STAT 847 - Final Project

QUESTION 1- Describe and justify two different topics or approaches you might want to consider for this dataset and task. You don't have to use these tasks in the actual analysis.

The approaches that would be considered for this dataset are:

1. Multiple Imputation:

This dataset contains multiple instances of Null values across different variable types like continuous and categorical variables. This method can be an effective method to handle the missing data by taking the predictions and adding random noise to the data. Creating 3-10 copies of the imputed dataset and combining the results of all the datasets in accordance to Rubin's rule can help identify the uncertainty added by the imputed values.

2. Quantile Regression

For features like total income or the credit score of the customer, the data is widely spread such that there are extreme or outlier values in the data. In general, some of the features are also skewed towards a certain class. Quantile regression has the ability to handle outlier values and skewness to give the conditional distribution of the response variable, providing the quantile distributions. Since many of the observations are having zero values with no normal distribution, quantile regression might be a great approach.

QUESTION 2- Describe and show the code used to clean and collect the data. (Optional)

```
library(stringr)
library(tidyverse)
```

```
## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.2 ──
## ✔ ggplot2 3.3.6      ✔ purrr   0.3.5
## ✔ tibble  3.1.8      ✔ dplyr   1.0.10
## ✔ tidyr   1.2.1      ✔ forcats 0.5.2
## ✔ readr   2.1.3
```

```
## Warning: package 'tibble' was built under R version 4.2.2
```

```
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
```

```
library(lubridate)
```

```
## Warning: package 'lubridate' was built under R version 4.2.2
```

```
## Loading required package: timechange
```

```
## Warning: package 'timechange' was built under R version 4.2.2

##
## Attaching package: 'lubridate'
##
## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union

library(dplyr)
library(ggplot2)
library(zoo)

##
## Attaching package: 'zoo'
##
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric

library(GGally)

## Warning: package 'GGally' was built under R version 4.2.3

## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2

library(rpart)
library(caret)

## Warning: package 'caret' was built under R version 4.2.2

## Loading required package: lattice
##
## Attaching package: 'caret'
##
## The following object is masked from 'package:purrr':
##
##     lift

library(naniar)

## Warning: package 'naniar' was built under R version 4.2.3

library(DescTools)

## Warning: package 'DescTools' was built under R version 4.2.3

##
## Attaching package: 'DescTools'
##
## The following objects are masked from 'package:caret':
```

```
##
##     MAE, RMSE

library(FactoMineR)

## Warning: package 'FactoMineR' was built under R version 4.2.3

#Reading the data
data = read_csv("C:/Users/User/Documents/Winter 2023/STAT 847/Final
Project/Loan Credit Kaggle/application_data.csv")

## Rows: 307511 Columns: 122
## ── Column specification
──────────────────────────────────────────────────────────────────────
## Delimiter: ","
## chr  (16): NAME_CONTRACT_TYPE, CODE_GENDER, FLAG_OWN_CAR, FLAG_OWN_REALTY,
N...
## dbl (106): SK_ID_CURR, TARGET, CNT_CHILDREN, AMT_INCOME_TOTAL, AMT_CREDIT,
A...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this
message.

#Data before cleaning
dim(data)

## [1] 307511    122

summary(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
```

```
#Data Cleaning
data = data[, -c(42:86)]
data = select(data, -c("TOTALAREA_MODE", "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",
                       "OWN_CAR_AGE"))

#Checking NA values in categorical features
colSums(is.na(data[, c("NAME_CONTRACT_TYPE", "FLAG_OWN_CAR",
                       "FLAG_OWN_REALTY",
"NAME_TYPE_SUITE","NAME_INCOME_TYPE",

"NAME_EDUCATION_TYPE","NAME_FAMILY_STATUS","NAME_HOUSING_TYPE",
```

```
"OCCUPATION_TYPE","WEEKDAY_APPR_PROCESS_START","ORGANIZATION_TYPE",
                        "FONDKAPREMONT_MODE", "HOUSETYPE_MODE",
"WALLSMATERIAL_MODE",
                        "EMERGENCYSTATE_MODE" )]))
```

```
##          NAME_CONTRACT_TYPE                FLAG_OWN_CAR
##                           0                           0
##             FLAG_OWN_REALTY             NAME_TYPE_SUITE
##                           0                        1292
##           NAME_INCOME_TYPE         NAME_EDUCATION_TYPE
##                           0                           0
##          NAME_FAMILY_STATUS           NAME_HOUSING_TYPE
##                           0                           0
##             OCCUPATION_TYPE WEEKDAY_APPR_PROCESS_START
##                       96391                           0
##           ORGANIZATION_TYPE          FONDKAPREMONT_MODE
##                           0                      210295
##              HOUSETYPE_MODE          WALLSMATERIAL_MODE
##                      154297                      156341
##         EMERGENCYSTATE_MODE
##                      145755
```

```
#Dropping features
data = select(data, -c("FONDKAPREMONT_MODE",
