BUSINESS ANALYTICS

FINAL PROJECT REPORT

PREDICTING THE CHANCES OF A BORROWER TO DEFAULT - FROM BANK'S PERSPECTIVE

Group 4

M. Huzaifa

Ali Kumayl

A. R. Shamsi

M. Asim

Faran Maood

Loan Defaults and their Effects on Financial Performance of Commercial Banks

Literature Review

The major impact of poor loans on banks is generally that they restrict banks' ability to develop financially (Karim, Chan & Hassan, 2010; Kuo et al., 2010). This effect results from poor loans depriving banks of the necessary liquidity, which restricts their capacity to support other potentially viable firms and provide credit to consumers. According to Karim et al. (2010), there are several alternative profitable ventures the bank cannot pursue since its resources are ensnared in subpar loans. Faced with these repercussions, the bank encounters a deficiency in produced revenues (Banking Survey, 2013), which results in diminished financial performance. (Karim et al., 2010; Nawaz et al. 2012; Banking Survey, 2013).

A bank will encounter several issues and unfavourable outcomes if a poor credit risk management strategy is not put in place. The most fundamental issues, which are the most evident and interconnected, focus around wanting to be profitable, solvent, and liquid. There may be numerous associated issues. As we know how banks like other businesses are there to make profits and money, therefore losses by way of defaults must be kept to a minimum for this to be accomplished. Before revising office operations, it makes sense to advise on losses experienced by banks as the same fundamental causes frequently arise.

Bloem and Gorter (2001) assert that although problems with non-performing loans may touch all industries, they have the most negative effects on financial organisations with sizable loan portfolios, such commercial banks and home finance companies. In addition, the significant default loans will have an impact on banks' capacity to extend credit. Massive non-performing loans might cause depositors and international investors to lose faith in banks, which could spark a bank run and cause liquidity issues. According to Caprio and Klingebiel (2002), non-performing loans made up nearly 75% of all loan assets during the 1997 banking crisis in Indonesia, which caused the failure of more than sixty banks. Therefore, banks with significant amounts of default loans on their records might

The most significant of these hazards is credit risk since it results from borrowers' inability to make payments on their loans or on their tardiness in fulfilling their commitments. Regarding its effects on bank performance, credit risk has been noted as the primary concern (Sinkey, 1992). Less capital adequacy results from this risk since the bank will seek for alternative sources of funding to replace the loss. Additionally, it will create a delay in the company or perhaps bankruptcy, which will result in decreased profitability due to an inability to supply other customers' needs.

Defaults have a bad impact on a bank's profitability. Revenues from positive loans are directly deducted to provide for provisions for bad and uncertain obligations. Credit paperwork that is true, perfect, and expressly binding lowers the likelihood of wilful default and improves bank performance. The credit and recovery processes are linearly correlated

with a bank's performance (Asari et al, 2011). Banks are not able to make money off of defaulted credits, as Asari et al. (2011) convincingly established. The veracity of credit documentation is a method used to prevent defaults, hence the study has direct bearing on how well a bank performs. Banks' overall loan portfolio is reduced by loan default provisions, which has an impact on the interest earned on such assets.

Problem/Research Question

We utilised our dataset to build a **binary choice model** that might predict whether or not a borrower may repay the loan. In order to enable banks to apply our model, we used a few variables from our dataset that are readily available to them.

Using this approach, we try to answer the Research Question:

RQ: Is there a relationship between a borrower **defaulting** (not paying back the loan to the bank) and some **characteristics** of the borrower that the bank already has access to (family status, job, etc.), which could predict whether a borrower is going to pay back the loan or not?

Data Processing

1. Install the required libraries and the corresponding libraries

```
library(dplyr)

install.packages("caret")
library(caret)

library(ggplot2)

install.packages("caTools")
library(caTools)

install.packages("Amelia")
library(Amelia)

install.packages("MASS")
library(MASS)
```

2. Importing the Dataset and Studying it

```
### LOADING DATASET
initial_data <- read.csv("application_data.csv")
```

Upon looking at the data, we can see that the second column “TARGET” tells us whether the borrower of the loan paid the loan back on time or not. It is a categorical variable where 1 means that the borrower had payment difficulties and was unable to payback the loan in time.

3. Cleaning the data

The imported data in its raw form has a lot of information and requires thorough analysis before we can infer any meaningful insight from this to prove our research question or otherwise. As such, the next step in processing the data was to clean the initial raw data set.

This was also done to ensure that our model is accurate with little or no effect of any outlier from any false values. As such, there are **two** main things to be cleaned from the raw data set.

