Bank Loan Case Study

Submitted By: Miss. Desai aishwarya Ganapat

Contents

- Project Description
- Approach
- Tech-Stack Used
- Insights
- Result

Project Description

- This case study attempts to analyze patterns in the data using Exploratory Data Analysis(EDA) and ensure that capable applicants are not rejected. The main Aim of study is to identify patterns that indicates if a customer will have difficulty paying their installments which can be used to make decision such as denying the loan, reducing the amount of loan or lending at a higher interest rate to risky applicants

Approach

- **Download the Dataset:** This is the first step while starting the analysis, to download and converting the data in to suitable format.
- **Understanding the data:** After successfully downloading or converting the data, one need to understand the labels given in the dataset. Also needs to understand problem statement.
- **Cleaning the data:** This is first step of analysis where we need to identify missing values, duplicate values from the data & clean it with proper process. Also it involves to delete unnecessary column which will not be used during analysis
- **Analyzing the data:** After the cleaning next is to explore the dataset and find answers to the question.
- **Representation:** This is last but important step which will make analysis representative and easy to understand to people.

Tech-Stack Used

- Microsoft Excel 2016 (for working, analysis purpose)
- Microsoft Word 2016 (for presentation purpose)

Data Description

- The Dataset provided for analysis contains information about loan application of finance company which includes two types of scenarios:
- 1. Customer with payment difficulties: These are customers who had a late payment of more than x days on at least one of the first y installments of loan.
- 2. All other cases: These are cases where the payment was made on time.
- previous_application.csv: Contains information about previous loan applications. application_data.csv: Provides details about the current loan applications.
- columns_description.csv: Describes the columns present in the other datasets, explaining what each column represents.

- The following Hyperlink shows the Analysis done on the Excel Sheet :
- https://docs.google.com/spreadsheets/d/1Uez3ng1J_Wm-BHbn8dvIoxlbfU4Lz5PW/edit?usp=sharing&ouid=112713524583719045365&rtpof=true&sd=true

Insights

Task : A

To Identify Missing Data & Deal with it appropriately.

To Identify missing data in the dataset and decide on an appropriate method to deal with it using Excel built-in functions and features.

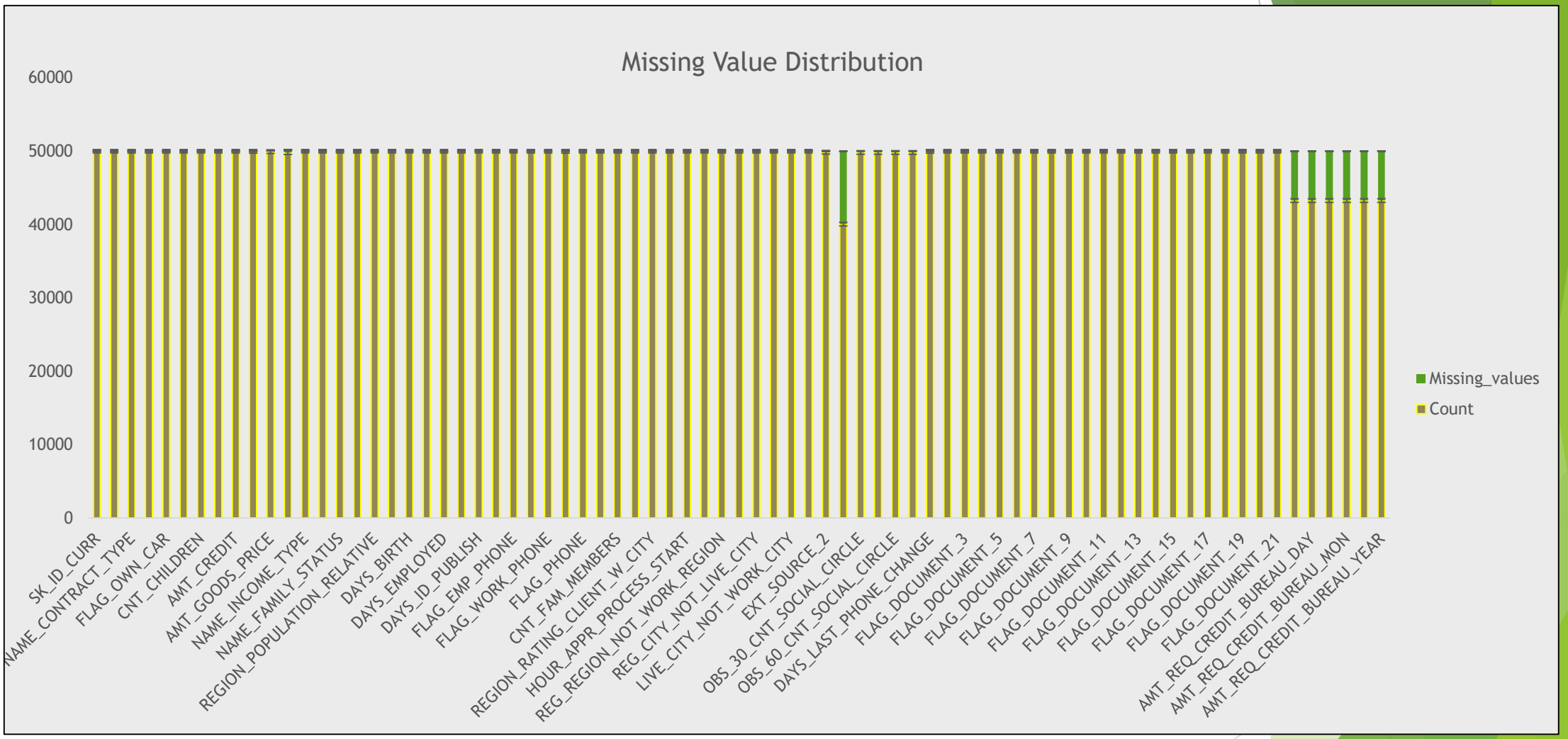
Ans:

Initially there were 123 columns and 49999 rows in the data. To deal with missing data I firstly calculate null values percentage on the basis of that I removed the column having null value percentage greater than 30%, also there were some undesirable columns so I delete those columns too. Then for the remaining columns I used median method to fill missing values for numerical columns. Also for better analysis I converted DAYS_BIRTH as DAYS_BIRTH(YEAR), DAYS_EMPLOYED as DAYS_EMPLOYED(YEAR), DAYS_REGISTRATION as DAYS_REGISTRATION (YEAR), DAYS_ID_PUBLISH as DAYS_ID_PUBLISH(YEAR).

