Bank Loan Case Study

By Hardi Palan

- **Project Description:** The Bank Loan Case Study project, My aim is to use Exploratory Data Analysis (EDA) to analyze patterns in the data and ensure that capable applicants are not rejected. My task is to use Exploratory Data Analysis (EDA) to analyze patterns in the data and ensure that capable applicants are not rejected. Through in-depth data analysis using Excel, Data Visualization and Statistics techniques this project seeks to extract valuable insights and to identify patterns that indicate if a customer will have difficulty paying their installments.
- **Approach:** I have gone through the dataset and understood all the given columns. Then I have observed that there are a total of 128 Columns and 49999 Rows. This dataset consists of unwanted columns, Null values and Blank rows. So, I have decided to Clean this dataset thoroughly using the formulas of COUNTA, COUNTIF, AVERAGE, MEDIAN, MODE and more. I have find the correlations of various columns and from that analyze pattern.
- **Tech-Stack Used:** Used Microsoft Excel for data cleaning and visualization.

1) Identify Missing Data and Deal with it Appropriately:

As a data analyst, you come across missing data in the loan application dataset. It is essential to handle missing data effectively to ensure the accuracy of the analysis.

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

Results: Before Cleaning

(yellow)	Indicate Null values
(red)	Indicate Null values Greater than 25%

1. Identify missing data and deal with appropriately

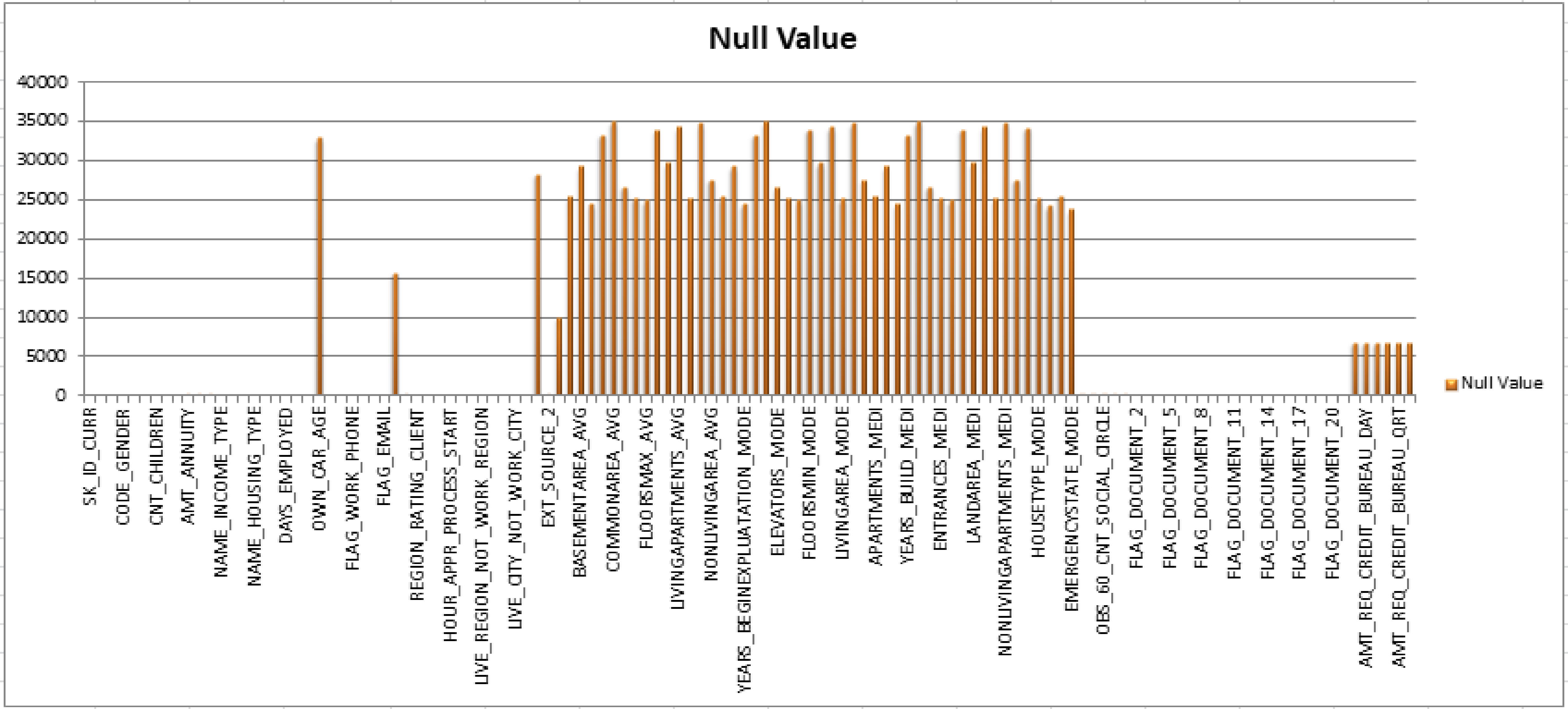
Column	Null Value
SK_ID_CURR	0
TARGET	0
NAME_CONTRACT_TYPE	0
CODE_GENDER	0
FLAG_OWN_CAR	0
FLAG_OWN_REALTY	0
CNT_CHILDREN	0
AMT_INCOME_TOTAL	0
AMT_CREDIT	0
AMT_ANNUITY	1
AMT_GOODS_PRICE	38
NAME_TYPE_SUITE	192
NAME_INCOME_TYPE	0
NAME_EDUCATION_TYPE	0
NAME_FAMILY_STATUS	0
NAME_HOUSING_TYPE	0
REGION_POPULATION_RELATIVE	0
DAYS_BIRTH	0
DAYS_EMPLOYED	0
DAYS_REGISTRATION	0
DAYS_ID_PUBLISH	0

OWN_CAR_AGE	32950
FLAG_MOBIL	0
FLAG_EMP_PHONE	0
FLAG_WORK_PHONE	0
FLAG_CONT_MOBILE	0
FLAG_PHONE	0
FLAG_EMAIL	0
OCCUPATION_TYPE	15654
CNT_FAM_MEMBERS	1
REGION_RATING_CLIENT	0
REGION_RATING_CLIENT_W_CITY	0
WEEKDAY_APPR_PROCESS_START	0
HOUR_APPR_PROCESS_START	0
REG_REGION_NOT_LIVE_REGION	0
REG_REGION_NOT_WORK_REGION	0
LIVE_REGION_NOT_WORK_REGION	0
REG_CITY_NOT_LIVE_CITY	0
REG_CITY_NOT_WORK_CITY	0
LIVE_CITY_NOT_WORK_CITY	0
ORGANIZATION_TYPE	0
EXT_SOURCE_1	28172

EXT_SOURCE_2	126
EXT_SOURCE_3	9944
APARTMENTS_AVG	25385
BASEMENTAREA_AVG	29199
YEARS_BEGINEXPLUATATION_AVG	24394
YEARS_BUILD_AVG	33239
COMMONAREA_AVG	34960
ELEVATORS_AVG	26651
ENTRANCES_AVG	25195
FLOORSMAX_AVG	24875
FLOORSMIN_AVG	33894
LANDAREA_AVG	29721
LIVINGAPARTMENTS_AVG	34226
LIVINGAREA_AVG	25137
NONLIVINGAPARTMENTS_AVG	34714
NONLIVINGAREA_AVG	27572
APARTMENTS_MODE	25385
BASEMENTAREA_MODE	29199
YEARS_BEGINEXPLUATATION_MODE	24394
YEARS_BUILD_MODE	33239