→ **Columns: (Variables)**

We first had to eliminate some variables to make the dataset easier to work with. The initial raw data set had a total of 122 variables. As such, we had to eliminate those variables that would significantly affect the target variable. Doing this would help us get a more accurate model along with making sure that the model does not overfit for the given data. To decide which columns to keep and which to eliminate, we have a look at the data again:

```
summary(initial_data)
```

```
> summary(initial_data)
   SK_ID_CURR           TARGET        NAME_CONTRACT_TYPE CODE_GENDER        FLAG_OWN_CAR
 Min.   :     0.07   Min.   :0.00000   Length:9038        Length:9038        Length:9038
 1st Qu.:102646.25   1st Qu.:0.00000   Class :character   Class :character   Class :character
 Median :105285.50   Median :0.00000   Mode  :character   Mode  :character   Mode  :character
 Mean   :114621.44   Mean   :0.07779
 3rd Qu.:107904.75   3rd Qu.:0.00000
 Max.   :219805.00   Max.   :1.00000

 FLAG_OWN_REALTY     CNT_CHILDREN    AMT_INCOME_TOTAL    AMT_CREDIT       AMT_ANNUITY
 Length:9038        Min.   :0.0000   Min.   :      0   Min.   :      0   Min.   :     0
 Class :character   1st Qu.:0.0000   1st Qu.: 112500   1st Qu.: 270000   1st Qu.: 16330
 Mode  :character   Median :0.0000   Median : 144000   Median : 509501   Median : 24939
                    Mean   :0.4152   Mean   : 167859   Mean   : 600702   Mean   : 27049
                    3rd Qu.:1.0000   3rd Qu.: 202500   3rd Qu.: 810000   3rd Qu.: 34587
                    Max.   :7.0000   Max.   :4500000   Max.   :2925000   Max.   :225000

 AMT_GOODS_PRICE    NAME_TYPE_SUITE    NAME_INCOME_TYPE   NAME_EDUCATION_TYPE NAME_FAMILY_STATUS
 Min.   :      1   Length:9038        Length:9038        Length:9038         Length:9038
 1st Qu.: 238500   Class :character   Class :character   Class :character    Class :character
 Median : 450000   Mode  :character   Mode  :character   Mode  :character    Mode  :character
 Mean   : 540415
 3rd Qu.: 679500
 Max.   :2925000
 NA's   :7
 NAME_HOUSING_TYPE  REGION_POPULATION_RELATIVE   DAYS_BIRTH         DAYS_EMPLOYED
 Length:9038        Min.   :0.000938          Min.   :-25160.000   Min.   :-15837.0
 Class :character   1st Qu.:0.010032          1st Qu.:-19641.750   1st Qu.: -2815.0
 Mode  :character   Median :0.018850          Median :-15841.500   Median : -1221.5
                    Mean   :0.021156          Mean   :-16055.726   Mean   : 63845.4
                    3rd Qu.:0.028663          3rd Qu.:-12434.000   3rd Qu.:  -295.2
                    Max.   :0.083300          Max.   :     0.047   Max.   :365243.0
```

```
 DAYS_REGISTRATION    DAYS_ID_PUBLISH  OWN_CAR_AGE     FLAG_MOBIL        FLAG_EMP_PHONE
 Min.   :-20981.000   Min.   :-6228   Min.   : 0.00   Length:9038        Length:9038
 1st Qu.: -7489.250   1st Qu.:-4300   1st Qu.: 5.00   Class :character   Class :character
 Median : -4503.500   Median :-3208   Median : 9.00   Mode  :character   Mode  :character
 Mean   : -5001.295   Mean   :-2973   Mean   :11.82
 3rd Qu.: -2019.250   3rd Qu.:-1675   3rd Qu.:15.00
 Max.   :     0.015   Max.   :    0   Max.   :65.00
                                      NA's   :6003
 FLAG_WORK_PHONE  FLAG_CONT_MOBILE    FLAG_PHONE          FLAG_EMAIL      OCCUPATION_TYPE
 Min.   :0.0000   Length:9038        Length:9038        Min.   :0.00000   Length:9038
 1st Qu.:0.0000   Class :character   Class :character   1st Qu.:0.00000   Class :character
 Median :0.0000   Mode  :character   Mode  :character   Median :0.00000   Mode  :character
 Mean   :0.2013                                        Mean   :0.05532
 3rd Qu.:0.0000                                        3rd Qu.:0.00000
 Max.   :1.0000                                        Max.   :9.00000

 CNT_FAM_MEMBERS REGION_RATING_CLIENT REGION_RATING_CLIENT_W_CITY WEEKDAY_APPR_PROCESS_START
 Min.   :1.000   Min.   :0.000        Min.   :-535.000           Length:9038
 1st Qu.:2.000   1st Qu.:2.000        1st Qu.:   2.000           Class :character
 Median :2.000   Median :2.000        Median :   2.000           Mode  :character
 Mean   :2.155   Mean   :2.042        Mean   :   1.962
 3rd Qu.:3.000   3rd Qu.:2.000        3rd Qu.:   2.000
 Max.   :9.000   Max.   :3.000        Max.   :   3.000

 HOUR_APPR_PROCESS_START REG_REGION_NOT_LIVE_REGION REG_REGION_NOT_WORK_REGION
 Min.   : 1.00           Min.   :0.00000            Min.   :0.00000
 1st Qu.:10.00           1st Qu.:0.00000            1st Qu.:0.00000
 Median :12.00           Median :0.00000            Median :0.00000
 Mean   :12.08           Mean   :0.01693            Mean   :0.04979
 3rd Qu.:14.00           3rd Qu.:0.00000            3rd Qu.:0.00000
 Max.   :22.00           Max.   :1.00000            Max.   :1.00000

 LIVE_REGION_NOT_WORK_REGION REG_CITY_NOT_LIVE_CITY REG_CITY_NOT_WORK_CITY LIVE_CITY_NOT_WORK_CITY
 Min.   :0.00000             Min.   :0.00000        Min.   :0.0000         Min.   :0.0000
 1st Qu.:0.00000             1st Qu.:0.00000        1st Qu.:0.0000         1st Qu.:0.0000
 Median :0.00000             Median :0.00000        Median :0.0000         Median :0.0000
 Mean   :0.03806             Mean   :0.08154        Mean   :0.2308         Mean   :0.1757
 3rd Qu.:0.00000             3rd Qu.:0.00000        3rd Qu.:0.0000         3rd Qu.:0.0000
 Max.   :1.00000             Max.   :1.00000        Max.   :1.0000         Max.   :1.0000

 ORGANIZATION_TYPE   EXT_SOURCE_1     EXT_SOURCE_2     EXT_SOURCE_3     APARTMENTS_AVG
 Length:9038        Min.   :0.000   Min.   :0.0000   Min.   :0.0000   Min.   :0.000
 Class :character   1st Qu.:0.336   1st Qu.:0.3911   1st Qu.:0.3728   1st Qu.:0.059
 Mode  :character   Median :0.512   Median :0.5660   Median :0.5353   Median :0.088
                    Mean   :0.505   Mean   :0.5149   Mean   :0.5117   Mean   :0.119
                    3rd Qu.:0.676   3rd Qu.:0.6654   3rd Qu.:0.6707   3rd Qu.:0.148
                    Max.   :0.929   Max.   :0.8550   Max.   :0.8825   Max.   :1.000
                    NA's   :5121    NA's   :28       NA's   :1788     NA's   :4581
 BASEMENTAREA_AVG YEARS_BEGINEXPLUATATION_AVG YEARS_BUILD_AVG COMMONAREA_AVG  ELEVATORS_AVG
 Min.   :0.000    Min.   :0.000               Min.   :0.000   Min.   :0.000   Min.   :0.000
 1st Qu.:0.044    1st Qu.:0.977               1st Qu.:0.687   1st Qu.:0.007   1st Qu.:0.000
 Median :0.076    Median :0.982               Median :0.755   Median :0.019   Median :0.000
 Mean   :0.089    Mean   :0.979               Mean   :0.751   Mean   :0.045   Mean   :0.079
 3rd Qu.:0.113    3rd Qu.:0.987               3rd Qu.:0.823   3rd Qu.:0.052   3rd Qu.:0.120
 Max.   :1.000    Max.   :1.000               Max.   :1.000   Max.   :1.000   Max.   :1.000
 NA's   :5267     NA's   :4403                NA's   :5992    NA's   :6298    NA's   :4786
 ENTRANCES_AVG   FLOORSMAX_AVG   FLOORSMIN_AVG    LANDAREA_AVG   LIVINGAPARTMENTS_AVG
 Min.   :0.000   Min.   :0.000   Min.   :0.000   Min.   :0.000   Min.   :0.000
 1st Qu.:0.069   1st Qu.:0.167   1st Qu.:0.083   1st Qu.:0.019   1st Qu.:0.050
 Median :0.138   Median :0.167   Median :0.208   Median :0.047   Median :0.074
 Mean   :0.151   Mean   :0.228   Mean   :0.235   Mean   :0.068   Mean   :0.100
 3rd Qu.:0.207   3rd Qu.:0.333   3rd Qu.:0.375   3rd Qu.:0.088   3rd Qu.:0.121
 Max.   :1.000   Max.   :1.000   Max.   :1.000   Max.   :1.000   Max.   :1.000
 NA's   :4519    NA's   :4469    NA's   :6119    NA's   :5343    NA's   :6172
```

```
 ENTRANCES_AVG    FLOORSMAX_AVG    FLOORSMIN_AVG     LANDAREA_AVG    LIVINGAPARTMENTS_AVG
 Min.   :0.000    Min.   :0.000    Min.   :0.000    Min.   :0.000    Min.   :0.000
 1st Qu.:0.069    1st Qu.:0.167    1st Qu.:0.083    1st Qu.:0.019    1st Qu.:0.050
 Median :0.138    Median :0.167    Median :0.208    Median :0.047    Median :0.074
 Mean   :0.151    Mean   :0.228    Mean   :0.235    Mean   :0.068    Mean   :0.100
 3rd Qu.:0.207    3rd Qu.:0.333    3rd Qu.:0.375    3rd Qu.:0.088    3rd Qu.:0.121
 Max.   :1.000    Max.   :1.000    Max.   :1.000    Max.   :1.000    Max.   :1.000
 NA's   :4519     NA's   :4469     NA's   :6119     NA's   :5343     NA's   :6172
 LIVINGAREA_AVG   NONLIVINGAPARTMENTS_AVG NONLIVINGAREA_AVG APARTMENTS_MODE BASEMENTAREA_MODE
 Min.   :0.000    Min.   :0.000           Min.   :0.000     Min.   :0.000   Min.   :0.000
 1st Qu.:0.047    1st Qu.:0.000           1st Qu.:0.000     1st Qu.:0.052   1st Qu.:0.040
 Median :0.075    Median :0.000           Median :0.004     Median :0.084   Median :0.074
 Mean   :0.109    Mean   :0.009           Mean   :0.029     Mean   :0.115   Mean   :0.088
 3rd Qu.:0.128    3rd Qu.:0.004           3rd Qu.:0.027     3rd Qu.:0.143   3rd Qu.:0.113
 Max.   :1.000    Max.   :2.000           Max.   :2.000     Max.   :1.000   Max.   :1.000
 NA's   :4531     NA's   :6263            NA's   :4951      NA's   :4582    NA's   :5268
 YEARS_BEGINEXPLUATATION_MODE YEARS_BUILD_MODE COMMONAREA_MODE ELEVATORS_MODE  ENTRANCES_MODE
 Min.   :0.000                Min.   :0.000    Min.   :0.000   Min.   :0.000   Min.   :0.000
 1st Qu.:0.977                1st Qu.:0.693    1st Qu.:0.007   1st Qu.:0.000   1st Qu.:0.069
 Median :0.982                Median :0.758    Median :0.017   Median :0.000   Median :0.138
 Mean   :0.978                Mean   :0.758    Mean   :0.042   Mean   :0.075   Mean   :0.146
 3rd Qu.:0.987                3rd Qu.:0.824    3rd Qu.:0.049   3rd Qu.:0.121   3rd Qu.:0.207
 Max.   :1.000                Max.   :1.000    Max.   :1.000   Max.   :1.000   Max.   :1.000
 NA's   :4404                 NA's   :5993     NA's   :6299    NA's   :4787    NA's   :4520
 FLOORSMAX_MODE   FLOORSMIN_MODE  LANDAREA_MODE    LIVINGAPARTMENTS_MODE LIVINGAREA_MODE
 Min.   :0.000    Min.   :0.000   Min.   :0.000    Min.   :0.000         Min.   :0.000
 1st Qu.:0.167    1st Qu.:0.083   1st Qu.:0.017    1st Qu.:0.054         1st Qu.:0.044
 Median :0.167    Median :0.208   Median :0.045    Median :0.074         Median :0.073
 Mean   :0.224    Mean   :0.231   Mean   :0.067    Mean   :0.104         Mean   :0.107
 3rd Qu.:0.333    3rd Qu.:0.375   3rd Qu.:0.086    3rd Qu.:0.128         3rd Qu.:0.125
 Max.   :1.000    Max.   :1.000   Max.   :1.000    Max.   :1.000         Max.   :1.000
 NA's   :4470     NA's   :6120    NA's   :5344     NA's   :6173          NA's   :4533
 NONLIVINGAPARTMENTS_MODE NONLIVINGAREA_MODE APARTMENTS_MEDI    BASEMENTAREA_MEDI
 Length:9038              Length:9038        Length:9038        Length:9038
 Class :character         Class :character   Class :character   Class :character
 Mode  :character         Mode  :character   Mode  :character   Mode  :character
```

```
 YEARS_BEGINEXPLUATATION_MEDI YEARS_BUILD_MEDI     COMMONAREA_MEDI    ELEVATORS_MEDI
 Min.   :0.000                Min.   :-14915.000   Min.   :-585.000   Min.   :-7324.000
 1st Qu.:0.977                1st Qu.:     0.691   1st Qu.:   0.007   1st Qu.:    0.000
 Median :0.982                Median :     0.758   Median :   0.019   Median :    0.000
 Mean   :0.979                Mean   :    -4.144   Mean   :  -0.169   Mean   :   -1.644
 3rd Qu.:0.987                3rd Qu.:     0.826   3rd Qu.:   0.051   3rd Qu.:    0.120
 Max.   :1.000                Max.   :     1.000   Max.   :   1.000   Max.   :    1.000
 NA's   :4404                 NA's   :5993         NA's   :6299       NA's   :4787
 ENTRANCES_MEDI      FLOORSMAX_MEDI  FLOORSMIN_MEDI  LANDAREA_MEDI   LIVINGAPARTMENTS_MEDI
 Min.   :-4501.000   Min.   :0.000   Min.   :0.000   Min.   :0.000   Min.   :0.000
 1st Qu.:    0.069   1st Qu.:0.167   1st Qu.:0.083   1st Qu.:0.019   1st Qu.:0.051
 Median :    0.138   Median :0.167   Median :0.208   Median :0.048   Median :0.075
 Mean   :   -0.846   Mean   :0.228   Mean   :0.235   Mean   :0.070   Mean   :0.101
 3rd Qu.:    0.207   3rd Qu.:0.333   3rd Qu.:0.375   3rd Qu.:0.089   3rd Qu.:0.123
 Max.   :    1.000   Max.   :1.000   Max.   :1.000   Max.   :1.000   Max.   :1.000
 NA's   :4520        NA's   :4471    NA's   :6120    NA's   :5344    NA's   :6173
 LIVINGAREA_MEDI NONLIVINGAPARTMENTS_MEDI NONLIVINGAREA_MEDI FONDKAPREMONT_MODE HOUSETYPE_MODE
 Min.   :0.000   Min.   :0.000            Min.   :0.000      Length:9038        Length:9038
 1st Qu.:0.047   1st Qu.:0.000            1st Qu.:0.000      Class :character   Class :character
 Median :0.075   Median :0.000            Median :0.003      Mode  :character   Mode  :character
 Mean   :0.110   Mean   :0.008            Mean   :0.028
 3rd Qu.:0.130   3rd Qu.:0.004            3rd Qu.:0.026
 Max.   :1.000   Max.   :1.000            Max.   :1.000
 NA's   :4532    NA's   :6264             NA's   :4952
```

```
 TOTALAREA_MODE  WALLSMATERIAL_MODE EMERGENCYSTATE_MODE OBS_30_CNT_SOCIAL_CIRCLE
 Min.   :0.000   Length:9038        Length:9038         Min.   : 0.000
 1st Qu.:0.042   Class :character   Class :character    1st Qu.: 0.000
 Median :0.068   Mode  :character   Mode  :character    Median : 0.000
 Mean   :0.103                                         Mean   : 1.413
 3rd Qu.:0.126                                         3rd Qu.: 2.000
 Max.   :2.000                                         Max.   :25.000
 NA's   :4348                                          NA's   :40
 DEF_30_CNT_SOCIAL_CIRCLE OBS_60_CNT_SOCIAL_CIRCLE DEF_60_CNT_SOCIAL_CIRCLE DAYS_LAST_PHONE_CHANGE
 Min.   :0.0000           Min.   : 0.000           Min.   :0.0000           Min.   :-3943.0
 1st Qu.:0.0000           1st Qu.: 0.000           1st Qu.:0.0000           1st Qu.:-1550.0
 Median :0.0000           Median : 0.000           Median :0.0000           Median : -747.0
 Mean   :0.1469           Mean   : 1.396           Mean   :0.1029           Mean   : -955.2
 3rd Qu.:0.0000           3rd Qu.: 2.000           3rd Qu.:0.0000           3rd Qu.: -261.0
 Max.   :5.0000           Max.   :25.000           Max.   :5.0000           Max.   :    0.0
 NA's   :40               NA's   :40               NA's   :40               NA's   :1
 FLAG_DOCUMENT_2 FLAG_DOCUMENT_3  FLAG_DOCUMENT_4    FLAG_DOCUMENT_5   FLAG_DOCUMENT_6
 Min.   :0       Min.   :0.0000   Length:9038        Min.   :0.00000   Min.   :0.00000
 1st Qu.:0       1st Qu.:0.0000   Class :character   1st Qu.:0.00000   1st Qu.:0.00000
 Median :0       Median :1.0000   Mode  :character   Median :0.00000   Median :0.00000
 Mean   :0       Mean   :0.7054                      Mean   :0.01541   Mean   :0.08693
 3rd Qu.:0       3rd Qu.:1.0000                      3rd Qu.:0.00000   3rd Qu.:0.00000
 Max.   :0       Max.   :1.0000                      Max.   :1.00000   Max.   :1.00000
 NA's   :1       NA's   :1                           NA's   :2         NA's   :2
 FLAG_DOCUMENT_7     FLAG_DOCUMENT_8   FLAG_DOCUMENT_9    FLAG_DOCUMENT_10    FLAG_DOCUMENT_11
 Min.   :0.0000000   Min.   :0.00000   Min.   :0.000000   Min.   :0.0000000   Min.   :0.000000
 1st Qu.:0.0000000   1st Qu.:0.00000   1st Qu.:0.000000   1st Qu.:0.0000000   1st Qu.:0.000000
 Median :0.0000000   Median :0.00000   Median :0.000000   Median :0.0000000   Median :0.000000
 Mean   :0.0000379   Mean   :0.07847   Mean   :0.003984   Mean   :0.0001073   Mean   :0.003938
 3rd Qu.:0.0000000   3rd Qu.:0.00000   3rd Qu.:0.000000   3rd Qu.:0.0000000   3rd Qu.:0.000000
 Max.   :0.3425289   Max.   :1.00000   Max.   :1.000000   Max.   :0.9697000   Max.   :1.000000
 NA's   :2           NA's   :2         NA's   :2          NA's   :2           NA's   :2


 FLAG_DOCUMENT_12    FLAG_DOCUMENT_13   FLAG_DOCUMENT_14   FLAG_DOCUMENT_15   FLAG_DOCUMENT_16
 Min.   :0.0000000   Min.   :0.000000   Min.   :0.000000   Min.   :0.000000   Min.   :0.000000
 1st Qu.:0.0000000   1st Qu.:0.000000   1st Qu.:0.000000   1st Qu.:0.000000   1st Qu.:0.000000
 Median :0.0000000   Median :0.000000   Median :0.000000   Median :0.000000   Median :0.000000
 Mean   :0.0000019   Mean   :0.002103   Mean   :0.002885   Mean   :0.000558   Mean   :0.009969
 3rd Qu.:0.0000000   3rd Qu.:0.000000   3rd Qu.:0.000000   3rd Qu.:0.000000   3rd Qu.:0.000000
 Max.   :0.0176000   Max.   :1.000000   Max.   :1.000000   Max.   :1.000000   Max.   :1.000000
 NA's   :2           NA's   :2          NA's   :2          NA's   :2          NA's   :2
 FLAG_DOCUMENT_17    FLAG_DOCUMENT_18   FLAG_DOCUMENT_19   FLAG_DOCUMENT_20    FLAG_DOCUMENT_21
 Min.   :0.0000000   Min.   :0.000000   Min.   :0.000000   Min.   :0.0000000   Min.   :0.0000000
 1st Qu.:0.0000000   1st Qu.:0.000000   1st Qu.:0.000000   1st Qu.:0.0000000   1st Qu.:0.0000000
 Median :0.0000000   Median :0.000000   Median :0.000000   Median :0.0000000   Median :0.0000000
 Mean   :0.0003372   Mean   :0.007416   Mean   :0.000223   Mean   :0.0004427   Mean   :0.0002213
 3rd Qu.:0.0000000   3rd Qu.:0.000000   3rd Qu.:0.000000   3rd Qu.:0.0000000   3rd Qu.:0.0000000
 Max.   :1.0000000   Max.   :1.000000   Max.   :1.000000   Max.   :1.0000000   Max.   :1.0000000
 NA's   :2           NA's   :2          NA's   :2          NA's   :2           NA's   :2
 AMT_REQ_CREDIT_BUREAU_HOUR AMT_REQ_CREDIT_BUREAU_DAY AMT_REQ_CREDIT_BUREAU_WEEK
 Min.   :0.0000             Min.   :0.0000            Min.   :0.0000
 1st Qu.:0.0000             1st Qu.:0.0000            1st Qu.:0.0000
 Median :0.0000             Median :0.0000            Median :0.0000
 Mean   :0.0068             Mean   :0.0079            Mean   :0.0324
 3rd Qu.:0.0000             3rd Qu.:0.0000            3rd Qu.:0.0000
 Max.   :2.0000             Max.   :4.0000            Max.   :4.0000
 NA's   :1223               NA's   :1223              NA's   :1223
 AMT_REQ_CREDIT_BUREAU_MON AMT_REQ_CREDIT_BUREAU_QRT AMT_REQ_CREDIT_BUREAU_YEAR
 Min.   : 0.0000           Min.   :0.0000            Min.   : 0.000
 1st Qu.: 0.0000           1st Qu.:0.0000            1st Qu.: 0.000
 Median : 0.0000           Median :0.0000            Median : 1.000
 Mean   : 0.2767           Mean   :0.2706            Mean   : 1.897
 3rd Qu.: 0.0000           3rd Qu.:0.0000            3rd Qu.: 3.000
 Max.   :16.0000           Max.   :8.0000            Max.   :13.000
 NA's   :1223              NA's   :1223              NA's   :1223
```