Column	Count	Missing_values
SK_ID_CURR	49999	0
TARGET	49999	0
NAME_CONTRACT_TYPE	49999	0
CODE_GENDER	49999	0
FLAG_OWN_CAR	49999	0
FLAG_OWN_REALTY	49999	0
CNT_CHILDREN	49999	0
AMT_INCOME_TOTAL	49999	0
AMT_CREDIT	49999	0
AMT_ANNUITY	49998	1
AMT_GOODS_PRICE	49961	76
NAME_TYPE_SUITE	49807	384
NAME_INCOME_TYPE	49999	0
NAME_EDUCATION_TYPE	49999	0
NAME_FAMILY_STATUS	49999	0
NAME_HOUSING_TYPE	49999	0
REGION_POPULATION_RELATIVE	49999	0
REGION_POPULATION_RELATIVE	49999	0
DAYS_BIRTH	49999	0
DAYS_EMPLOYED	49999	0
DAYS_EMPLOYED	49999	0
DAYS_REGISTRATION	49999	0
DAYS_ID_PUBLISH	49999	0
FLAG_MOBIL	49999	0
FLAG_EMP_PHONE	49999	0
FLAG_EMP_PHONE	49999	0
FLAG_WORK_PHONE	49999	0
FLAG_CONT_MOBILE	49998	1
FLAG_PHONE	49999	0
FLAG_EMAIL	49999	0
CNT_FAM_MEMBERS	49999	0
REGION_RATING_CLIENT	49999	0
REGION_RATING_CLIENT_W_CITY	49999	0
WEEKDAY_APPR_PROCESS_START	49999	0
HOUR_APPR_PROCESS_START	49999	0

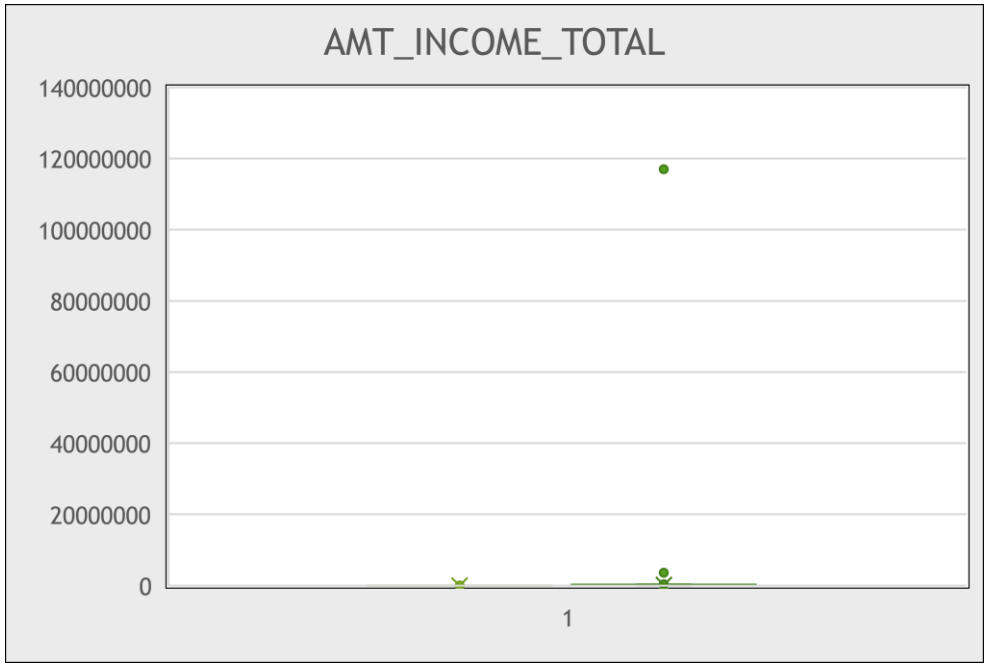
REG_REGION_NOT_LIVE_REGION	49999	0
REG_REGION_NOT_WORK_REGION	49999	0
LIVE_REGION_NOT_WORK_REGION	49999	0
REG_CITY_NOT_LIVE_CITY	49999	0
REG_CITY_NOT_WORK_CITY	49999	0
LIVE_CITY_NOT_WORK_CITY	49999	0
ORGANIZATION_TYPE	49999	0
EXT_SOURCE_2	49873	126
EXT_SOURCE_3	40055	9944
OBS_30_CNT_SOCIAL_CIRCLE	49831	186
DEF_30_CNT_SOCIAL_CIRCLE	49831	186
OBS_60_CNT_SOCIAL_CIRCLE	49831	186
DEF_60_CNT_SOCIAL_CIRCLE	49831	186
DAYS_LAST_PHONE_CHANGE	49998	1
FLAG_DOCUMENT_2	49999	0
FLAG_DOCUMENT_3	49999	0
FLAG_DOCUMENT_4	49999	0
FLAG_DOCUMENT_5	49999	0
FLAG_DOCUMENT_6	49999	0
FLAG_DOCUMENT_7	49999	0
FLAG_DOCUMENT_8	49999	0
FLAG_DOCUMENT_9	49999	0
FLAG_DOCUMENT_10	49999	0

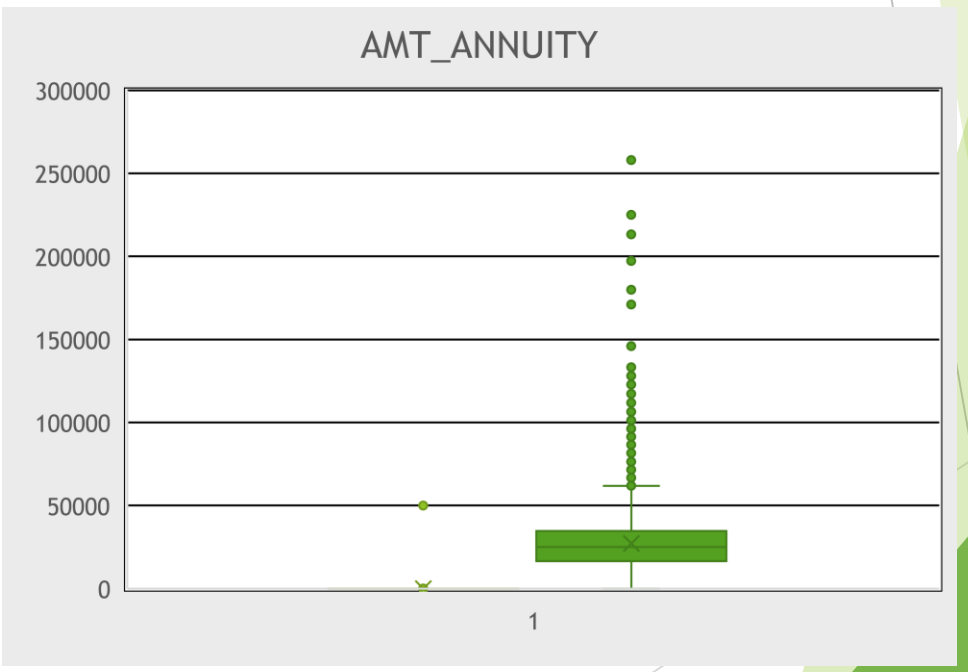
FLAG_DOCUMENT_11	49999	0
FLAG_DOCUMENT_12	49999	0
FLAG_DOCUMENT_13	49999	0
FLAG_DOCUMENT_14	49999	0
FLAG_DOCUMENT_15	49999	0
FLAG_DOCUMENT_16	49999	0
FLAG_DOCUMENT_17	49999	0
FLAG_DOCUMENT_18	49999	0
FLAG_DOCUMENT_19	49999	0
FLAG_DOCUMENT_20	49999	0
FLAG_DOCUMENT_21	49999	0
AMT_REQ_CREDIT_BUREAU_HOUR	43265	6734
AMT_REQ_CREDIT_BUREAU_DAY	43265	6734
AMT_REQ_CREDIT_BUREAU_WEEK	43265	6734
AMT_REQ_CREDIT_BUREAU_MON	43265	6734
AMT_REQ_CREDIT_BUREAU_QRT	43265	6734
AMT_REQ_CREDIT_BUREAU_YEAR	43265	6734