COMMONAREA_MODE	34960
ELEVATORS_MODE	26651
ENTRANCES_MODE	25195
FLOORSMAX_MODE	24875
FLOORSMIN_MODE	33894
LANDAREA_MODE	29721
LIVINGAPARTMENTS_MODE	34226
LIVINGAREA_MODE	25137
NONLIVINGAPARTMENTS_MODE	34714
NONLIVINGAREA_MODE	27572
APARTMENTS_MEDI	25385
BASEMENTAREA_MEDI	29199
YEARS_BEGINEXPLUATATION_MEDI	24394
YEARS_BUILD_MEDI	33239
COMMONAREA_MEDI	34960
ELEVATORS_MEDI	26651
ENTRANCES_MEDI	25195
FLOORSMAX_MEDI	24875
FLOORSMIN_MEDI	33894
LANDAREA_MEDI	29721
LIVINGAPARTMENTS_MEDI	34226
LIVINGAREA_MEDI	25137
NONLIVINGAPARTMENTS_MEDI	34714
NONLIVINGAREA_MEDI	27572
FONDKAPREMONT_MODE	34191
HOUSETYPE_MODE	25075
TOTALAREA_MODE	24148
WALLSMATERIAL_MODE	25459
EMERGENCYSTATE_MODE	23698
OBS_30_CNT_SOCIAL_CIRCLE	168
DEF_30_CNT_SOCIAL_CIRCLE	168

OBS_60_CNT_SOCIAL_CIRCLE	168
DEF_60_CNT_SOCIAL_CIRCLE	168
DAYS_LAST_PHONE_CHANGE	1
FLAG_DOCUMENT_2	0
FLAG_DOCUMENT_3	0
FLAG_DOCUMENT_4	0
FLAG_DOCUMENT_5	0
FLAG_DOCUMENT_6	0
FLAG_DOCUMENT_7	0
FLAG_DOCUMENT_8	0
FLAG_DOCUMENT_9	0
FLAG_DOCUMENT_10	0
FLAG_DOCUMENT_11	0
FLAG_DOCUMENT_12	0
FLAG_DOCUMENT_13	0
FLAG_DOCUMENT_14	0
FLAG_DOCUMENT_15	0
FLAG_DOCUMENT_16	0
FLAG_DOCUMENT_17	0
FLAG_DOCUMENT_18	0
FLAG_DOCUMENT_19	0
FLAG_DOCUMENT_20	0
FLAG_DOCUMENT_21	0
AMT_REQ_CREDIT_BUREAU_HOUR	6734
AMT_REQ_CREDIT_BUREAU_DAY	6734
AMT_REQ_CREDIT_BUREAU_WEEK	6734
AMT_REQ_CREDIT_BUREAU_MON	6734
AMT_REQ_CREDIT_BUREAU_QRT	6734
AMT_REQ_CREDIT_BUREAU_YEAR	6734


Null Value
40000
35000
30000
25000
20000
15000
10000
5000
0
SK_ID_CURR
CODE_GENDER
CNT_CHILDREN
AMT_ANNUITY
NAME_INCOME_TYPE
NAME_HOUSING_TYPE
DAYS_EMPLOYED
OWN_CAR_AGE
FLAG_WORK_PHONE
FLAG_EMAIL
REGION_RATING_CLIENT
HOUR_APPR_PROCESS_START
LIVE_REGION_NOT_WORK_REGION
LIVE_CITY_NOT_WORK_CITY
EXT_SOURCE_2
BASEMENTAREA_AVG
COMMONAREA_AVG
FLOORSMAX_AVG
LIVINGAPARTMENTS_AVG
NONLIVINGAREA_AVG
YEARS_BEGINEXPLUATATION_MODE
ELEVATORS_MODE
FLOORSMIN_MODE
LIVINGAREA_MODE
APARTMENTS_MEDI
YEARS_BUILD_MEDI
ENTRANCES_MEDI
LANDAREA_MEDI
NONLIVINGAPARTMENTS_MEDI
HOUSETYPE_MODE
EMERGENCYSTATE_MODE
OBS_60_CNT_SOCIAL_CIRCLE
FLAG_DOCUMENT_2
FLAG_DOCUMENT_5
FLAG_DOCUMENT_8
FLAG_DOCUMENT_11
FLAG_DOCUMENT_14
FLAG_DOCUMENT_17
FLAG_DOCUMENT_20
AMT_REQ_CREDIT_BUREAU_DAY
AMT_REQ_CREDIT_BUREAU_QRT
Null Value

- **Results: After Cleaning**

1. Removed and Replaced missing data with Median and Mode

Column	Median
AMT_ANNUITY	24939
AMT_GOODS_PRICE	450000
CNT_FAM_MEMBERS	2
EXT_SOURCE_2	0.5655854
EXT_SOURCE_3	0.5352763
OBS_30_CNT_SOCIAL_CIRCLE	0
DEF_30_CNT_SOCIAL_CIRCLE	0
OBS_60_CNT_SOCIAL_CIRCLE	0
DEF_60_CNT_SOCIAL_CIRCLE	0
DAYS_LAST_PHONE_CHANGE	-755
AMT_REQ_CREDIT_BUREAU_HOUR	0
AMT_REQ_CREDIT_BUREAU_DAY	0
AMT_REQ_CREDIT_BUREAU_WEEK	0
AMT_REQ_CREDIT_BUREAU_MON	0
AMT_REQ_CREDIT_BUREAU_QRT	0
AMT_REQ_CREDIT_BUREAU_YEAR	1

- I have used these values to replace the null values in the columns which has null values less than 25%.

- **Insights:** There are many missing values in the dataset. The columns having null values above 25% are deleted and the missing values are replaced using median and mode.

2) Identify Outliers in the Dataset:

Outliers can significantly impact the analysis and distort the results. You need to identify outliers in the loan application dataset.

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

2. Identifying Outliers in the dataset

A. CNT_CHILDREN

Calculation	Values
QUARTILE Q1	0
QUARTILE Q3	1
Inter Quartile Range IQR	1
Lower Bound	-1.5
Upper Bound	2.5

B. AMT_INCOME_TOTAL

Calculation	Values
QUARTILE Q1	112500
QUARTILE Q3	202500
Inter Quartile Range IQR	90000
Lower Bound	-22500
Upper Bound	337500

