EDA - Practice Project

Joseph Pushnam

2024-09-29

1 **Introduction: Statement of the Business and Analytic Problems**

The primary objective of this project is to predict loan default risk using data from loan applicants, helping the business to make informed decisions and mitigate financial risk associated with loan approvals. The dataset includes various demographic, financial, and loan-specific features, and the target variable (`TARGET`) indicates whether an applicant has defaulted on their loan (`1` for default, `0` for no default).

To understand the data and its potential for predictive modeling, the following key questions are addressed in the Exploratory Data Analysis (EDA):

1.1 **Questions Addressed in the EDA:**

Is the target variable (loan default) imbalanced, and what would be the accuracy of a simple majority class classifier?

```
-   This question explores whether the target variable is imbalanced, which is important for deciding whether any balancing
techniques (e.g., oversampling or undersampling) will be needed. We also calculate the accuracy of a simple model that alway
s predicts the majority class, serving as a baseline for future modeling efforts.
```

What is the relationship between the target and the predictors? Are there potentially strong predictors that could be included in a model?

```
-   We investigate correlations and other relationships between the target and key predictors, looking for features that may
strongly influence loan default. Identifying strong predictors at this stage helps in feature selection and engineering for
future modeling.
```

How can the `skimr` and `janitor` packages in R assist with data exploration and cleaning?

```
-   The `skimr` package provides useful data summaries, including information on data types, missing values, and basic stati
stics. The `janitor` package helps streamline data cleaning by removing empty rows/columns and renaming columns. These tools
are used throughout the EDA to simplify and enhance data exploration and preparation.
```

What is the scope of missing data, and what are the possible solutions? Should we remove rows, remove columns, or impute missing values?

```
-   Missing data can impact model accuracy if not handled appropriately. We explore the extent of missing data across the da
taset and consider different strategies to address it, such as removing rows or columns, or using imputation methods like me
an/median imputation for numeric variables or mode imputation for categorical variables.
```

2 **Brief Exploratory Data Analysis**

2.1 **Checking for Class Imbalance:**

```r
# Set CRAN mirror
options(repos = c(CRAN = "https://cran.rstudio.com/"))

# Load libraries
library(skimr)
library(janitor)
```

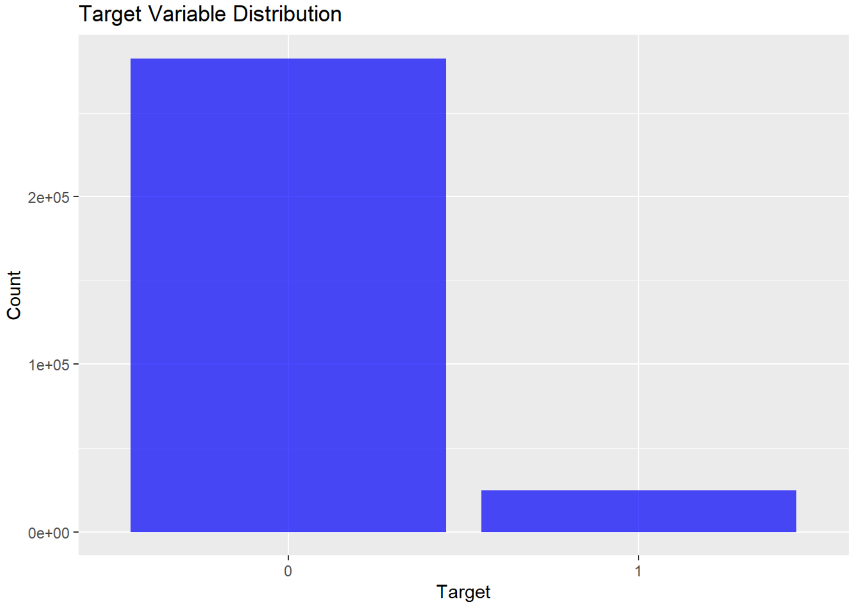
```
##
## Attaching package: 'janitor'
```

```
## The following objects are masked from 'package:stats':
##
##     chisq.test, fisher.test
```

```r
library(ggplot2)

# Set working directory
setwd("C:\\Users\\Joseph\\OneDrive\\Desktop\\UoU\\DB_R\\3. Fall 2024\\Practice Project")

# Read the dataset
train_data <- read.csv("application_train.csv")

# Count of target classes
table(train_data$TARGET)
```

```
##
##      0      1
## 282686  24825
```

```
# Plotting the distribution
ggplot(train_data, aes(x = as.factor(TARGET))) +
  geom_bar(fill = "blue", alpha = 0.7) +
  labs(title = "Target Variable Distribution", x = "Target", y = "Count")
```



```
# Summary of the data
summary(train_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
```

2.2 **Majority Class Classifier Accuracy:**

2.2.1 Correlation with Numeric Variables: You can use cor() for a quick look at correlations between the numeric features and the target.

```
majority_class <- max(table(train_data$TARGET)) / nrow(train_data)
majority_class # This will give the accuracy of a simple majority class model
```

```
## [1] 0.9192712
```

2.3 **Explore the Relationship Between Target and Predictors**

2.3.1 Correlation with Numeric Variables: You can use cor() for a quick look at correlations between the numeric features and the target.

```
numeric_columns <- sapply(train_data, is.numeric)
cor_matrix <- cor(train_data[, numeric_columns], use = "complete.obs")
```