                        "HOUSETYPE_MODE",
"WALLSMATERIAL_MODE","EMERGENCYSTATE_MODE", "OCCUPATION_TYPE",
                        "NAME_TYPE_SUITE"))

#Checking XNA values
colSums(data == "XNA")
```

```
##                   SK_ID_CURR                       TARGET
##                            0                            0
##           NAME_CONTRACT_TYPE                  CODE_GENDER
##                            0                            4
##                 FLAG_OWN_CAR              FLAG_OWN_REALTY
##                            0                            0
##                 CNT_CHILDREN             AMT_INCOME_TOTAL
##                            0                            0
##                   AMT_CREDIT                 AMT_ANNUITY
##                            0                           NA
##               AMT_GOODS_PRICE             NAME_INCOME_TYPE
##                           NA                            0
##          NAME_EDUCATION_TYPE           NAME_FAMILY_STATUS
##                            0                            0
##            NAME_HOUSING_TYPE  REGION_POPULATION_RELATIVE
##                            0                            0
##                   DAYS_BIRTH                DAYS_EMPLOYED
##                            0                            0
```

```
##           DAYS_REGISTRATION             DAYS_ID_PUBLISH
##                           0                           0
##                 FLAG_MOBIL              FLAG_EMP_PHONE
##                           0                           0
##             FLAG_WORK_PHONE            FLAG_CONT_MOBILE
##                           0                           0
##                  FLAG_PHONE                  FLAG_EMAIL
##                           0                           0
##             CNT_FAM_MEMBERS        REGION_RATING_CLIENT
##                          NA                           0
## REGION_RATING_CLIENT_W_CITY  WEEKDAY_APPR_PROCESS_START
##                           0                           0
##     HOUR_APPR_PROCESS_START  REG_REGION_NOT_LIVE_REGION
##                           0                           0
##  REG_REGION_NOT_WORK_REGION LIVE_REGION_NOT_WORK_REGION
##                           0                           0
##      REG_CITY_NOT_LIVE_CITY      REG_CITY_NOT_WORK_CITY
##                           0                           0
##     LIVE_CITY_NOT_WORK_CITY           ORGANIZATION_TYPE
##                           0                       55374
##    OBS_30_CNT_SOCIAL_CIRCLE    DEF_30_CNT_SOCIAL_CIRCLE
##                          NA                          NA
##    OBS_60_CNT_SOCIAL_CIRCLE    DEF_60_CNT_SOCIAL_CIRCLE
##                          NA                          NA
##      DAYS_LAST_PHONE_CHANGE             FLAG_DOCUMENT_2
##                          NA                           0
##             FLAG_DOCUMENT_3             FLAG_DOCUMENT_4
##                           0                           0
##             FLAG_DOCUMENT_5             FLAG_DOCUMENT_6
##                           0                           0
##             FLAG_DOCUMENT_7             FLAG_DOCUMENT_8
##                           0                           0
##             FLAG_DOCUMENT_9            FLAG_DOCUMENT_10
##                           0                           0
##            FLAG_DOCUMENT_11            FLAG_DOCUMENT_12
##                           0                           0
##            FLAG_DOCUMENT_13            FLAG_DOCUMENT_14
##                           0                           0
##            FLAG_DOCUMENT_15            FLAG_DOCUMENT_16
##                           0                           0
##            FLAG_DOCUMENT_17            FLAG_DOCUMENT_18
##                           0                           0
##            FLAG_DOCUMENT_19            FLAG_DOCUMENT_20
##                           0                           0
##            FLAG_DOCUMENT_21
##                           0
```

```
#Remove XNA values in CODE_GENDER
data = subset(data, CODE_GENDER!="XNA")
```

```r
#Dropping ORGANIZATION_TYPE (55374 XNA values)
data = select(data, -c("ORGANIZATION_TYPE"))

#Replace NA values of Count of family members with 0 (2 Null values)
data$CNT_FAM_MEMBERS = replace(data$CNT_FAM_MEMBERS,
is.na(data$CNT_FAM_MEMBERS),0)

#Replace 12 NA values of AMT_ANNUITY with the mean
data$AMT_ANNUITY = replace(data$AMT_ANNUITY, is.na(data$AMT_ANNUITY),
mean(data$AMT_ANNUITY, na.rm=TRUE))

#Replace 278 NA values of AMT_GOODS_PRICE with the mean
data$AMT_GOODS_PRICE = replace(data$AMT_GOODS_PRICE,
is.na(data$AMT_GOODS_PRICE), mean(data$AMT_GOODS_PRICE, na.rm=TRUE))


#Changing the days of birth in years
data$AGE = as.integer(trunc(abs(data$DAYS_BIRTH/365)))
data$YEARS_EMPLOYED = as.integer(trunc(abs(data$DAYS_EMPLOYED/365)))

#Removing rows that has employment years as 1000
check_data = subset(data, YEARS_EMPLOYED!=1000)


#Removing other rows that contain NA values
new_data = na.omit(check_data)
dim(new_data)
```

```
## [1] 251283     64
```

```r
#Converting it into factors/categorical features
new_data$TARGET = as.factor(new_data$TARGET)
new_data$FLAG_DOCUMENT_3 = as.factor(new_data$FLAG_DOCUMENT_3)
```