From the above summary of data we observe that our data can be divided into 2 sections.

A. *Information where the client lives*:
   There are a lot of columns that show normalised 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 NA's as it is.

```
> summary(initial_data$NONLVINGAREA_MODE)
Length  Class   Mode
     0   NULL   NULL
> summary(initial_data$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
```

The values involved in the second case, for example, contain a lot of NA's (179943) with repetitions of 0. Therefore, we will remove these columns from our data.

```
### CLEANING DATASET
clean <- subset(initial_data, select = -c( NONLIVINGAREA_MODE, OWN_CAR_AGE, EXT_SOURCE_1, APARTMENTS_AVG, BASEMENTAREA_AVG,
                                           YEARS_BEGINEXPLUATATION_AVG, YEARS_BUILD_AVG, COMMONAREA_AVG, ELEVATORS_AVG,
                                           ENTRANCES_AVG, FLOORSMAX_AVG, FLOORSMIN_AVG, LANDAREA_AVG, LIVINGAPARTMENTS_AVG,
                                           LIVINGAREA_AVG, NONLIVINGAPARTMENTS_AVG, NONLIVINGAREA_AVG, APARTMENTS_MODE,
                                           BASEMENTAREA_MODE, YEARS_BEGINEXPLUATATION_MODE, YEARS_BUILD_MODE, COMMONAREA_MODE,
                                           ELEVATORS_MODE, ENTRANCES_MODE, FLOORSMAX_MODE, FLOORSMIN_MODE, LANDAREA_MODE,
                                           LIVINGAPARTMENTS_MODE, LIVINGAREA_MODE, NONLIVINGAPARTMENTS_MODE, APARTMENTS_MEDI,
                                           BASEMENTAREA_MEDI, YEARS_BEGINEXPLUATATION_MEDI, YEARS_BUILD_MEDI, COMMONAREA_MEDI,
                                           ELEVATORS_MEDI, ENTRANCES_MEDI, FLOORSMAX_MEDI, FLOORSMIN_MEDI, LANDAREA_MEDI,
                                           LIVINGAPARTMENTS_MEDI, LIVINGAREA_MEDI, NONLIVINGAPARTMENTS_MEDI, NONLIVINGAREA_MEDI,
                                           FONDKAPREMONT_MODE, HOUSETYPE_MODE, TOTALAREA_MODE, WALLSMATERIAL_MODE, TOTALAREA_MODE,
                                           WALLSMATERIAL_MODE, EMERGENCYSTATE_MODE))
```

B. *Variables with no variation:*

There are a few columns named "FLAG_DOCUMENTS", that are supposed to record the information provided by the client. These flag variables represent whether the particular documents have been provided by the loan borrower or not. However we can see that these variables have very little variation and therefore should be removed from our data set as well. The following code is used:

```
clean <- subset(clean, select = -c( FLAG_DOCUMENT_2, FLAG_DOCUMENT_3, FLAG_DOCUMENT_4, FLAG_DOCUMENT_5, FLAG_DOCUMENT_6,
                                    FLAG_DOCUMENT_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 variables based on insights and research question:*

  On the basis of reason, we were able to eliminate some more factors. The removal of these variables still provides us with sufficient predictors to create a sound model because it is apparent that they would not have a substantial impact on the TARGET value. Additionally, some of these factors were used to eliminate potential bias. For instance, the variable REGION RATING CLIENT provided a rating for the client's home region. Due to the fact that the bank itself established this rating, there may be prejudice.