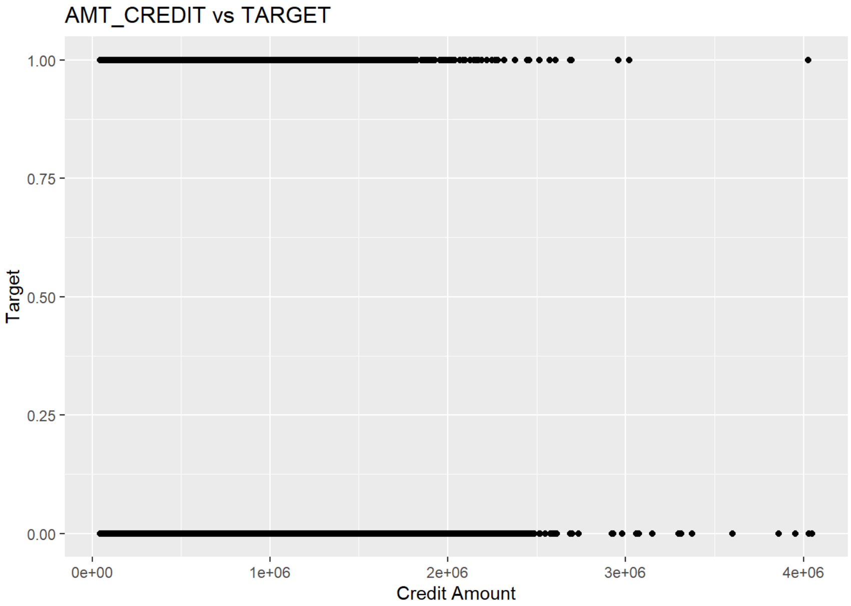
```
## Warning in cor(train_data[, numeric_columns], use = "complete.obs"): the
## standard deviation is zero
```

```
cor_target <- cor_matrix[, "TARGET"]
print(cor_target[order(-abs(cor_target))])  # Sorted by strength of correlation
```

```
##                       TARGET                 EXT_SOURCE_3
##                 1.000000e+00                -1.586054e-01
##                 EXT_SOURCE_2                 EXT_SOURCE_1
##                -1.371169e-01                -1.351815e-01
##  REGION_RATING_CLIENT_W_CITY         REGION_RATING_CLIENT
##                 6.426384e-02                 5.623161e-02
##              FLAG_DOCUMENT_3                FLOORSMAX_AVG
##                 5.239061e-02                -4.934060e-02
##               FLOORSMAX_MODE               FLOORSMAX_MEDI
##                -4.927142e-02                -4.880812e-02
##                   DAYS_BIRTH             AMT_INCOME_TOTAL
##                 4.618023e-02                -4.599431e-02
##                  OWN_CAR_AGE               TOTALAREA_MODE
##                 3.837052e-02                -3.311888e-02
##   AMT_REQ_CREDIT_BUREAU_YEAR               FLOORSMIN_MEDI
##                 3.279824e-02                -3.254174e-02
##                FLOORSMIN_AVG                ELEVATORS_AVG
##                -3.252648e-02                -3.203765e-02
##               APARTMENTS_AVG         LIVINGAPARTMENTS_AVG
##                -3.185354e-02                -3.156523e-02
##               FLOORSMIN_MODE               LIVINGAREA_AVG
##                -3.123269e-02                -3.108370e-02
##               ELEVATORS_MEDI              APARTMENTS_MEDI
##                -3.057312e-02                -3.052623e-02
##   REGION_POPULATION_RELATIVE        LIVINGAPARTMENTS_MEDI
##                -3.027419e-02                -3.015265e-02
##              LIVINGAREA_MEDI              AMT_GOODS_PRICE
##                -3.013481e-02                -3.011508e-02
##              DAYS_ID_PUBLISH              LIVINGAREA_MODE
##                 2.987001e-02                -2.915797e-02
##             FLAG_DOCUMENT_13        LIVINGAPARTMENTS_MODE
##                -2.890926e-02                -2.792125e-02
##     OBS_30_CNT_SOCIAL_CIRCLE               ELEVATORS_MODE
##                 2.788881e-02                -2.751939e-02
##     OBS_60_CNT_SOCIAL_CIRCLE              APARTMENTS_MODE
##                 2.750033e-02                -2.665493e-02
##           NONLIVINGAREA_MEDI            NONLIVINGAREA_AVG
##                -2.654524e-02                -2.613666e-02
##              YEARS_BUILD_AVG             YEARS_BUILD_MEDI
##                -2.561819e-02                -2.512312e-02
##     DEF_60_CNT_SOCIAL_CIRCLE           NONLIVINGAREA_MODE
##                 2.412805e-02                -2.364367e-02
##             YEARS_BUILD_MODE                   AMT_CREDIT
##                -2.284402e-02                -2.213439e-02
##             BASEMENTAREA_AVG             FLAG_DOCUMENT_16
##                -2.087791e-02                -1.965151e-02
##     DEF_30_CNT_SOCIAL_CIRCLE              CNT_FAM_MEMBERS
##                 1.943179e-02                -1.929619e-02
##            BASEMENTAREA_MEDI       DAYS_LAST_PHONE_CHANGE
##                -1.921669e-02                 1.790208e-02
##               COMMONAREA_AVG              COMMONAREA_MEDI
##                -1.776508e-02                -1.727416e-02
##                 CNT_CHILDREN      HOUR_APPR_PROCESS_START
##                -1.633035e-02                -1.625102e-02
##   AMT_REQ_CREDIT_BUREAU_WEEK    AMT_REQ_CREDIT_BUREAU_DAY
##                 1.616949e-02                 1.515155e-02
##              FLAG_DOCUMENT_7              FLAG_WORK_PHONE
##                 1.495330e-02                 1.464951e-02
##            BASEMENTAREA_MODE              COMMONAREA_MODE
##                -1.460262e-02                -1.452439e-02
##                   SK_ID_CURR YEARS_BEGINEXPLUATATION_MEDI
##                 1.451369e-02                -1.439711e-02
##              FLAG_DOCUMENT_8               FLAG_EMP_PHONE
##                -1.439409e-02                 1.428056e-02
##    AMT_REQ_CREDIT_BUREAU_QRT                   FLAG_PHONE
##                 1.326654e-02                -1.310217e-02
##              FLAG_DOCUMENT_5                DAYS_EMPLOYED
##                -1.282136e-02                -1.259943e-02
##       REG_CITY_NOT_WORK_CITY   REG_REGION_NOT_LIVE_REGION
##                 1.211448e-02                -1.100620e-02
##  YEARS_BEGINEXPLUATATION_AVG                ENTRANCES_AVG
##                -1.063228e-02                -1.058016e-02
##               ENTRANCES_MEDI             FLAG_DOCUMENT_18
##                -1.039053e-02                -9.594059e-03
##                 LANDAREA_AVG                LANDAREA_MEDI
##                -9.462830e-03                -9.342955e-03
##               ENTRANCES_MODE             FLAG_DOCUMENT_19
##                -9.135605e-03                 8.607177e-03
##                LANDAREA_MODE             FLAG_CONT_MOBILE
##                -8.097740e-03                 8.001784e-03
##   REG_REGION_NOT_WORK_REGION              FLAG_DOCUMENT_6
##                -7.993280e-03                -7.917716e-03
##                  AMT_ANNUITY   AMT_REQ_CREDIT_BUREAU_HOUR
##                -6.698003e-03                 6.329620e-03
##             FLAG_DOCUMENT_11     NONLIVINGAPARTMENTS_MODE
##                -6.267669e-03                 5.963353e-03
##    AMT_REQ_CREDIT_BUREAU_MON                   FLAG_EMAIL
##                -5.736402e-03                 5.313036e-03
##             FLAG_DOCUMENT_14      LIVE_CITY_NOT_WORK_CITY
##                 4.620775e-03                 4.403652e-03
##            DAYS_REGISTRATION             FLAG_DOCUMENT_21
##                 4.309767e-03                -4.120458e-03
##     NONLIVINGAPARTMENTS_MEDI             FLAG_DOCUMENT_17
##                 3.795607e-03                -3.364191e-03
##      NONLIVINGAPARTMENTS_AVG             FLAG_DOCUMENT_15
##                 2.617824e-03                -2.595100e-03
##             FLAG_DOCUMENT_10              FLAG_DOCUMENT_4
##                -2.378738e-03                -2.378738e-03
##             FLAG_DOCUMENT_12             FLAG_DOCUMENT_20
##                -2.378738e-03                 2.357059e-03
## YEARS_BEGINEXPLUATATION_MODE       REG_CITY_NOT_LIVE_CITY
##                -1.659073e-03                 8.729674e-04
##  LIVE_REGION_NOT_WORK_REGION              FLAG_DOCUMENT_9
##                -8.081022e-04                -9.695645e-05
##                   FLAG_MOBIL              FLAG_DOCUMENT_2
##                           NA                           NA
```