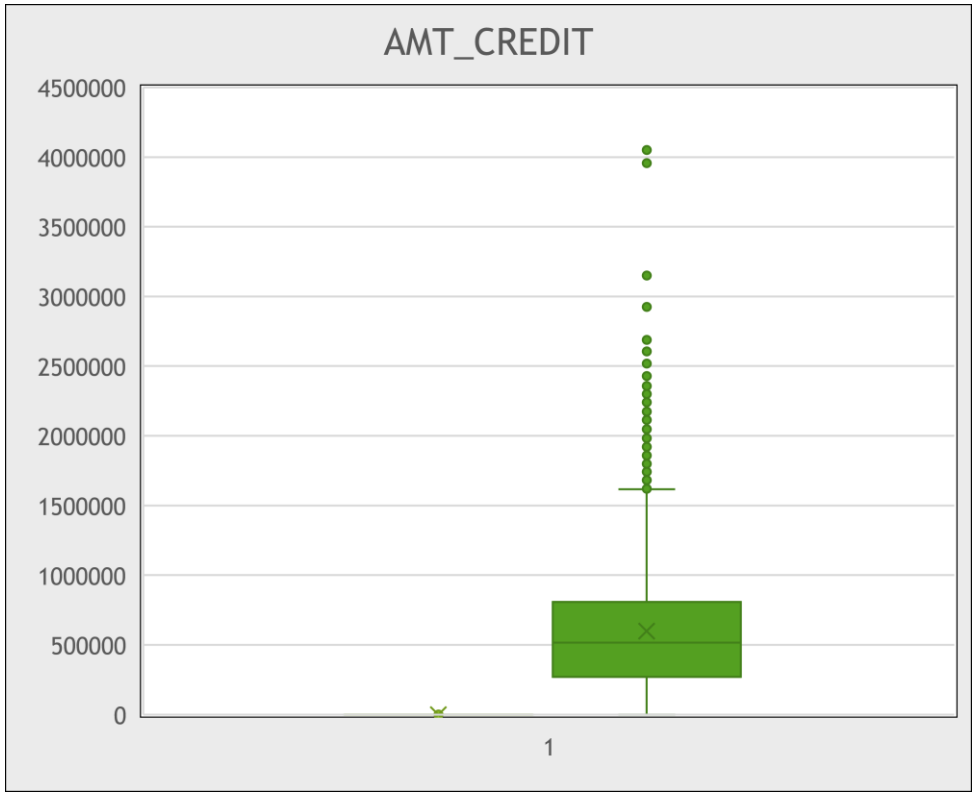

Missing Value Distribution
60000
50000
40000
30000
20000
10000
0
SK_ID_CURR
NAME_CONTRACT_TYPE
FLAG_OWN_CAR
CNT_CHILDREN
AMT_CREDIT
AMT_GOODS_PRICE
NAME_INCOME_TYPE
NAME_FAMILY_STATUS
REGION_POPULATION_RELATIVE
DAYS_BIRTH
DAYS_EMPLOYED
DAYS_ID_PUBLISH
FLAG_EMP_PHONE
FLAG_WORK_PHONE
FLAG_PHONE
CNT_FAM_MEMBERS
REGION_RATING_CLIENT_W_CITY
HOUR_APPR_PROCESS_START
REG_REGION_NOT_WORK_REGION
REG_CITY_NOT_LIVE_CITY
LIVE_CITY_NOT_WORK_CITY
EXT_SOURCE_2
OBS_30_CNT_SOCIAL_CIRCLE
OBS_60_CNT_SOCIAL_CIRCLE
DAYS_LAST_PHONE_CHANGE
FLAG_DOCUMENT_3
FLAG_DOCUMENT_5
FLAG_DOCUMENT_7
FLAG_DOCUMENT_9
FLAG_DOCUMENT_11
FLAG_DOCUMENT_13
FLAG_DOCUMENT_15
FLAG_DOCUMENT_17
FLAG_DOCUMENT_19
FLAG_DOCUMENT_21
AMT_REQ_CREDIT_BUREAU_DAY
AMT_REQ_CREDIT_BUREAU_MON
AMT_REQ_CREDIT_BUREAU_YEAR
Missing_values
Count

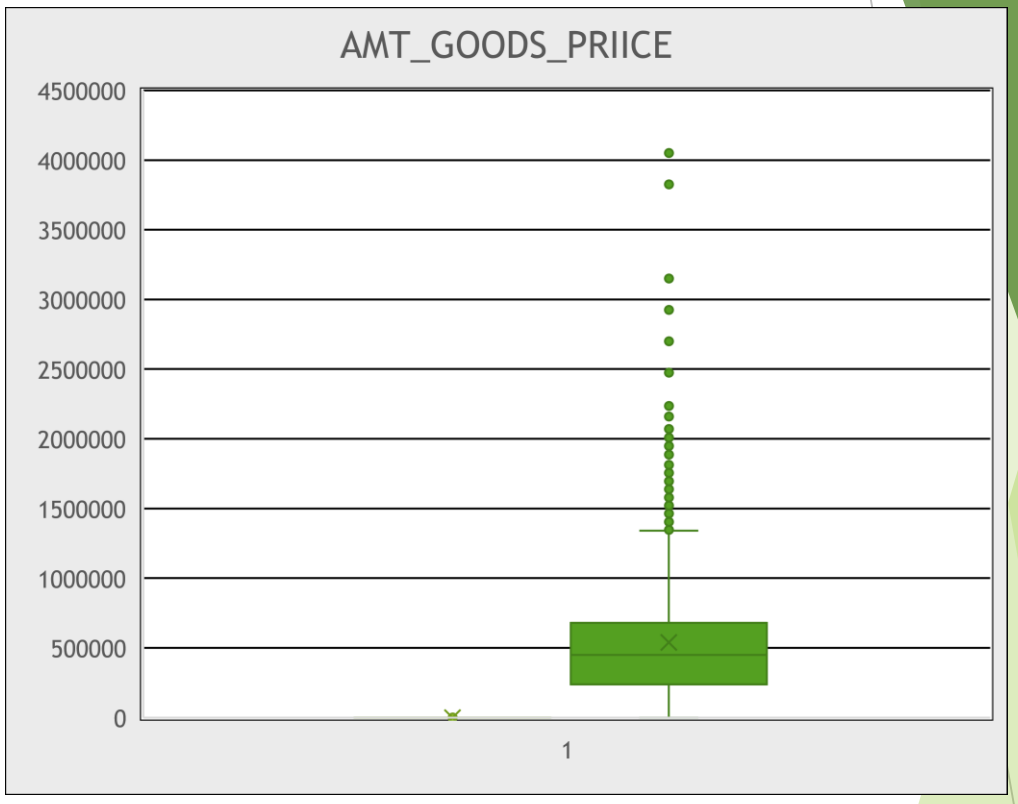
Task 2: Identify Outliers in the Dataset: Detect and identify outliers in the dataset using Excel statistical functions and features, focusing on numerical variables.

- Box plots of Target column vs
- 1. AMT_INCOME_TOTAL 2. AMT_ANNUITY 3. AMT_CREDIT 4. AMT_GOODS_PRICE 5. CNT_FAM_MEMBERS






