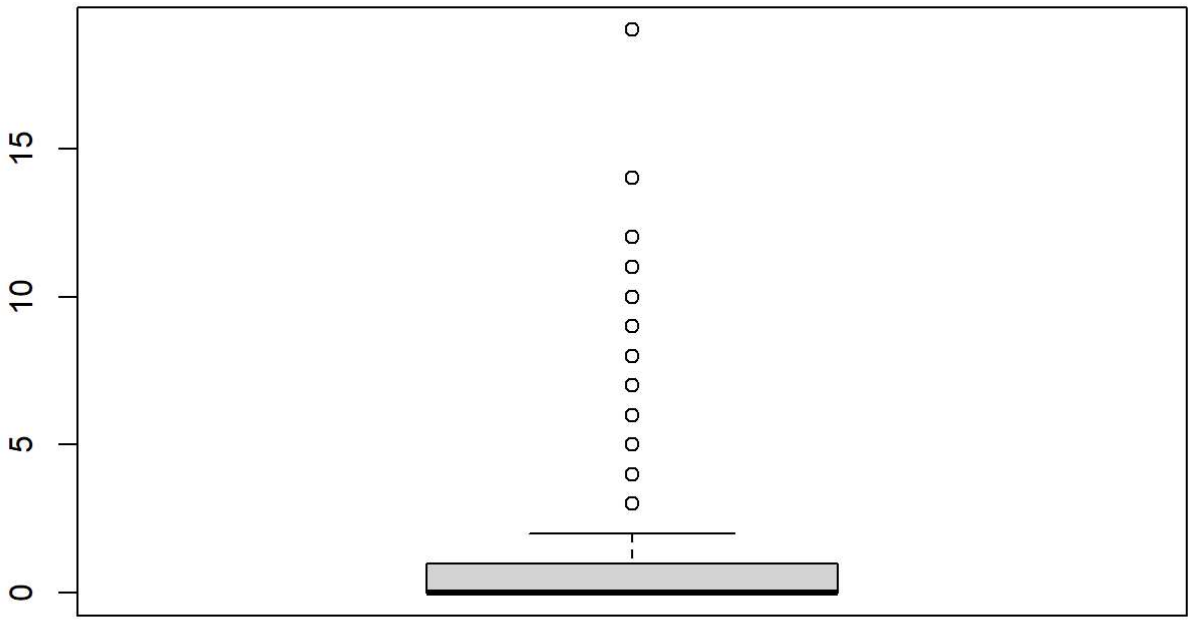
- To handle this problem, we were going to simply remove the rows with NULL values. However, we realized that in some cases, the NULL (or NA) values were justified, such as 'EXT_SOURCE_2' or 'ORGANIZATION TYPE'. This is because these values are optional and can be left empty, and having these values as NULL actually might have a strong effect on the output of the 'TARGET'. Therefore, we changed the NA value for these to "Unknown" for intuitive sense. For the rest of the variables, we used the deletion method and eliminated the rows with NULL values to make sure that there is no incomplete rows that could corrupt our data.

```
# Replace NULL VALUES:

clean$ORGANIZATION_TYPE[clean$ORGANIZATION_TYPE == "XNA"] <- "Unknown"
clean$OCCUPATION_TYPE[clean$OCCUPATION_TYPE == ""] <- "UNKNOWN"


# Removing rows with NULL values:
clean_data <- clean[!(is.na(clean$AMT_ANNUITY)|is.na(clean$AMT_GOODS_PRICE)|is.na(clean$CNT_FAM_MEMBERS)|is.na(clean$DAYS_LA
ST_PHONE_CHANGE)|clean$CODE_GENDER == "XNA"),]
summary(clean_data)
```

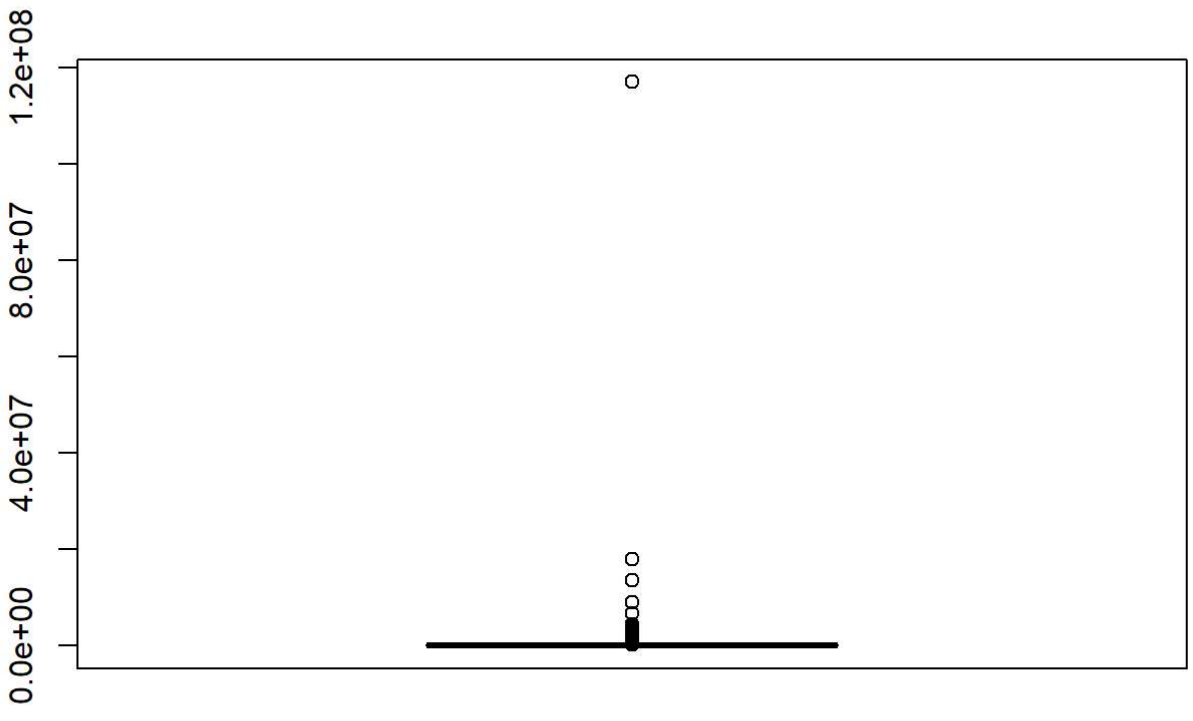
```
##      TARGET        NAME_CONTRACT_TYPE CODE_GENDER        FLAG_OWN_CAR
##  Min.   :0.00000   Length:307216      Length:307216      Length:307216
##  1st Qu.:0.00000   Class :character   Class :character   Class :character
##  Median :0.00000   Mode  :character   Mode  :character   Mode  :character
##  Mean   :0.08074
##  3rd Qu.:0.00000
##  Max.   :1.00000
##
##  FLAG_OWN_REALTY     CNT_CHILDREN    AMT_INCOME_TOTAL      AMT_CREDIT
##  Length:307216      Min.   : 0.000   Min.   :    25650   Min.   :  45000
##  Class :character   1st Qu.: 0.000   1st Qu.:   112500   1st Qu.: 270000
##  Mode  :character   Median : 0.000   Median :   148500   Median : 514602
##                     Mean   : 0.417   Mean   :   168832   Mean   : 599320
##                     3rd Qu.: 1.000   3rd Qu.:   202500   3rd Qu.: 808650
##                     Max.   :19.000   Max.   :117000000   Max.   :4050000
##
##   AMT_ANNUITY     AMT_GOODS_PRICE   NAME_INCOME_TYPE   NAME_EDUCATION_TYPE
##  Min.   :  1616   Min.   :  40500   Length:307216      Length:307216
##  1st Qu.: 16551   1st Qu.: 238500   Class :character   Class :character
##  Median : 24917   Median : 450000   Mode  :character   Mode  :character
##  Mean   : 27121   Mean   : 538400
##  3rd Qu.: 34596   3rd Qu.: 679500
##  Max.   :258026   Max.   :4050000
##
##  NAME_FAMILY_STATUS NAME_HOUSING_TYPE    DAYS_BIRTH    DAYS_EMPLOYED
##  Length:307216      Length:307216      Min.   : 7489   Min.   :     0
##  Class :character   Class :character   1st Qu.:12415   1st Qu.:   933
##  Mode  :character   Mode  :character   Median :15753   Median :  2219
##                                        Mean   :16039   Mean   : 67762
##                                        3rd Qu.:19684   3rd Qu.:  5711
##                                        Max.   :25229   Max.   :365243
##
##  DAYS_REGISTRATION DAYS_ID_PUBLISH   FLAG_MOBIL   FLAG_EMAIL
##  Min.   :    0     Min.   :   0    Min.   :1    Min.   :0.00000
##  1st Qu.: 2010     1st Qu.:1720    1st Qu.:1    1st Qu.:0.00000
##  Median : 4504     Median :3255    Median :1    Median :0.00000
##  Mean   : 4986     Mean   :2994    Mean   :1    Mean   :0.05671
##  3rd Qu.: 7480     3rd Qu.:4299    3rd Qu.:1    3rd Qu.:0.00000
##  Max.   :24672     Max.   :7197    Max.   :1    Max.   :1.00000
##
##  OCCUPATION_TYPE    CNT_FAM_MEMBERS  ORGANIZATION_TYPE   EXT_SOURCE_2
##  Length:307216      Min.   : 1.000   Length:307216      Min.   :0.0000
##  Class :character   1st Qu.: 2.000   Class :character   1st Qu.:0.3924
##  Mode  :character   Median : 2.000   Mode  :character   Median :0.5660
##                     Mean   : 2.153                      Mean   :0.5144
##                     3rd Qu.: 3.000                      3rd Qu.:0.6636
##                     Max.   :20.000                      Max.   :0.8550
##                                                         NA's   :658
##   EXT_SOURCE_3   DAYS_LAST_PHONE_CHANGE
##  Min.   :0.00    Min.   :   0.0
##  1st Qu.:0.37    1st Qu.: 274.0
##  Median :0.54    Median : 757.0
##  Mean   :0.51    Mean   : 962.9
##  3rd Qu.:0.67    3rd Qu.:1570.0
##  Max.   :0.90    Max.   :4292.0
##  NA's   :60895
```

To further increase the integrity of our data, we used box plots to identify outliers and removed them from our data space. We made box plots only for numeric data as trying to find outliers using box plots in character data will give erroneous results.

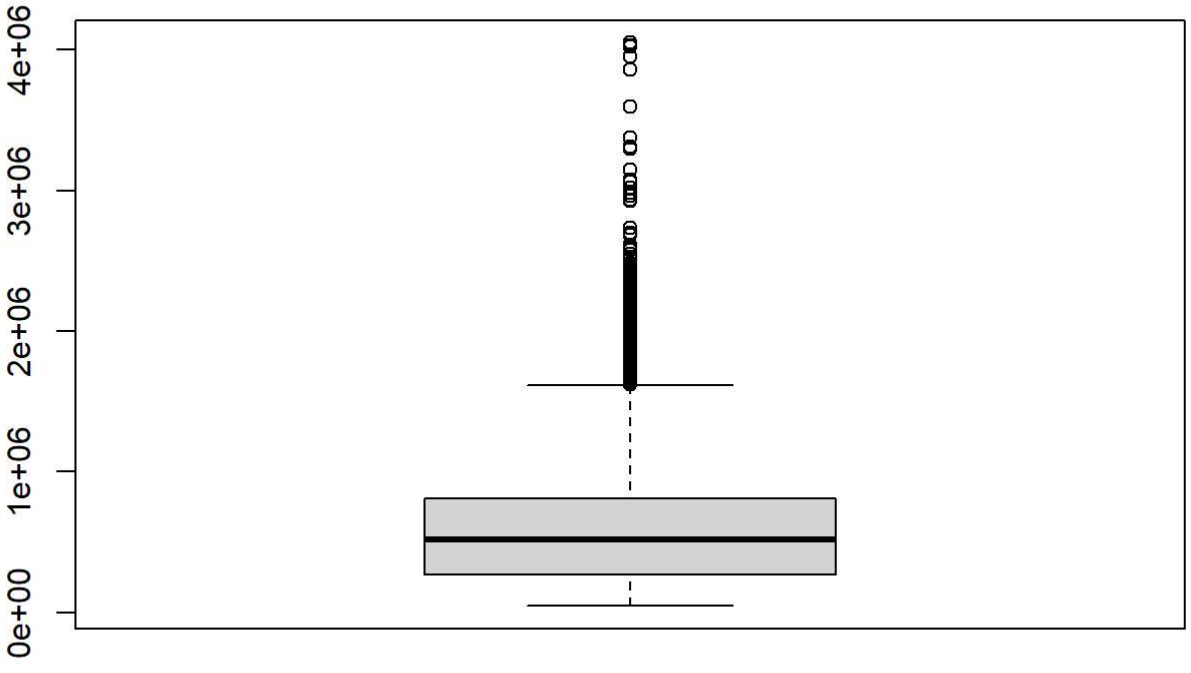
```
#Making Boxplots for numeric data to identify outliers
boxplot(clean_data$CNT_CHILDREN)
```