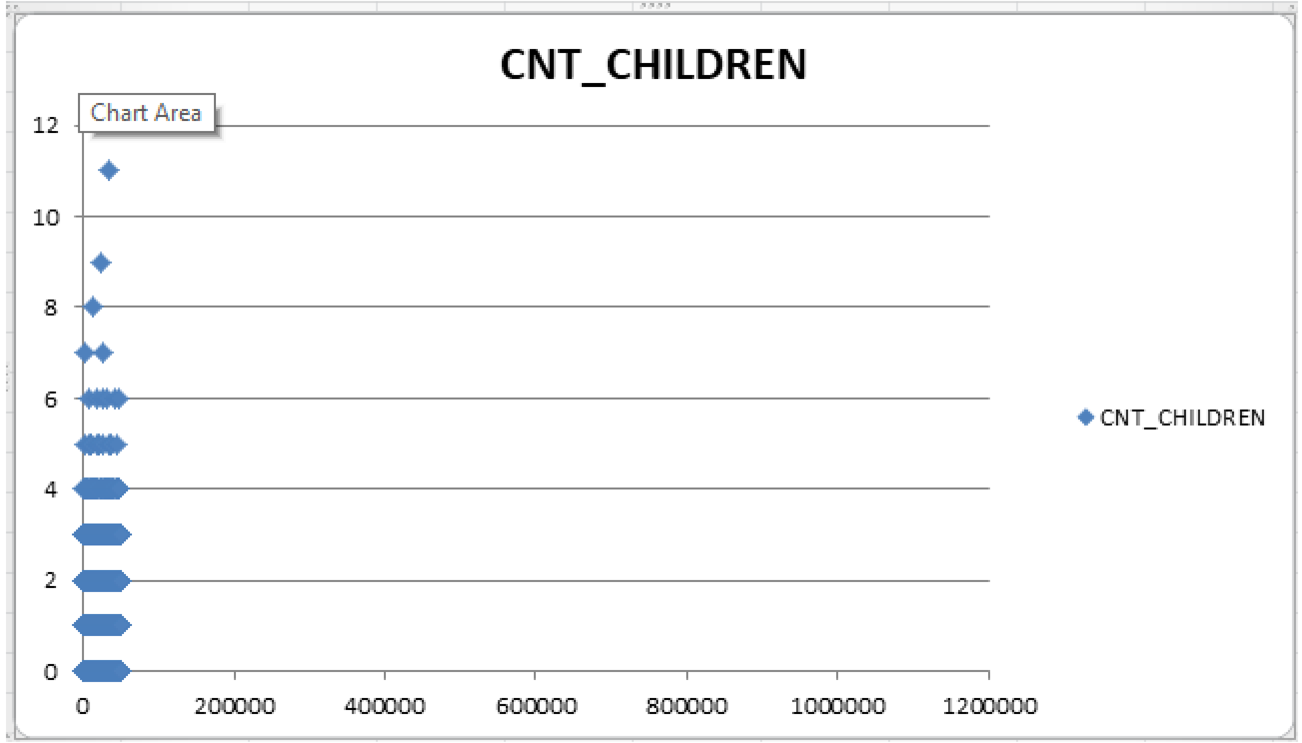
C. AMT_CREDIT

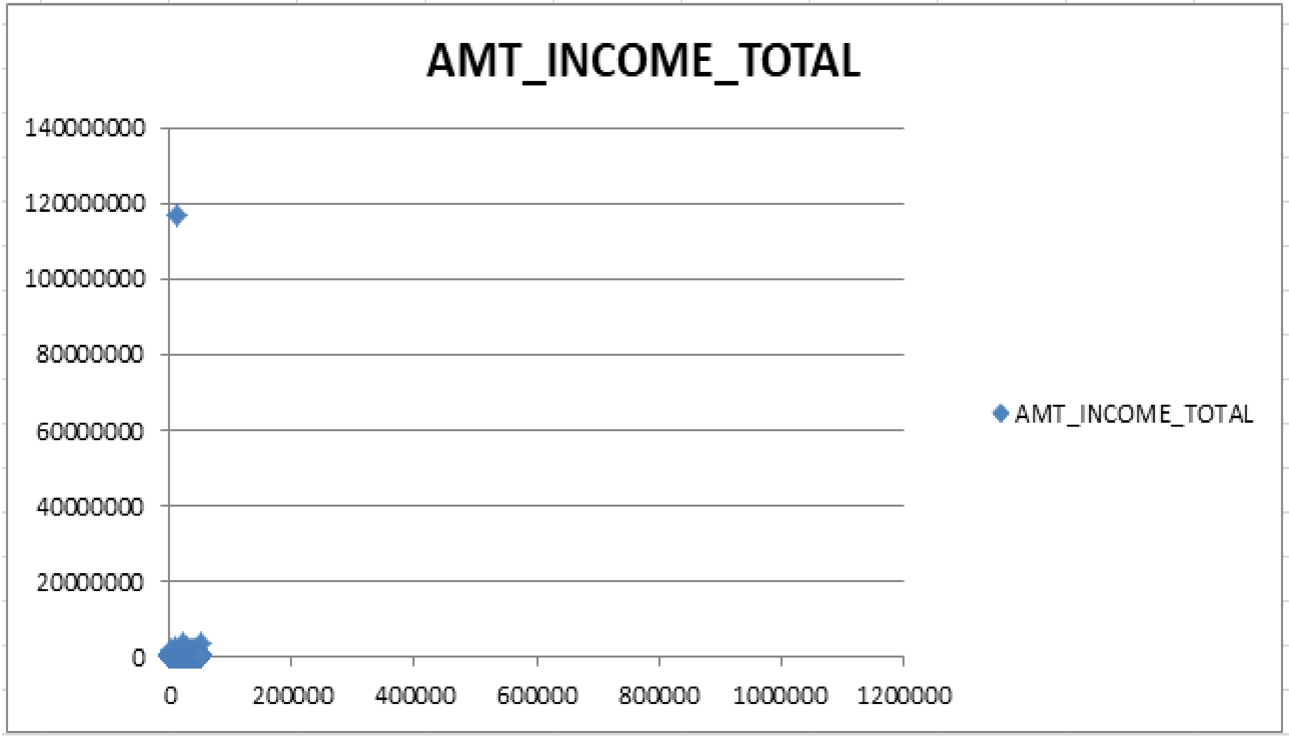
Calculation	Values
QUARTILE Q1	270000
QUARTILE Q3	808650
Inter Quartile Range IQR	538650
Lower Bound	-537975
Upper Bound	1616625

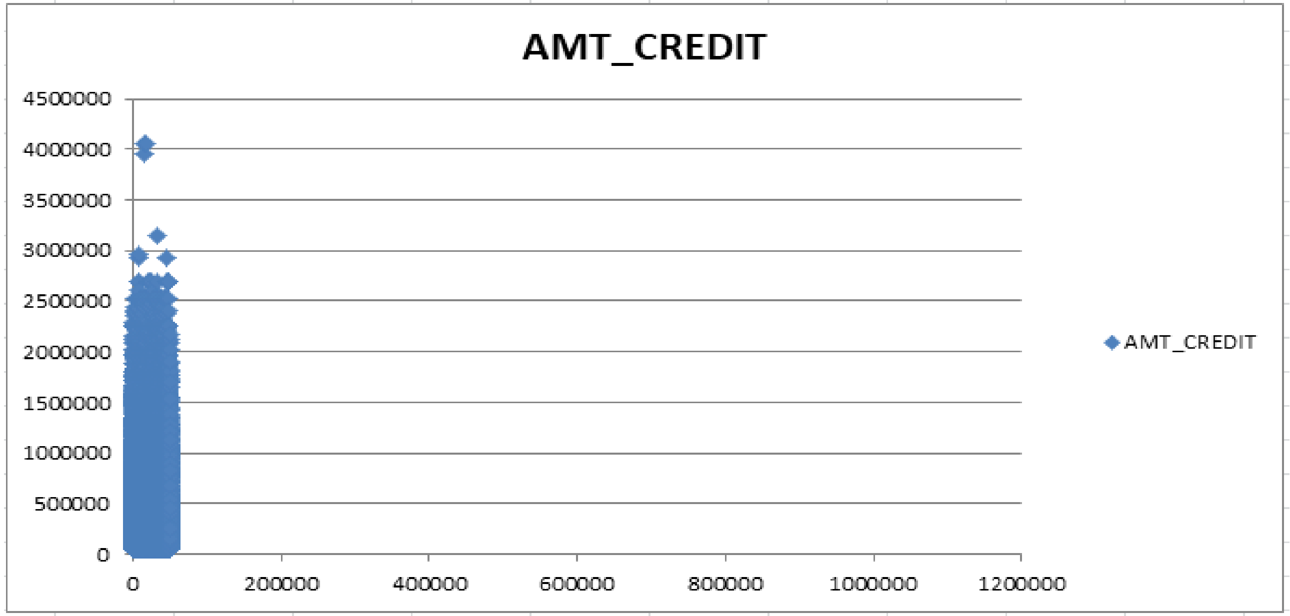
D. AMT_ANNUITY	
Calculation	Values
QUARTILE Q1	16456.5
QUARTILE Q3	34596
Inter Quartile Range IQR	18139.5
Lower Bound	-10752.75
Upper Bound	61805.25