AMT_CREDIT
4500000
4000000
3500000
3000000
2500000
2000000
1500000
1000000
500000
0
1

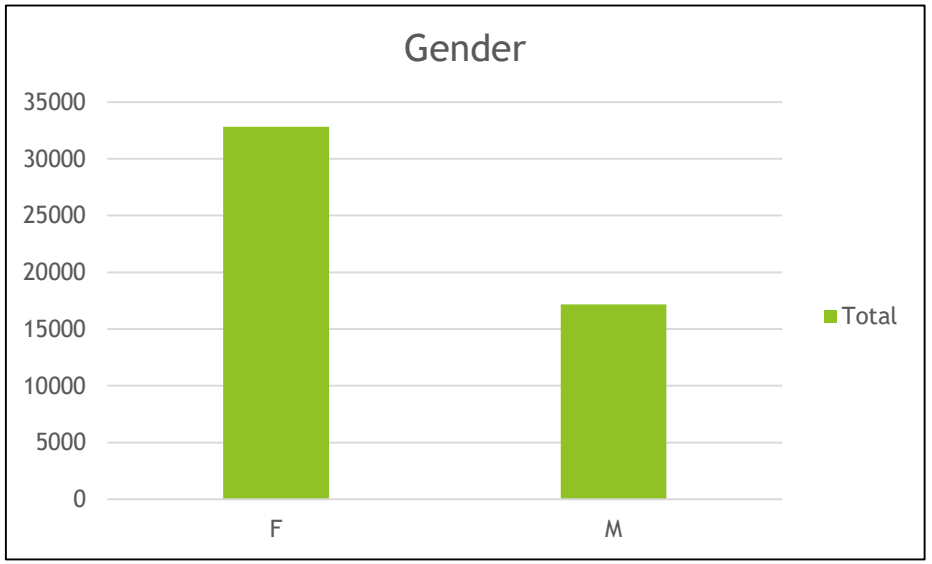

AMT_GOODS_PRIICE
4500000
4000000
3500000
3000000
2500000
2000000
1500000
1000000
500000
0
1

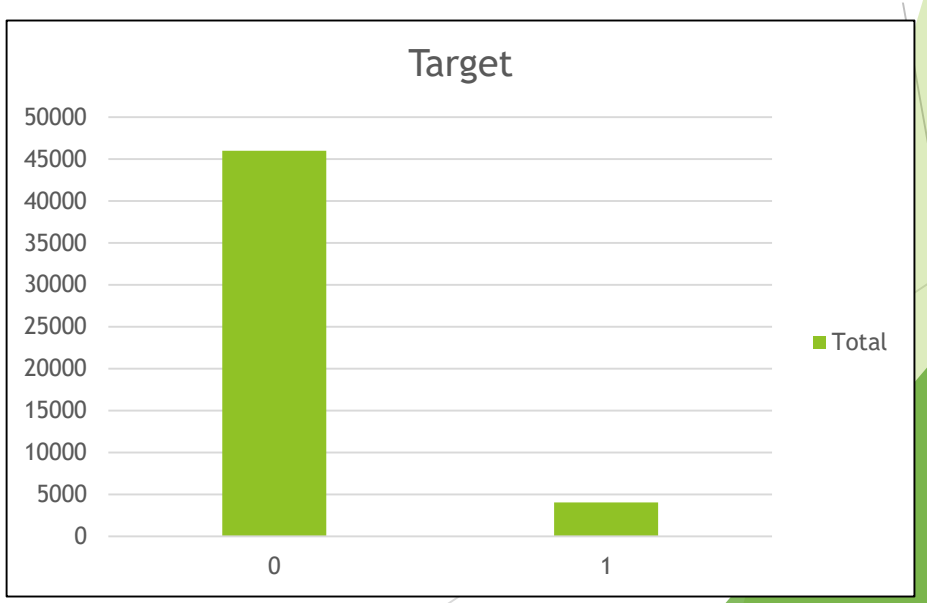
C. Analyse Data Imbalance Task: Determine if there is data imbalance in the loan application and calculate ratio of data imbalance using excel function

We can see the data imbalance using following Pivot tables and charts:

Row Labels	Count of CODE_GENDER
0	45973
1	4026
Grand Total	49999

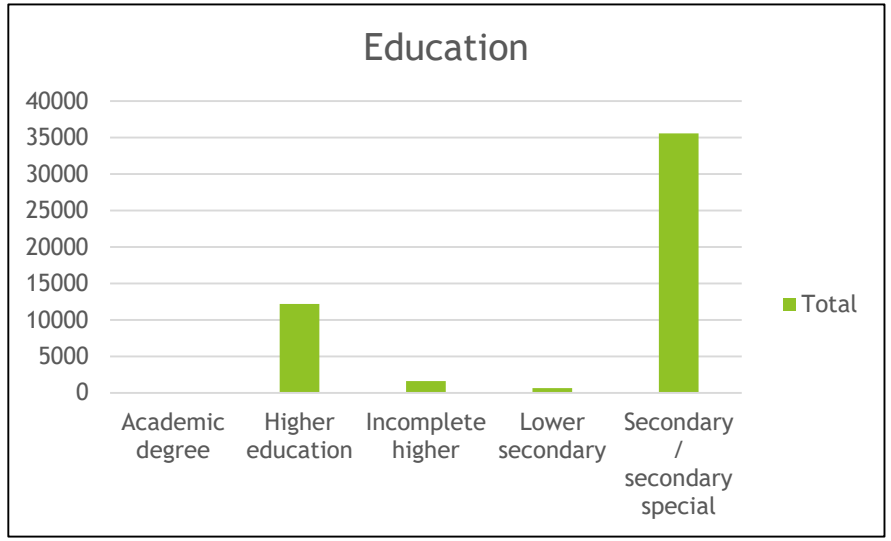
Row Labels	Count of TARGET
F	32823
M	17174
Grand Total	49997

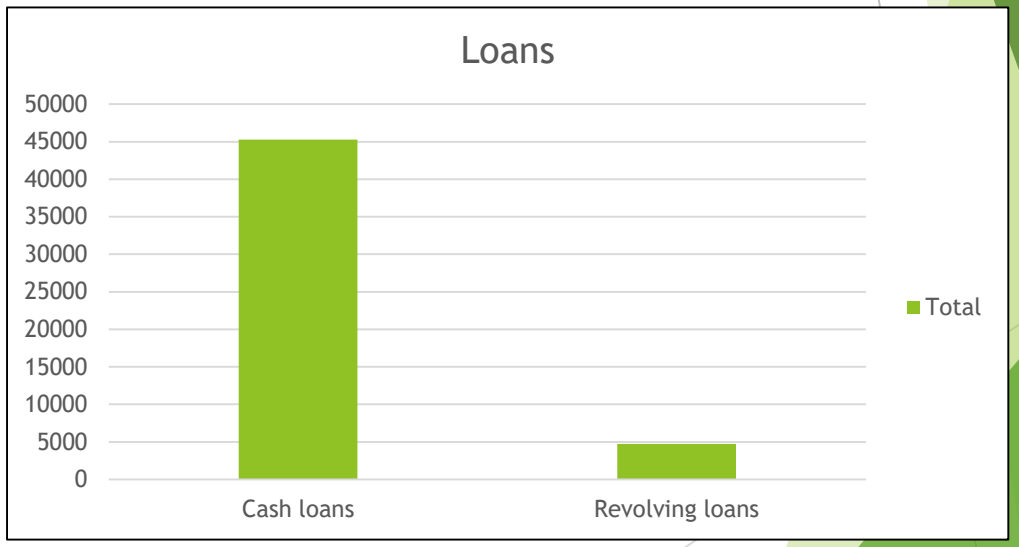




Row Labels	Count of TARGET
Academic degree	20
Higher education	12167
Incomplete higher	1620
Lower secondary	620
Secondary / secondary special	35572
Grand Total	49999