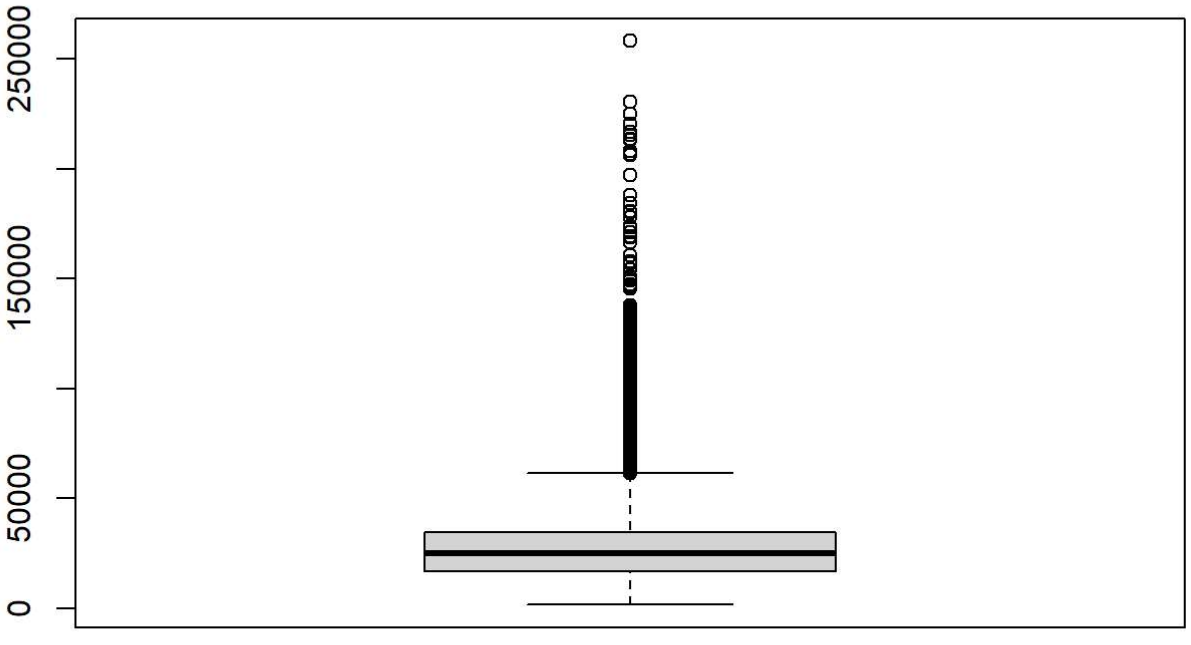

```
boxplot(clean_data$AMT_INCOME_TOTAL)
```



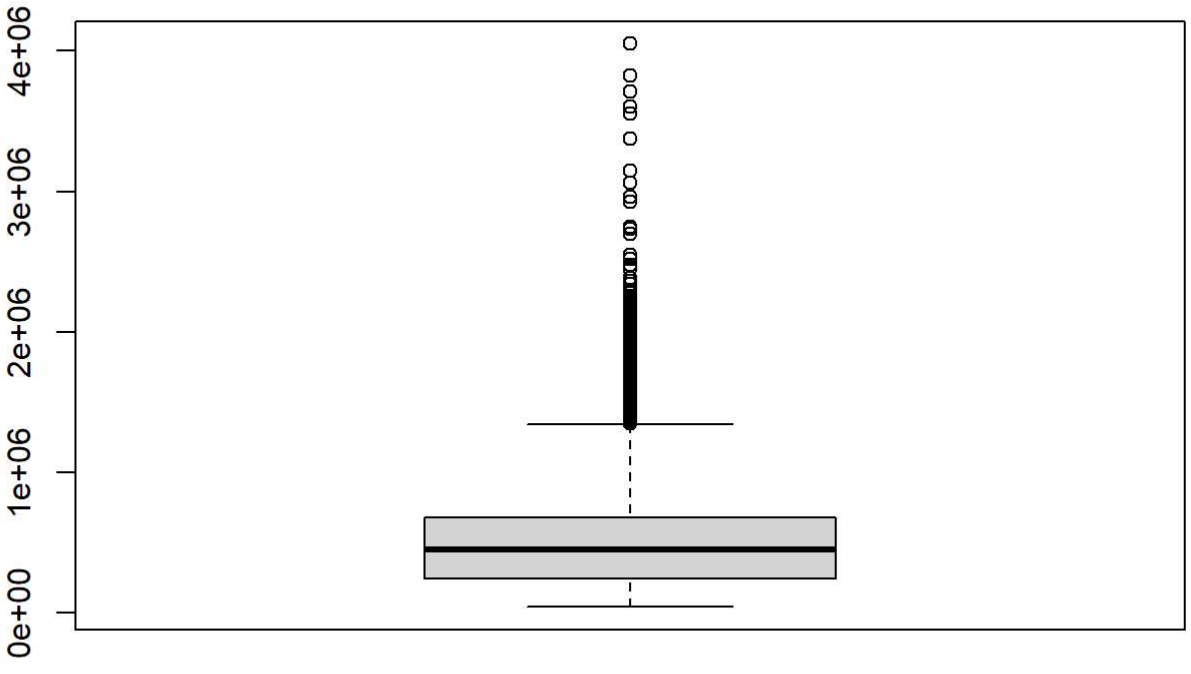

```
boxplot(clean_data$AMT_CREDIT)
```



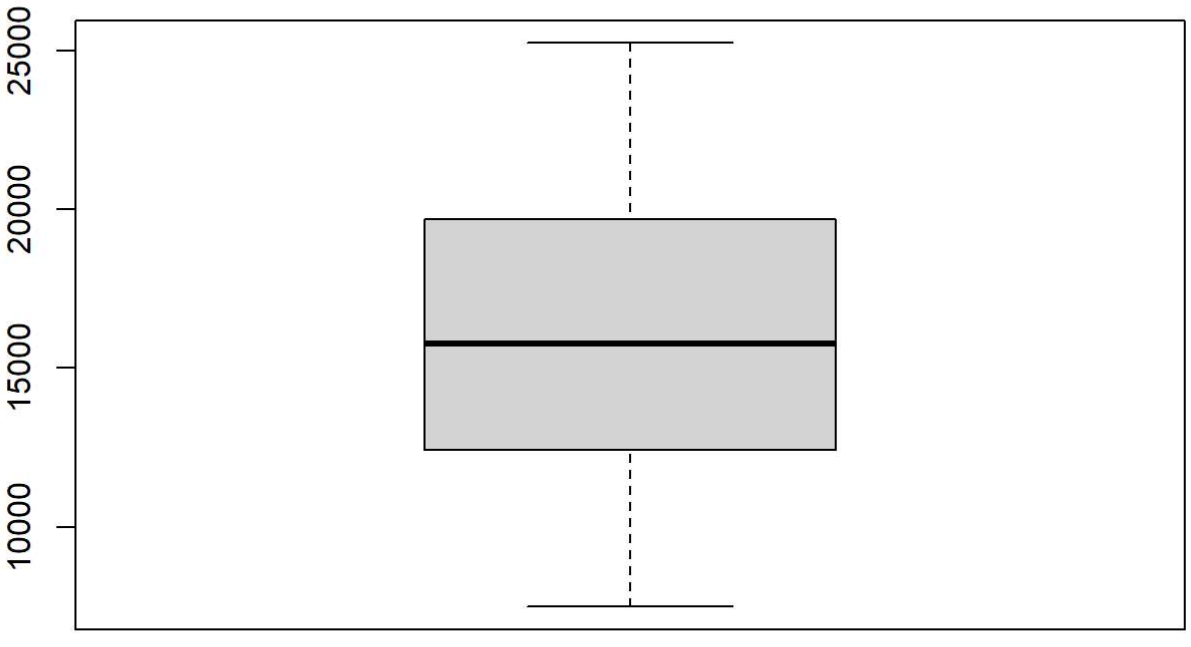

```
boxplot(clean_data$AMT_ANNUITY)
```



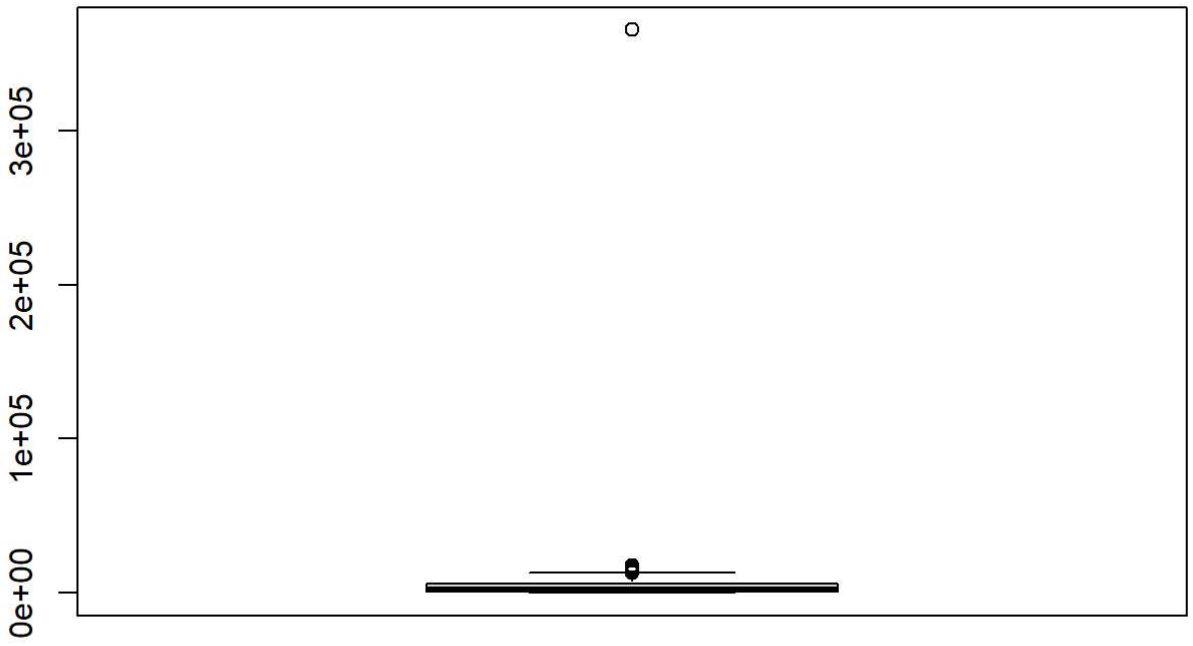

```
boxplot(clean_data$AMT_GOODS_PRICE)
```