The dataset has been collected from Kaggle website(https://www.kaggle.com/datasets/kamleshatara/credit-eda-assignment) named Loan Credit. It contains 307511 rows with 122 features initially. After analysing the dataset, it is observed that considerable data cleaning is required. Features such as "EXT_SOURCE_1", "EXT_SOURCE_2", and "EXT_SOURCE_3" are not relevant to the data analysis and no information is provided on their importance to the loan study. Hence these features are removed. Moreover, features like "BASEMENTAREA_AVG", "ELEVATOR_AVG", etc contain the scores of the basement and elevator of the customer's current residence. Since these features contain around 50% NA values in them and removing those particular rows will significantly reduce the data size, therefore those features are dropped. Other features like AMT_REQ_CREDIT_BUREAU" ,"OWN_CAR_AGE" are also removed due to their less importance and the presence of significant NA values.

Now we check the NA values in categorical variables. Features like "FONDKAPREMOUNT_MODE", "HOUSETYPE_MODE", "WALLSMATERIAL_MODE", "EMERGENCYSTATE_MODE" also contain more than 50% NA values in them. Also, their description is not present in the metafile, hence these features are dropped. "OCCUPATION_TYPE" showcasing the occupation of the customer contains 96391 NA values. Since we can't replace the NA values with the most frequent value in this feature due to bias, this column is also dropped. It is also observed that there are XNA values in the categorical features which is another indication of NA values. There are only 4 XNA values in "CODE_GENDER" which showcases the gender of the customer, so those particular rows are dropped, but "ORGANIZATION_TYPE" contains 55374 XNA values hence this feature is not considered and dropped. Some continuous features like "AMT_ANNUITY", "AMY_GOOD_PRICE" contain less than 1% of NA values, those values are replaced by their mean value via mean imputation.

"DAYS_BIRTH" and "DAY_EMPLOYED" contain age and employment status of the customer in days. It is transformed into years for proper analysis and presentation of data. One interesting observation for DAY_EMPLOYED features was that after transformation, there were values that showed the customers have a 1000-year employment status. Since this is impossible in nature, those rows are removed. Other NA values are removed, and the data contains 251283 rows with 64 features after cleaning.

QUESTION 3- Give a ggpairs plot of what you think are the six most important variables. At least one must be categorical, and one continuous. Explain your choice of variables and the trends between them.

From the ggpairs plot the 6 most important variables chosen are: 1. NAME_CONTRACT_TYPE - Type of Loan 2. AMT_INCOME_TOTAL - Total Income of the customer 3. AMT_ANNUITY - Loan Annuity value 4. AMT_GOOD_PRICE - Price of the goods 5. FLAG_DOCUMENT_3 - If the customer provided document 3 or not 6. AMT_CREDIT - Credit Score of the Customer

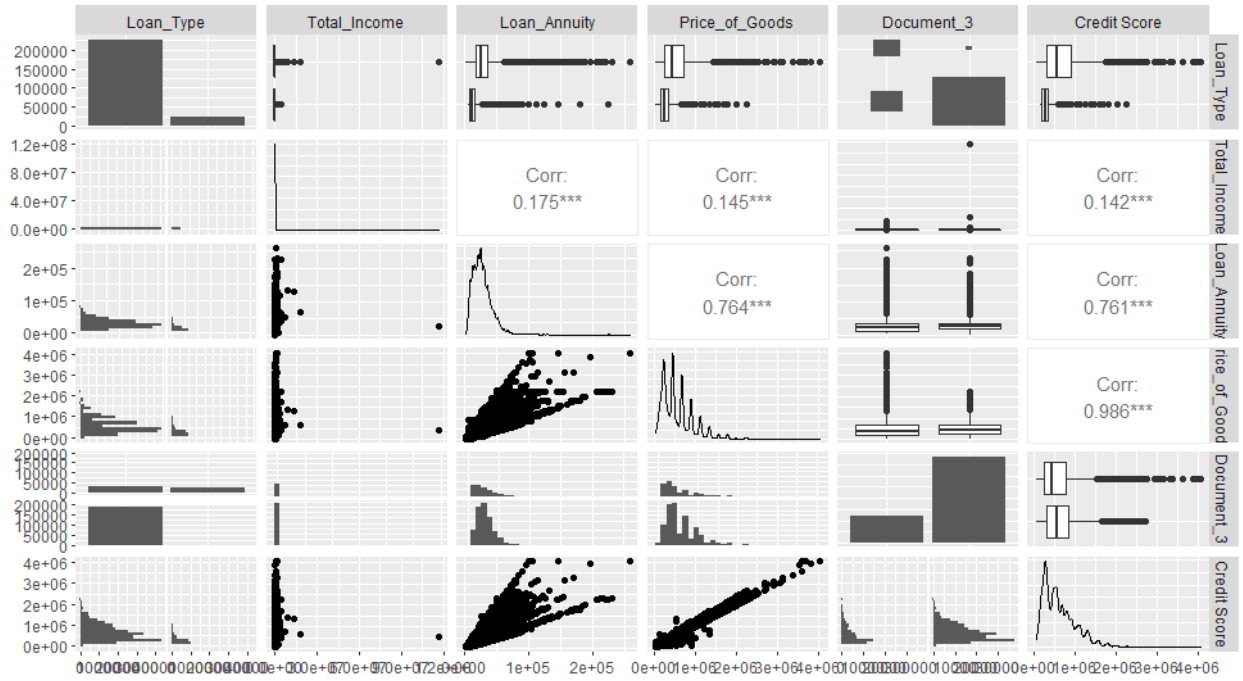
Variables like Credit Score, Price _of_Goods, and Loan_Annuity have a positive correlation, with those features being the most significant(***) among the other 64 features. Document_3 also has a good distribution with the Credit_Score, Laon_Annuity and the Price_of_Goods variables and the Loan_Type variable is included to predict the type of loan in accordance with the other variables.

```
#Creating the ggpairs plot
vars = c("NAME_CONTRACT_TYPE", "AMT_INCOME_TOTAL","AMT_ANNUITY",
"AMT_GOODS_PRICE", "FLAG_DOCUMENT_3", "AMT_CREDIT")

ggpairs(new_data, columns = vars, columnLabels =
c("Loan_Type","Total_Income", "Loan_Annuity",
                                                    "Price_of_Goods",
"Document_3", "Credit Score"))
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

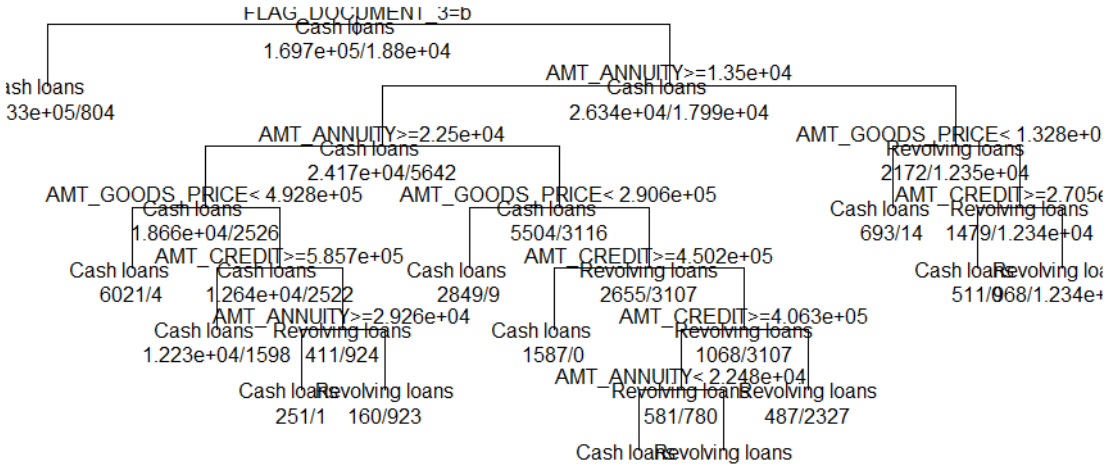


QUESTION 4- Build a classification tree of one of the six variables from the last part as a function of the other five, and any other explanatory variables you think are necessary. Show code, explain reasoning, and show the tree as a simple (ugly) plot. Show the confusion matrix. Give two example predictions and follow them down the tree