2.3.2 Exploring Strong Predictors: You can also visualize potential strong predictors with scatter plots:

```
ggplot(train_data, aes(x = AMT_CREDIT, y = TARGET)) +
  geom_point() +
  labs(title = "AMT_CREDIT vs TARGET", x = "Credit Amount", y = "Target")
```



2.4 **Missing Data Exploration**

2.4.1 Setup to discover missing values

```
# Skim for a general overview
skim(train_data)
```

Data summary

Name	train_data
Number of rows	307511
Number of columns	122
________________________	
Column type frequency:	
character	16
numeric	106
________________________	
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
NAME_CONTRACT_TYPE	0	1	10	15	0	2	0
CODE_GENDER	0	1	1	3	0	3	0
FLAG_OWN_CAR	0	1	1	1	0	2	0
FLAG_OWN_REALTY	0	1	1	1	0	2	0
NAME_TYPE_SUITE	0	1	0	15	1292	8	0
NAME_INCOME_TYPE	0	1	7	20	0	8	0
NAME_EDUCATION_TYPE	0	1	15	29	0	5	0
NAME_FAMILY_STATUS	0	1	5	20	0	6	0
NAME_HOUSING_TYPE	0	1	12	19	0	6	0
OCCUPATION_TYPE	0	1	0	21	96391	19	0
WEEKDAY_APPR_PROCESS_START	0	1	6	9	0	7	0
ORGANIZATION_TYPE	0	1	3	22	0	58	0
FONDKAPREMONT_MODE	0	1	0	21	210295	5	0
HOUSETYPE_MODE	0	1	0	16	154297	4	0
WALLSMATERIAL_MODE	0	1	0	12	156341	8	0
EMERGENCYSTATE_MODE	0	1	0	3	145755	3	0