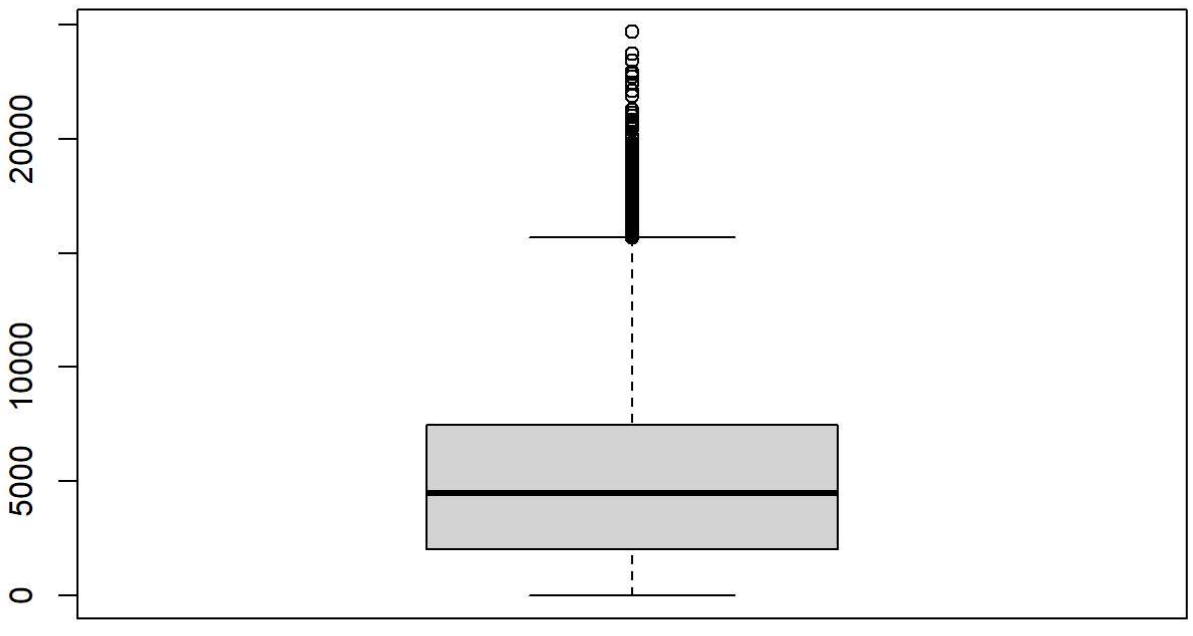

```
boxplot(clean_data$DAYS_BIRTH)
```



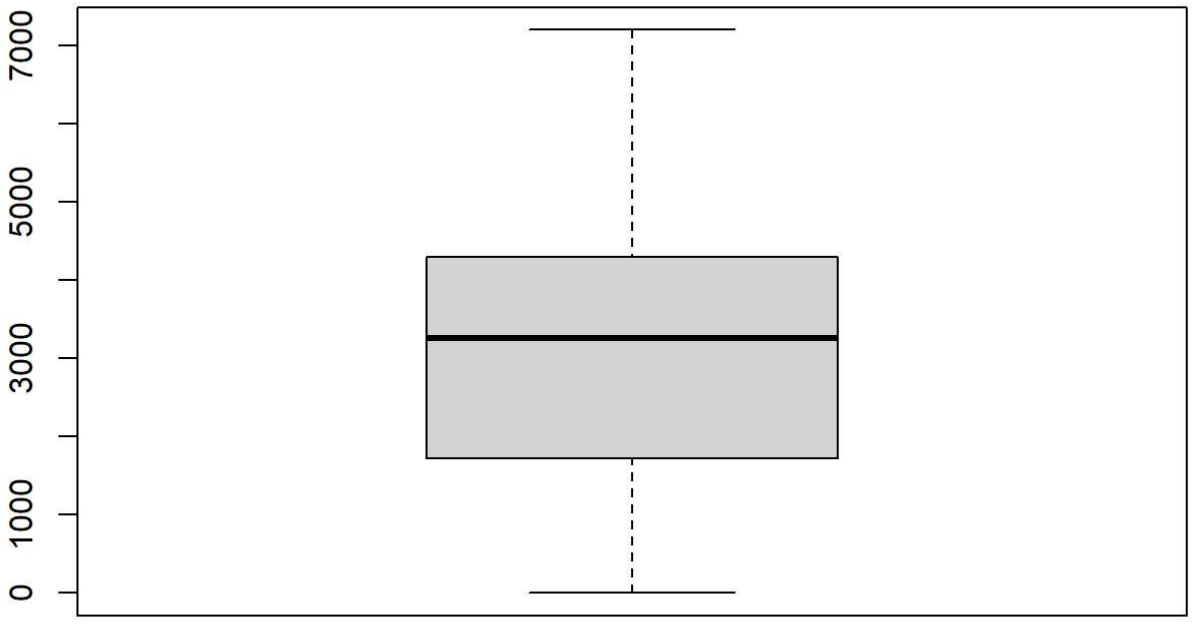

```
boxplot(clean_data$DAYS_EMPLOYED)
```



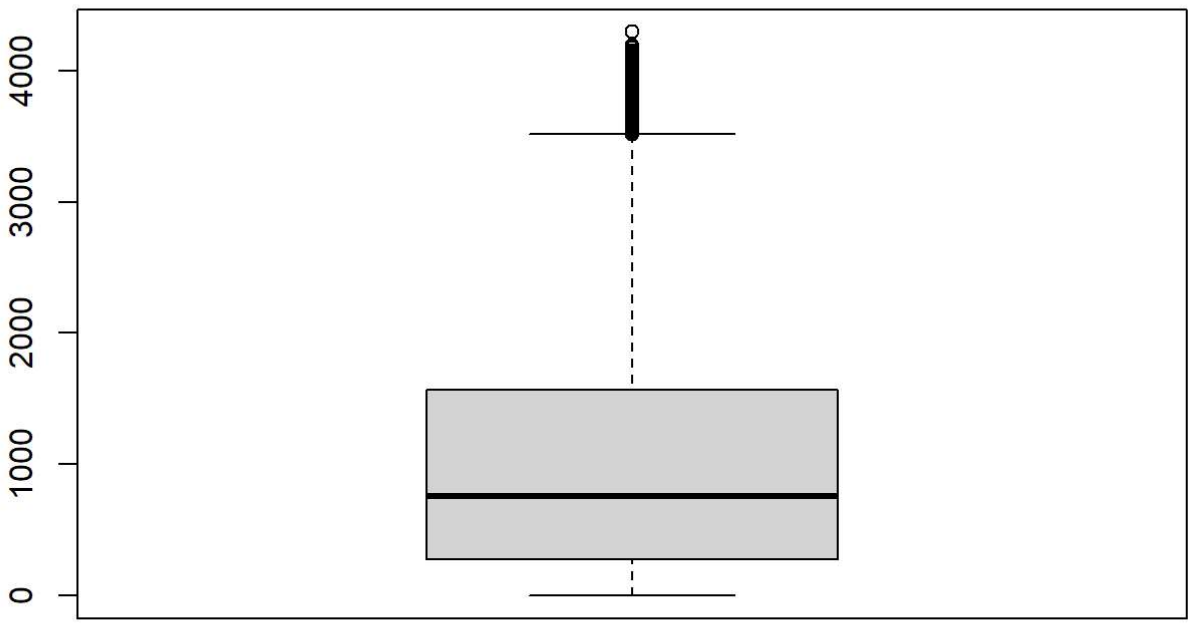

```
boxplot(clean_data$DAYS_REGISTRATION)
```



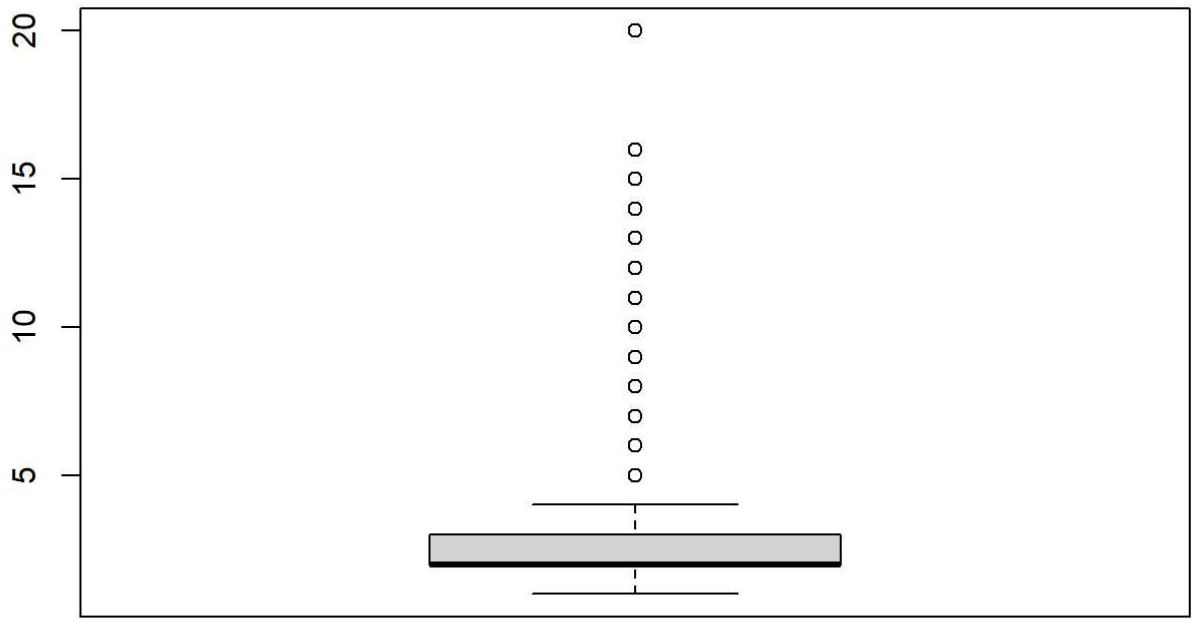

```
boxplot(clean_data$DAYS_ID_PUBLISH)
```




```
boxplot(clean_data$DAYS_LAST_PHONE_CHANGE)
```




```
boxplot(clean_data$CNT_FAM_MEMBERS)
```




```
#Removing Outliers
outliers <- boxplot(clean_data$CNT_CHILDREN, plot = FALSE)$out
clean_data<-clean_data[!(clean_data$CNT_CHILDREN %in% outliers), ]

outliers <- boxplot(clean_data$AMT_INCOME_TOTAL, plot = FALSE)$out
clean_data<-clean_data[!(clean_data$AMT_INCOME_TOTAL %in% outliers), ]

outliers <- boxplot(clean_data$AMT_CREDIT, plot = FALSE)$out
clean_data<-clean_data[!(clean_data$AMT_CREDIT %in% outliers), ]

outliers <- boxplot(clean_data$AMT_ANNUITY, plot = FALSE)$out
clean_data<-clean_data[!(clean_data$AMT_ANNUITY %in% outliers), ]

outliers <- boxplot(clean_data$AMT_GOODS_PRICE, plot = FALSE)$out
clean_data<-clean_data[!(clean_data$AMT_GOODS_PRICE %in% outliers), ]

outliers <- boxplot(clean_data$DAYS_BIRTH, plot = FALSE)$out
clean_data<-clean_data[!(clean_data$DAYS_BIRTH %in% outliers), ]

outliers <- boxplot(clean_data$DAYS_EMPLOYED, plot = FALSE)$out
clean_data<-clean_data[!(clean_data$DAYS_EMPLOYED %in% outliers), ]

outliers <- boxplot(clean_data$DAYS_REGISTRATION, plot = FALSE)$out
clean_data<-clean_data[!(clean_data$DAYS_REGISTRATION %in% outliers), ]

outliers <- boxplot(clean_data$DAYS_ID_PUBLISH, plot = FALSE)$out
clean_data<-clean_data[!(clean_data$DAYS_ID_PUBLISH %in% outliers), ]

outliers <- boxplot(clean_data$CNT_FAM_MEMBERS, plot = FALSE)$out
clean_data<-clean_data[!(clean_data$CNT_FAM_MEMBERS %in% outliers), ]
```

4. Dividing the data

Since we are making a model, we need test data on which we will be able to run our model to check its accuracy.Therefore, we will now split our data into two random parts, the testing and the training data. Since we are going to tran our model on the training data, it will be significantly larger than the testing data.

```
# We are going to use 80% of the data as training data

clean_data$TARGET<-as.factor(clean_data$TARGET) # Explicitly define the dependent variable as being categorical
split_data = sample.split(clean_data$TARGET, SplitRatio = 0.8)
train = subset(clean_data, split_data==TRUE)
test = subset(clean_data, split_data==FALSE)
```

When we get our data, there is one thing we have to keep in mind: since we split the data, there might be some levels for the atrributes with factor data types that are not sufficiently available in both training and testing dataset. This problem luckily only occurs in one variable: ‘NAME_INCOME_TYPE’.

```
table(clean_data$NAME_INCOME_TYPE)
```

```
##
##          Businessman Commercial associate      Maternity leave
##                    3                60049                    3
##            Pensioner        State servant              Student
##                    8                18715                   16
##              Working
##               144763
```

```
table(train$NAME_INCOME_TYPE)
```

```
##
##          Businessman Commercial associate      Maternity leave
##                    3                48061                    2
##            Pensioner        State servant              Student
##                    6                14962                   12
##              Working
##               115799
```

```
table(test$NAME_INCOME_TYPE)
```