```
#Splitting the data into training and testing data
split_data = createDataPartition(y = new_data$NAME_CONTRACT_TYPE, p=0.75,
list=FALSE)
train = new_data[split_data,]
test = new_data[-split_data,]

#Creating the classification tree
fit = rpart(NAME_CONTRACT_TYPE ~ AMT_INCOME_TOTAL+AMT_GOODS_PRICE+
              AMT_ANNUITY+FLAG_DOCUMENT_3+AMT_CREDIT,
            method = "class",data=train)

#Plotting the tree and checking the summary of the tree
plot(fit, uniform=TRUE)
text(fit, use.n=TRUE, all=TRUE, cex=1)
```




```
printcp(fit)

##
## Classification tree:
## rpart(formula = NAME_CONTRACT_TYPE ~ AMT_INCOME_TOTAL + AMT_GOODS_PRICE +
##     AMT_ANNUITY + FLAG_DOCUMENT_3 + AMT_CREDIT, data = train,
##     method = "class")
##
## Variables actually used in tree construction:
## [1] AMT_ANNUITY     AMT_CREDIT      AMT_GOODS_PRICE FLAG_DOCUMENT_3
##
## Root node error: 18796/188463 = 0.099733
##
## n= 188463
##
##         CP nsplit rel error  xerror      xstd
## 1 0.270962      0   1.00000 1.00000 0.0069207
## 2 0.035859      2   0.45808 0.45808 0.0048226
## 3 0.035699      3   0.42222 0.39663 0.0045019
## 4 0.025910      6   0.31512 0.31512 0.0040297
## 5 0.014205      7   0.28921 0.28921 0.0038656
## 6 0.013647      9   0.26080 0.25963 0.0036682
## 7 0.010000     12   0.21967 0.21973 0.0033814

summary(fit)

## Call:
## rpart(formula = NAME_CONTRACT_TYPE ~ AMT_INCOME_TOTAL + AMT_GOODS_PRICE +
##     AMT_ANNUITY + FLAG_DOCUMENT_3 + AMT_CREDIT, data = train,
```

```
##     method = "class")
##   n= 188463
##
##           CP nsplit rel error    xerror        xstd
## 1 0.27096191      0 1.0000000 1.0000000 0.006920746
## 2 0.03585869      2 0.4580762 0.4580762 0.004822608
## 3 0.03569908      3 0.4222175 0.3966269 0.004501883
## 4 0.02590977      6 0.3151202 0.3151202 0.004029688
## 5 0.01420515      7 0.2892105 0.2892105 0.003865617
## 6 0.01364652      9 0.2608002 0.2596297 0.003668155
## 7 0.01000000     12 0.2196744 0.2197276 0.003381412
##
## Variable importance
##       AMT_CREDIT      AMT_ANNUITY  FLAG_DOCUMENT_3  AMT_GOODS_PRICE
##               26               24               23               22
## AMT_INCOME_TOTAL
##                4
##
## Node number 1: 188463 observations,    complexity param=0.2709619
##   predicted class=Cash loans       expected loss=0.0997331  P(node) =1
##     class counts: 169667 18796
##    probabilities: 0.900 0.100
##   left son=2 (144175 obs) right son=3 (44288 obs)
##   Primary splits:
##       FLAG_DOCUMENT_3  splits as  RL,           improve=10865.4800, (0
missing)
##       AMT_ANNUITY      < 13502.25 to the right, improve= 7657.6940, (0
missing)
##       AMT_CREDIT       < 270022.5 to the right, improve= 4092.1020, (0
missing)
##       AMT_GOODS_PRICE  < 405112.5 to the right, improve= 2497.5300, (0
missing)
##       AMT_INCOME_TOTAL < 112701.8 to the right, improve=  170.2675, (0
missing)
##   Surrogate splits:
##       AMT_ANNUITY      < 9002.25  to the right, agree=0.774, adj=0.038, (0
split)
##       AMT_GOODS_PRICE  < 2252250  to the left,  agree=0.765, adj=0.002, (0
split)
##       AMT_INCOME_TOTAL < 593624.2 to the left,  agree=0.765, adj=0.001, (0
split)
##       AMT_CREDIT       < 2697750  to the left,  agree=0.765, adj=0.001, (0
split)
##
## Node number 2: 144175 observations
##   predicted class=Cash loans       expected loss=0.005632044  P(node)
=0.7650043
##     class counts: 143363   812
##    probabilities: 0.994 0.006
##
```