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
SK_ID_CURR	0	1.00	278180.52	102790.18	100002.00	189145.50	278202.00	367142.50	456255.00	▇▇▇▇▇
TARGET	0	1.00	0.08	0.27	0.00	0.00	0.00	0.00	1.00	▇▁▁▁▁
CNT_CHILDREN	0	1.00	0.42	0.72	0.00	0.00	0.00	1.00	19.00	▇▁▁▁▁
AMT_INCOME_TOTAL	0	1.00	168797.92	237123.15	25650.00	112500.00	147150.00	202500.00	117000000.00	▇▁▁▁▁
AMT_CREDIT	0	1.00	599026.00	402490.78	45000.00	270000.00	513531.00	808650.00	4050000.00	▇▂▁▁▁
AMT_ANNUITY	12	1.00	27108.57	14493.74	1615.50	16524.00	24903.00	34596.00	258025.50	▇▁▁▁▁
AMT_GOODS_PRICE	278	1.00	538396.21	369446.46	40500.00	238500.00	450000.00	679500.00	4050000.00	▇▂▁▁▁
REGION_POPULATION_RELATIVE	0	1.00	0.02	0.01	0.00	0.01	0.02	0.03	0.07	▇▇▇▇▇
DAYS_BIRTH	0	1.00	-16037.00	4363.99	-25229.00	-19682.00	-15750.00	-12413.00	-7489.00	▃▃▇▇▇
DAYS_EMPLOYED	0	1.00	63815.05	141275.77	-17912.00	-2760.00	-1213.00	-289.00	365243.00	▇▁▁▁▁
DAYS_REGISTRATION	0	1.00	-4986.12	3522.89	-24672.00	-7479.50	-4504.00	-2010.00	0.00	▁▁▁▁▂
DAYS_ID_PUBLISH	0	1.00	-2994.20	1509.45	-7197.00	-4299.00	-3254.00	-1720.00	0.00	▁▅▅▇▇
OWN_CAR_AGE	202929	0.34	12.06	11.94	0.00	5.00	9.00	15.00	91.00	▇▁▁▁▁
FLAG_MOBIL	0	1.00	1.00	0.00	0.00	1.00	1.00	1.00	1.00	▁▁▁▁▁
FLAG_EMP_PHONE	0	1.00	0.82	0.38	0.00	1.00	1.00	1.00	1.00	▂▁▁▁▁
FLAG_WORK_PHONE	0	1.00	0.20	0.40	0.00	0.00	0.00	0.00	1.00	▇▁▁▁▁
FLAG_CONT_MOBILE	0	1.00	1.00	0.04	0.00	1.00	1.00	1.00	1.00	▁▁▁▁▁
FLAG_PHONE	0	1.00	0.28	0.45	0.00	0.00	0.00	1.00	1.00	▇▁▁▁▁
FLAG_EMAIL	0	1.00	0.06	0.23	0.00	0.00	0.00	0.00	1.00	▇▁▁▁▁
CNT_FAM_MEMBERS	2	1.00	2.15	0.91	1.00	2.00	2.00	3.00	20.00	▇▁▁▁▁
REGION_RATING_CLIENT	0	1.00	2.05	0.51	1.00	2.00	2.00	2.00	3.00	▁▁▁▁▇
REGION_RATING_CLIENT_W_CITY	0	1.00	2.03	0.50	1.00	2.00	2.00	2.00	3.00	▁▁▁▁▇
HOUR_APPR_PROCESS_START	0	1.00	12.06	3.27	0.00	10.00	12.00	14.00	23.00	▁▁▃▇▇
REG_REGION_NOT_LIVE_REGION	0	1.00	0.02	0.12	0.00	0.00	0.00	0.00	1.00	▇▁▁▁▁
REG_REGION_NOT_WORK_REGION	0	1.00	0.05	0.22	0.00	0.00	0.00	0.00	1.00	▇▁▁▁▁
LIVE_REGION_NOT_WORK_REGION	0	1.00	0.04	0.20	0.00	0.00	0.00	0.00	1.00	▇▁▁▁▁
REG_CITY_NOT_LIVE_CITY	0	1.00	0.08	0.27	0.00	0.00	0.00	0.00	1.00	▇▁▁▁▁
REG_CITY_NOT_WORK_CITY	0	1.00	0.23	0.42	0.00	0.00	0.00	0.00	1.00	▇▁▁▁▁