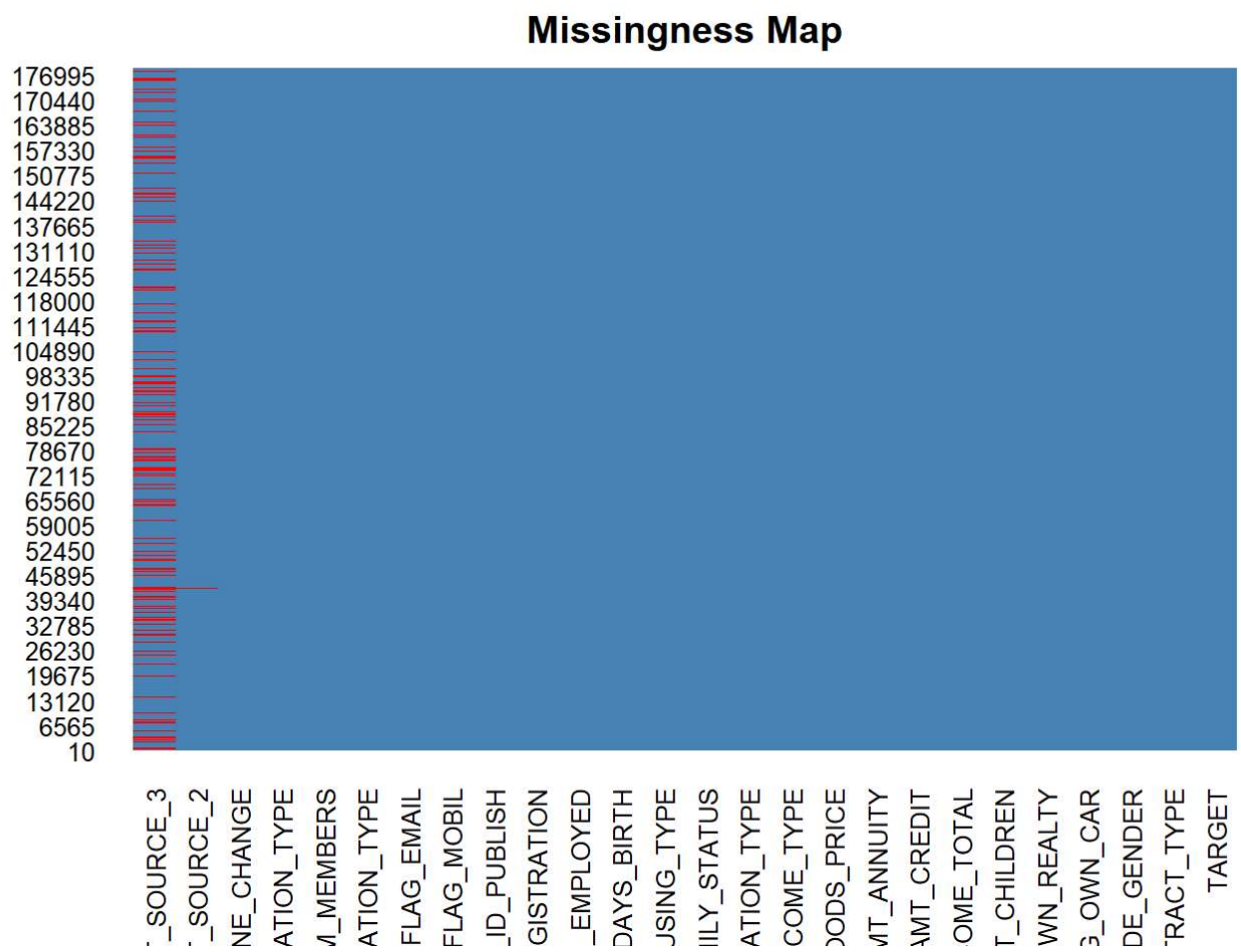
```
##
## Commercial associate      Maternity leave            Pensioner
##                11988                    1                    2
##        State servant              Student              Working
##                 3753                    4                28964
```

Some of the levels in this variable, like Businessman, are in small numbers and hence can be removed from the training and test data set. Otherwise, they might all be in only one of the two sets and that will cause problems. In the above code snipppet, we can see this happening for businessman, where they are all in the train data only. However, we cannot only remove businessman as the test and train data is randomly allocated and so we have to cater all such levels. Furthermore, since they are very small in number, removing them does not effect the accuracy of our model.

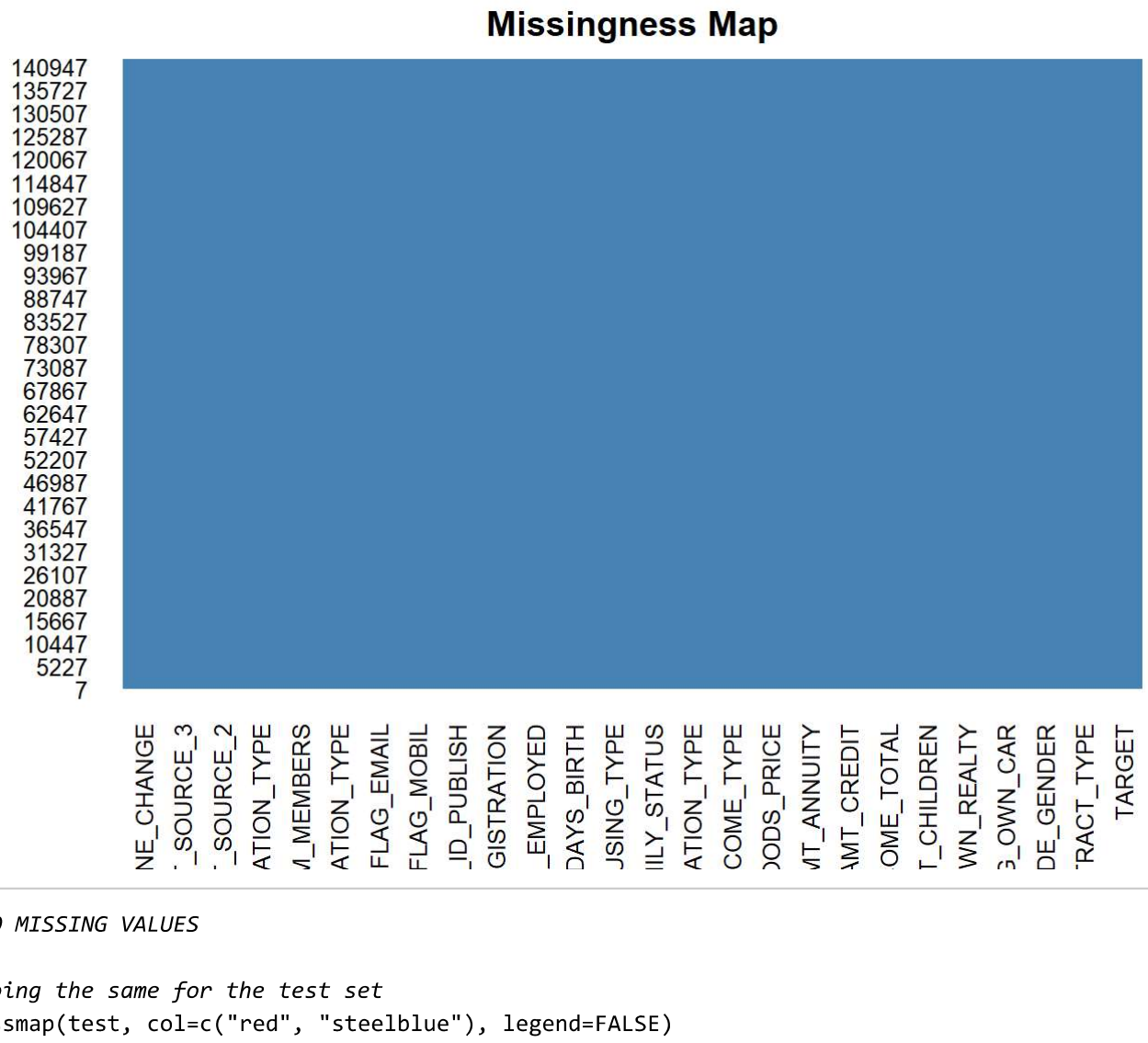
```
train <- train[!(train$NAME_INCOME_TYPE=="Maternity leave"|train$NAME_INCOME_TYPE=="Businessman"),]
test <- test[!(test$NAME_INCOME_TYPE=="Maternity leave"|test$NAME_INCOME_TYPE=="Businessman"),]
```

Before running the model, there is one final thing that we have to handle, and that is checking for missing data. Sometimes, we might have missing values in our observable data and it is important to remove them. Missing values also casue problems in plotting graphs later on using the model. First we will check if we have missing models using the missing plot and then purify the data.

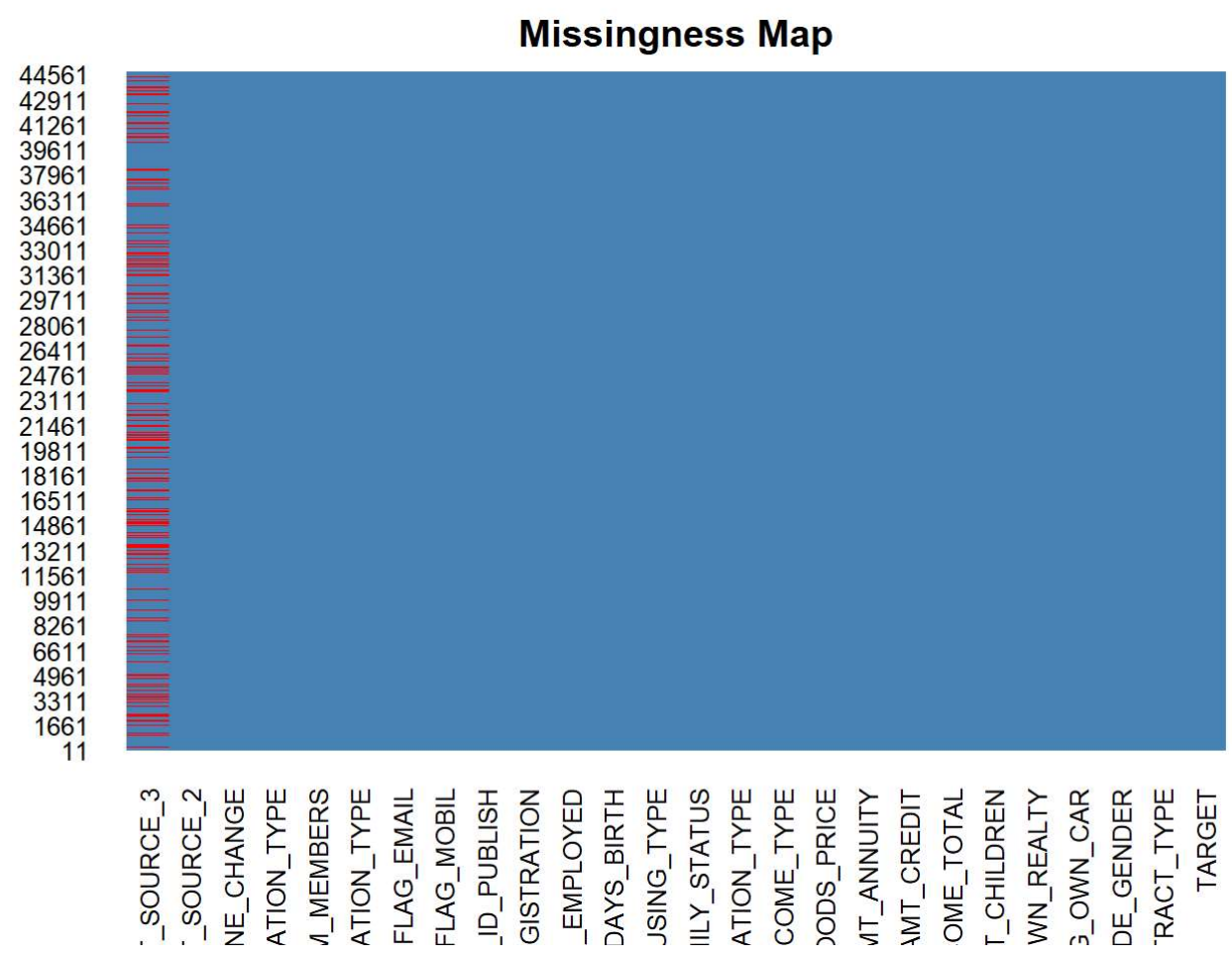
```
#Checking the missing plot for the training data:
missmap(train, col=c("red", "steelblue"), legend=FALSE)
```