```
## Node number 3: 44288 observations,    complexity param=0.2709619
##   predicted class=Cash loans       expected loss=0.4060694  P(node)
=0.2349957
##     class counts: 26304 17984
##    probabilities: 0.594 0.406
##   left son=6 (29824 obs) right son=7 (14464 obs)
##   Primary splits:
##       AMT_ANNUITY      < 13502.25 to the right, improve=8546.698, (0
missing)
##       AMT_CREDIT       < 405798.8 to the right, improve=5834.667, (0
missing)
##       AMT_GOODS_PRICE  < 405112.5 to the right, improve=5002.111, (0
missing)
##       AMT_INCOME_TOTAL < 135186.8 to the right, improve=1439.613, (0
missing)
##   Surrogate splits:
##       AMT_CREDIT       < 270022.5 to the right, agree=0.926, adj=0.773, (0
split)
##       AMT_GOODS_PRICE  < 272250   to the right, agree=0.897, adj=0.684, (0
split)
##       AMT_INCOME_TOTAL < 112529.2 to the right, agree=0.743, adj=0.212, (0
split)
##
## Node number 6: 29824 observations,    complexity param=0.03569908
##   predicted class=Cash loans       expected loss=0.1897465  P(node)
=0.1582486
##     class counts: 24165  5659
##    probabilities: 0.810 0.190
##   left son=12 (21270 obs) right son=13 (8554 obs)
##   Primary splits:
##       AMT_ANNUITY      < 22509    to the right, improve=684.44690, (0
missing)
##       AMT_CREDIT       < 675191.2 to the right, improve=431.44310, (0
missing)
##       AMT_GOODS_PRICE  < 290250   to the left,  improve=354.95660, (0
missing)
##       AMT_INCOME_TOTAL < 447750   to the left,  improve= 39.10069, (0
missing)
##   Surrogate splits:
##       AMT_CREDIT       < 450036   to the right, agree=0.846, adj=0.463, (0
split)
##       AMT_GOODS_PRICE  < 407362.5 to the right, agree=0.824, adj=0.385, (0
split)
##       AMT_INCOME_TOTAL < 114606   to the right, agree=0.727, adj=0.049, (0
split)
##
## Node number 7: 14464 observations,    complexity param=0.03585869
##   predicted class=Revolving loans  expected loss=0.1478844  P(node)
=0.07674716
##     class counts:  2139 12325
```

```
##    probabilities: 0.148 0.852
##   left son=14 (700 obs) right son=15 (13764 obs)
##   Primary splits:
##       AMT_GOODS_PRICE  < 132750   to the left,  improve=1022.184000, (0
missing)
##       AMT_CREDIT       < 134887.5 to the left,  improve= 943.858700, (0
missing)
##       AMT_ANNUITY      < 6747.75  to the left,  improve= 330.410200, (0
missing)
##       AMT_INCOME_TOTAL < 136012.5 to the right, improve=   9.312806, (0
missing)
##   Surrogate splits:
##       AMT_CREDIT  < 134887.5 to the left,  agree=0.995, adj=0.889, (0
split)
##       AMT_ANNUITY < 6684.75  to the left,  agree=0.959, adj=0.159, (0
split)
##
## Node number 12: 21270 observations,    complexity param=0.01364652
##   predicted class=Cash loans       expected loss=0.1218148  P(node)
=0.1128603
##     class counts: 18679  2591
##    probabilities: 0.878 0.122
##   left son=24 (6064 obs) right son=25 (15206 obs)
##   Primary splits:
##       AMT_GOODS_PRICE  < 492750   to the left,  improve=248.33750, (0
missing)
##       AMT_CREDIT       < 494775   to the left,  improve=133.75860, (0
missing)
##       AMT_INCOME_TOTAL < 179635.5 to the left,  improve=111.89100, (0
missing)
##       AMT_ANNUITY      < 24747.75 to the left,  improve= 51.65921, (0
missing)
##   Surrogate splits:
##       AMT_CREDIT  < 577662.8 to the left,  agree=0.932, adj=0.760, (0
split)
##       AMT_ANNUITY < 26896.5  to the left,  agree=0.750, adj=0.123, (0
split)
##
## Node number 13: 8554 observations,    complexity param=0.03569908
##   predicted class=Cash loans       expected loss=0.3586626  P(node)
=0.04538822
##     class counts:  5486  3068
##    probabilities: 0.641 0.359
##   left son=26 (2843 obs) right son=27 (5711 obs)
##   Primary splits:
##       AMT_GOODS_PRICE  < 290250   to the left,  improve=1078.43700, (0
missing)
##       AMT_CREDIT       < 292207.5 to the left,  improve= 729.24590, (0
missing)
##       AMT_ANNUITY      < 22497.75 to the left,  improve= 623.68430, (0
```