```
clean <- subset(clean, select = -c( FLAG_EMP_PHONE, FLAG_WORK_PHONE, FLAG_CONT_MOBILE, FLAG_PHONE, REGION_RATING_CLIENT,
                                    REGION_RATING_CLIENT_W_CITY, OBS_30_CNT_SOCIAL_CIRCLE, DEF_30_CNT_SOCIAL_CIRCLE,
                                    OBS_60_CNT_SOCIAL_CIRCLE, DEF_60_CNT_SOCIAL_CIRCLE, AMT_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_PROCESS_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       FLAG_OWN_REALTY     CNT_CHILDREN    AMT_INCOME_TOTAL    AMT_CREDIT
 Min.   :0.00000   Length:9038        Length:9038        Length:9038        Length:9038        Min.   :0.0000   Min.   :      0   Min.   :      0
 1st Qu.:0.00000   Class :character   Class :character   Class :character   Class :character   1st Qu.:0.0000   1st Qu.: 112500   1st Qu.: 270000
 Median :0.00000   Mode  :character   Mode  :character   Mode  :character   Mode  :character   Median :0.0000   Median : 144000   Median : 509501
 Mean   :0.07779                                                           Mean   :0.4152   Mean   : 167859   Mean   : 600702
 3rd Qu.:0.00000                                                           3rd Qu.:1.0000   3rd Qu.: 202500   3rd Qu.: 810000
 Max.   :1.00000                                                           Max.   :7.0000   Max.   :4500000   Max.   :2925000

  AMT_ANNUITY     AMT_GOODS_PRICE   NAME_INCOME_TYPE   NAME_EDUCATION_TYPE NAME_FAMILY_STATUS NAME_HOUSING_TYPE    DAYS_BIRTH
 Min.   :     0   Min.   :      1   Length:9038        Length:9038         Length:9038        Length:9038        Min.   :-25160.000
 1st Qu.: 16330   1st Qu.: 238500   Class :character   Class :character    Class :character   Class :character   1st Qu.:-19641.750
 Median : 24939   Median : 450000   Mode  :character   Mode  :character    Mode  :character   Mode  :character   Median :-15841.500
 Mean   : 27049   Mean   : 540415                                                                              Mean   :-16055.726
 3rd Qu.: 34587   3rd Qu.: 679500                                                                              3rd Qu.:-12434.000
 Max.   :225000   Max.   :2925000                                                                              Max.   :     0.047
                  NA's   :7
 DAYS_EMPLOYED      DAYS_REGISTRATION    DAYS_ID_PUBLISH  FLAG_MOBIL          FLAG_EMAIL      OCCUPATION_TYPE    CNT_FAM_MEMBERS ORGANIZATION_TYPE
 Min.   :-15837.0   Min.   :-20981.000   Min.   :-6228   Length:9038        Min.   :0.00000   Length:9038        Min.   :1.000   Length:9038
 1st Qu.: -2815.0   1st Qu.: -7489.250   1st Qu.:-4300   Class :character   1st Qu.:0.00000   Class :character   1st Qu.:2.000   Class :character
 Median : -1221.5   Median : -4503.500   Median :-3208   Mode  :character   Median :0.00000   Mode  :character   Median :2.000   Mode  :character
 Mean   :  63845.4  Mean   : -5001.295   Mean   :-2973                      Mean   :0.05532                      Mean   :2.155
 3rd Qu.:  -295.2   3rd Qu.: -2019.250   3rd Qu.:-1675                      3rd Qu.:0.00000                      3rd Qu.:3.000
 Max.   :365243.0   Max.   :     0.015   Max.   :    0                      Max.   :9.00000                      Max.   :9.000

  EXT_SOURCE_2     EXT_SOURCE_3     DAYS_LAST_PHONE_CHANGE
 Min.   :0.0000   Min.   :0.0000   Min.   :-3943.0
 1st Qu.:0.3911   1st Qu.:0.3728   1st Qu.:-1550.0
 Median :0.5660   Median :0.5353   Median : -747.0
 Mean   :0.5149   Mean   :0.5117   Mean   : -955.2
 3rd Qu.:0.6654   3rd Qu.:0.6707   3rd Qu.: -261.0
 Max.   :0.8550   Max.   :0.8825   Max.   :    0.0
 NA's   :28       NA's   :1788     NA's   :1
```

As a result, we now have a collection of data with columns that are statistically significant and can be used as predictors in our model. Labels were eliminated since they would have not improved our model and would have merely increased noise. We took care not to introduce any bias due to the missing variables when doing this. This bias may occur if we exclude factors that strongly influence the dependent variable, which might lead to an overestimation of the impact of the predictors we kept.

→ **Rows**

We have access to a lot of data, which we have now cleansed to create a useful data collection. Let's visualise what the rows include:

```
> str(clean)
'data.frame':   9038 obs. of  26 variables:
 $ TARGET                : num  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          : num  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_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" ...
 $ DAYS_BIRTH            : num  9461 16765 19046 19005 19932 ...
 $ DAYS_EMPLOYED         : num  637 1188 225 3039 3038 ...
 $ DAYS_REGISTRATION     : num  3648 1186 4260 9833 4311 ...
 $ DAYS_ID_PUBLISH       : num  2120 291 2531 2437 3458 ...
 $ FLAG_MOBIL            : chr  "1" "1" "1" "1" ...
 $ FLAG_EMAIL            : num  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 ...
 $ ORGANIZATION_TYPE     : chr  "Business Entity Type 3" "School" "Government" "Business Entity Type 3" ...
 $ 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 ...
 $ DAYS_LAST_PHONE_CHANGE: num  1134 828 815 617 1106 ...
```

We identified a few issues here:

1. The variables that had category or factor values were converted to integer or char data types.

- To do this, we used R's **as.factor** function to transform these numbers to factors. We eventually decided to let the model handle it and maintain these values in the already allocated data types after realising that the model automatically treats this when being trained.

2. The variables that tracked days had negative values.

- We were able to overcome this issue with the aid of a few basic lines.

```
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 a significant number of NULL values.

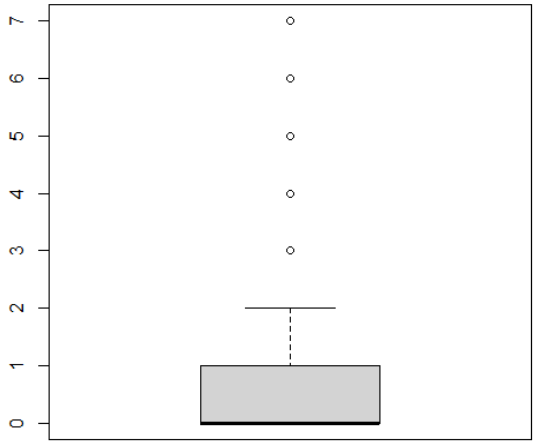
- We were only going to delete the rows with NULL values to solve this issue. We soon discovered that several parameters, such as "EXT_SOURCE 2" or "ORGANISATION_TYPE," were justified with NULL (or NA) values. This is due to the fact that these variables are optional and can be left empty, and having them set to NULL really may have a significant impact on the 'TARGET's' output. In order to make more intuitive sense, we altered the NA value for these to "Unknown". To ensure that there were no incomplete rows that may taint our data, we performed the deletion approach for the remaining variables and removed the rows with NULL values.

```
### REPLACING NULL VALUES
clean$ORGANIZATION_TYPE[clean$ORGANIZATION_TYPE == "XNA"] <- "Unknown"
clean$OCCUPATION_TYPE[clean$OCCUPATION_TYPE == ""] <- "UNKNOWN"

### REMOVING 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_LAST_PHONE_CHANGE)|clean$CODE_GENDER == "XNA"),]

summary(clean_data)
```

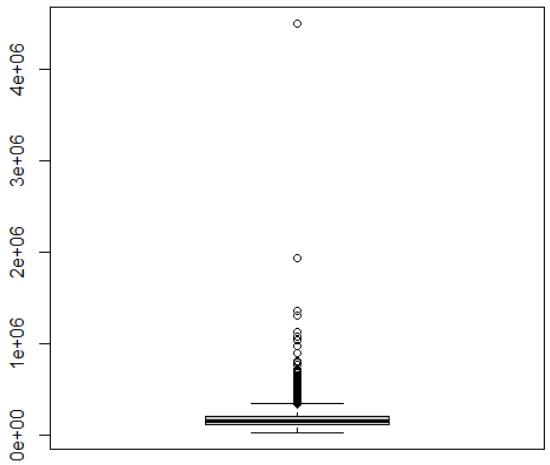
```
> summary(clean_data)
     TARGET        NAME_CONTRACT_TYPE CODE_GENDER        FLAG_OWN_CAR       FLAG_OWN_REALTY     CNT_CHILDREN    AMT_INCOME_TOTAL     AMT_CREDIT
 Min.   :0.00000   Length:9030        Length:9030        Length:9030        Length:9030        Min.   :0.0000   Min.   :  25650   Min.   :  45000
 1st Qu.:0.00000   Class :character   Class :character   Class :character   Class :character   1st Qu.:0.0000   1st Qu.: 112500   1st Qu.: 270000
 Median :0.00000   Mode  :character   Mode  :character   Mode  :character   Mode  :character   Median :0.0000   Median : 144000   Median : 509762
 Mean   :0.07774                                                           Mean   :0.4153   Mean   : 167918   Mean   : 601027
 3rd Qu.:0.00000                                                           3rd Qu.:1.0000   3rd Qu.: 202500   3rd Qu.: 810000
 Max.   :1.00000                                                           Max.   :7.0000   Max.   :4500000   Max.   :2925000

  AMT_ANNUITY     AMT_GOODS_PRICE   NAME_INCOME_TYPE   NAME_EDUCATION_TYPE NAME_FAMILY_STATUS NAME_HOUSING_TYPE    DAYS_BIRTH    DAYS_EMPLOYED
 Min.   :  2596   Min.   :  45000   Length:9030        Length:9030         Length:9030        Length:9030        Min.   : 7705   Min.   :    17.0
 1st Qu.: 16360   1st Qu.: 238500   Class :character   Class :character    Class :character   Class :character   1st Qu.:12437   1st Qu.:   933.8
 Median : 24939   Median : 450000   Mode  :character   Mode  :character    Mode  :character   Mode  :character   Median :15848   Median :  2266.0
 Mean   : 27063   Mean   : 540475                                                                              Mean   :16061   Mean   : 67875.5
 3rd Qu.: 34587   3rd Qu.: 679500                                                                              3rd Qu.:19642   3rd Qu.:  5906.5
 Max.   :225000   Max.   :2925000                                                                              Max.   :25160   Max.   :365243.0

 DAYS_REGISTRATION DAYS_ID_PUBLISH  FLAG_MOBIL          FLAG_EMAIL      OCCUPATION_TYPE    CNT_FAM_MEMBERS ORGANIZATION_TYPE   EXT_SOURCE_2
 Min.   :    0     Min.   :   0    Length:9030        Min.   :0.00000   Length:9030        Min.   :1.000   Length:9030        Min.   :0.000074
 1st Qu.: 2020     1st Qu.:1675    Class :character   1st Qu.:0.00000   Class :character   1st Qu.:2.000   Class :character   1st Qu.:0.391168
 Median : 4504     Median :3210    Mode  :character   Median :0.00000   Mode  :character   Median :2.000   Mode  :character   Median :0.566144
 Mean   : 5003     Mean   :2974                       Mean   :0.05426                      Mean   :2.155                      Mean   :0.514987
 3rd Qu.: 7494     3rd Qu.:4301                       3rd Qu.:0.00000                      3rd Qu.:3.000                      3rd Qu.:0.665465
 Max.   :20981     Max.   :6228                       Max.   :1.00000                      Max.   :9.000                      Max.   :0.855000
                                                                                                                          NA's   :28

  EXT_SOURCE_3    DAYS_LAST_PHONE_CHANGE
 Min.   :0.0005   Min.   :   0.0
 1st Qu.:0.3736   1st Qu.: 261.0
 Median :0.5353   Median : 747.0
 Mean   :0.5117   Mean   : 955.5
 3rd Qu.:0.6707   3rd Qu.:1550.0
 Max.   :0.8825   Max.   :3943.0
 NA's   :1786
```