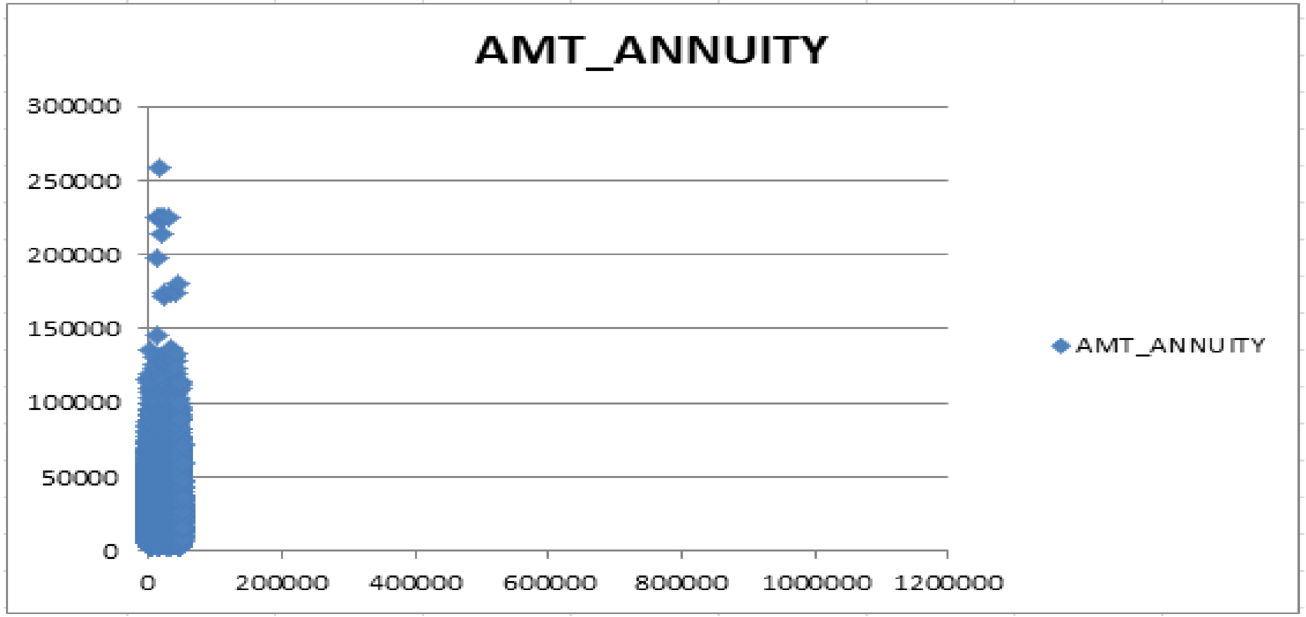
E. AMT_GOODS_PRICE	
Calculation	Values
QUARTILE Q1	238500
QUARTILE Q3	679500
Inter Quartile Range IQR	441000
Lower Bound	-423000
Upper Bound	1341000

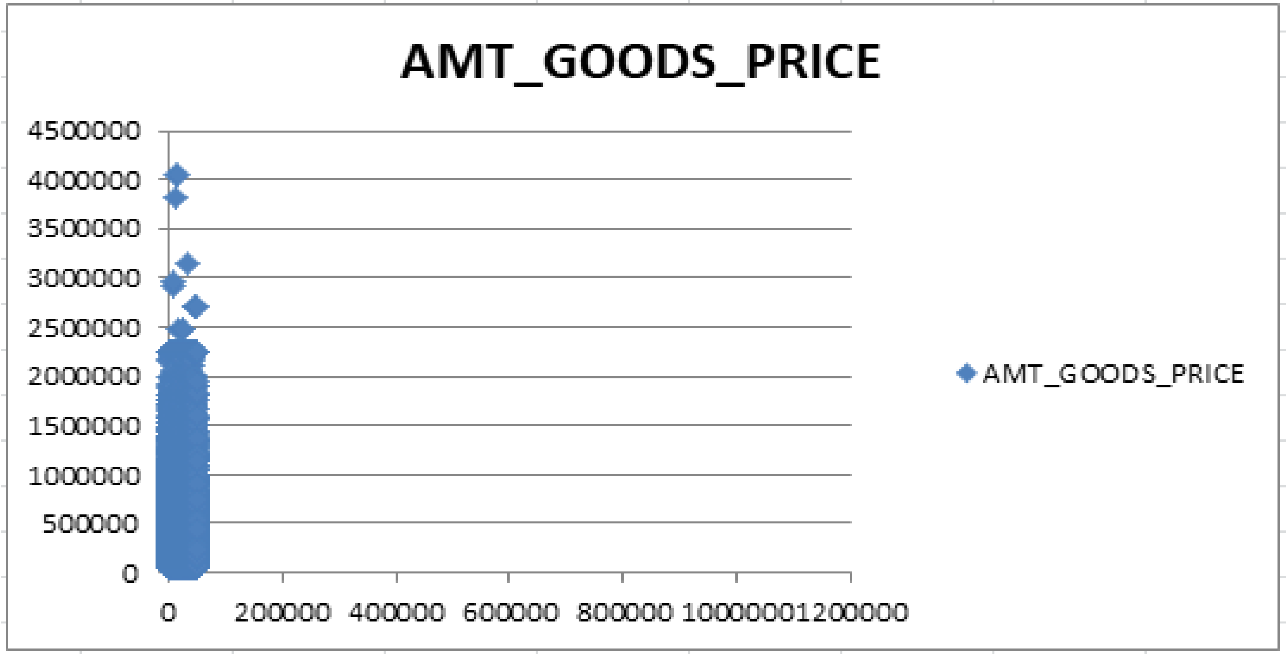
F. DAYS_BIRTH	
Calculation	Values
QUARTILE Q1	-19644
QUARTILE Q3	-12378
Inter Quartile Range IQR	7266
Lower Bound	-30543
Upper Bound	-1479

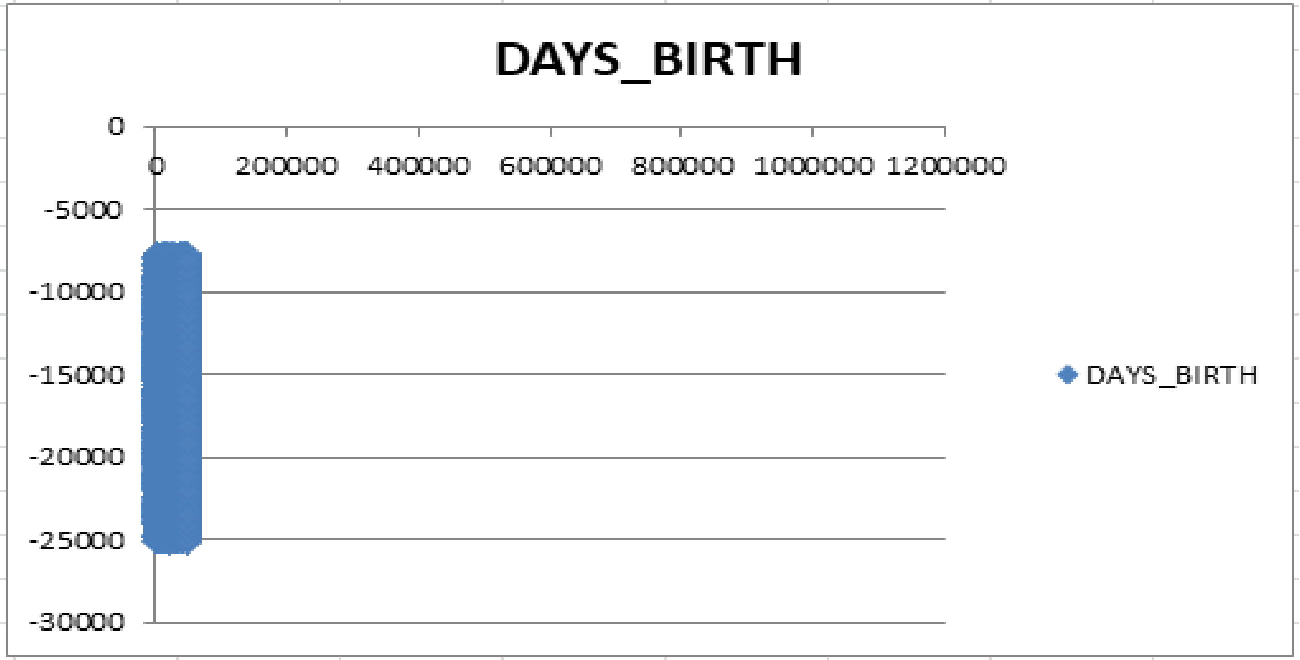





AMT_CREDIT
4500000
4000000
3500000
3000000
2500000
2000000
1500000
1000000
500000
0
0
200000
400000
600000
800000
1000000
1200000
AMT_CREDIT