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
LIVE_CITY_NOT_WORK_CITY	0	1.00	0.18	0.38	0.00	0.00	0.00	0.00	1.00	▇▁
EXT_SOURCE_1	173378	0.44	0.50	0.21	0.01	0.33	0.51	0.68	0.96	▂▇
EXT_SOURCE_2	660	1.00	0.51	0.19	0.00	0.39	0.57	0.66	0.85	▂▂
EXT_SOURCE_3	60965	0.80	0.51	0.19	0.00	0.37	0.54	0.67	0.90	▂▃
APARTMENTS_AVG	156061	0.49	0.12	0.11	0.00	0.06	0.09	0.15	1.00	▇▁
BASEMENTAREA_AVG	179943	0.41	0.09	0.08	0.00	0.04	0.08	0.11	1.00	▇▁
YEARS_BEGINEXPLUATATION_AVG	150007	0.51	0.98	0.06	0.00	0.98	0.98	0.99	1.00	▁▁
YEARS_BUILD_AVG	204488	0.34	0.75	0.11	0.00	0.69	0.76	0.82	1.00	▁▁
COMMONAREA_AVG	214865	0.30	0.04	0.08	0.00	0.01	0.02	0.05	1.00	▇▁
ELEVATORS_AVG	163891	0.47	0.08	0.13	0.00	0.00	0.00	0.12	1.00	▇▁
ENTRANCES_AVG	154828	0.50	0.15	0.10	0.00	0.07	0.14	0.21	1.00	▇▃
FLOORSMAX_AVG	153020	0.50	0.23	0.14	0.00	0.17	0.17	0.33	1.00	▇▃
FLOORSMIN_AVG	208642	0.32	0.23	0.16	0.00	0.08	0.21	0.38	1.00	▅▇
LANDAREA_AVG	182590	0.41	0.07	0.08	0.00	0.02	0.05	0.09	1.00	▇▁
LIVINGAPARTMENTS_AVG	210199	0.32	0.10	0.09	0.00	0.05	0.08	0.12	1.00	▇▁
LIVINGAREA_AVG	154350	0.50	0.11	0.11	0.00	0.05	0.07	0.13	1.00	▇▁
NONLIVINGAPARTMENTS_AVG	213514	0.31	0.01	0.05	0.00	0.00	0.00	0.00	1.00	▇▁
NONLIVINGAREA_AVG	169682	0.45	0.03	0.07	0.00	0.00	0.00	0.03	1.00	▇▁
APARTMENTS_MODE	156061	0.49	0.11	0.11	0.00	0.05	0.08	0.14	1.00	▇▁
BASEMENTAREA_MODE	179943	0.41	0.09	0.08	0.00	0.04	0.07	0.11	1.00	▇▁
YEARS_BEGINEXPLUATATION_MODE	150007	0.51	0.98	0.06	0.00	0.98	0.98	0.99	1.00	▁▁
YEARS_BUILD_MODE	204488	0.34	0.76	0.11	0.00	0.70	0.76	0.82	1.00	▁▁
COMMONAREA_MODE	214865	0.30	0.04	0.07	0.00	0.01	0.02	0.05	1.00	▇▁
ELEVATORS_MODE	163891	0.47	0.07	0.13	0.00	0.00	0.00	0.12	1.00	▇▁
ENTRANCES_MODE	154828	0.50	0.15	0.10	0.00	0.07	0.14	0.21	1.00	▇▃
FLOORSMAX_MODE	153020	0.50	0.22	0.14	0.00	0.17	0.17	0.33	1.00	▇▃
FLOORSMIN_MODE	208642	0.32	0.23	0.16	0.00	0.08	0.21	0.38	1.00	▅▇
LANDAREA_MODE	182590	0.41	0.06	0.08	0.00	0.02	0.05	0.08	1.00	▇▁
LIVINGAPARTMENTS_MODE	210199	0.32	0.11	0.10	0.00	0.05	0.08	0.13	1.00	▇▁
LIVINGAREA_MODE	154350	0.50	0.11	0.11	0.00	0.04	0.07	0.13	1.00	▇▁
NONLIVINGAPARTMENTS_MODE	213514	0.31	0.01	0.05	0.00	0.00	0.00	0.00	1.00	▇▁
NONLIVINGAREA_MODE	169682	0.45	0.03	0.07	0.00	0.00	0.00	0.02	1.00	▇▁
APARTMENTS_MEDI	156061	0.49	0.12	0.11	0.00	0.06	0.09	0.15	1.00	▇▁
BASEMENTAREA_MEDI	179943	0.41	0.09	0.08	0.00	0.04	0.08	0.11	1.00	▇▁
YEARS_BEGINEXPLUATATION_MEDI	150007	0.51	0.98	0.06	0.00	0.98	0.98	0.99	1.00	▁▁
YEARS_BUILD_MEDI	204488	0.34	0.76	0.11	0.00	0.69	0.76	0.83	1.00	▁▁
COMMONAREA_MEDI	214865	0.30	0.04	0.08	0.00	0.01	0.02	0.05	1.00	▇▁
ELEVATORS_MEDI	163891	0.47	0.08	0.13	0.00	0.00	0.00	0.12	1.00	▇▁
ENTRANCES_MEDI	154828	0.50	0.15	0.10	0.00	0.07	0.14	0.21	1.00	▇▃
FLOORSMAX_MEDI	153020	0.50	0.23	0.15	0.00	0.17	0.17	0.33	1.00	▇▃
FLOORSMIN_MEDI	208642	0.32	0.23	0.16	0.00	0.08	0.21	0.38	1.00	▅▇
LANDAREA_MEDI	182590	0.41	0.07	0.08	0.00	0.02	0.05	0.09	1.00	▇▁
LIVINGAPARTMENTS_MEDI	210199	0.32	0.10	0.09	0.00	0.05	0.08	0.12	1.00	▇▁
LIVINGAREA_MEDI	154350	0.50	0.11	0.11	0.00	0.05	0.07	0.13	1.00	▇▁
NONLIVINGAPARTMENTS_MEDI	213514	0.31	0.01	0.05	0.00	0.00	0.00	0.00	1.00	▇▁
NONLIVINGAREA_MEDI	169682	0.45	0.03	0.07	0.00	0.00	0.00	0.03	1.00	▇▁
TOTALAREA_MODE	148431	0.52	0.10	0.11	0.00	0.04	0.07	0.13	1.00	▇▁
OBS_30_CNT_SOCIAL_CIRCLE	1021	1.00	1.42	2.40	0.00	0.00	0.00	2.00	348.00	▇▁