To further increase the integrity of our data, we used box plots in order to identify outlier values which would have otherwise affected our model, and thus removed them. It is important to note that we only made box plots for numeric variables as using boxplots for data in character form will give erroneous results.

```
boxplot(clean_data$CNT_CHILDREN)
```

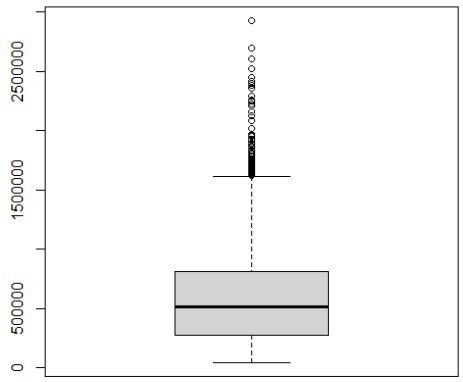



```
boxplot(clean_data$AMT_INCOME_TOTAL)
```

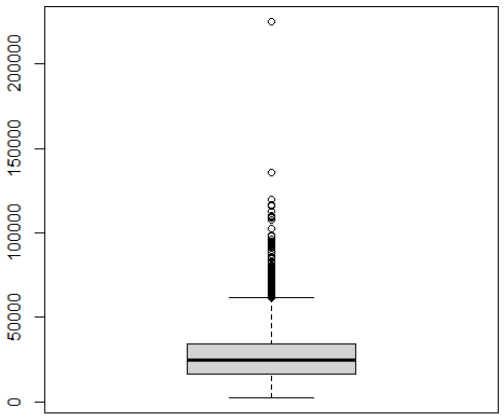



```
boxplot(clean_data$AMT_CREDIT)
```

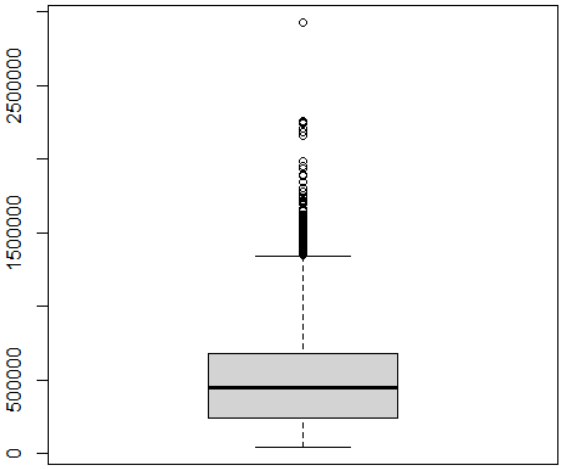
```
boxplot(clean_data$AMT_ANNUITY)
```



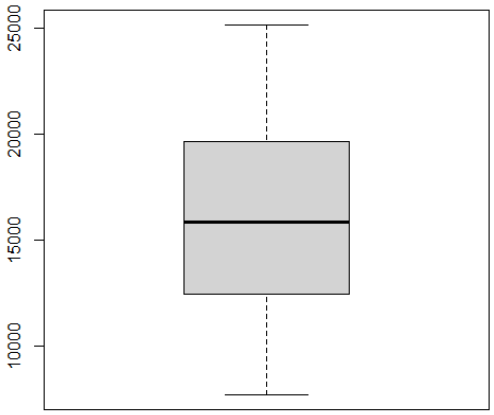

```
boxplot(clean_data$AMT_GOODS_PRICE)
```



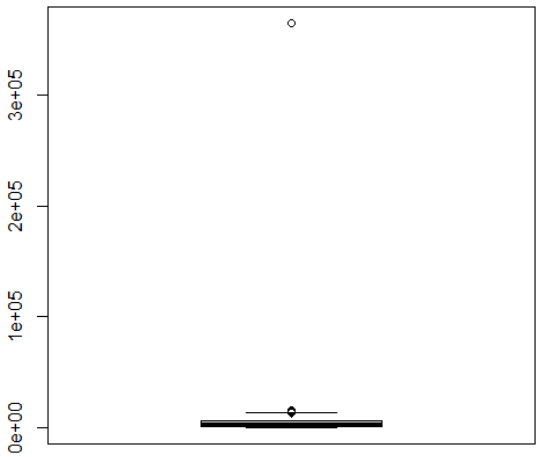

```
boxplot(clean_data$DAYS_BIRTH)
```



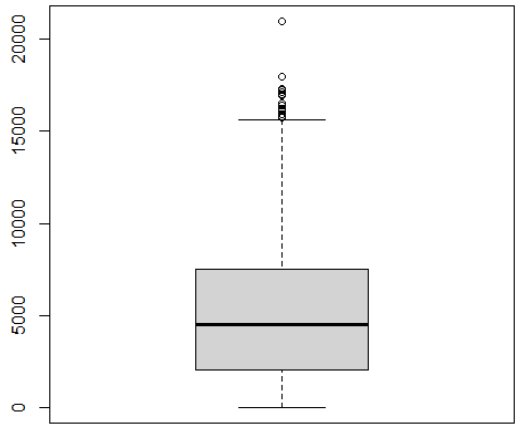

```
boxplot(clean_data$DAYS_EMPLOYED)
```



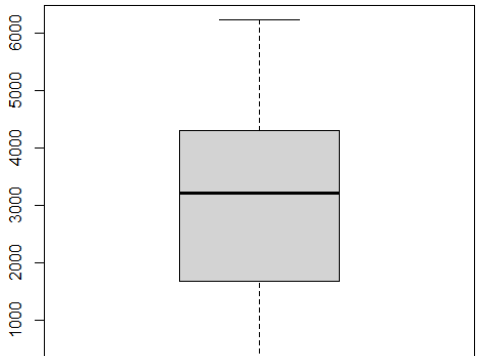

```
boxplot(clean_data$DAYS_REGISTRATION)
```



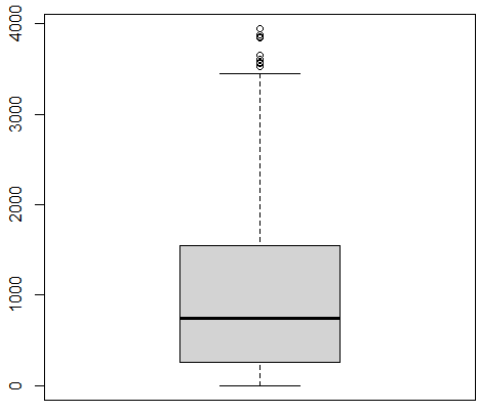

```
boxplot(clean_data$DAYS_ID_PUBLISH)
```

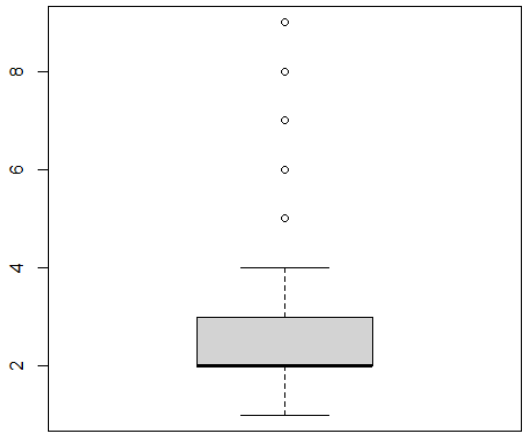



```
boxplot(clean_data$DAYS_LAST_PHONE_CHANGE)
```