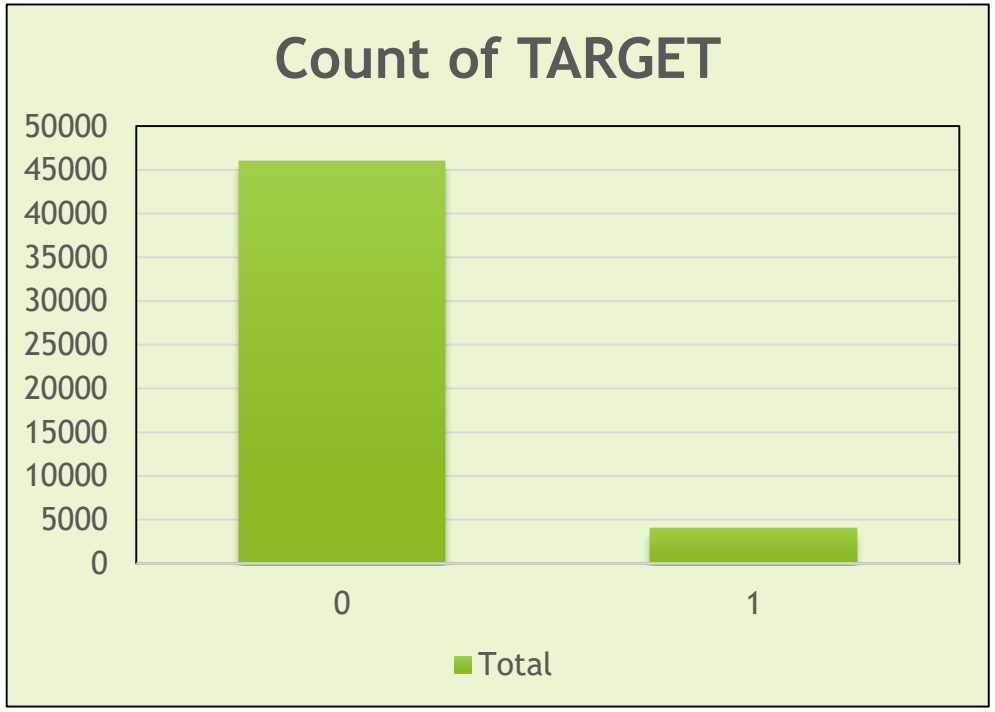
Row Labels	Count of TARGET
Cash loans	45276
Revolving loans	4723
Grand Total	49999

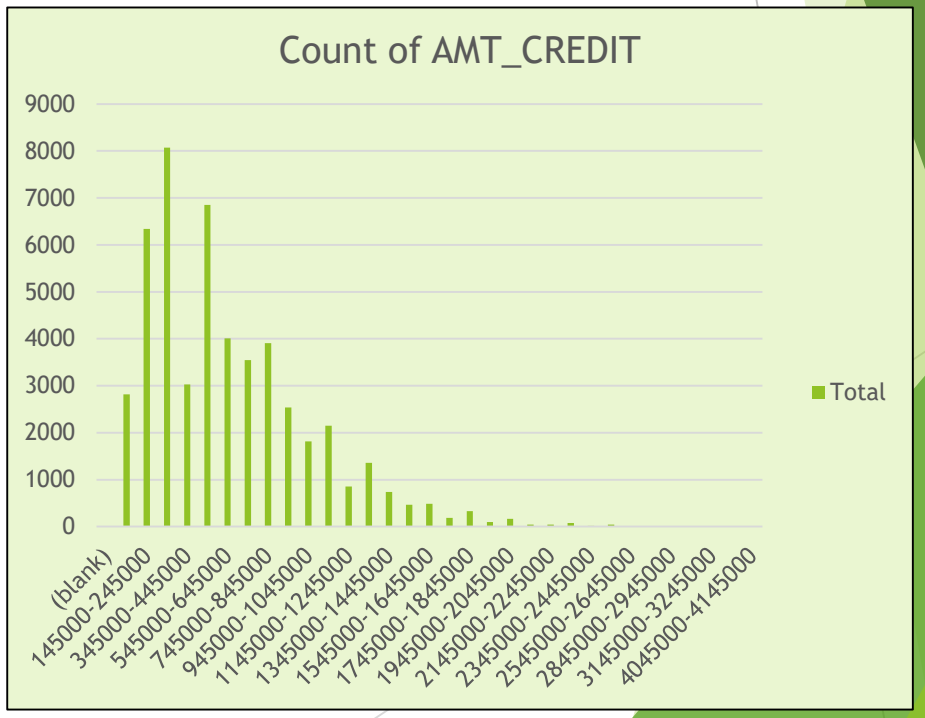


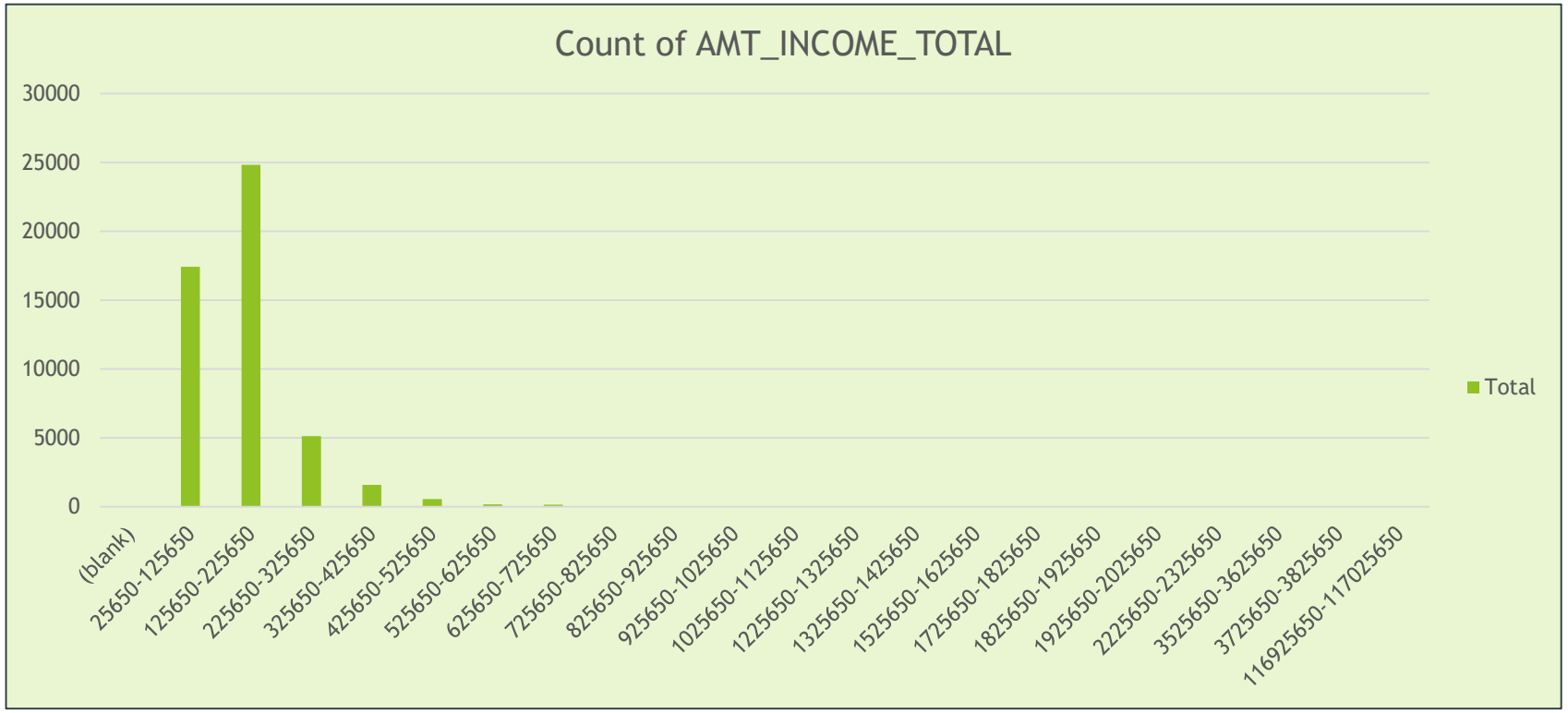


D. Perform Univariate, Segmented Univariate and Bivariate Analysis Task: Perform univariate analysis to understand distribution of individual variables. Segmented univariate analysis to compare variable distributions for different scenarios, and bivariate analysis to explore relationship between variables and target variable

- **Univariate Analysis:** The Univariate Analysis focuses on examining and describing the individual variables in isolation. The summary and analysis for the single variable.
- The Plots obtained while performing the Univariate Analysis are as follows.