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
DEF_30_CNT_SOCIAL_CIRCLE	1021	1.00	0.14	0.45	0.00	0.00	0.00	0.00	34.00	▇▁
OBS_60_CNT_SOCIAL_CIRCLE	1021	1.00	1.41	2.38	0.00	0.00	0.00	2.00	344.00	▇▁
DEF_60_CNT_SOCIAL_CIRCLE	1021	1.00	0.10	0.36	0.00	0.00	0.00	0.00	24.00	▇▁
DAYS_LAST_PHONE_CHANGE	1	1.00	-962.86	826.81	-4292.00	-1570.00	-757.00	-274.00	0.00	▁▂
FLAG_DOCUMENT_2	0	1.00	0.00	0.01	0.00	0.00	0.00	0.00	1.00	▇▁
FLAG_DOCUMENT_3	0	1.00	0.71	0.45	0.00	0.00	1.00	1.00	1.00	▃▁
FLAG_DOCUMENT_4	0	1.00	0.00	0.01	0.00	0.00	0.00	0.00	1.00	▇▁
FLAG_DOCUMENT_5	0	1.00	0.02	0.12	0.00	0.00	0.00	0.00	1.00	▇▁
FLAG_DOCUMENT_6	0	1.00	0.09	0.28	0.00	0.00	0.00	0.00	1.00	▇▁
FLAG_DOCUMENT_7	0	1.00	0.00	0.01	0.00	0.00	0.00	0.00	1.00	▇▁
FLAG_DOCUMENT_8	0	1.00	0.08	0.27	0.00	0.00	0.00	0.00	1.00	▇▁
FLAG_DOCUMENT_9	0	1.00	0.00	0.06	0.00	0.00	0.00	0.00	1.00	▇▁
FLAG_DOCUMENT_10	0	1.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	▇▁
FLAG_DOCUMENT_11	0	1.00	0.00	0.06	0.00	0.00	0.00	0.00	1.00	▇▁
FLAG_DOCUMENT_12	0	1.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	▇▁
FLAG_DOCUMENT_13	0	1.00	0.00	0.06	0.00	0.00	0.00	0.00	1.00	▇▁
FLAG_DOCUMENT_14	0	1.00	0.00	0.05	0.00	0.00	0.00	0.00	1.00	▇▁
FLAG_DOCUMENT_15	0	1.00	0.00	0.03	0.00	0.00	0.00	0.00	1.00	▇▁
FLAG_DOCUMENT_16	0	1.00	0.01	0.10	0.00	0.00	0.00	0.00	1.00	▇▁
FLAG_DOCUMENT_17	0	1.00	0.00	0.02	0.00	0.00	0.00	0.00	1.00	▇▁
FLAG_DOCUMENT_18	0	1.00	0.01	0.09	0.00	0.00	0.00	0.00	1.00	▇▁
FLAG_DOCUMENT_19	0	1.00	0.00	0.02	0.00	0.00	0.00	0.00	1.00	▇▁
FLAG_DOCUMENT_20	0	1.00	0.00	0.02	0.00	0.00	0.00	0.00	1.00	▇▁
FLAG_DOCUMENT_21	0	1.00	0.00	0.02	0.00	0.00	0.00	0.00	1.00	▇▁
AMT_REQ_CREDIT_BUREAU_HOUR	41519	0.86	0.01	0.08	0.00	0.00	0.00	0.00	4.00	▇▁
AMT_REQ_CREDIT_BUREAU_DAY	41519	0.86	0.01	0.11	0.00	0.00	0.00	0.00	9.00	▇▁
AMT_REQ_CREDIT_BUREAU_WEEK	41519	0.86	0.03	0.20	0.00	0.00	0.00	0.00	8.00	▇▁
AMT_REQ_CREDIT_BUREAU_MON	41519	0.86	0.27	0.92	0.00	0.00	0.00	0.00	27.00	▇▁
AMT_REQ_CREDIT_BUREAU_QRT	41519	0.86	0.27	0.79	0.00	0.00	0.00	0.00	261.00	▇▁
AMT_REQ_CREDIT_BUREAU_YEAR	41519	0.86	1.90	1.87	0.00	0.00	1.00	3.00	25.00	▇▁

```
# Checking missing values
library(dplyr)
```