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



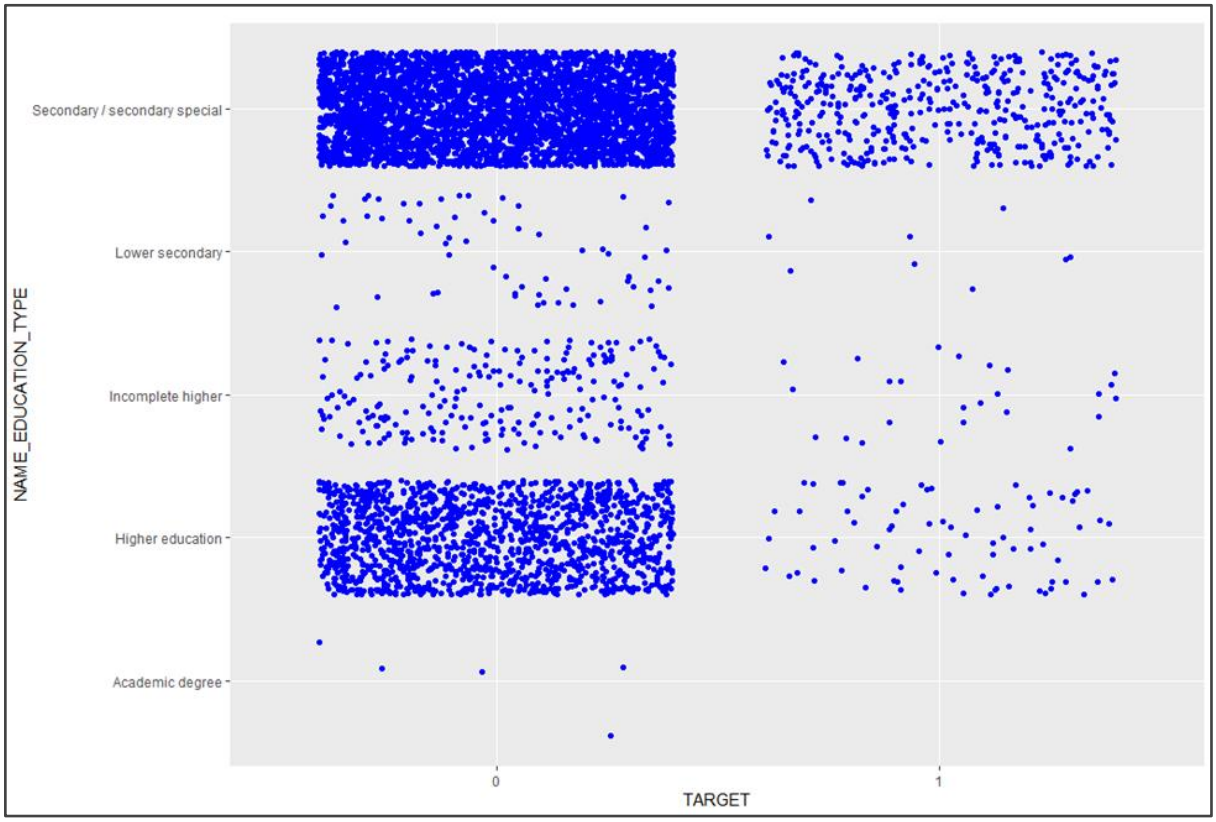


```
### REMOVING OUTLIERS
outliers <- boxplot(clean_data$CNT_CHILDREN, plot = FALSE)$out
clean_data <- clean_data[!(clean_data$CNT_CHILDREN %in% outliers),]

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

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

4. Preliminary Analysis

To understand our data and to derive patterns for the defaults through other factors, we chose a few of the variables in our dataset to compare against the loan defaults to notice trends.

Education vs Default




```
ggplot(clean_data,aes(x=TARGET,y=NAME_EDUCATION_TYPE)) +
  geom_jitter(colour='blue')
```

A very natural and intuitive comparison can be made on defaults with comparison to the education that a person receives. It might be argued that a more educated person would know the consequences of defaulting on a loan, or perhaps that the person would be better suited to understanding their financial capacity and need and how they would be able to service a loan properly.

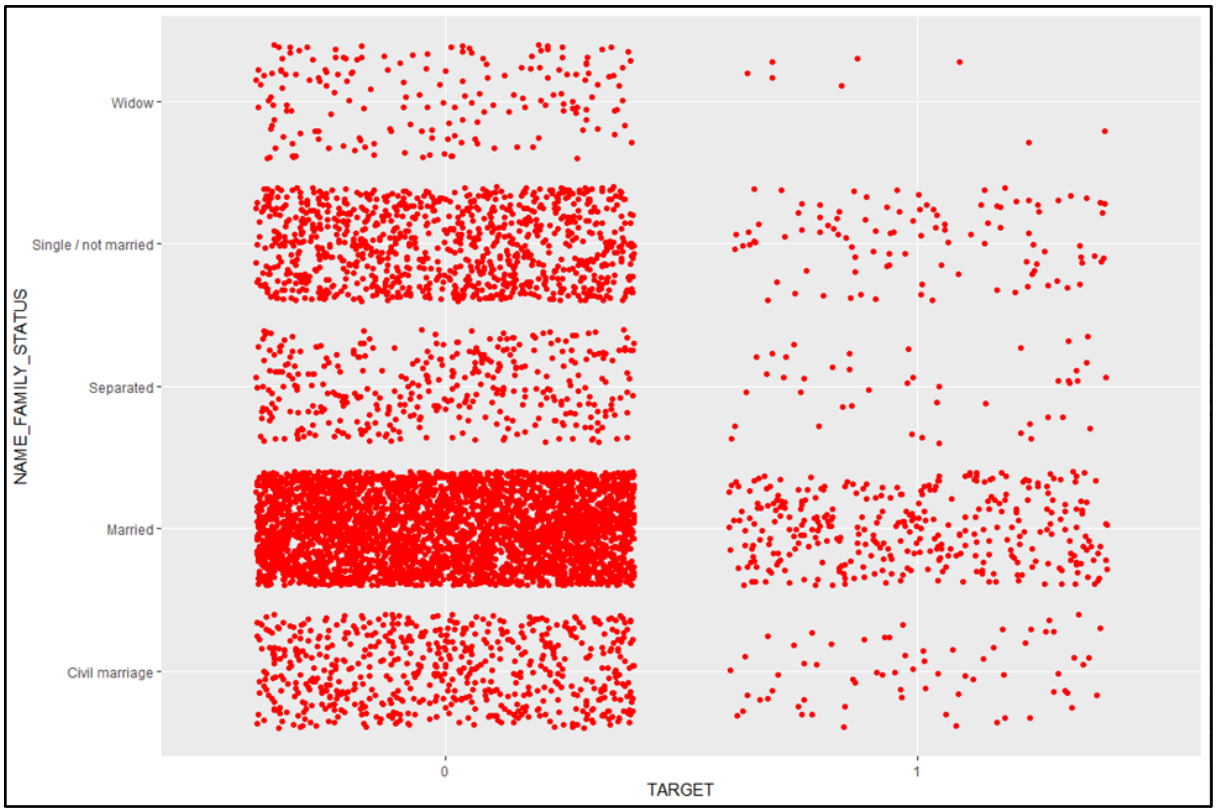
This assumption is supported by the graph to some extent. The graph mentions the education levels on the y-axis and the 0/1 denoting yes/no in terms of the loan default of our target variable on the x-axis.

For the loan defaulters i.e. x-axis = 1, the loan defaulters decrease as the education level increases. This gives a hint that perhaps the more educated people do not default as much as the less educated.

On the non-defaulters end, however, there is a disparity where non-defaults increase in a the higher education region and fall on both ends towards academic degree and incomplete

higher education. This can be attributed to either a higher financial need during the higher education degree and starting out in life or a more rebellious attitude during a certain phase of life. However, a conclusion cannot be drawn through just this limited representation.

Family Status vs Default




```
ggplot(clean_data,aes(x=TARGET,y=NAME_FAMILY_STATUS)) +
  geom_jitter(colour='red')
```

When comparing the family status of a person to defaulting, no trend can be observed in the absolute manner. Especially when comparing the two different defaulting states together, due to the greater number of the non-defaulting people, the graph seems skewed. However, if we look closer towards the non-defaulters, the married couples are the ones in majority in the non-defaulting category.

In retrospect though, conclusions from this graph would be ineffective because the higher population demographics skew the overall results quite significantly,

Even after looking at many other similar comparisons, no conclusive decision could be made as to what has a detrimental effect on the default of a loan. This could be due to a number of reasons. Either there are multiple factors that are having an effect on the default of a loan or that a single factor's effect on the overall default cannot be measured through such a simple one-to-one comparison.

Hence, in order to draw a more substantial conclusion on the factors that have an effect on the default of a loan and to predict these factors better, we chose to go forward with making a model. The steps in achieving this model are shown below.

5. Dividing the Data:

When working on the model we need a test data on which we will be able to run our model to check for accuracy. Therefore, we divide our data into two random parts, testing and the training. Since we are going to train our model on the training data, it will be significantly larger than the testing data.

```
### DIVIDE THE DATA INTO TRAIN AND TEST SETS
clean_data$TARGET <- as.factor(clean_data$TARGET)
divide = sample.split(clean_data$TARGET, SplitRatio = 0.75)

train = subset(clean_data, divide == TRUE)
test = subset(clean_data, divide == FALSE)
```

When we get our data, it is also important to note that there might be some levels for the attributes with factor data types that are not sufficiently available in both training and testing dataset. We have highlighted this in one variable: “NAME_INCOME_TYPE”

```
> table(clean_data$NAME_INCOME_TYPE)

        Businessman Commercial associate      Maternity leave            Pensioner        State servant
                  3                59923                    3                    8                18690
            Student              Working
                 16               144591
```

```
> table(train$NAME_INCOME_TYPE)

Commercial associate            Pensioner        State servant              Student              Working
               35430                    5                11876                    6                86297
> table(test$NAME_INCOME_TYPE)

Commercial associate            Pensioner        State servant              Student              Working
               11723                    2                 3958                    6                28801
```

Some of the levels in this variable 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 snippet, we can see this happening for businessmen, 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 affect the accuracy of our model.

```
### REMOVING UNNECESSARY DATA
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 cause 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.

```
### REMOVING THE MISSING DATA FROM TRAIN DATA
train <- train[complete.cases(train),]
missmap(train, col = c("red", "blue"), legend = FALSE)

### CHECK MISSING DATA ON TEST DATA
missmap(test, col = c("red", "blue"), legend = FALSE)

### REMOVING THE MISSING DATA FROM TEST DATA
test <- test[complete.cases(test),]
missmap(test, col = c("red", "blue"), legend = FALSE)
```