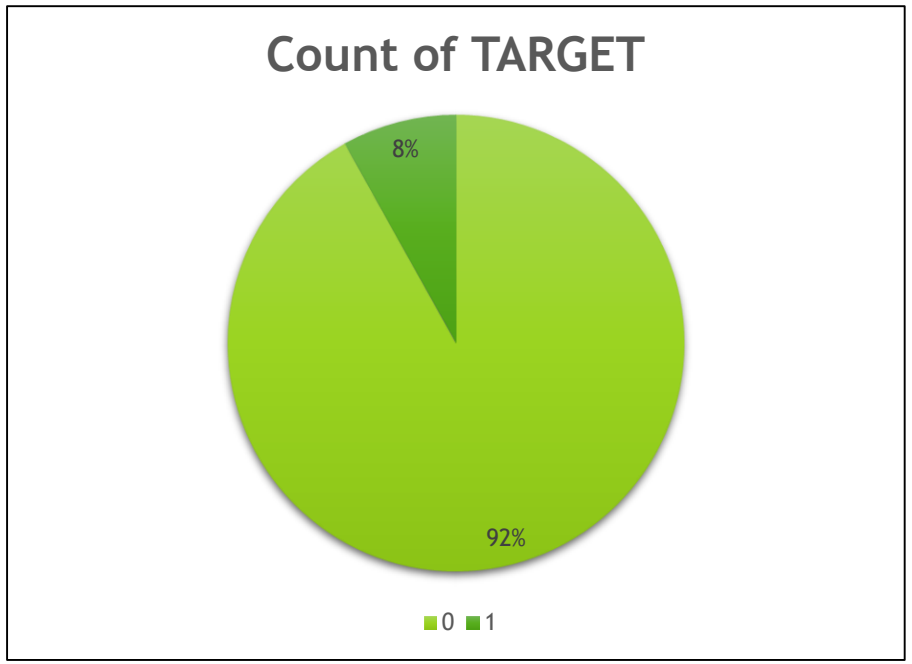


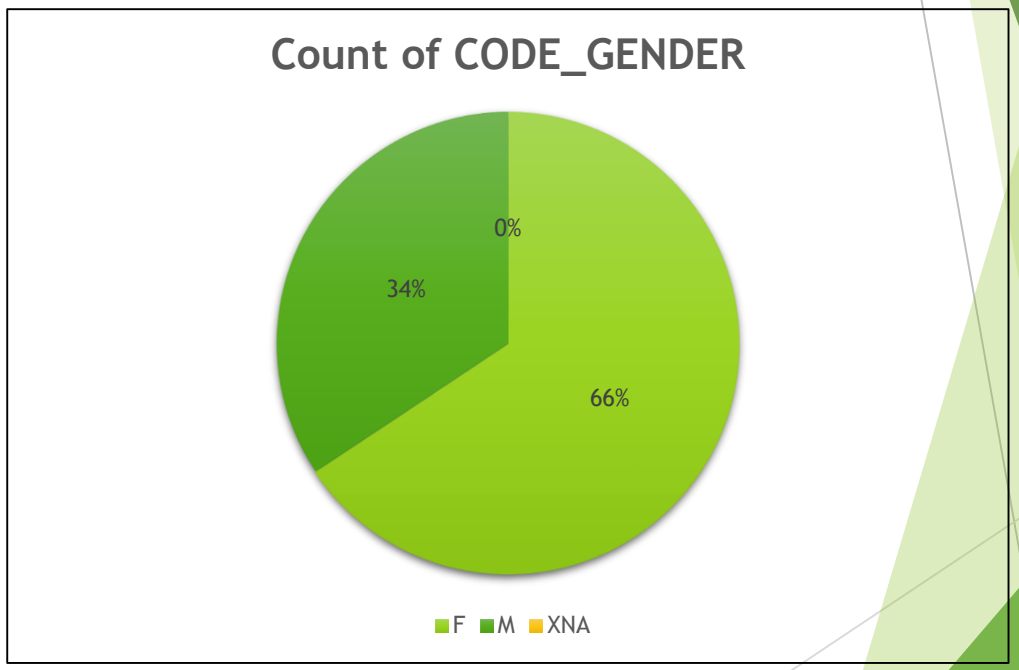


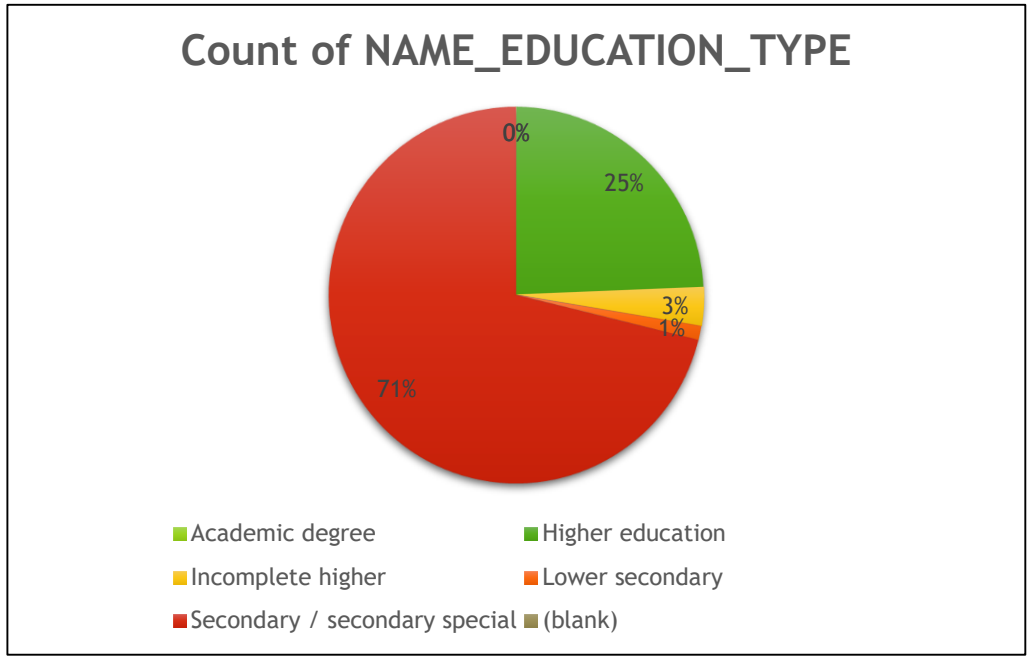

Count of AMT_INCOME_TOTAL
30000
25000
20000
15000
10000
5000
0
(blank)
25650-125650
125650-225650
225650-325650
325650-425650
425650-525650
525650-625650
625650-725650
725650-825650
825650-925650
925650-1025650
1025650-1125650
1225650-1325650
1325650-1425650
1525650-1625650
1725650-1825650
1825650-1925650
1925650-2025650
2225650-2325650
3525650-3625650
3725650-3825650
11692650-117025650
Total