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

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

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

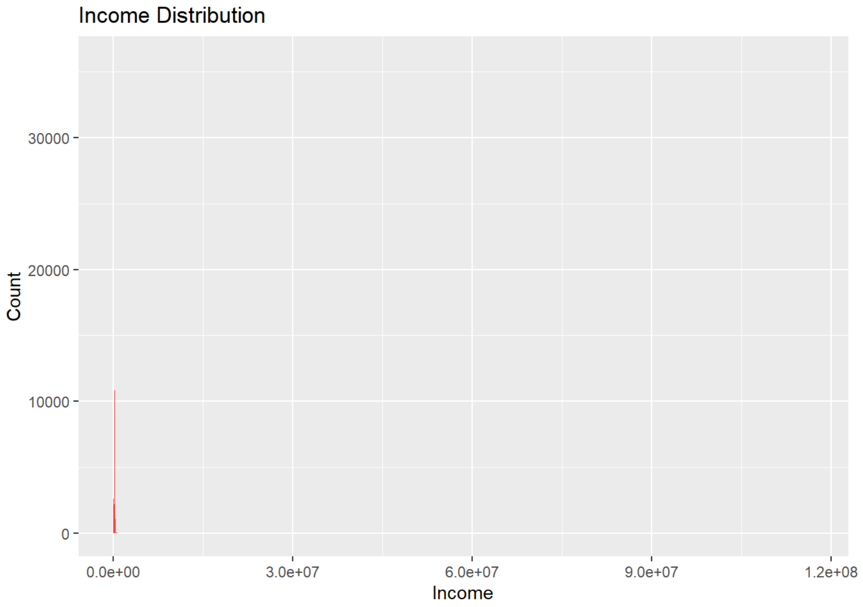
```
train_data %>%
  summarise(across(everything(), ~ sum(is.na(.))))
```

```
##   SK_ID_CURR TARGET NAME_CONTRACT_TYPE CODE_GENDER FLAG_OWN_CAR FLAG_OWN_REALTY
## 1          0      0                  0           0            0               0
##   CNT_CHILDREN AMT_INCOME_TOTAL AMT_CREDIT AMT_ANNUITY AMT_GOODS_PRICE
## 1            0                0          0          12             278
##   NAME_TYPE_SUITE NAME_INCOME_TYPE NAME_EDUCATION_TYPE NAME_FAMILY_STATUS
## 1               0                0                   0                  0
##   NAME_HOUSING_TYPE REGION_POPULATION_RELATIVE DAYS_BIRTH DAYS_EMPLOYED
## 1                 0                          0          0             0
##   DAYS_REGISTRATION DAYS_ID_PUBLISH OWN_CAR_AGE FLAG_MOBIL FLAG_EMP_PHONE
## 1                 0               0      202929          0              0
##   FLAG_WORK_PHONE FLAG_CONT_MOBILE FLAG_PHONE FLAG_EMAIL OCCUPATION_TYPE
## 1               0                0          0          0               0
##   CNT_FAM_MEMBERS REGION_RATING_CLIENT REGION_RATING_CLIENT_W_CITY
## 1               2                    0                           0
##   WEEKDAY_APPR_PROCESS_START HOUR_APPR_PROCESS_START REG_REGION_NOT_LIVE_REGION
## 1                          0                       0                          0
##   REG_REGION_NOT_WORK_REGION LIVE_REGION_NOT_WORK_REGION REG_CITY_NOT_LIVE_CITY
## 1                          0                           0                      0
##   REG_CITY_NOT_WORK_CITY LIVE_CITY_NOT_WORK_CITY ORGANIZATION_TYPE EXT_SOURCE_1
## 1                      0                       0                 0       173378
##   EXT_SOURCE_2 EXT_SOURCE_3 APARTMENTS_AVG BASEMENTAREA_AVG
## 1          660        60965         156061           179943
##   YEARS_BEGINEXPLUATATION_AVG YEARS_BUILD_AVG COMMONAREA_AVG ELEVATORS_AVG
## 1                      150007          204488         214865        163891
##   ENTRANCES_AVG FLOORSMAX_AVG FLOORSMIN_AVG LANDAREA_AVG LIVINGAPARTMENTS_AVG
## 1        154828        153020        208642       182590               210199
##   LIVINGAREA_AVG NONLIVINGAPARTMENTS_AVG NONLIVINGAREA_AVG APARTMENTS_MODE
## 1         154350                  213514            169682          156061
##   BASEMENTAREA_MODE YEARS_BEGINEXPLUATATION_MODE YEARS_BUILD_MODE
## 1            179943                       150007           204488
##   COMMONAREA_MODE ELEVATORS_MODE ENTRANCES_MODE FLOORSMAX_MODE FLOORSMIN_MODE
## 1          214865         163891         154828         153020         208642
##   LANDAREA_MODE LIVINGAPARTMENTS_MODE LIVINGAREA_MODE NONLIVINGAPARTMENTS_MODE
## 1        182590                210199          154350                   213514
##   NONLIVINGAREA_MODE APARTMENTS_MEDI BASEMENTAREA_MEDI
## 1             169682          156061            179943
##   YEARS_BEGINEXPLUATATION_MEDI YEARS_BUILD_MEDI COMMONAREA_MEDI ELEVATORS_MEDI
## 1                       150007           204488          214865         163891
##   ENTRANCES_MEDI FLOORSMAX_MEDI FLOORSMIN_MEDI LANDAREA_MEDI
## 1         154828         153020         208642        182590
##   LIVINGAPARTMENTS_MEDI LIVINGAREA_MEDI NONLIVINGAPARTMENTS_MEDI
## 1                210199          154350                   213514
##   NONLIVINGAREA_MEDI FONDKAPREMONT_MODE HOUSETYPE_MODE TOTALAREA_MODE
## 1             169682                  0              0         148431
##   WALLSMATERIAL_MODE EMERGENCYSTATE_MODE OBS_30_CNT_SOCIAL_CIRCLE
## 1                  0                   0                     1021
##   DEF_30_CNT_SOCIAL_CIRCLE OBS_60_CNT_SOCIAL_CIRCLE DEF_60_CNT_SOCIAL_CIRCLE
## 1                     1021                     1021                     1021
##   DAYS_LAST_PHONE_CHANGE FLAG_DOCUMENT_2 FLAG_DOCUMENT_3 FLAG_DOCUMENT_4
## 1                      1               0               0               0
##   FLAG_DOCUMENT_5 FLAG_DOCUMENT_6 FLAG_DOCUMENT_7 FLAG_DOCUMENT_8
## 1               0               0               0               0
##   FLAG_DOCUMENT_9 FLAG_DOCUMENT_10 FLAG_DOCUMENT_11 FLAG_DOCUMENT_12
## 1               0                0                0                0
##   FLAG_DOCUMENT_13 FLAG_DOCUMENT_14 FLAG_DOCUMENT_15 FLAG_DOCUMENT_16
## 1                0                0                0                0
##   FLAG_DOCUMENT_17 FLAG_DOCUMENT_18 FLAG_DOCUMENT_19 FLAG_DOCUMENT_20
## 1                0                0                0                0
##   FLAG_DOCUMENT_21 AMT_REQ_CREDIT_BUREAU_HOUR AMT_REQ_CREDIT_BUREAU_DAY
## 1                0                      41519                     41519
##   AMT_REQ_CREDIT_BUREAU_WEEK AMT_REQ_CREDIT_BUREAU_MON
## 1                      41519                     41519
##   AMT_REQ_CREDIT_BUREAU_QRT AMT_REQ_CREDIT_BUREAU_YEAR
## 1                     41519                      41519
```

```
# Using janitor for a cleaner view of missing data
train_data_clean <- train_data %>% remove_empty("cols")
```

```
# Removing columns with more than 50% missing values
train_data <- train_data[, colMeans(is.na(train_data)) < 0.5]

# Imputing with median
train_data$AMT_INCOME_TOTAL[is.na(train_data$AMT_INCOME_TOTAL)] <- median(train_data$AMT_INCOME_TOTAL, na.rm = TRUE)
```

```
ggplot(train_data, aes(x = AMT_INCOME_TOTAL)) +
  geom_histogram(binwidth = 5000, fill = "red", alpha = 0.7) +
  labs(title = "Income Distribution", x = "Income", y = "Count")
```



```
# Converting categorical variables into factors
train_data$CODE_GENDER <- as.factor(train_data$CODE_GENDER)

# For models requiring dummy variables
train_data <- model.matrix(~ CODE_GENDER + FLAG_OWN_CAR, data = train_data)
```