6. Running the Model

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

Regression:
Now that we have made our model, we can start interpreting it. Simple 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 centred 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 the form of some other variables might be showing a pattern 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.

```
> ### 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
-1.4774  -0.4496  -0.3247  -0.2323   3.2615

Coefficients: (2 not defined because of singularities)
                                              Estimate Std. Error z value Pr(>|z|)
(Intercept)                                 -1.124e+01  6.293e+01  -0.179 0.858291
NAME_CONTRACT_TYPERevolving loans           -3.254e-01  4.298e-02  -7.570 3.74e-14 ***
CODE_GENDERM                                 3.039e-01  2.687e-02  11.310  < 2e-16 ***
FLAG_OWN_CARY                               -2.899e-01  2.344e-02 -12.367  < 2e-16 ***
FLAG_OWN_REALTYY                             5.010e-02  2.256e-02   2.221 0.026337 *
CNT_CHILDREN                                 7.899e-03  1.589e-02   0.497 0.619171
AMT_INCOME_TOTAL                            -1.016e-06  1.983e-07  -5.127 2.95e-07 ***
AMT_CREDIT                                   2.225e-06  1.643e-07  13.538  < 2e-16 ***
AMT_ANNUITY                                  1.489e-05  1.479e-06  10.069  < 2e-16 ***
AMT_GOODS_PRICE                             -2.817e-06  1.884e-07 -14.955  < 2e-16 ***
NAME_INCOME_TYPEPensioner                   -1.075e+01  2.311e+02  -0.047 0.962883
NAME_INCOME_TYPEState servant               -2.978e-03  5.064e-02  -0.059 0.953102
NAME_INCOME_TYPEStudent                     -1.145e+01  1.978e+02  -0.058 0.953829
NAME_INCOME_TYPEWorking                      1.065e-01  2.455e-02   4.338 1.44e-05 ***
NAME_EDUCATION_TYPEHigher education          1.098e+01  6.293e+01   0.175 0.861420
NAME_EDUCATION_TYPEIncomplete higher         1.109e+01  6.293e+01   0.176 0.860141
NAME_EDUCATION_TYPELower secondary           1.140e+01  6.293e+01   0.181 0.856200
NAME_EDUCATION_TYPESecondary / secondary special  1.126e+01  6.293e+01   0.179 0.857964
NAME_FAMILY_STATUSMarried                   -1.263e-01  3.289e-02  -3.840 0.000123 ***
NAME_FAMILY_STATUSSeparated                 -3.758e-02  4.989e-02  -0.753 0.451286
NAME_FAMILY_STATUSSingle / not married      -8.314e-02  3.904e-02  -2.129 0.033216 *
NAME_FAMILY_STATUSWidow                     -2.502e-01  7.792e-02  -3.211 0.001321 **
NAME_HOUSING_TYPEHouse / apartment          -3.813e-02  1.560e-01  -0.244 0.806883
NAME_HOUSING_TYPEMunicipal apartment         4.724e-02  1.647e-01   0.287 0.774247
```

(Refer to the code in R file for the complete model)

```
ORGANIZATION_TYPETransport: type 3            6.815e-01  3.054e-01   2.231 0.025662 *
ORGANIZATION_TYPETransport: type 4            8.507e-02  2.890e-01   0.294 0.768450
ORGANIZATION_TYPEUniversity                   3.063e-02  3.266e-01   0.094 0.925266
EXT_SOURCE_2                                 -2.104e+00  5.113e-02 -41.140  < 2e-16 ***
EXT_SOURCE_3                                 -2.861e+00  5.267e-02 -54.320  < 2e-16 ***
DAYS_LAST_PHONE_CHANGE                       -4.691e-05  1.367e-05  -3.431 0.000601 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 78839  on 133613  degrees of freedom
Residual deviance: 70191  on 133505  degrees of freedom
AIC: 70409

Number of Fisher Scoring iterations: 12
```

6.1 Use the model to check accuracy

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

### ACCURACY VISUALIZATION
t <- table(glm.pred, test$TARGET)
t
```

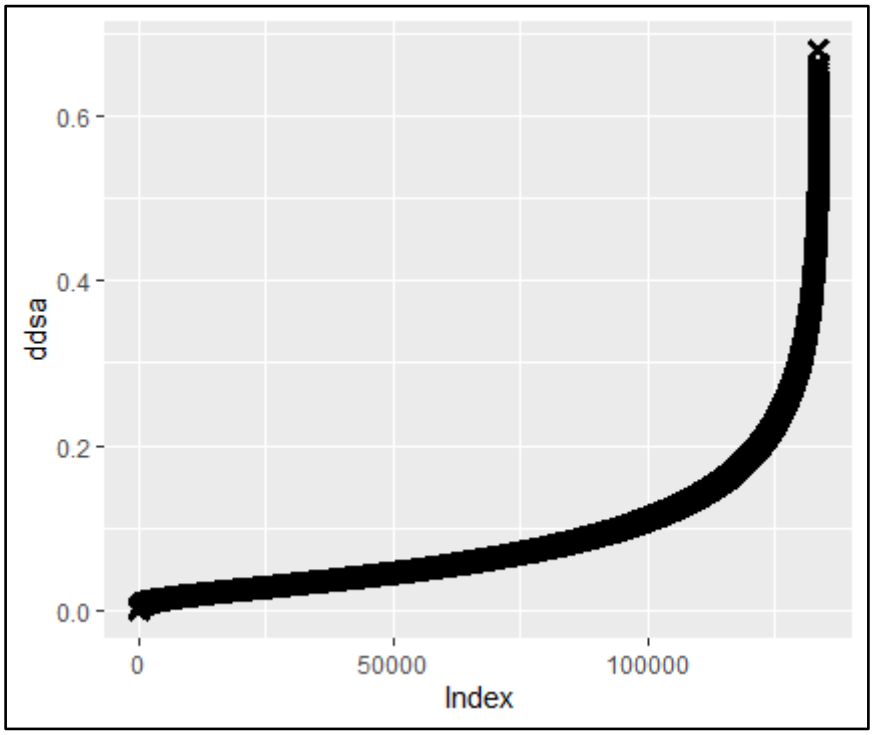
```
glm.pred       0     1
  Not Paid    40    35
  Paid     40630  3785
```

```
> ### ACCURACY INCLUDING TYPE 1 & 2 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] 91.40256
```

6.2 Using model to make a graph

As mentioned earlier, the model helps us get a value between 0 and 1 which helps us decide where to classify the output.

```
### USING MODEL TO PLOT THE 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")
```



7. Making the model better

Our model gives us 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 I errors, our accuracy turns out to be very bad.

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 **down sampling**. This will make sure that there are

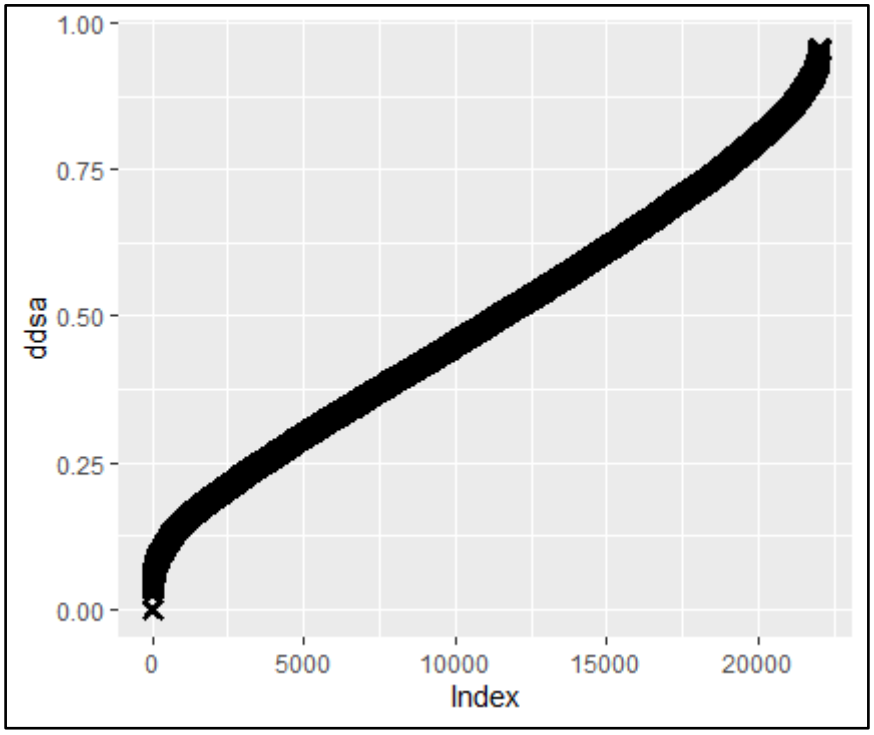
an equal number of cases for both the client paying the loan and not paying it. After down sampling our training data, we are going to use the same method as above to make a new model and then train it to see its results.

```
### DOWNSAMPLING AND MAKING THE MODEL BETTER ###
`%notin%` <- Negate(`%in%`)
options(scipen = 999)

clean_data$TARGET <- as.factor(clean_data$TARGET)
set.seed(100)

#Dividing the data set
trainDataIndex <- createDataPartition(clean_data$TARGET, p = 0.7, list = F)
trainData <- clean_data[trainDataIndex, ]
testData <- clean_data[-trainDataIndex, ]

down_train <- downSample(x = trainData[, colnames(trainData) %notin% "TARGET"], y = trainData$TARGET)

down_train <- down_train[!(down_train$NAME_INCOME_TYPE == "Student"|down_train$NAME_INCOME_TYPE == "Mat
testData <- testData[!(testData$NAME_INCOME_TYPE == "Student"|testData$NAME_INCOME_TYPE == "Maternity l

#Removing missing values
down_train <- down_train[complete.cases(down_train),]
testData <- testData[complete.cases(testData),]

#Building & 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)
```

```
> summary(down_glm.pred)
   Length     Class      Mode
    53445 character character
```

```
> #Finding accuracy for both 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] 68.97745
```

```
#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 that 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. 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. 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 R^2 in linear regression. Similarly, the lesser the number of Fisher scoring iterations our model requires, the better our model is, as the Fisher Scoring iterations tell us how quickly our gl() 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 down sampled model

When refining a model, the objective is to minimise the deviance of the model. Since we have a lot of independent predictors, we now run a function which will help us identify for us which predictors are actually not helping and are thus causing a higher deviance and AIC and consequently need to be removed from the data set.