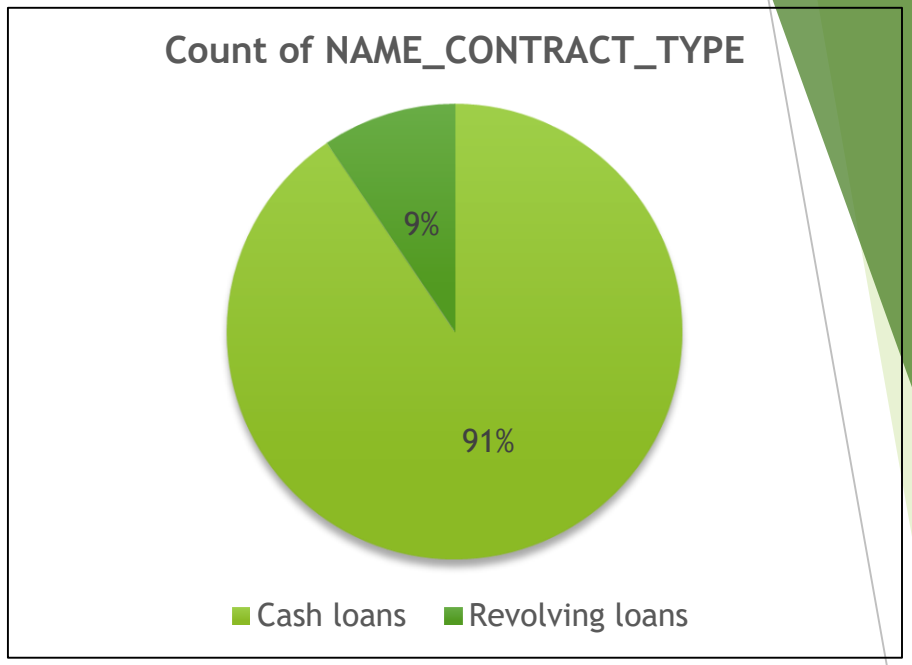
Segmented Univariate Analysis:

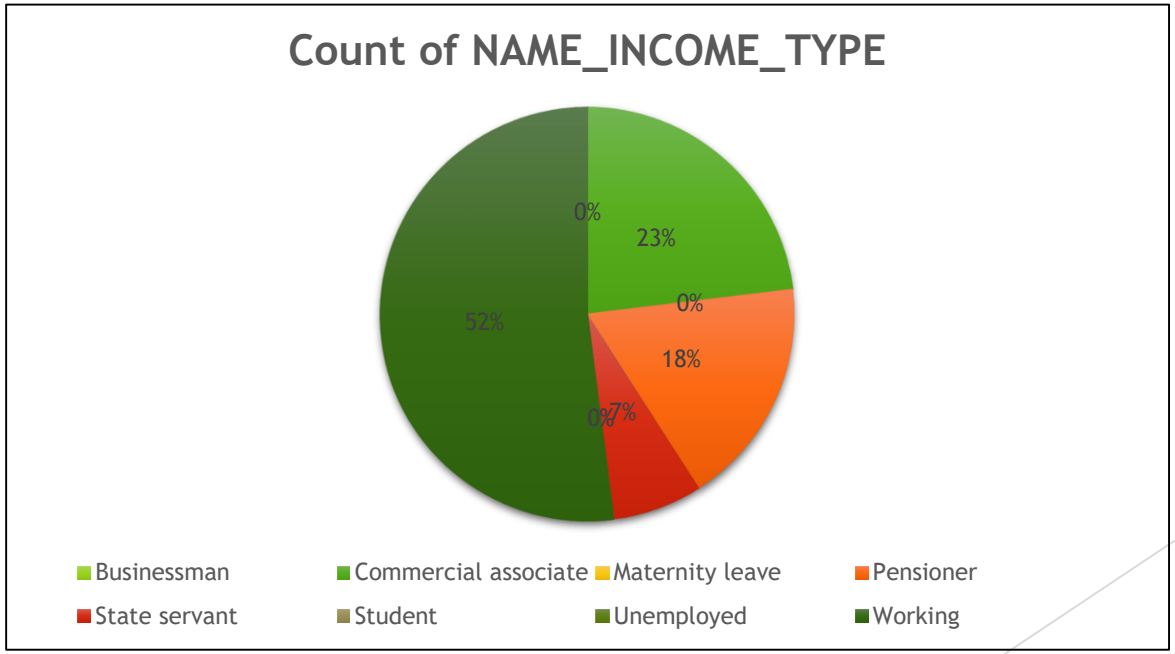
- It is the extension of Univariate Analysis which involves the splitting of data into specific segments or the groups contained.



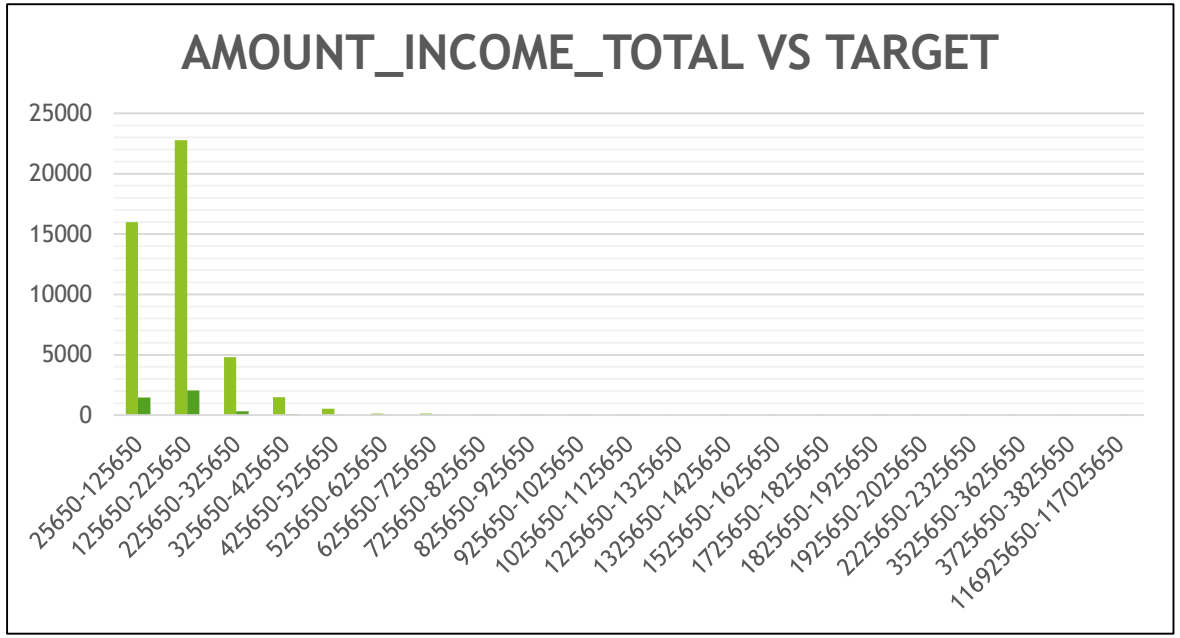


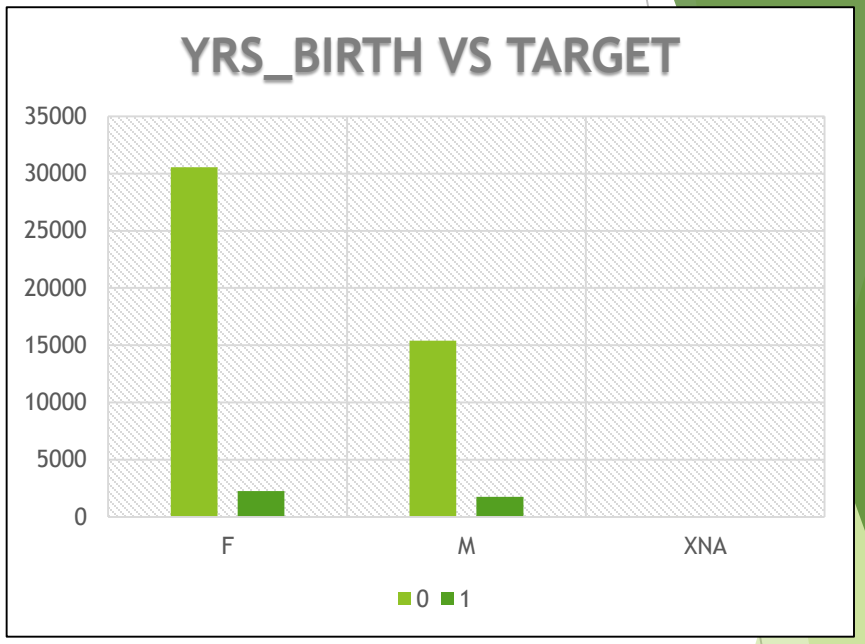

Count of NAME_EDUCATION_TYPE
0%
25%
3%
1%
71%
Academic degree
Higher education
Incomplete higher
Lower secondary
Secondary / secondary special
(blank)