AMT_ANNUITY
300000
250000
200000
150000
100000
50000
0
0
200000
400000
600000
800000
1000000
1200000
AMT_ANNUITY


AMT_GOODS_PRICE
4500000
4000000
3500000
3000000
2500000
2000000
1500000
1000000
500000
0
0
200000
400000
600000
800000
10000001200000
AMT_GOODS_PRICE


DAYS_BIRTH
0
0
200000
400000
600000
800000
1000000
1200000
-5000
-10000
-15000
-20000
-25000
-30000
DAYS_BIRTH

3) Analyze Data Imbalance:

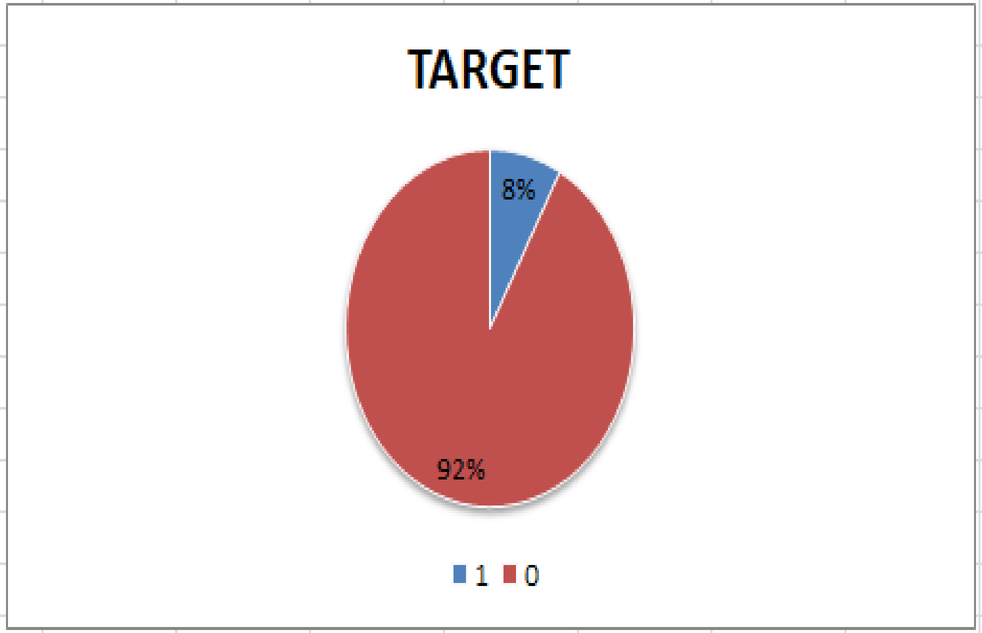
Data imbalance can affect the accuracy of the analysis, especially for binary classification problems. Understanding the data distribution is crucial for building reliable models.

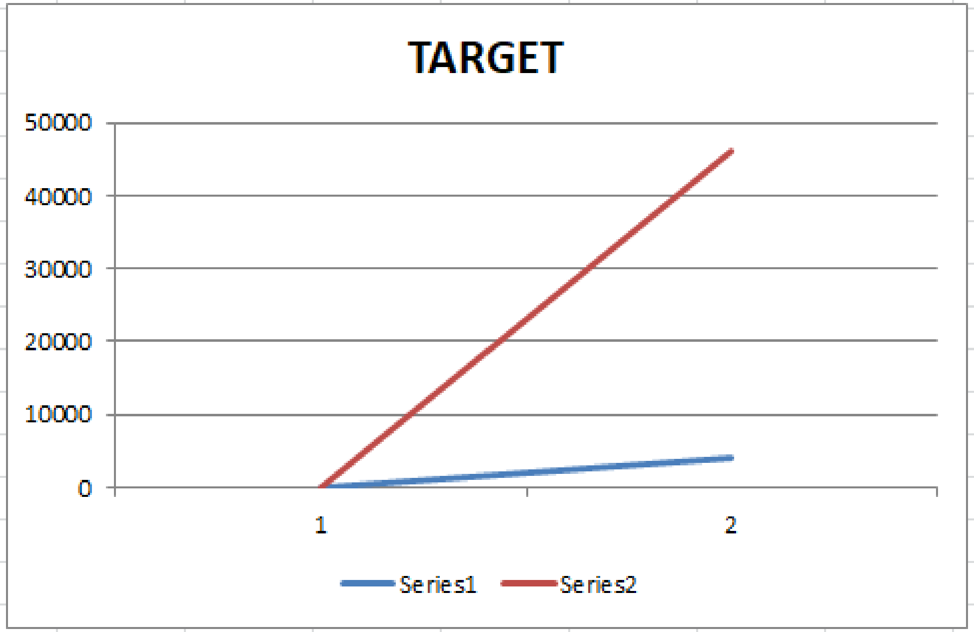
Task: Determine if there is data imbalance in the loan application dataset and calculate the ratio of data imbalance using Excel functions.

Results:

5. Analyze Data Imbalance

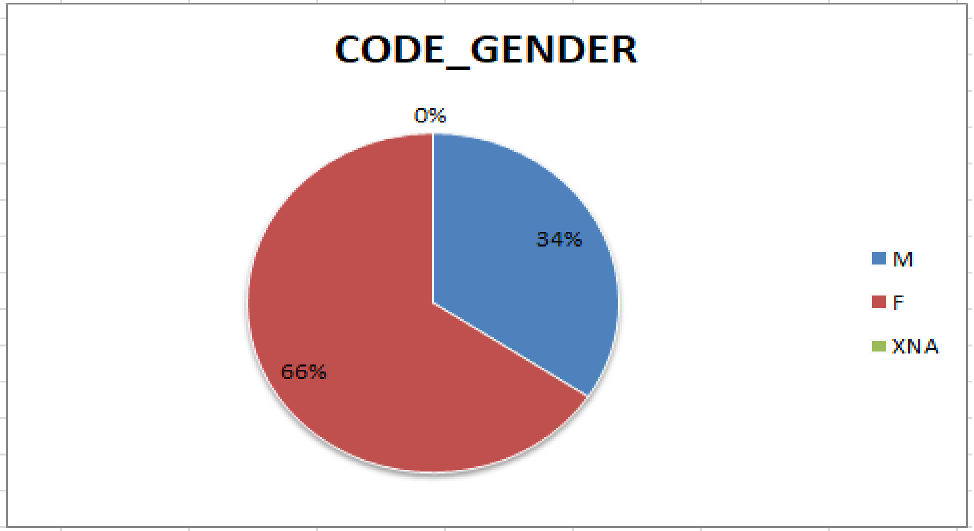
Target	Occurrence
1	4026
0	45973

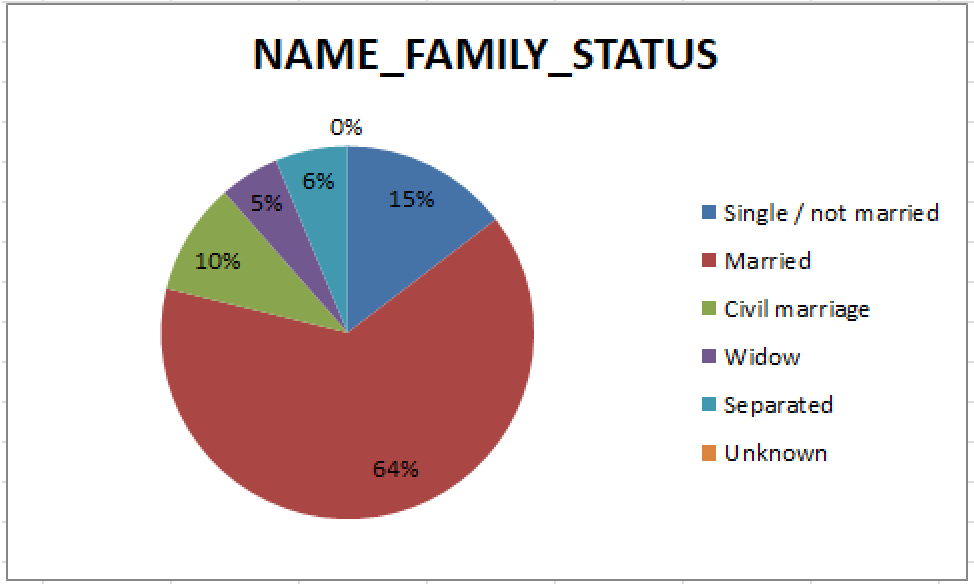




CODE_GENDER	Occurrence
M	17174
F	32823
XNA	2

NAME_FAMILY_STATUS	Occurrence
Single / not married	7306
Married	32094
Civil marriage	4859
Widow	2597
Separated	3142
Unknown	1





- **Insights:** People who has low income, Married, Working and has age 38-39 years have taken the loan mostly and also they are most likely to default the loan.