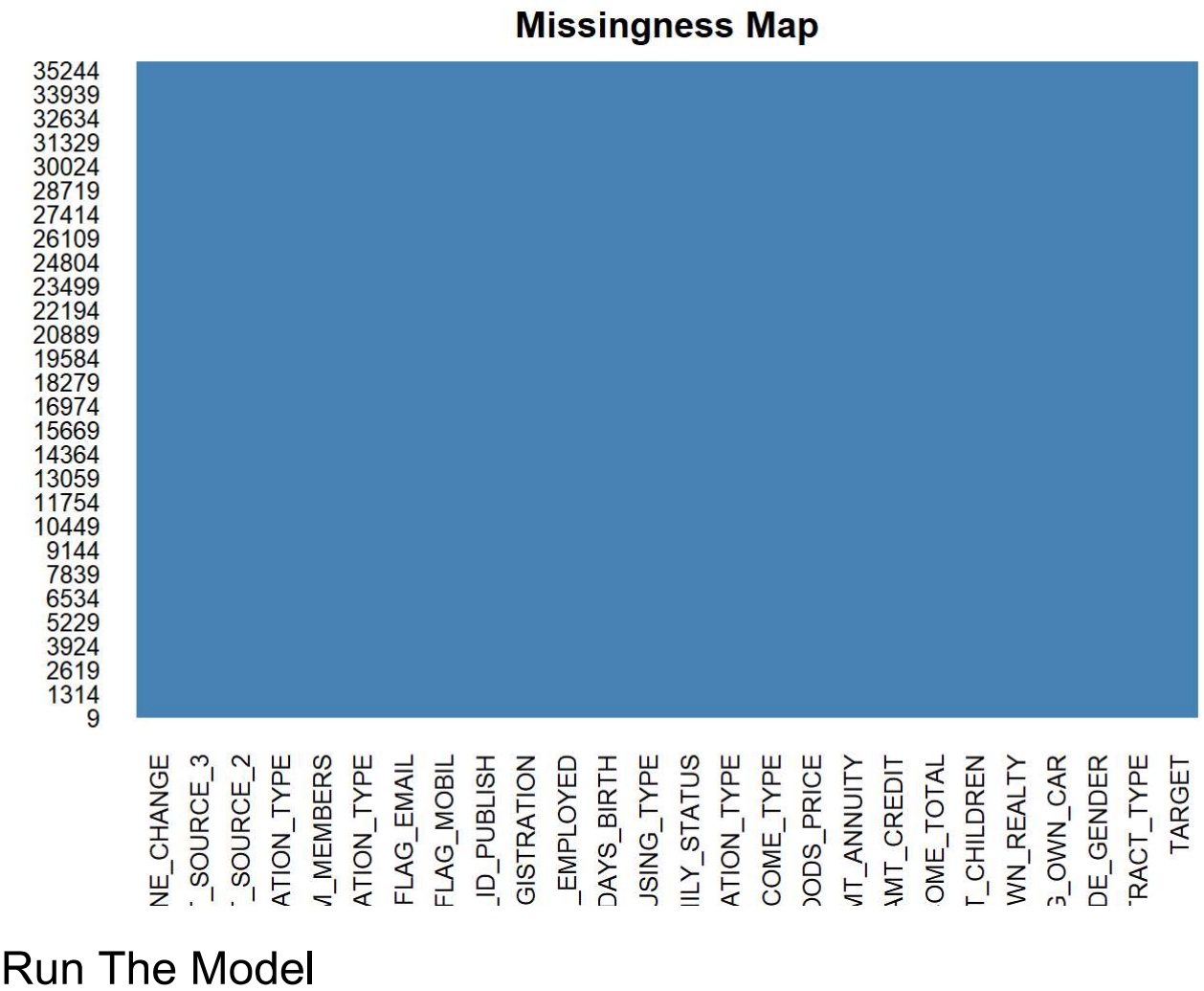

```
#Missing data exists.

#Solving the problem
train<- train[complete.cases(train),]

#Checking again
missmap(train, col=c("red", "steelblue"), legend=FALSE)
```




```
#NO MISSING VALUES

#Doing the same for the test set
missmap(test, col=c("red", "steelblue"), legend=FALSE)
```




```
test <- test[complete.cases(test),]
missmap(test, col=c("red", "steelblue"), legend=FALSE)
```