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

```
Call:
glm(formula = Class ~ NAME_CONTRACT_TYPE + CODE_GENDER + FLAG_OWN_CAR +
    AMT_INCOME_TOTAL + AMT_CREDIT + AMT_ANNUITY + AMT_GOODS_PRICE +
    NAME_INCOME_TYPE + NAME_EDUCATION_TYPE + NAME_FAMILY_STATUS +
    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
-2.5093  -0.9796  -0.4437   0.9949   2.4746

Coefficients:
                                                Estimate    Std. Error z value             Pr(>|z|)
(Intercept)                                -9.0735135838 72.5668811442  -0.125             0.900495
NAME_CONTRACT_TYPERevolving loans          -0.3174455508  0.0584322617  -5.433  0.00000005550447609 ***
CODE_GENDERM                                0.3393371953  0.0388870708   8.726 < 0.0000000000000002 ***
FLAG_OWN_CARY                              -0.2693052129  0.0342594750  -7.861  0.00000000000000382 ***
AMT_INCOME_TOTAL                           -0.0000016685  0.0000002820  -5.917  0.00000000328774047 ***
AMT_CREDIT                                  0.0000024855  0.0000002444  10.169 < 0.0000000000000002 ***
AMT_ANNUITY                                 0.0000158758  0.0000022024   7.208  0.00000000000056623 ***
AMT_GOODS_PRICE                            -0.0000030788  0.0000002736 -11.253 < 0.0000000000000002 ***
NAME_INCOME_TYPEState servant              -0.0539786831  0.0652873461  -0.827             0.408358
NAME_INCOME_TYPEWorking                     0.0907489633  0.0355107920   2.556             0.010603 *
NAME_EDUCATION_TYPEHigher education        11.3785221055 72.5667794859   0.157             0.875402
NAME_EDUCATION_TYPEIncomplete higher       11.5573002528 72.5668152625   0.159             0.873461
NAME_EDUCATION_TYPELower secondary         11.7624931435 72.5669226873   0.162             0.871234
NAME_EDUCATION_TYPESecondary / secondary special 11.7176144914 72.5667790631   0.161             0.871720
NAME_FAMILY_STATUSMarried                  -0.1644470106  0.0494507965  -3.325             0.000883 ***
NAME_FAMILY_STATUSSeparated                 0.0016366452  0.0740874403   0.022             0.982376
NAME_FAMILY_STATUSSingle / not married     -0.1054588836  0.0588397892  -1.792             0.073084 .
NAME_FAMILY_STATUSWidow                    -0.3403132388  0.1088180426  -3.127             0.001764 **
DAYS_EMPLOYED                              -0.0000722038  0.0000076226  -9.472 < 0.0000000000000002 ***
DAYS_REGISTRATION                          -0.0000136508  0.0000048721  -2.802             0.005082 **
DAYS_ID_PUBLISH                            -0.0000245841  0.0000101920  -2.412             0.015861 *
FLAG_EMAIL                                 -0.1212432551  0.0649525448  -1.867             0.061951 .
OCCUPATION_TYPECleaning staff               0.4103441786  0.1390215287   2.952             0.003161 **
OCCUPATION_TYPECooking staff                0.2589148366  0.1272923886   2.034             0.041950 *
OCCUPATION_TYPECore staff                  -0.0423633486  0.0998922775  -0.424             0.671500
OCCUPATION_TYPEDrivers                      0.2999068456  0.1077488520   2.783             0.005379 **
OCCUPATION_TYPEHigh skill tech staff        0.0409088976  0.1160794532   0.352             0.724522
OCCUPATION_TYPEHR staff                     0.0763164631  0.3922566359   0.195             0.845739
OCCUPATION_TYPEIT staff                    -0.2794251845  0.3797139311  -0.736             0.461802
OCCUPATION_TYPELaborers                     0.2098858683  0.0952640717   2.203             0.027581 *
OCCUPATION_TYPELow-skill Laborers           0.5787438374  0.1777924567   3.255             0.001133 **
OCCUPATION_TYPEManagers                     0.0483261103  0.1063733998   0.454             0.649608
OCCUPATION_TYPEMedicine staff              -0.0012908899  0.1217602224  -0.011             0.991541
OCCUPATION_TYPEPrivate service staff        0.0058312018  0.1773867326   0.033             0.973776
OCCUPATION_TYPERealty agents                0.1474241499  0.2832772494   0.520             0.602768
OCCUPATION_TYPESales staff                  0.1876156636  0.0965281798   1.944             0.051939 .
OCCUPATION_TYPESecretaries                  0.1356815166  0.2302829630   0.589             0.555731
OCCUPATION_TYPESecurity staff               0.2163515746  0.1256713320   1.722             0.085148 .
OCCUPATION_TYPEUNKNOWN                      0.1055473492  0.0956902561   1.103             0.270023
OCCUPATION_TYPEWaiters/barmen staff         0.5547839432  0.2273638352   2.440             0.014684 *
EXT_SOURCE_2                               -2.1949221268  0.0786708599 -27.900 < 0.0000000000000002 ***
EXT_SOURCE_3                               -2.9978552698  0.0776174563 -38.623 < 0.0000000000000002 ***
DAYS_LAST_PHONE_CHANGE                     -0.0000340008  0.0000196138  -1.734             0.083004 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 30534  on 22030  degrees of freedom
Residual deviance: 26104  on 21988  degrees of freedom
AIC: 26190

Number of Fisher Scoring iterations: 10
```

```r
#Testing Model
e <- as.data.frame(exp(coef(final_model)))
final_pred <- predict(final_model, newdata = testData, type = "response")
final_glm.pred <- ifelse(final_pred > 0.5, "Not Paid", "Paid")

#Checking Accuracy
final_t <- table(final_glm.pred, testData$TARGET)
final_t
```

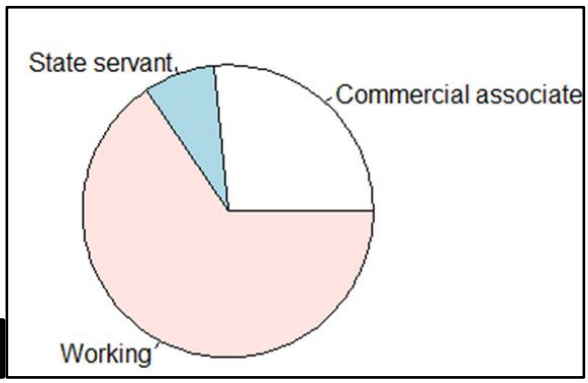
```
> #Testing Model
> e <- as.data.frame(exp(coef(final_model)))
> final_pred <- predict(final_model, newdata = testData, type = "response")
> final_glm.pred <- ifelse(final_pred > 0.5, "Not Paid", "Paid")
>
> #Checking Accuracy
> final_t <- table(final_glm.pred, testData$TARGET)
> final_t

final_glm.pred     0     1
      Not Paid 15173  3084
      Paid     33671  1517
```

Additional visualisations

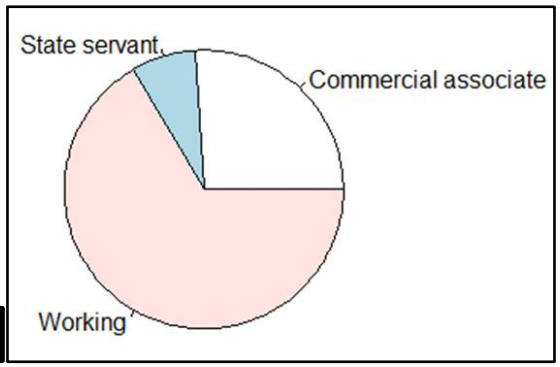
After dividing our dataset into test and train sets, it is imperative for us to ensure that the test and train sets are representative of the actual data set. This means that if 3 variables A, B, and C were present in the ratio 2:3:2 in an actual data set, their representation in the train and test data sets should also be in a similar ratio.

To better represent these figures, we used the pie chart in the ggplot2 library to visualise the representation of our NAME_INCOME_TYPE variable in our original data, the test data, and the train data.

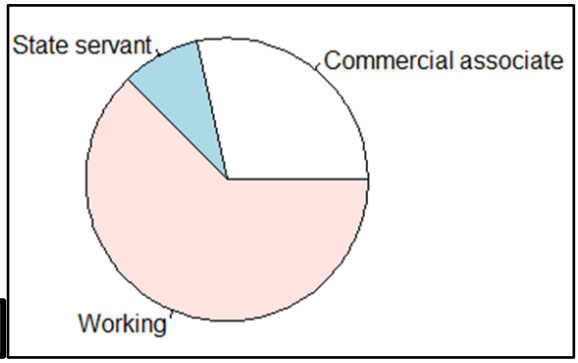
```
pie(table(clean_data$NAME_INCOME_TYPE))
```




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



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



As can be seen in all 3 graphs, the proportions of each of the Income types is similar which makes it safe for us to assume that the test and train data sets are representative of the actual data.

Recommendation/Conclusion:

We were able to test our logistic model, respond positively to our research question, and demonstrate that **there does exist a correlation** between a client defaulting on their obligations and the parameters we selected. To further assess and improve our model, we were able to test it on data. The AIC value dropped thanks to our efforts. Additionally, we observed a difference between the null and residual deviances, demonstrating the model's efficacy. We were also able to reduce the residual deviance as well.

For Pakistan Commercial Banks:

Such a model is the need of time in Pakistan considering the high consumer default rates. Although a few such models exist, there are doubts about how conclusive and rigid their results are. This is the reason most commercial banks are reluctant to provide loans to the masses. Banks are sustaining themselves effectively by lending loans to the Government instead of the actual consumer. Most of the population is not financially inclusive, and hence is not able to take loans.

Furthermore, we recommend that there is a need for more variables which are linked with the mass population rather than only those people who already have bank accounts. What could be possible as such variables could relate to data points concerning how much top ups are done by certain individuals, or the amount of online shopping done. Various other points could also be considered that removes the constraint of restricting credit score to only those people who already have bank accounts.

Lastly, to reduce the default rate of overall Pakistan Commercial Banks, there is a need for a central credit risk score of each customer which is automatically updated to all the lending platforms. All banks should have an interconnected data sharing system, through which they can effectively assess the true credit score of a customer. As we know that recently, there has been a flux of new startups which revolve around lending verticals. If a central data system is created, the success rate of all such startups would sky rocket.

Bibliography:

Partovi, Elmira, and Roman Matousek. “Bank Efficiency and Non-Performing Loans: Evidence from Turkey.” *Research in International Business and Finance*, vol. 48, 2019, pp. 287–309., https://doi.org/10.1016/j.ribaf.2018.12.011.

“Figure 1: Systems Design Schematics from: (A) Son Et Al. (2010); (b) Lau Et Al. (2010); (c) Fujiwara Et Al. (2011); and (d) Gjerlufsen Et Al. (2011).” https://doi.org/10.7717/peerjcs.88/fig-1.

Gorter, C., and Adriaan M. Bloem. “The Treatment of Nonperforming Loans in Macroeconomic Statistics.” *IMF Working Papers*, vol. 01, no. 209, 2001, p. 1., https://doi.org/10.5089/9781451874754.001.

Caprio, Gerard, and Daniela Klingebiel. “Bank Insolvencies: Cross-Country Experience.” *Policy Research Working Papers*, 1999, https://doi.org/10.1596/1813-9450-1620.

Winder, James. “A Survey Article on International Banking.” *Encyclopedia of Finance*, 2012, pp. 607–620., https://doi.org/10.1007/978-1-4614-5360-4_53.

Presented to:
Miss Maheen Syed

Muhammad Asim
Muhammad Huzaifa
Ali Kumayl
Abdur Rehman Shamsi
Faran Maood

LOAN DEFAULTS

AND THEIR EFFECTS ON FINANCIAL PERFORMANCE OF COMMERCIAL BANKS