```
missing)
##       AMT_INCOME_TOTAL < 94263.75 to the left,  improve=  75.32765, (0
missing)
##   Surrogate splits:
##       AMT_CREDIT       < 314550   to the left,  agree=0.938, adj=0.813, (0
split)
##       AMT_ANNUITY      < 15747.75 to the left,  agree=0.718, adj=0.151, (0
split)
##       AMT_INCOME_TOTAL < 59271.75 to the left,  agree=0.669, adj=0.005, (0
split)
##
## Node number 14: 700 observations
##   predicted class=Cash loans       expected loss=0.01857143  P(node)
=0.003714257
##     class counts:   687    13
##    probabilities: 0.981 0.019
##
## Node number 15: 13764 observations,    complexity param=0.02590977
##   predicted class=Revolving loans  expected loss=0.1054926  P(node)
=0.0730329
##     class counts:  1452 12312
##    probabilities: 0.105 0.895
##   left son=30 (487 obs) right son=31 (13277 obs)
##   Primary splits:
##       AMT_CREDIT       < 270533.2 to the right, improve=807.925900, (0
missing)
##       AMT_GOODS_PRICE  < 272250   to the right, improve=205.852500, (0
missing)
##       AMT_ANNUITY      < 13497.75 to the left,  improve=107.813700, (0
missing)
##       AMT_INCOME_TOTAL < 136012.5 to the right, improve=  6.525694, (0
missing)
##   Surrogate splits:
##       AMT_GOODS_PRICE < 272250   to the right, agree=0.97, adj=0.162, (0
split)
##
## Node number 24: 6064 observations
##   predicted class=Cash loans       expected loss=0.0008245383  P(node)
=0.03217608
##     class counts:  6059     5
##    probabilities: 0.999 0.001
##
## Node number 25: 15206 observations,    complexity param=0.01364652
##   predicted class=Cash loans       expected loss=0.1700644  P(node)
=0.08068427
##     class counts: 12620  2586
##    probabilities: 0.830 0.170
##   left son=50 (13849 obs) right son=51 (1357 obs)
##   Primary splits:
##       AMT_CREDIT       < 585666   to the right, improve=802.5398, (0
```

```
missing)
##       AMT_GOODS_PRICE  < 587250   to the right, improve=453.9985, (0
missing)
##       AMT_ANNUITY      < 33754.5  to the right, improve=159.1287, (0
missing)
##       AMT_INCOME_TOTAL < 179635.5 to the left,  improve=101.2714, (0
missing)
##   Surrogate splits:
##       AMT_GOODS_PRICE < 587250   to the right, agree=0.963, adj=0.589, (0
split)
##
## Node number 26: 2843 observations
##   predicted class=Cash loans       expected loss=0.002813929  P(node)
=0.01508519
##     class counts:  2835     8
##    probabilities: 0.997 0.003
##
## Node number 27: 5711 observations,    complexity param=0.03569908
##   predicted class=Revolving loans  expected loss=0.4641919  P(node)
=0.03030303
##     class counts:  2651  3060
##    probabilities: 0.464 0.536
##   left son=54 (1604 obs) right son=55 (4107 obs)
##   Primary splits:
##       AMT_CREDIT       < 450173.2 to the right, improve=1280.67900, (0
missing)
##       AMT_GOODS_PRICE  < 452250   to the right, improve= 880.87630, (0
missing)
##       AMT_ANNUITY      < 22497.75 to the left,  improve= 341.26240, (0
missing)
##       AMT_INCOME_TOTAL < 98579.25 to the left,  improve=  75.02297, (0
missing)
##   Surrogate splits:
##       AMT_GOODS_PRICE  < 452250   to the right, agree=0.937, adj=0.777, (0
split)
##       AMT_INCOME_TOTAL < 63085.5  to the left,  agree=0.720, adj=0.003, (0
split)
##       AMT_ANNUITY      < 13637.25 to the left,  agree=0.719, adj=0.001, (0
split)
##
## Node number 30: 487 observations
##   predicted class=Cash loans       expected loss=0  P(node) =0.002584062
##     class counts:   487     0
##    probabilities: 1.000 0.000
##
## Node number 31: 13277 observations
##   predicted class=Revolving loans  expected loss=0.07268208  P(node)
=0.07044884
##     class counts:   965 12312
##    probabilities: 0.073 0.927
```

```
##
## Node number 50: 13849 observations
##   predicted class=Cash loans       expected loss=0.1192144  P(node)
=0.07348392
##     class counts: 12198  1651
##    probabilities: 0.881 0.119
##
## Node number 51: 1357 observations,    complexity param=0.01364652
##   predicted class=Revolving loans  expected loss=0.3109801  P(node)
=0.007200352
##     class counts:   422   935
##    probabilities: 0.311 0.689
##   left son=102 (266 obs) right son=103 (1091 obs)
##   Primary splits:
##       AMT_ANNUITY      < 29263.5  to the right, improve=303.94510, (0
missing)
##       AMT_GOODS_PRICE  < 536948.9 to the left,  improve= 72.30940, (0
missing)
##       AMT_INCOME_TOTAL < 159750   to the left,  improve= 45.65358, (0
missing)
##       AMT_CREDIT       < 584552.2 to the left,  improve= 29.31413, (0
missing)
##
## Node number 54: 1604 observations
##   predicted class=Cash loans       expected loss=0  P(node) =0.008510954
##     class counts:  1604     0
##    probabilities: 1.000 0.000
##
## Node number 55: 4107 observations,    complexity param=0.01420515
##   predicted class=Revolving loans  expected loss=0.2549306  P(node)
=0.02179208
##     class counts:  1047  3060
##    probabilities: 0.255 0.745
##   left son=110 (1338 obs) right son=111 (2769 obs)
##   Primary splits:
##       AMT_CREDIT       < 406665   to the right, improve=132.97290, (0
missing)
##       AMT_ANNUITY      < 22484.25 to the left,  improve=108.83070, (0
missing)
##       AMT_GOODS_PRICE  < 335250   to the left,  improve= 61.63652, (0
missing)
##       AMT_INCOME_TOTAL < 94263.75 to the left,  improve= 39.61998, (0
missing)
##   Surrogate splits:
##       AMT_GOODS_PRICE  < 407250   to the right, agree=0.949, adj=0.843, (0
split)
##       AMT_ANNUITY      < 20850.75 to the right, agree=0.923, adj=0.763, (0
split)
##       AMT_INCOME_TOTAL < 362250   to the right, agree=0.682, adj=0.025, (0
split)
```