5. Run The Model

```
model<-glm( TARGET ~ . , family = "binomial" , data=train )
summary(model)
```

```
##
## Call:
## glm(formula = TARGET ~ ., family = "binomial", data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.4956  -0.4487  -0.3233  -0.2310   3.2747
##
## Coefficients: (2 not defined because of singularities)
##                                               Estimate Std. Error z value
## (Intercept)                                 -1.088e+01  6.005e+01  -0.181
## NAME_CONTRACT_TYPERevolving loans           -3.153e-01  4.120e-02  -7.654
## CODE_GENDERM                                 3.065e-01  2.601e-02  11.781
## FLAG_OWN_CARY                               -2.900e-01  2.268e-02 -12.785
## FLAG_OWN_REALTYY                             6.091e-02  2.188e-02   2.784
## CNT_CHILDREN                                -8.401e-03  1.546e-02  -0.544
## AMT_INCOME_TOTAL                            -8.755e-07  1.912e-07  -4.578
## AMT_CREDIT                                   2.058e-06  1.606e-07  12.815
## AMT_ANNUITY                                  1.515e-05  1.429e-06  10.597
## AMT_GOODS_PRICE                             -2.648e-06  1.838e-07 -14.408
## NAME_INCOME_TYPEPensioner                   -1.071e+01  2.280e+02  -0.047
## NAME_INCOME_TYPEState servant                1.547e-04  4.893e-02   0.003
## NAME_INCOME_TYPEStudent                     -1.116e+01  1.609e+02  -0.069
## NAME_INCOME_TYPEWorking                      9.618e-02  2.375e-02   4.049
## NAME_EDUCATION_TYPEHigher education          1.075e+01  6.005e+01   0.179
## NAME_EDUCATION_TYPEIncomplete higher         1.082e+01  6.005e+01   0.180
## NAME_EDUCATION_TYPELower secondary           1.123e+01  6.005e+01   0.187
## NAME_EDUCATION_TYPESecondary / secondary special  1.104e+01  6.005e+01   0.184
## NAME_FAMILY_STATUSMarried                   -1.477e-01  3.161e-02  -4.671
## NAME_FAMILY_STATUSSeparated                 -6.380e-02  4.807e-02  -1.327
## NAME_FAMILY_STATUSSingle / not married      -1.054e-01  3.760e-02  -2.802
## NAME_FAMILY_STATUSWidow                     -3.615e-01  7.709e-02  -4.689
## NAME_HOUSING_TYPEHouse / apartment           1.276e-01  1.632e-01   0.782
## NAME_HOUSING_TYPEMunicipal apartment         2.279e-01  1.710e-01   1.332
## NAME_HOUSING_TYPEOffice apartment           -2.409e-02  1.981e-01  -0.122
## NAME_HOUSING_TYPERented apartment            2.955e-01  1.744e-01   1.694
## NAME_HOUSING_TYPEWith parents                1.681e-01  1.668e-01   1.008
## DAYS_BIRTH                                  -8.955e-06  3.369e-06  -2.658
## DAYS_EMPLOYED                               -5.913e-05  5.695e-06 -10.383
## DAYS_REGISTRATION                           -1.325e-05  3.335e-06  -3.973
## DAYS_ID_PUBLISH                             -2.271e-05  6.835e-06  -3.323
## FLAG_MOBIL                                          NA         NA      NA
## FLAG_EMAIL                                  -1.244e-01  4.234e-02  -2.938
## OCCUPATION_TYPECleaning staff                2.276e-01  9.527e-02   2.389
## OCCUPATION_TYPECooking staff                 2.861e-01  8.737e-02   3.274
## OCCUPATION_TYPECore staff                    6.060e-02  7.349e-02   0.825
## OCCUPATION_TYPEDrivers                       3.367e-01  7.543e-02   4.463
## OCCUPATION_TYPEHigh skill tech staff         1.366e-01  8.156e-02   1.674
## OCCUPATION_TYPEHR staff                      1.433e-02  2.512e-01   0.057
## OCCUPATION_TYPEIT staff                     -1.963e-01  2.739e-01  -0.717
## OCCUPATION_TYPELaborers                      2.706e-01  6.823e-02   3.966
## OCCUPATION_TYPELow-skill Laborers            4.156e-01  1.099e-01   3.781
## OCCUPATION_TYPEManagers                      5.981e-02  7.689e-02   0.778
## OCCUPATION_TYPEMedicine staff                1.076e-01  9.561e-02   1.126
## OCCUPATION_TYPEPrivate service staff        -2.798e-02  1.354e-01  -0.207
## OCCUPATION_TYPERealty agents                -1.998e-03  2.144e-01  -0.009
## OCCUPATION_TYPESales staff                   2.043e-01  6.980e-02   2.926
## OCCUPATION_TYPESecretaries                   1.108e-01  1.620e-01   0.684
## OCCUPATION_TYPESecurity staff                2.755e-01  9.312e-02   2.959
## OCCUPATION_TYPEUNKNOWN                       1.646e-01  6.835e-02   2.408
## OCCUPATION_TYPEWaiters/barmen staff          2.415e-01  1.371e-01   1.761
## CNT_FAM_MEMBERS                                     NA         NA      NA
## ORGANIZATION_TYPEAgriculture                 3.007e-03  2.622e-01   0.011
## ORGANIZATION_TYPEBank                       -5.172e-01  2.750e-01  -1.881
## ORGANIZATION_TYPEBusiness Entity Type 1     -1.868e-01  2.528e-01  -0.739
## ORGANIZATION_TYPEBusiness Entity Type 2     -1.430e-01  2.491e-01  -0.574
## ORGANIZATION_TYPEBusiness Entity Type 3     -1.191e-01  2.453e-01  -0.486
## ORGANIZATION_TYPECleaning                   -1.911e-01  3.820e-01  -0.500
## ORGANIZATION_TYPEConstruction                8.207e-02  2.507e-01   0.327
## ORGANIZATION_TYPECulture                    -3.884e-01  4.115e-01  -0.944
## ORGANIZATION_TYPEElectricity                -3.854e-01  3.015e-01  -1.278
## ORGANIZATION_TYPEEmergency                  -2.482e-01  3.253e-01  -0.763
## ORGANIZATION_TYPEGovernment                 -2.400e-01  2.502e-01  -0.959
## ORGANIZATION_TYPEHotel                      -2.605e-01  2.957e-01  -0.881
## ORGANIZATION_TYPEHousing                    -1.612e-01  2.616e-01  -0.616
## ORGANIZATION_TYPEIndustry: type 1            1.257e-01  2.759e-01   0.455
## ORGANIZATION_TYPEIndustry: type 10          -5.498e-01  5.993e-01  -0.917
## ORGANIZATION_TYPEIndustry: type 11          -1.750e-01  2.617e-01  -0.669
## ORGANIZATION_TYPEIndustry: type 12          -6.739e-01  4.121e-01  -1.635
## ORGANIZATION_TYPEIndustry: type 13          -8.578e-01  7.825e-01  -1.096
## ORGANIZATION_TYPEIndustry: type 2           -6.795e-01  3.567e-01  -1.905
## ORGANIZATION_TYPEIndustry: type 3           -1.136e-01  2.572e-01  -0.442
## ORGANIZATION_TYPEIndustry: type 4           -1.544e-01  2.904e-01  -0.532
```

```
## ORGANIZATION_TYPEIndustry: type 5               -2.715e-01  3.150e-01  -0.862
## ORGANIZATION_TYPEIndustry: type 6               -1.215e+00  7.724e-01  -1.573
## ORGANIZATION_TYPEIndustry: type 7               -1.418e-01  2.773e-01  -0.511
## ORGANIZATION_TYPEIndustry: type 8               -5.559e-02  1.159e+00  -0.048
## ORGANIZATION_TYPEIndustry: type 9               -4.011e-01  2.612e-01  -1.535
## ORGANIZATION_TYPEInsurance                      -1.780e-01  3.410e-01  -0.522
## ORGANIZATION_TYPEKindergarten                   -1.844e-01  2.529e-01  -0.729
## ORGANIZATION_TYPELegal Services                  3.571e-01  3.993e-01   0.894
## ORGANIZATION_TYPEMedicine                       -1.371e-01  2.528e-01  -0.542
## ORGANIZATION_TYPEMilitary                       -5.503e-01  2.733e-01  -2.013
## ORGANIZATION_TYPEMobile                         -6.715e-02  3.607e-01  -0.186
## ORGANIZATION_TYPEOther                          -1.276e-01  2.479e-01  -0.515
## ORGANIZATION_TYPEPolice                         -4.209e-01  2.765e-01  -1.522
## ORGANIZATION_TYPEPostal                         -3.885e-02  2.668e-01  -0.146
## ORGANIZATION_TYPERealtor                         3.321e-01  3.511e-01   0.946
## ORGANIZATION_TYPEReligion                        9.049e-03  5.987e-01   0.015
## ORGANIZATION_TYPERestaurant                      1.104e-01  2.649e-01   0.417
## ORGANIZATION_TYPESchool                         -3.082e-01  2.523e-01  -1.221
## ORGANIZATION_TYPESecurity                       -1.418e-01  2.620e-01  -0.541
## ORGANIZATION_TYPESecurity Ministries            -4.523e-01  2.807e-01  -1.611
## ORGANIZATION_TYPESelf-employed                  -1.255e-02  2.459e-01  -0.051
## ORGANIZATION_TYPEServices                       -1.460e-01  2.866e-01  -0.509
## ORGANIZATION_TYPETelecom                        -1.871e-02  3.204e-01  -0.058
## ORGANIZATION_TYPETrade: type 1                  -5.289e-01  3.836e-01  -1.379
## ORGANIZATION_TYPETrade: type 2                  -5.317e-01  2.738e-01  -1.942
## ORGANIZATION_TYPETrade: type 3                  -1.969e-01  2.567e-01  -0.767
## ORGANIZATION_TYPETrade: type 4                  -9.894e-01  7.816e-01  -1.266
## ORGANIZATION_TYPETrade: type 5                  -7.585e-01  1.063e+00  -0.713
## ORGANIZATION_TYPETrade: type 6                  -6.487e-01  3.587e-01  -1.808
## ORGANIZATION_TYPETrade: type 7                  -1.152e-01  2.511e-01  -0.459
## ORGANIZATION_TYPETransport: type 1              -5.832e-01  4.932e-01  -1.183
## ORGANIZATION_TYPETransport: type 2              -2.017e-01  2.664e-01  -0.757
## ORGANIZATION_TYPETransport: type 3               3.471e-01  2.707e-01   1.282
## ORGANIZATION_TYPETransport: type 4              -8.222e-02  2.530e-01  -0.325
## ORGANIZATION_TYPEUniversity                     -3.860e-01  2.992e-01  -1.290
## EXT_SOURCE_2                                    -2.067e+00  4.954e-02 -41.730
## EXT_SOURCE_3                                    -2.915e+00  5.104e-02 -57.103
## DAYS_LAST_PHONE_CHANGE                          -4.538e-05  1.319e-05  -3.441
##                                                Pr(>|z|)
## (Intercept)                                    0.856251
## NAME_CONTRACT_TYPERevolving loans              1.95e-14 ***
## CODE_GENDERM                                    < 2e-16 ***
## FLAG_OWN_CARY                                   < 2e-16 ***
## FLAG_OWN_REALTYY                               0.005376 **
## CNT_CHILDREN                                   0.586734
## AMT_INCOME_TOTAL                               4.70e-06 ***
## AMT_CREDIT                                      < 2e-16 ***
## AMT_ANNUITY                                     < 2e-16 ***
## AMT_GOODS_PRICE                                 < 2e-16 ***
## NAME_INCOME_TYPEPensioner                      0.962535
## NAME_INCOME_TYPEState servant                  0.997477
## NAME_INCOME_TYPEStudent                        0.944670
## NAME_INCOME_TYPEWorking                        5.14e-05 ***
## NAME_EDUCATION_TYPEHigher education            0.857975
## NAME_EDUCATION_TYPEIncomplete higher           0.857065
## NAME_EDUCATION_TYPELower secondary             0.851631
## NAME_EDUCATION_TYPESecondary / secondary special 0.854179
## NAME_FAMILY_STATUSMarried                      3.00e-06 ***
## NAME_FAMILY_STATUSSeparated                    0.184459
## NAME_FAMILY_STATUSSingle / not married         0.005074 **
## NAME_FAMILY_STATUSWidow                        2.74e-06 ***
## NAME_HOUSING_TYPEHouse / apartment             0.434174
## NAME_HOUSING_TYPEMunicipal apartment           0.182800
## NAME_HOUSING_TYPEOffice apartment              0.903235
## NAME_HOUSING_TYPERented apartment              0.090297 .
## NAME_HOUSING_TYPEWith parents                  0.313478
## DAYS_BIRTH                                     0.007862 **
## DAYS_EMPLOYED                                   < 2e-16 ***
## DAYS_REGISTRATION                              7.10e-05 ***
## DAYS_ID_PUBLISH                                0.000892 ***
## FLAG_MOBIL                                           NA
## FLAG_EMAIL                                     0.003307 **
## OCCUPATION_TYPECleaning staff                  0.016902 *
## OCCUPATION_TYPECooking staff                   0.001059 **
## OCCUPATION_TYPECore staff                      0.409617
## OCCUPATION_TYPEDrivers                         8.08e-06 ***
## OCCUPATION_TYPEHigh skill tech staff           0.094048 .
## OCCUPATION_TYPEHR staff                        0.954491
## OCCUPATION_TYPEIT staff                        0.473565
## OCCUPATION_TYPELaborers                        7.31e-05 ***
## OCCUPATION_TYPELow-skill Laborers              0.000156 ***
## OCCUPATION_TYPEManagers                        0.436671
## OCCUPATION_TYPEMedicine staff                  0.260233
```

```
## OCCUPATION_TYPEPrivate service staff            0.836317
## OCCUPATION_TYPERealty agents                    0.992564
## OCCUPATION_TYPESales staff                      0.003429 **
## OCCUPATION_TYPESecretaries                      0.494069
## OCCUPATION_TYPESecurity staff                   0.003090 **
## OCCUPATION_TYPEUNKNOWN                          0.016024 *
## OCCUPATION_TYPEWaiters/barmen staff             0.078248 .
## CNT_FAM_MEMBERS                                       NA
## ORGANIZATION_TYPEAgriculture                    0.990853
## ORGANIZATION_TYPEBank                           0.060040 .
## ORGANIZATION_TYPEBusiness Entity Type 1         0.459779
## ORGANIZATION_TYPEBusiness Entity Type 2         0.566120
## ORGANIZATION_TYPEBusiness Entity Type 3         0.627185
## ORGANIZATION_TYPECleaning                       0.616936
## ORGANIZATION_TYPEConstruction                   0.743397
## ORGANIZATION_TYPECulture                        0.345327
## ORGANIZATION_TYPEElectricity                    0.201220
## ORGANIZATION_TYPEEmergency                      0.445474
## ORGANIZATION_TYPEGovernment                     0.337356
## ORGANIZATION_TYPEHotel                          0.378369
## ORGANIZATION_TYPEHousing                        0.537670
## ORGANIZATION_TYPEIndustry: type 1               0.648792
## ORGANIZATION_TYPEIndustry: type 10              0.358959
## ORGANIZATION_TYPEIndustry: type 11              0.503618
## ORGANIZATION_TYPEIndustry: type 12              0.102032
## ORGANIZATION_TYPEIndustry: type 13              0.272930
## ORGANIZATION_TYPEIndustry: type 2               0.056802 .
## ORGANIZATION_TYPEIndustry: type 3               0.658835
## ORGANIZATION_TYPEIndustry: type 4               0.595042
## ORGANIZATION_TYPEIndustry: type 5               0.388736
## ORGANIZATION_TYPEIndustry: type 6               0.115678
## ORGANIZATION_TYPEIndustry: type 7               0.609021
## ORGANIZATION_TYPEIndustry: type 8               0.961756
## ORGANIZATION_TYPEIndustry: type 9               0.124722
## ORGANIZATION_TYPEInsurance                      0.601629
## ORGANIZATION_TYPEKindergarten                   0.465912
## ORGANIZATION_TYPELegal Services                 0.371082
## ORGANIZATION_TYPEMedicine                       0.587703
## ORGANIZATION_TYPEMilitary                       0.044073 *
## ORGANIZATION_TYPEMobile                         0.852291
## ORGANIZATION_TYPEOther                          0.606650
## ORGANIZATION_TYPEPolice                         0.127948
## ORGANIZATION_TYPEPostal                         0.884235
## ORGANIZATION_TYPERealtor                        0.344231
## ORGANIZATION_TYPEReligion                       0.987941
## ORGANIZATION_TYPERestaurant                     0.676759
## ORGANIZATION_TYPESchool                         0.221903
## ORGANIZATION_TYPESecurity                       0.588334
## ORGANIZATION_TYPESecurity Ministries            0.107122
## ORGANIZATION_TYPESelf-employed                  0.959281
## ORGANIZATION_TYPEServices                       0.610579
## ORGANIZATION_TYPETelecom                        0.953418
## ORGANIZATION_TYPETrade: type 1                  0.167982
## ORGANIZATION_TYPETrade: type 2                  0.052142 .
## ORGANIZATION_TYPETrade: type 3                  0.442948
## ORGANIZATION_TYPETrade: type 4                  0.205560
## ORGANIZATION_TYPETrade: type 5                  0.475665
## ORGANIZATION_TYPETrade: type 6                  0.070543 .
## ORGANIZATION_TYPETrade: type 7                  0.646301
## ORGANIZATION_TYPETransport: type 1              0.236949
## ORGANIZATION_TYPETransport: type 2              0.449084
## ORGANIZATION_TYPETransport: type 3              0.199723
## ORGANIZATION_TYPETransport: type 4              0.745247
## ORGANIZATION_TYPEUniversity                     0.196968
## EXT_SOURCE_2                                     < 2e-16 ***
## EXT_SOURCE_3                                     < 2e-16 ***
## DAYS_LAST_PHONE_CHANGE                          0.000579 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 84091  on 142686  degrees of freedom
## Residual deviance: 74768  on 142578  degrees of freedom
## AIC: 74986
##
## Number of Fisher Scoring iterations: 12
```

***LOGISTIC REGRESSION*:** Now that we have made our model, we can start interpreting it. Simple logistic regression helps you classify your answer, rather than predict a value. It is called the logit model too, as we are using log to get a sigmoid model which tells us how likely we are to get the null hypothesis or vice versa. Here we can find some information regarding our model. The deviance residuals stats look good as they are close to being centered on 0 and are roughly symmetrical. Similarly, our **null deviance (the value without using the parameters and only the intercept)** is larger than our residual deviance, which means that our model helps us predict the output better. Lastly, the asterisks in from of some

variables represent that these predictors are statistically very significant and the other variables might be showing a patter created to randomness. We can also see that for these variables, the **z value probability is also much less than 0.05, showing statistical significance and a strong relation**. However, we do not remove those predictors completely as it could cause omitted variable bias. There are two more things in the summary, hte **AIC and the Fisher Scoring iterations**. These are talked about later on.