4) Perform Univariate, Segmented Univariate and Bivariate Analysis:

To gain insights into the driving factors of loan default, it is important to conduct various analyses on consumer and loan attributes.

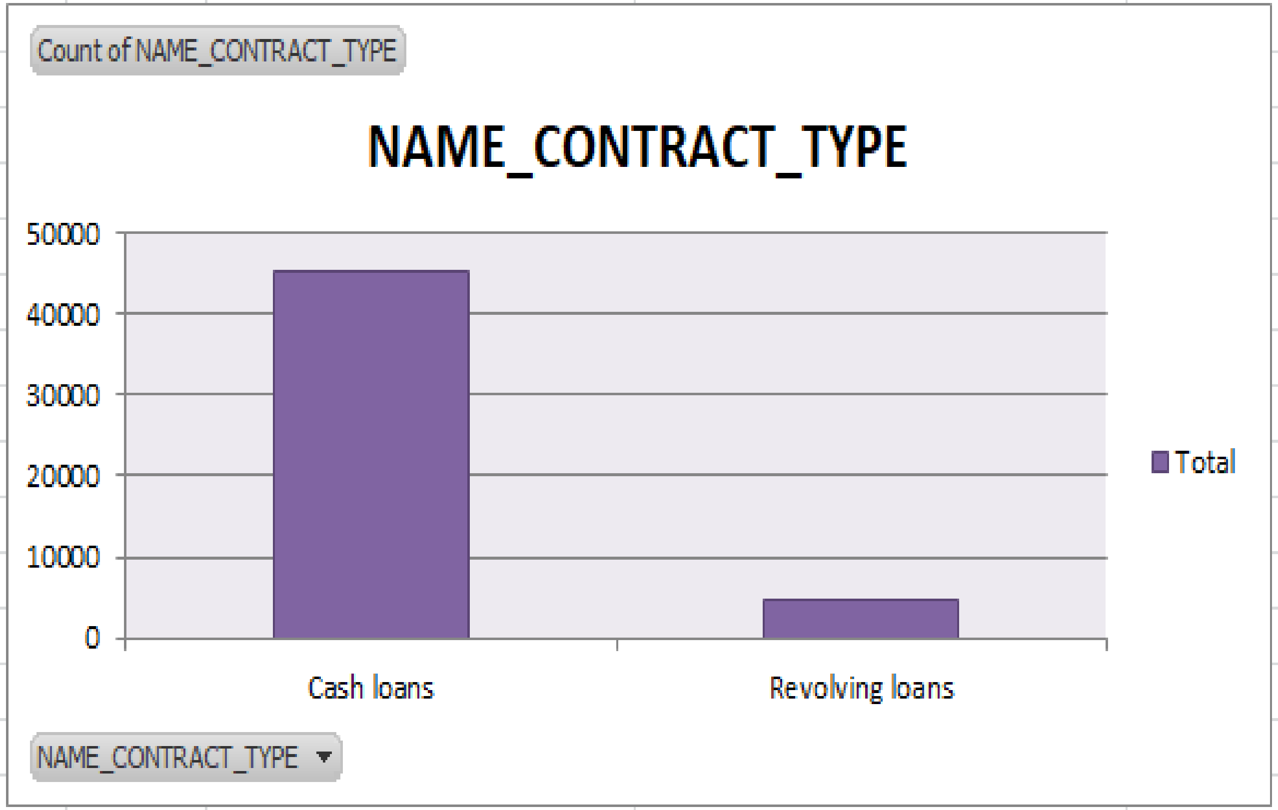
Task: Perform univariate analysis to understand the distribution of individual variables, segmented univariate analysis to compare variable distributions for different scenarios, and bivariate analysis to explore relationships between variables and the target variable using Excel functions and features.

Results:

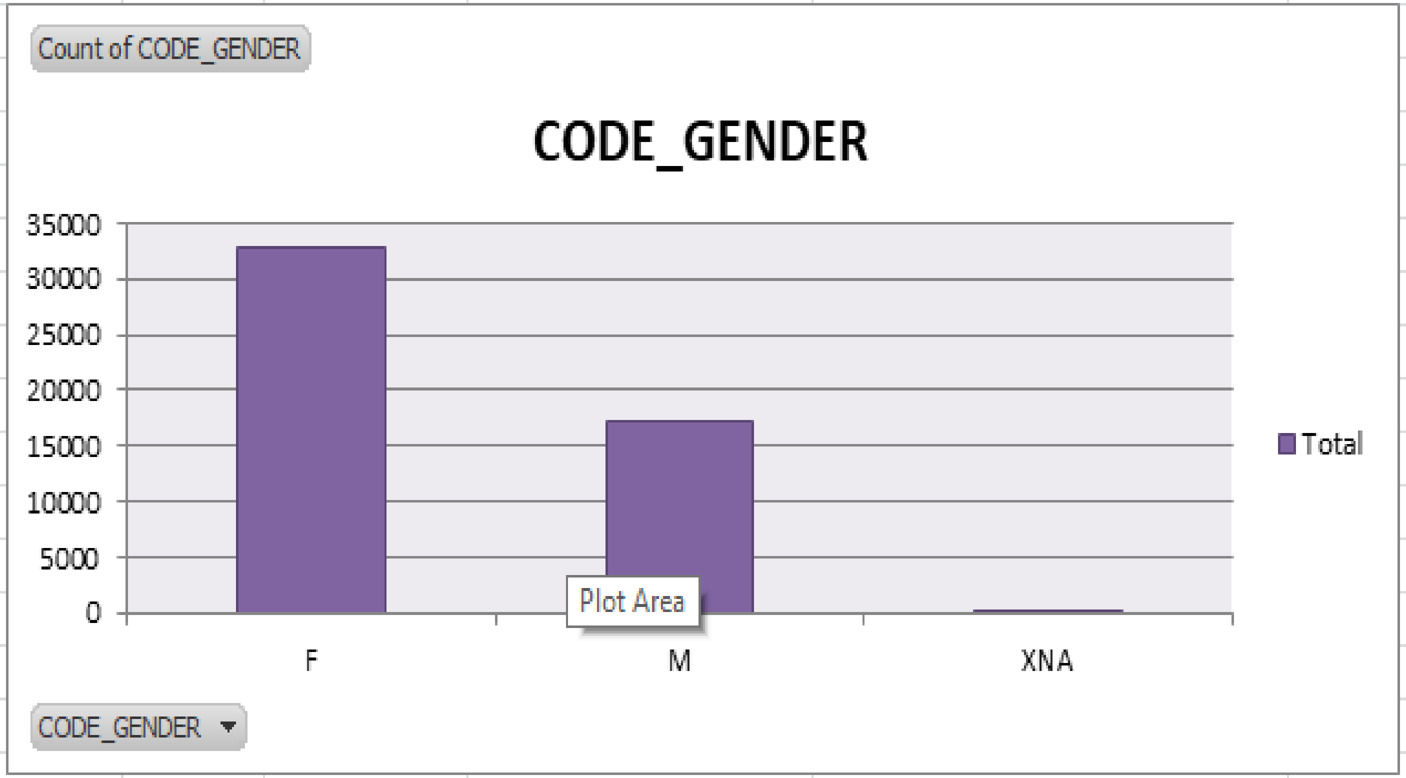
4. Perform Univariate Analysis:

Values	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY
Average	170767.5905	599700.5815	27107.33399
Median	145800	514777.5	24939
Mode	135000	450000	9000
StdDev	531819.0951	402415.4339	14562.80203

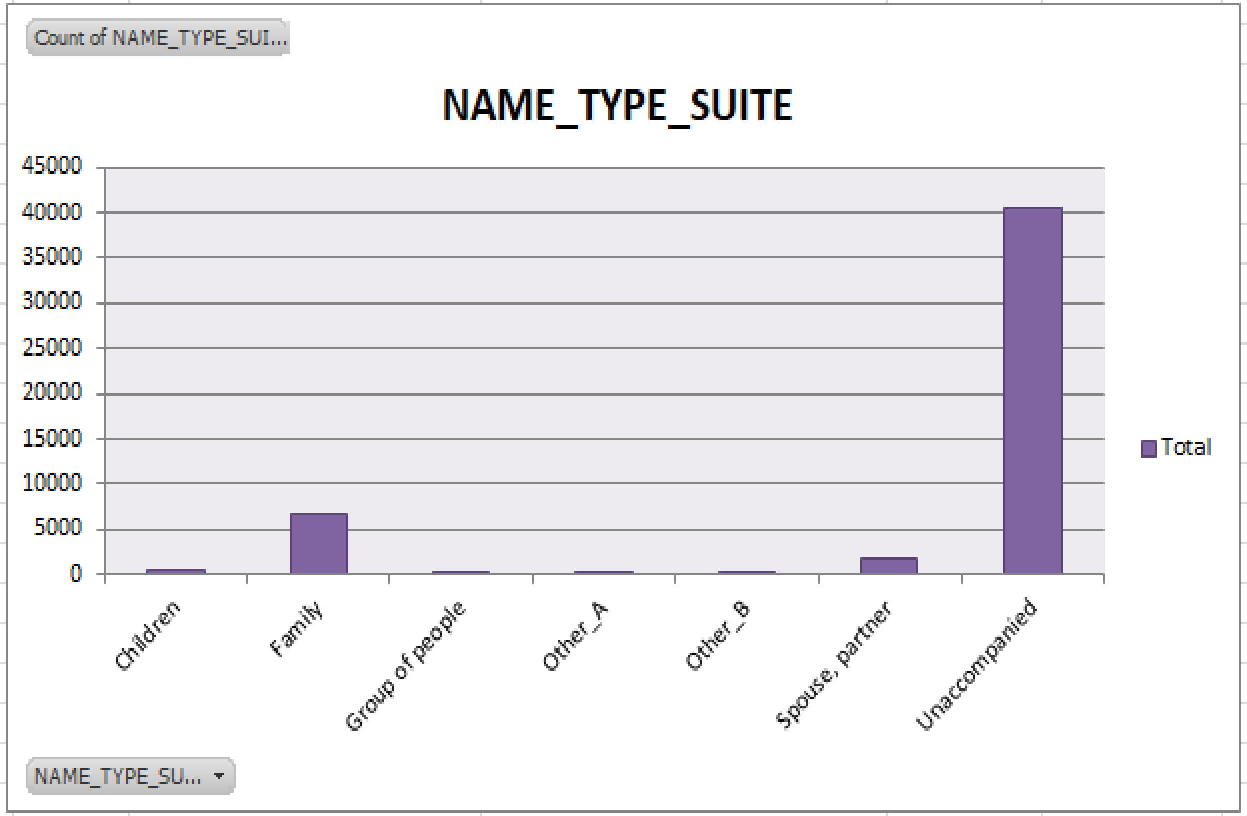
Row Labels	Count of NAME_CONTRACT_TYPE
Cash loans	45276
Revolving loans	4723
Grand Total	**49999**



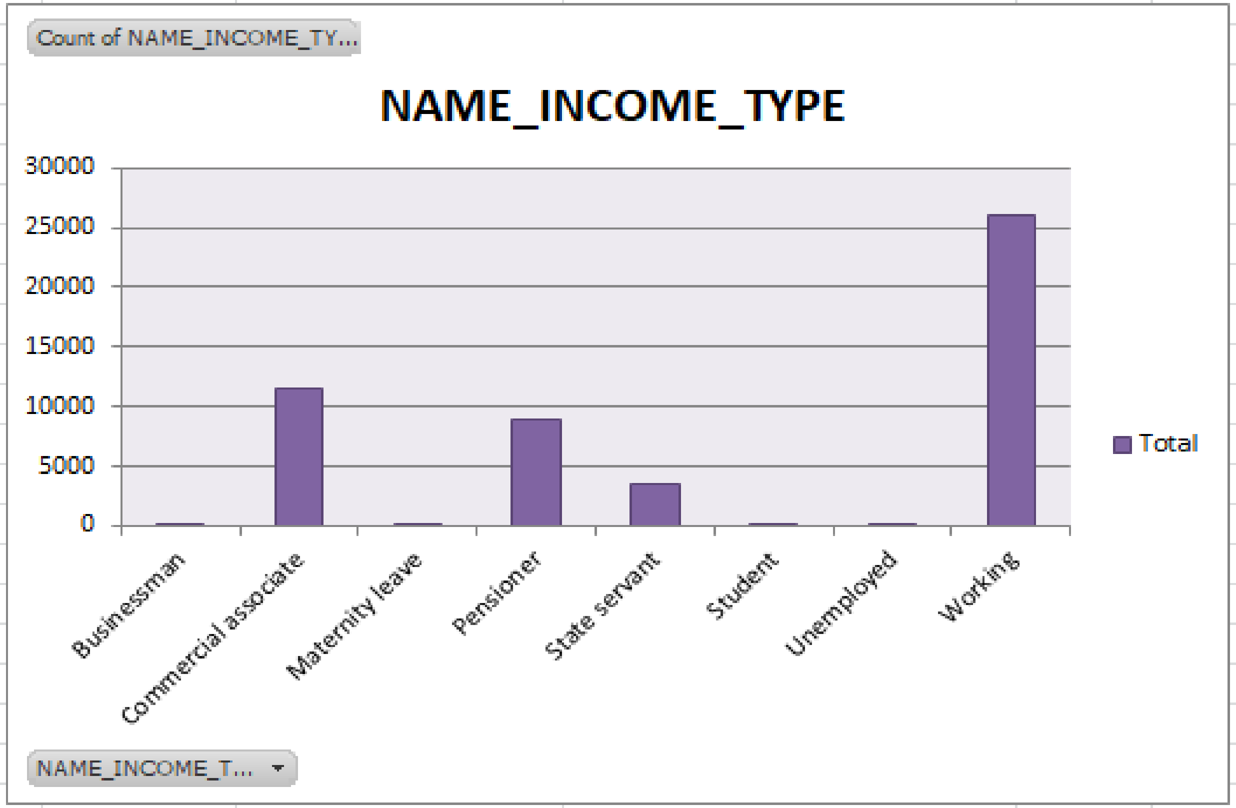
Row Labels	Count of CODE_GENDER
F	32823
M	17174
XNA	2
Grand Total	**49999**



Row Labels	Count of NAME_TYPE_SUITE
Children	546
Family	6577
Group of people	37
Other_A	139
Other_B	260
Spouse, partner	1855
Unaccompanied	40585
Grand Total	**49999**