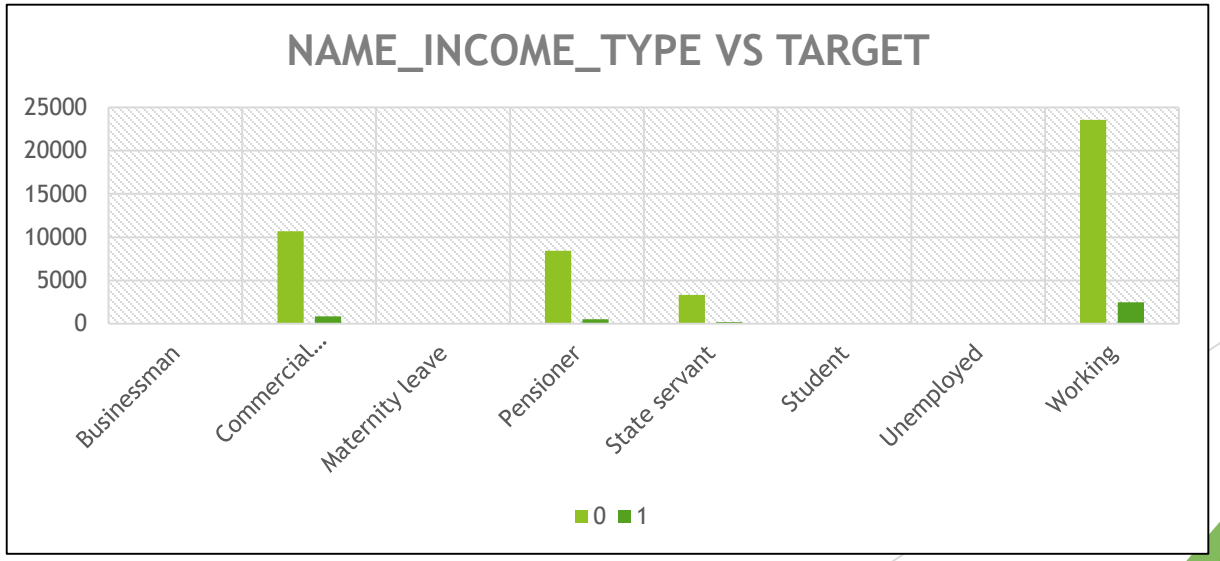

Count of NAME_CONTRACT_TYPE
9%
91%
Cash loans
Revolving loans


Count of NAME_INCOME_TYPE
0%
23%
0%
18%
7%
52%
Businessman
Commercial associate
Maternity leave
Pensioner
State servant
Student
Unemployed
Working

- **Bivariate Analysis**: These examines the relationship between the two variables. Following are the plots obtained while performing the Bivariate Analysis:







E. Identify Top Correlation for Different scenarios.
Task: To determine top correlation for each segmented data(client with payment difficulties and all other cases.

- **The correlation between the variables with TARGET 0**

	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRICE	REGION_POPULATION_RELATIVE	(YRS) Days_Birth	(YRS)Days_Employed_Years	(YRS) Days_registration	REGION_RATING_CLIENT
CNT_CHILDREN	1									
AMT_INCOME_TOTAL	0.00960733	1								
AMT_CREDIT	0.00504255	0.06931916	1							
AMT_ANNUITY	0.02628548	0.0830053	0.76947553	1						
AMT_GOODS_PRICE	0.00029971	0.06989797	0.98670511	0.774116398	1					
REGION_POPULATION_RELATIVE	-0.0255404	0.0298374	0.09506075	0.115076869	0.0991383	1				
(YRS) Days_Birth	-0.3293217	-0.01602	0.0593588	-0.007768753	0.05768474	0.0323967	1			
(YRS)Days_Employed_Years	-0.24155	-0.0315139	-0.0676851	-0.10869655	-0.0649475	-0.0041024	0.62181278	1		
(YRS) Days_registration	-0.1812323	-0.0099449	-0.003466	-0.033255105	-0.0061173	0.05949006	0.33367781	0.209058335	1	
REGION_RATING_CLIENT	0.02575705	-0.0381822	-0.100494	-0.125786534	-0.1036188	-0.5326613	-0.0166943	0.03456358	-0.08755381	1

Highly correlated variables are:

Var 1	Var2	Corr coeff
AMT_ANNUITY	AMT_CREDIT	0.76949879
AMT_GOODS_PRICE	AMT_CREDIT	0.98670439
AMT_GOODS_PRICE	AMT_ANNUITY	0.77413414
(YRS) DAYS_EMPLOYED	(YRS)DAYS_BIRTH	0.62148914

The correlation between the variables with TARGET 1

CNT_CHILDREN	1									
AMT_INCOME_TOTAL	0.00960733	1								
AMT_CREDIT	0.005042552	0.069319162	1							
AMT_ANNUITY	0.026285477	0.083005301	0.769475 53	1						
AMT_GOODS_PRICE	0.000299712	0.069897973	0.986705 11	0.77411639 8	1					
REGION_POPULATION_RELATIVE	-0.025540408	0.029837396	0.095060 75	0.11507686 9	0.099138301	1				
(YRS) Days_Birth	-0.329321666	-0.016019977	0.059358	-0.00776875 3	0.057684741	0.0323967	1			
(YRS)Days_Employed_Years	-0.24154996	-0.031513894	-0.067685 1	-0.10869655	-0.064947528	-0.0041024	0.621812778	1		
(YRS) Days_registration	-0.181232261	-0.00994491	-0.003466	-0.03325510 5	-0.006117338	-0.0594900 6	0.333677806	0.20905833 5	1	
REGION_RATING_CLIENT	0.02575705	-0.038182249	-0.100494	-0.12578653 4	-0.103618752	-0.5326613	-0.016694261	0.03456358	-0.0875538 1	1

The high Correlation coefficients are obtained as:

VAR1	VAR2	Correlation Coeff
AMT_CREDIT	AMT_GOODS_PRICE	0.982268
AMT_ANNUITY	AMT_CREDIT	0.749665
AMT_GOODS_PRICE	AMT_ANNUITY	0.749504
(YRS)DAYS_EMPLOYED	(YRS)DAYS_BIRTH	0.587858

RESULTS

- It is observed that most of the clients become default due to other cases.
- Cash Loans have much higher defaults than Revolving Loans.
- The Education Type Academic Degree has a less number of defaults.
- The Clients with Income ranging between 25000-1025000 has the highest defaults.
- Credit amount of the bank loan is generally falling in the range of 45000-1,45000.
- There are no defaults for clients who are Businessman and students.
- Males are less inclined towards the defaults than the females.

THANK YOU