6.1. Use the Model to check Accuracy

```
pred <- predict(model, newdata = test, type = "response")
glm.pred <- ifelse(pred > 0.5, "Not Paid", "Paid")

# Visualizing how many values were detected correctly.
t<-table(glm.pred, test$TARGET)
t
```

```
##
## glm.pred       0     1
##   Not Paid    30    43
##   Paid     32568  3023
```

```
# Finding accuracy, including both type I and type II errors:
accuracy_1 = (t[2,1] + t[1,2]) / (t[1,1] + t[2,2]+t[2,1] + t[1,2])
accuracy_1 = accuracy_1*100
accuracy_1
```

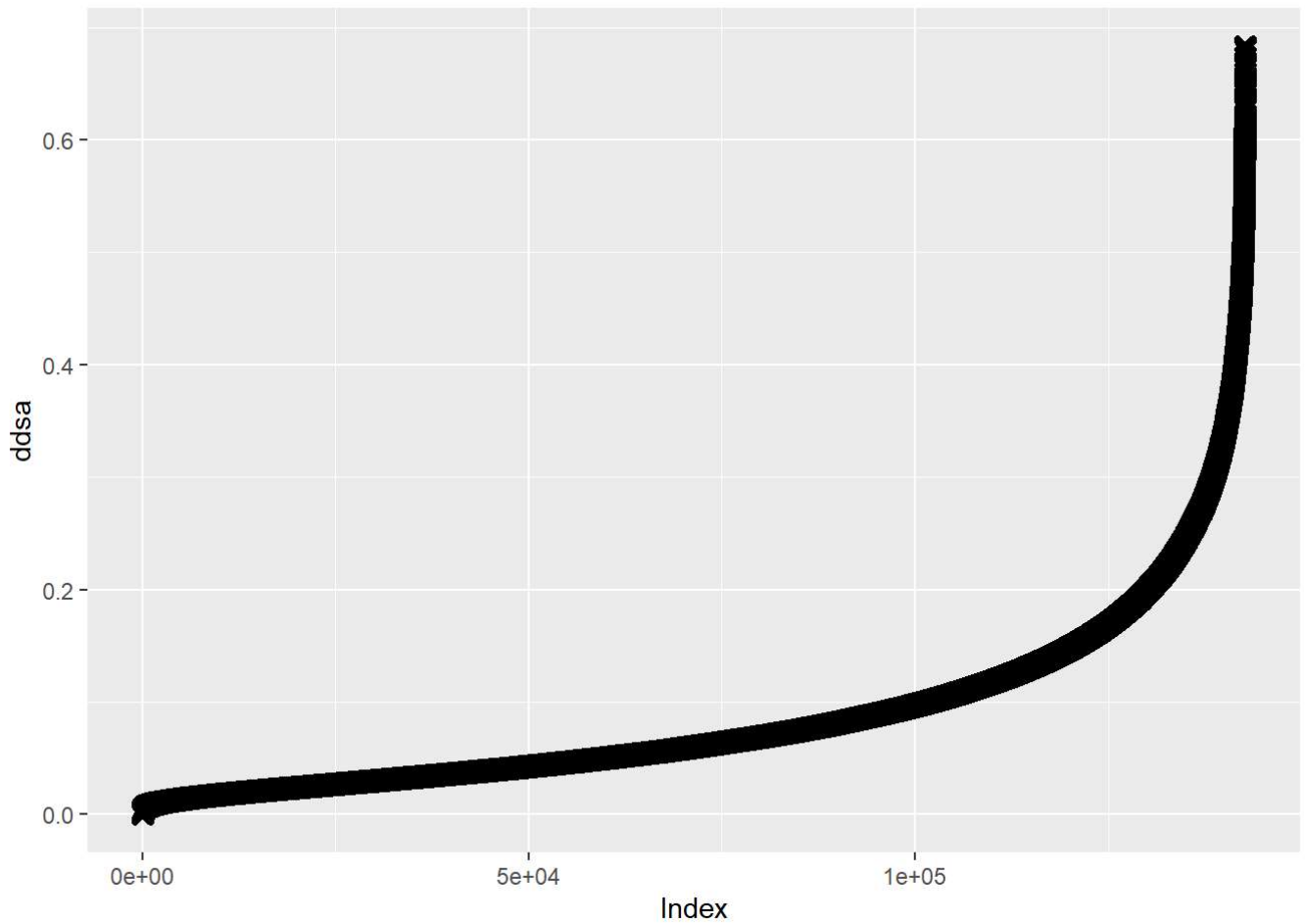
```
## [1] 91.43955
```

6.2. Using the Model to Make a Graph

As mentioned earlier, the logit model helps us to get a value between 0 and 1 which helps us decided where to classify the output. We are now going to make such a graph:

```
predicted.data <- data.frame(prob = model$fitted.values, def = train$TARGET)
predicted.data <- predicted.data[order(predicted.data$prob, decreasing = FALSE),]
predicted.data$rank <- 1:nrow(predicted.data)

ggplot(data = predicted.data, aes(x=rank, y= prob)) +
  geom_point(aplha=1, shape=4, stroke=2)+
  xlab("Index")+
  ylab("ddsa")
```

```
## Warning: Ignoring unknown parameters: aplha
```



7. Making the model better.

Our model gives us a very good accuracy. However, there are faults with this model. For example, while finding the accuracy, we considered both **type I and type II errors**. Type I errors are those where the null hypothesis (in this case the client not being a defaulter) is true but the model rejects it (returns false). If we were to consider only type II errors, our accuracy turns out to be very bad **(1.4024788%)**. In other words, our model is very good for predicting those who will pay the loan but not so good to predict those who will not repay the loan. We will now try to better our model.

One of the reasons for this problem could be that there is a severe imbalance in our dataset for the dependent variable. The 'TARGET' column has way more instances of 0s than 1s. To solve this problem, we are going to use **downsampling**. This will make sure that there are equal number of cases for both the client paying the loan and not paying it. After downsampling our training data, we are going to use the same method as above to make a new model and then train it to see its results.

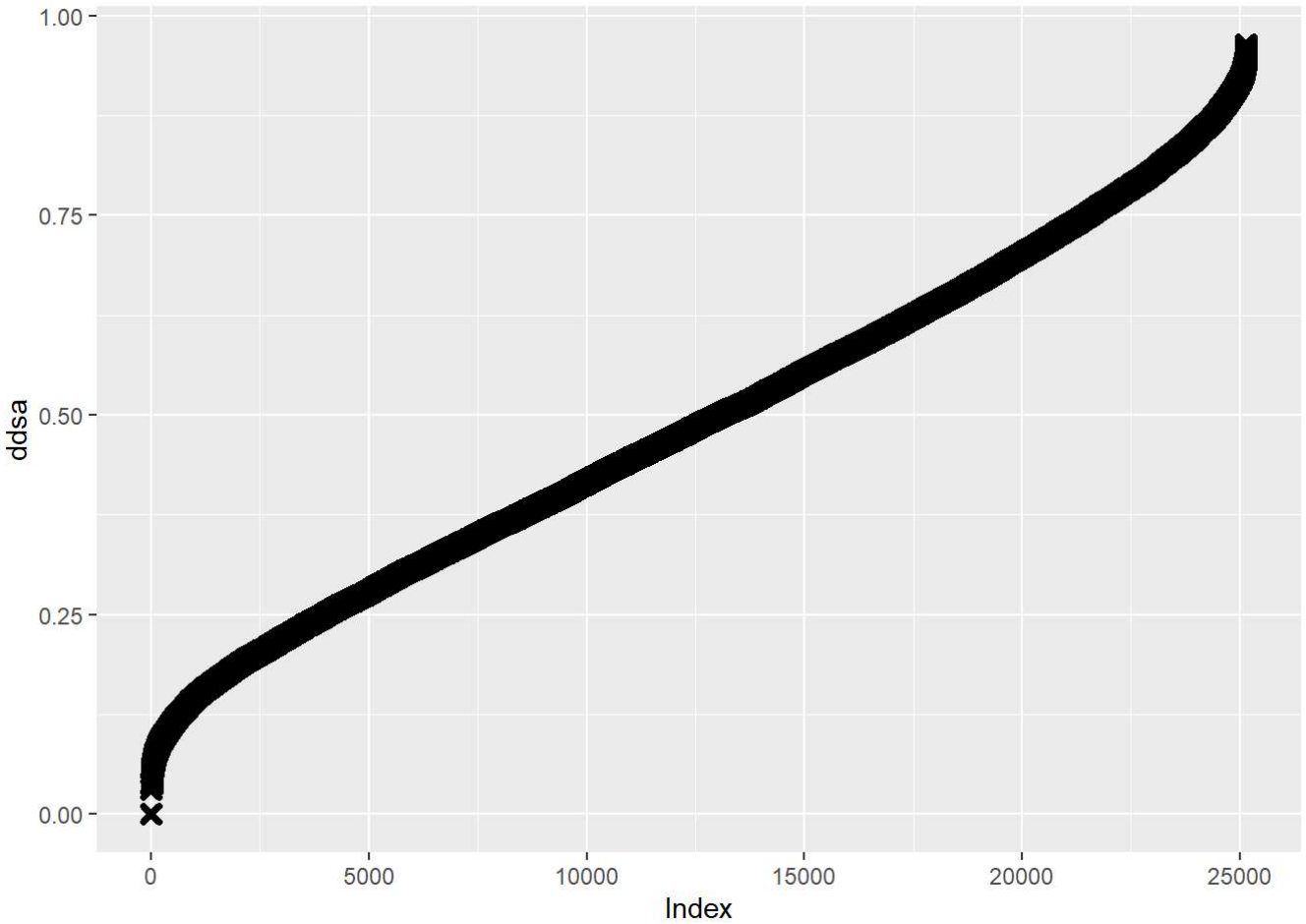
```
'%ni%' <- Negate('%in%')  # define 'not in' func
options(scipen=999)  # prevents printing scientific notations.

# Now we make the training and test data again. This time around, I have used a different function just for the sake of shwo
ing different ways. However, there is no difference between the two and the above mehtod can be used again to obtain the sam
e result.
clean_data$TARGET<-as.factor(clean_data$TARGET)
set.seed(100)

# Dividing the data into test and train sets
trainDataIndex <- createDataPartition(clean_data$TARGET, p=0.8, list = F)  # 80% training data
trainData <- clean_data[trainDataIndex, ]
testData <- clean_data[-trainDataIndex, ]

down_train <- downSample(x = trainData[, colnames(trainData) %ni% "TARGET"],
                         y = trainData$TARGET)
# An important thing to note at this point is that when we use this down sample function, our dependent variable, which in t
his case is 'Target', will change and be now identified by the keyword 'Class'.

down_train <- down_train[!(down_train$NAME_INCOME_TYPE=="Student"|down_train$NAME_INCOME_TYPE=="Maternity leave"|down_train
$NAME_INCOME_TYPE=="Pensioner"|down_train$NAME_INCOME_TYPE=="Businessman"),]
testData <- testData[!(testData$NAME_INCOME_TYPE=="Student"|testData$NAME_INCOME_TYPE=="Maternity leave"|testData$NAME_INCOM
E_TYPE=="Pensioner"|testData$NAME_INCOME_TYPE=="Businessman"),]
#We are removing more levels from this variable as in downsampled version, some more levels were reduced to have lesser numb
er of values than required to ensure participation of level in both test and train data.

# Now we remove missing values
down_train<-down_train[complete.cases(down_train),]
testData <- testData[complete.cases(testData),]

# Building and fitting a glm model
down_model<-glm( Class ~ . , family = "binomial" , data=down_train )
down_pred <- predict(down_model, newdata = testData, type = "response")
down_glm.pred <- ifelse(down_pred > 0.5, "Not Paid", "Paid")
summary(down_glm.pred)
```

```
##    Length     Class      Mode
##     35682 character character
```

```
# Finding accuracy for both type errors
dt <- table(down_glm.pred, testData$TARGET)
accuracy_2 = (dt[2,1] + dt[1,2]) / (dt[1,1] + dt[2,2]+dt[2,1] + dt[1,2])
accuracy_2*100
```

```
## [1] 69.03761
```

```
# Constructing graph
predicted.data <- data.frame(prob = down_model$fitted.values, def = down_train$Class)
predicted.data <- predicted.data[order(predicted.data$prob, decreasing = FALSE),]
predicted.data$rank <- 1:nrow(predicted.data)

ggplot(data = predicted.data, aes(x=rank, y= prob)) +
  geom_point(aplha=1, shape=4, stroke=2)+
  xlab("Index")+
  ylab("ddsa")
```



As we can see the accuracy of this model is lesser than the previous model, but this one is much better in dealing with **Type II error**, i.e. it is more accurate in detecting when someone will not return the money **(Accuracy is 67.3102094% in this case, which is a huge improvement)**. Both models might be used for different use cases and they have their advantages, but for the remaining refinement, we are going to use our second model in which we downsampled our dataset as that model is experimentally more sound. We are also using this model because of **AIC and Fisher Scoring**. Going to the end of the summary, we see that we have the AIC. The **AIC** is the measure of how good your model is and can be thought of as the alternative to the $\mathbf{R^2}$ in linear regression. Similarly, the lesser the number of Fisher scoring iteration our model requires, the better our model is, as **the Fisher Scoring iterations tell us how quickly our glm() function converged on the maximum likelihood estimates for the coefficients**. The AIC decreases by a lot in the downsampled model, and this change shows that this model is much better than our original one.