3 Additional Questions to Guide Exploration

Some questions you can explore during EDA include:

- Is the data balanced or imbalanced with respect to the target variable (loan default)?
- What are the relationships between key features (e.g., loan amount, income) and the target variable?
- Which features are the strongest predictors of loan default?
- Are there any significant data quality issues, such as missing or outlier values?
- Does the data need preprocessing, such as normalization or encoding of categorical variables?
- How are demographic variables (e.g., age, family status) related to loan default?
- What trends can be seen with credit amount, loan annuity, and income total?

4 Data Description

Your dataset includes information about loan applicants, such as demographic information (e.g., age, gender), financial information (e.g., income, credit amount), and loan application details. The file `application_train.csv` is the primary dataset, and additional data files (such as `bureau.csv`) can be used to enrich the analysis by joining based on `SK_ID_CURR`.

Summary of the Data:

- The dataset contains both numeric and categorical variables.
- Key variables include `AMT_CREDIT`, `AMT_INCOME_TOTAL`, `AMT_ANNUITY`, and `DAYS_BIRTH`.
- The target variable is `TARGET`, which indicates loan repayment status.

5 Missing Data

During the data exploration, you might encounter missing values in several columns. For instance, columns related to property or financial information may have missing values for some applicants. Addressing missing data is crucial for accurate model development.

Proposed Solutions:

- For columns with a significant portion of missing data (e.g., more than 50%), consider dropping these columns.
- For columns with fewer missing values, consider imputing with median or mean values (for numerical columns) or the most frequent value (for categorical columns).
- You can also consider more sophisticated imputation methods like k-nearest neighbors or predictive modeling.

6 Findings and Discussion

Based on the initial analysis:

- The target variable (`TARGET`) is imbalanced, with a majority of applicants repaying their loans on time and a smaller proportion defaulting.
- Certain variables, such as `AMT_CREDIT` and `AMT_INCOME_TOTAL`, show potential for being strong predictors of loan default, based on their distribution and correlation with the target variable.

- There are some data quality issues, including missing values in multiple columns, especially those related to property ownership and financial records.
- The exploratory visualizations suggest that applicants with lower income or higher loan amounts might be at greater risk of defaulting on their loans.