```
##
## Node number 102: 266 observations
##   predicted class=Cash loans       expected loss=0.0112782  P(node)
=0.001411418
##     class counts:   263     3
##    probabilities: 0.989 0.011
##
## Node number 103: 1091 observations
##   predicted class=Revolving loans  expected loss=0.1457379  P(node)
=0.005788935
##     class counts:   159   932
##    probabilities: 0.146 0.854
##
## Node number 110: 1338 observations,    complexity param=0.01420515
##   predicted class=Revolving loans  expected loss=0.4379671  P(node)
=0.007099537
##     class counts:   586   752
##    probabilities: 0.438 0.562
##   left son=220 (636 obs) right son=221 (702 obs)
##   Primary splits:
##       AMT_ANNUITY      < 22484.25 to the left,  improve=562.88460, (0
missing)
##       AMT_GOODS_PRICE  < 425250   to the left,  improve=135.62880, (0
missing)
##       AMT_CREDIT       < 449028   to the left,  improve= 84.86718, (0
missing)
##       AMT_INCOME_TOTAL < 143212.5 to the left,  improve= 70.81415, (0
missing)
##   Surrogate splits:
##       AMT_GOODS_PRICE  < 447750   to the left,  agree=0.716, adj=0.403, (0
split)
##       AMT_CREDIT       < 449028   to the left,  agree=0.716, adj=0.403, (0
split)
##       AMT_INCOME_TOTAL < 143212.5 to the left,  agree=0.649, adj=0.261, (0
split)
##
## Node number 111: 2769 observations
##   predicted class=Revolving loans  expected loss=0.1664861  P(node)
=0.01469254
##     class counts:   461  2308
##    probabilities: 0.166 0.834
##
## Node number 220: 636 observations
##   predicted class=Cash loans       expected loss=0.08018868  P(node)
=0.003374668
##     class counts:   585    51
##    probabilities: 0.920 0.080
##
## Node number 221: 702 observations
##   predicted class=Revolving loans  expected loss=0.001424501  P(node)
```

```
=0.003724869
##     class counts:     1   701
##    probabilities: 0.001 0.999

#Testing the classification tree on the testing set
predictions = predict(fit, newdata = test, type = "class")

#Creating the confustion Matrix
confusion_matrix = table(predictions, test$NAME_CONTRACT_TYPE)
print(confusion_matrix)

##
## predictions       Cash loans Revolving loans
##   Cash loans           56030             820
##   Revolving loans        525            5445

#First example
pred1 = predict(fit, data.frame(AMT_INCOME_TOTAL = 67500, AMT_GOODS_PRICE =
513000,
                                AMT_ANNUITY = 29000 , FLAG_DOCUMENT_3 = "1",
AMT_CREDIT=12000),
                type="class")
print(pred1)

##          1
## Cash loans
## Levels: Cash loans Revolving loans
```

A classification tree is built with NAME_CONTRACT_TYPE as a function of the other 5 variables. The data is split into training and testing sets and the classification tree is built. When analysing the tree, it is observed that the variables used for tree construction were "AMY_ANNUITY"," AMT_CREDIT"," AMT_GOODS_PRICE", AND "FLAG_DOCUMENT_3, with the other variables excluded. With further analysis, it is observed that as we move down the tree, the more complex the model becomes. At the end of all splits, the relative error is 0.221 which means the model is able to explain 78% of the total variations in the data. Moreover, the AMT_CREDIT, AMT_ANNUITY, FLAG_DOCUMENT_3, and AMT_GOODS_PRICE features have almost equal importance to the model with each contributing 26%, 24%, 24% and 22% respectively to the model.

When the model is tested on the tested data, a confusion matrix is created which showcased that the model is able to predict 51007 data correctly for the Cash Loans type with 5627 data identified incorrectly as Revolving Loans. For Revolving Loans, the model is only able to predict 638 values correctly with 5548 values falsely predicted due to the imbalance in the data. Moving forward 2 example prediction data are used to predict the type of loan.

QUESTION 6- Build another model using one of the continuous variables from your six most important. This time use your model selection and dimension reduction tools, and include at least one non-linear term.

```
#Selecting the data
data_6 = select(new_data, c("NAME_CONTRACT_TYPE",
"AMT_INCOME_TOTAL","AMT_ANNUITY","AMT_GOODS_PRICE", "FLAG_DOCUMENT_3",
"AMT_CREDIT"))
dim(data_6)

## [1] 251283      6

#Implementing Factor Analysis of Mixed Data
res = FAMD(data_6, graph=FALSE)
summary(res)

##
## Call:
## FAMD(base = data_6, graph = FALSE)
##
##
## Eigenvalues
##                        Dim.1  Dim.2  Dim.3  Dim.4  Dim.5
## Variance               2.844  1.476  0.951  0.422  0.295
## % of var.             47.401 24.592 15.844  7.026  4.918
## Cumulative % of var. 47.401 71.993 87.837 94.863 99.781
##
## Individuals (the 10 first)
##                         Dist    Dim.1    ctr   cos2    Dim.2    ctr   cos2
## 1                   |  1.006 | -0.518  0.000  0.265 | -0.774  0.000  0.591
|
## 2                   |  2.462 |  2.300  0.001  0.873 |  0.034  0.000  0.000
|
## 3                   |  4.141 | -3.098  0.001  0.560 |  2.591  0.002  0.391
|
## 4                   |  1.208 | -0.587  0.000  0.236 | -0.844  0.000  0.488
|
## 5                   |  1.911 | -0.597  0.000  0.098 |  0.936  0.000  0.240
|
## 6                   |  0.812 | -0.201  0.000  0.061 | -0.748  0.000  0.847
|
## 7                   |  3.847 |  2.888  0.001  0.564 |  1.898  0.001  0.244
|
## 8                   |  3.724 |  3.500  0.002  0.883 |  0.416  0.000  0.012
|
## 9                   |  3.604 | -1.806  0.000  0.251 |  2.959  0.002  0.674
|
## 10                  |  0.877 |  0.102  0.000  0.013 | -0.634  0.000  0.522
|
```