8. Further Refining the Downsampled Model:

To refine a model, the main goal is to decrease the deviance and the AIc of a model. Since we have a lot of independent predictors, we are going to now run a function which will identify for us which predictors are actually not helping and are causing a higher AIC and will remove them from our dataset:

```
final_model<-stepAIC(down_model,direction="backward",trace=FALSE)
summary(final_model)
```

```
##
## Call:
## glm(formula = Class ~ NAME_CONTRACT_TYPE + CODE_GENDER + FLAG_OWN_CAR +
##     FLAG_OWN_REALTY + AMT_INCOME_TOTAL + AMT_CREDIT + AMT_ANNUITY +
##     AMT_GOODS_PRICE + NAME_INCOME_TYPE + NAME_EDUCATION_TYPE +
##     NAME_FAMILY_STATUS + NAME_HOUSING_TYPE + DAYS_BIRTH + DAYS_EMPLOYED +
##     DAYS_REGISTRATION + DAYS_ID_PUBLISH + FLAG_EMAIL + OCCUPATION_TYPE +
##     EXT_SOURCE_2 + EXT_SOURCE_3 + DAYS_LAST_PHONE_CHANGE, family = "binomial",
##     data = down_train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4863  -0.9802  -0.4466   1.0015   2.4287
##
## Coefficients:
##                                                 Estimate     Std. Error
## (Intercept)                                -10.3802564327 105.8977568943
## NAME_CONTRACT_TYPERevolving loans           -0.2636087918   0.0555124445
## CODE_GENDERM                                 0.3177583870   0.0362916486
## FLAG_OWN_CARY                               -0.3213752821   0.0319060954
## FLAG_OWN_REALTYY                             0.0664799601   0.0310398932
## AMT_INCOME_TOTAL                            -0.0000011810   0.0000002660
## AMT_CREDIT                                   0.0000021518   0.0000002300
## AMT_ANNUITY                                  0.0000197939   0.0000020894
## AMT_GOODS_PRICE                             -0.0000027233   0.0000002576
## NAME_INCOME_TYPEState servant               -0.0261787543   0.0612097448
## NAME_INCOME_TYPEWorking                      0.0947508101   0.0332434080
## NAME_EDUCATION_TYPEHigher education         12.4116485908 105.8974796386
## NAME_EDUCATION_TYPEIncomplete higher        12.5937354216 105.8975011016
## NAME_EDUCATION_TYPELower secondary          12.9597379018 105.8975548279
## NAME_EDUCATION_TYPESecondary / secondary special 12.7397320548 105.8974766687
## NAME_FAMILY_STATUSMarried                   -0.1902631533   0.0460202323
## NAME_FAMILY_STATUSSeparated                 -0.0458434206   0.0686603246
## NAME_FAMILY_STATUSSingle / not married      -0.1117505138   0.0550206071
## NAME_FAMILY_STATUSWidow                     -0.3353930905   0.1025112805
## NAME_HOUSING_TYPEHouse / apartment           0.1129070950   0.2105521911
## NAME_HOUSING_TYPEMunicipal apartment         0.2350602348   0.2229714500
## NAME_HOUSING_TYPEOffice apartment           -0.2005620417   0.2599214662
## NAME_HOUSING_TYPERented apartment            0.3462859449   0.2311745719
## NAME_HOUSING_TYPEWith parents                0.2325062112   0.2172801349
## DAYS_BIRTH                                  -0.0000110526   0.0000046796
## DAYS_EMPLOYED                               -0.0000588987   0.0000072983
## DAYS_REGISTRATION                           -0.0000080162   0.0000046510
## DAYS_ID_PUBLISH                             -0.0000223997   0.0000095507
## FLAG_EMAIL                                  -0.1275532312   0.0601070370
## OCCUPATION_TYPECleaning staff                0.2783022554   0.1301638423
## OCCUPATION_TYPECooking staff                 0.4300016077   0.1213560733
## OCCUPATION_TYPECore staff                   -0.0563343733   0.0934487941
## OCCUPATION_TYPEDrivers                       0.3748469668   0.1010820343
## OCCUPATION_TYPEHigh skill tech staff         0.0063056470   0.1071658837
## OCCUPATION_TYPEHR staff                     -0.0534438600   0.3335103325
## OCCUPATION_TYPEIT staff                     -0.5614274674   0.3530135215
## OCCUPATION_TYPELaborers                      0.2584858858   0.0891935388
## OCCUPATION_TYPELow-skill Laborers            0.5169034322   0.1654077384
## OCCUPATION_TYPEManagers                      0.1316728792   0.1000413422
## OCCUPATION_TYPEMedicine staff                0.0540508752   0.1136641888
## OCCUPATION_TYPEPrivate service staff        -0.0277105739   0.1735855028
## OCCUPATION_TYPERealty agents                 0.3421924608   0.2644386345
## OCCUPATION_TYPESales staff                   0.2541990990   0.0906773134
## OCCUPATION_TYPESecretaries                   0.3845631255   0.2262093519
## OCCUPATION_TYPESecurity staff                0.2977347477   0.1177755110
## OCCUPATION_TYPEUNKNOWN                       0.1217295952   0.0894337485
## OCCUPATION_TYPEWaiters/barmen staff          0.3057007187   0.2020824041
## EXT_SOURCE_2                                -2.1839264286   0.0736812290
## EXT_SOURCE_3                                -2.9199170941   0.0721256938
## DAYS_LAST_PHONE_CHANGE                      -0.0000282259   0.0000181788
##                                            z value             Pr(>|z|)
## (Intercept)                                 -0.098             0.921915
## NAME_CONTRACT_TYPERevolving loans           -4.749 0.000002047858659168
## CODE_GENDERM                                 8.756 < 0.0000000000000002
## FLAG_OWN_CARY                              -10.073 < 0.0000000000000002
## FLAG_OWN_REALTYY                             2.142             0.032213
## AMT_INCOME_TOTAL                            -4.440 0.000009003891276638
## AMT_CREDIT                                   9.354 < 0.0000000000000002
## AMT_ANNUITY                                  9.473 < 0.0000000000000002
## AMT_GOODS_PRICE                            -10.570 < 0.0000000000000002
## NAME_INCOME_TYPEState servant               -0.428             0.668877
## NAME_INCOME_TYPEWorking                      2.850             0.004369
## NAME_EDUCATION_TYPEHigher education          0.117             0.906698
## NAME_EDUCATION_TYPEIncomplete higher         0.119             0.905336
## NAME_EDUCATION_TYPELower secondary           0.122             0.902598
## NAME_EDUCATION_TYPESecondary / secondary special  0.120             0.904244
## NAME_FAMILY_STATUSMarried                   -4.134 0.000035598066097276
```

```
## NAME_FAMILY_STATUSSeparated                    -0.668             0.504335
## NAME_FAMILY_STATUSSingle / not married         -2.031             0.042248
## NAME_FAMILY_STATUSWidow                        -3.272             0.001069
## NAME_HOUSING_TYPEHouse / apartment              0.536             0.591791
## NAME_HOUSING_TYPEMunicipal apartment            1.054             0.291784
## NAME_HOUSING_TYPEOffice apartment              -0.772             0.440336
## NAME_HOUSING_TYPERented apartment               1.498             0.134148
## NAME_HOUSING_TYPEWith parents                   1.070             0.284585
## DAYS_BIRTH                                     -2.362             0.018182
## DAYS_EMPLOYED                                  -8.070 0.00000000000000702
## DAYS_REGISTRATION                              -1.724             0.084787
## DAYS_ID_PUBLISH                                -2.345             0.019009
## FLAG_EMAIL                                     -2.122             0.033829
## OCCUPATION_TYPECleaning staff                   2.138             0.032509
## OCCUPATION_TYPECooking staff                    3.543             0.000395
## OCCUPATION_TYPECore staff                      -0.603             0.546617
## OCCUPATION_TYPEDrivers                          3.708             0.000209
## OCCUPATION_TYPEHigh skill tech staff            0.059             0.953080
## OCCUPATION_TYPEHR staff                        -0.160             0.872687
## OCCUPATION_TYPEIT staff                        -1.590             0.111748
## OCCUPATION_TYPELaborers                         2.898             0.003755
## OCCUPATION_TYPELow-skill Laborers               3.125             0.001778
## OCCUPATION_TYPEManagers                         1.316             0.188112
## OCCUPATION_TYPEMedicine staff                   0.476             0.634408
## OCCUPATION_TYPEPrivate service staff           -0.160             0.873167
## OCCUPATION_TYPERealty agents                    1.294             0.195654
## OCCUPATION_TYPESales staff                      2.803             0.005058
## OCCUPATION_TYPESecretaries                      1.700             0.089125
## OCCUPATION_TYPESecurity staff                   2.528             0.011472
## OCCUPATION_TYPEUNKNOWN                          1.361             0.173477
## OCCUPATION_TYPEWaiters/barmen staff             1.513             0.130342
## EXT_SOURCE_2                                  -29.640 < 0.0000000000000002
## EXT_SOURCE_3                                  -40.484 < 0.0000000000000002
## DAYS_LAST_PHONE_CHANGE                         -1.553             0.120500
##
## (Intercept)
## NAME_CONTRACT_TYPERevolving loans               ***
## CODE_GENDERM                                   ***
## FLAG_OWN_CARY                                  ***
## FLAG_OWN_REALTYY                               *
## AMT_INCOME_TOTAL                               ***
## AMT_CREDIT                                     ***
## AMT_ANNUITY                                    ***
## AMT_GOODS_PRICE                                ***
## NAME_INCOME_TYPEState servant
## NAME_INCOME_TYPEWorking                        **
## NAME_EDUCATION_TYPEHigher education
## NAME_EDUCATION_TYPEIncomplete higher
## NAME_EDUCATION_TYPELower secondary
## NAME_EDUCATION_TYPESecondary / secondary special
## NAME_FAMILY_STATUSMarried                      ***
## NAME_FAMILY_STATUSSeparated
## NAME_FAMILY_STATUSSingle / not married         *
## NAME_FAMILY_STATUSWidow                        **
## NAME_HOUSING_TYPEHouse / apartment
## NAME_HOUSING_TYPEMunicipal apartment
## NAME_HOUSING_TYPEOffice apartment
## NAME_HOUSING_TYPERented apartment
## NAME_HOUSING_TYPEWith parents
## DAYS_BIRTH                                     *
## DAYS_EMPLOYED                                  ***
## DAYS_REGISTRATION                              .
## DAYS_ID_PUBLISH                                *
## FLAG_EMAIL                                     *
## OCCUPATION_TYPECleaning staff                  *
## OCCUPATION_TYPECooking staff                   ***
## OCCUPATION_TYPECore staff
## OCCUPATION_TYPEDrivers                         ***
## OCCUPATION_TYPEHigh skill tech staff
## OCCUPATION_TYPEHR staff
## OCCUPATION_TYPEIT staff
## OCCUPATION_TYPELaborers                        **
## OCCUPATION_TYPELow-skill Laborers              **
## OCCUPATION_TYPEManagers
## OCCUPATION_TYPEMedicine staff
## OCCUPATION_TYPEPrivate service staff
## OCCUPATION_TYPERealty agents
## OCCUPATION_TYPESales staff                     **
## OCCUPATION_TYPESecretaries                     .
## OCCUPATION_TYPESecurity staff                  *
## OCCUPATION_TYPEUNKNOWN
## OCCUPATION_TYPEWaiters/barmen staff
## EXT_SOURCE_2                                   ***
```

```
## EXT_SOURCE_3                                 ***
## DAYS_LAST_PHONE_CHANGE
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 34857  on 25148  degrees of freedom
## Residual deviance: 29906  on 25099  degrees of freedom
## AIC: 30006
##
## Number of Fisher Scoring iterations: 11
```

```
e <- as.data.frame(exp(coef(final_model)))
# As mentioned earlier, factors are made when you run the glm model. Here, we get the likeliness for each predictor and this shows by how much the likeliness will increase for the output TARGET for a change in factor in the variables. This change is determined relative to a base factor level set by default in R for each variable. The more statistically significant variables cause a higher increase in likeliness.

# Testing model
final_pred <- predict(final_model, newdata = testData, type = "response")


final_glm.pred <- ifelse(final_pred > 0.5, "Not Paid", "Paid")
```

9. CONCLUSION:

We were able to test our logistic model and come up with an affirmative answer to our research question and showed that there indeed is a relation between a client being a defaulter and our chose parameters. We were further able to test our model on data to see how well it works and refine it. We were able to bring down the AIC value from 74986.0367358 to 30005.8025691. We also saw that there was a difference between the null and residual deviance, which shows the effectiveness of the model and we were also able to bring down the residual deviance from 74768.0367358 to 29905.8025691.

10. REFERENCES:

Dataset (https://www.kaggle.com/mishra5001/credit-card)