```
##                      Dim.3    ctr   cos2
## 1                    0.332  0.000  0.109 |
## 2                   -0.009  0.000  0.000 |
## 3                   -0.513  0.000  0.015 |
## 4                    0.114  0.000  0.009 |
## 5                   -0.282  0.000  0.022 |
## 6                   -0.128  0.000  0.025 |
## 7                   -0.803  0.000  0.044 |
## 8                    0.091  0.000  0.001 |
## 9                   -0.482  0.000  0.018 |
## 10                  -0.183  0.000  0.044 |
##
## Continuous variables
##                        Dim.1    ctr   cos2    Dim.2    ctr   cos2    Dim.3
## AMT_INCOME_TOTAL   |  0.228  1.824  0.052 |  0.221  3.304  0.049 |  0.948
## AMT_ANNUITY        |  0.876 26.956  0.767 |  0.129  1.127  0.017 | -0.040
## AMT_GOODS_PRICE    |  0.942 31.217  0.888 |  0.212  3.054  0.045 | -0.131
## AMT_CREDIT         |  0.949 31.644  0.900 |  0.176  2.103  0.031 | -0.127
##                        ctr   cos2
## AMT_INCOME_TOTAL   94.461  0.898 |
## AMT_ANNUITY         0.169  0.002 |
## AMT_GOODS_PRICE     1.799  0.017 |
## AMT_CREDIT          1.693  0.016 |
##
## Categories
##                          Dim.1      ctr     cos2   v.test      Dim.2
ctr
## Cash loans         |    0.240    0.643    0.349  214.683 |   -0.311
4.003
## Revolving loans    |   -2.170    5.806    0.349 -214.683 |    2.808
36.130
## 0                  |   -0.710    1.462    0.115 -116.838 |    1.890
38.483
## 1                  |    0.218    0.448    0.115  116.838 |   -0.579
11.796
##                       cos2   v.test      Dim.3      ctr     cos2   v.test
## Cash loans            0.585 -385.750 |    0.031    0.097    0.006   48.118
|
## Revolving loans       0.585  385.750 |   -0.281    0.873    0.006  -48.118
|
## 0                     0.818  431.767 |   -0.164    0.696    0.006  -46.622
|
## 1                     0.818 -431.767 |    0.050    0.213    0.006   46.622
|
```

```r
#Creating the model based on the importance of the variables obtained from
FAMD
model = glm(AMT_INCOME_TOTAL ~ I(log(AMT_GOODS_PRICE)) + I(AMT_ANNUITY^3)+
AMT_CREDIT + NAME_CONTRACT_TYPE, data = data_6)
summary(model)
```

```
##
## Call:
## glm(formula = AMT_INCOME_TOTAL ~ I(log(AMT_GOODS_PRICE)) +
I(AMT_ANNUITY^3) +
##     AMT_CREDIT + NAME_CONTRACT_TYPE, data = data_6)
##
## Deviance Residuals:
##       Min         1Q     Median         3Q        Max
##  -1646678     -53946     -16494      31963  116830448
##
## Coefficients:
##                                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      -1.474e+05  2.153e+04  -6.848 7.48e-12
***
## I(log(AMT_GOODS_PRICE))           2.233e+04  1.792e+03  12.462  < 2e-16
***
## I(AMT_ANNUITY^3)                  1.202e-10  3.245e-12  37.046  < 2e-16
***
## AMT_CREDIT                        4.257e-02  3.201e-03  13.296  < 2e-16
***
## NAME_CONTRACT_TYPERevolving loans  1.867e+04  1.755e+03  10.641  < 2e-16
***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 65328121071)
##
##     Null deviance: 1.6863e+16  on 251282  degrees of freedom
## Residual deviance: 1.6416e+16  on 251278  degrees of freedom
## AIC: 6970740
##
## Number of Fisher Scoring iterations: 2
```

Due to the presence of both continuous and categorical variables, Factor Analysis of Mixed Data is implemented as the dimension reduction method on the 6 variables. It is observed that dimension 2 is able to explain 72 % of the variance with further dimension signalling noise, therefore the variable contribution for dimension 2 is considered, with AMT_INCOME_TOTAL, AMT_ANNUITY, AMT_GOOD_PRICE, AMT_CREDIT contributing 3.304, 1.127, 3.054, 2.103 values respectively.

A linear model of AMT_INCOME_TOTAL as a function of the log of AMT_GOODS_PRICE + cube of AMT_ANNUITY + AMT_CREDIT + NAME_CONTRACT_TYPE is created with all the variables being significant for the model.

QUESTION 8- Discuss briefly any ethical concerns like residual disclosure that might arise from the use of your data set, possibly in combination with some additional data outside your dataset. (Option)

Residual disclosure is a phenomenon where a combination of anonymized pieces of info about a person or company identify the person and reveals information that wasn't intended to be open. This dataset has the potential to reveal the identity of the person. It contains information like income, credit score and the organization_type and occupation type of the customers. Moreover, it also contains personal information like the number of children and the status of the customer in the family. Knowing these data can pinpoint particular customers with ease, in addition to revealing other aspects of the customer's information present in the dataset like the phone number, address and age.