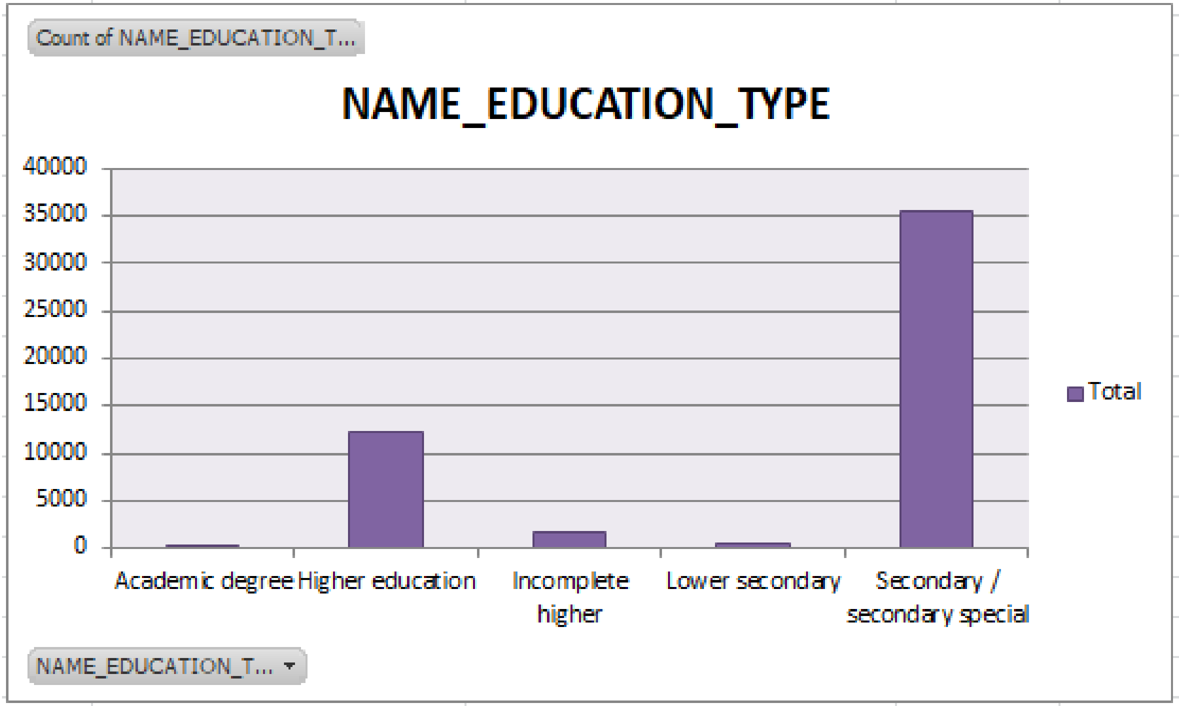


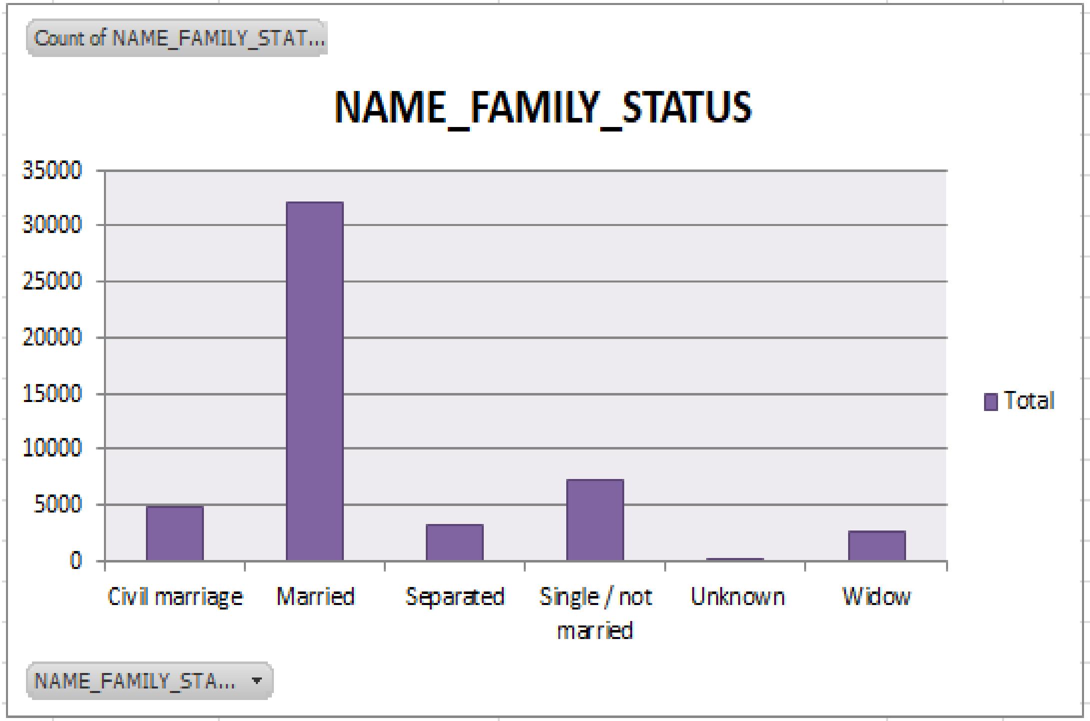
Row Labels	Count of NAME_INCOME_TYPE
Businessman	2
Commercial associate	11543
Maternity leave	1
Pensioner	8920
State servant	3512
Student	5
Unemployed	6
Working	26010
Grand Total	**49999**



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

Row Labels	Count of NAME_FAMILY_STATUS
Civil marriage	4859
Married	32094
Separated	3142
Single / not married	7306
Unknown	1
Widow	2597
Grand Total	**49999**



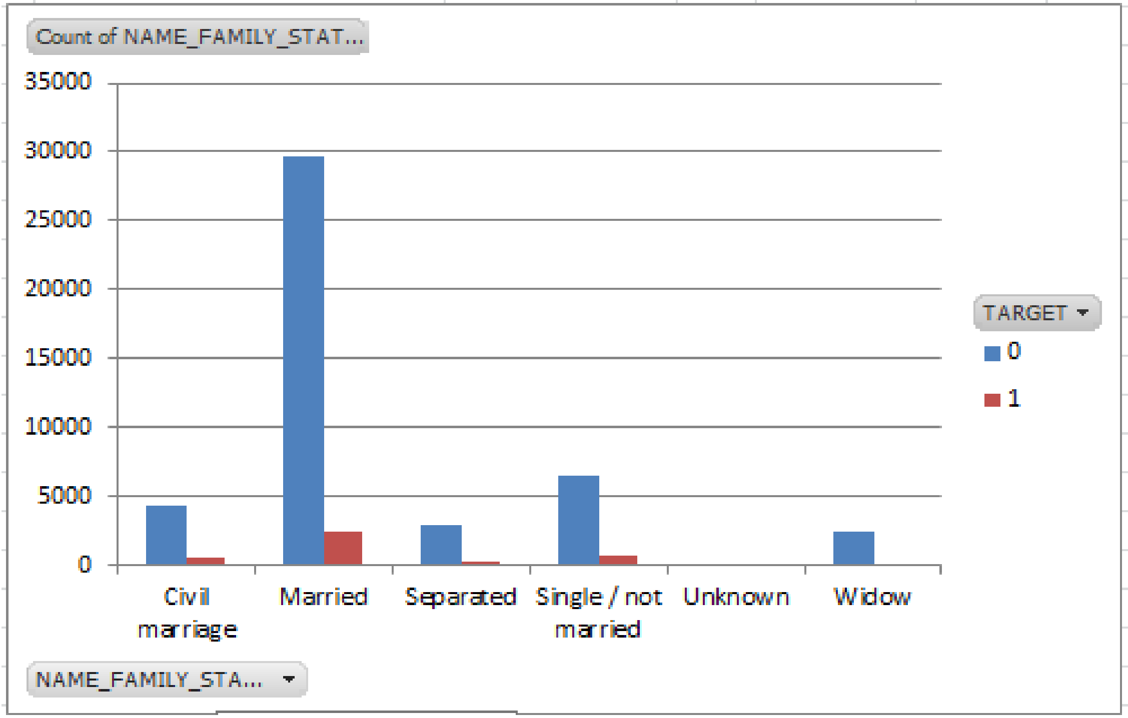


- **Insights:** Majority of the applicants were offered loans in the credit range of 9 Lacs and above.

4. Performing Segmented Univariate and Bivariate Analysis:

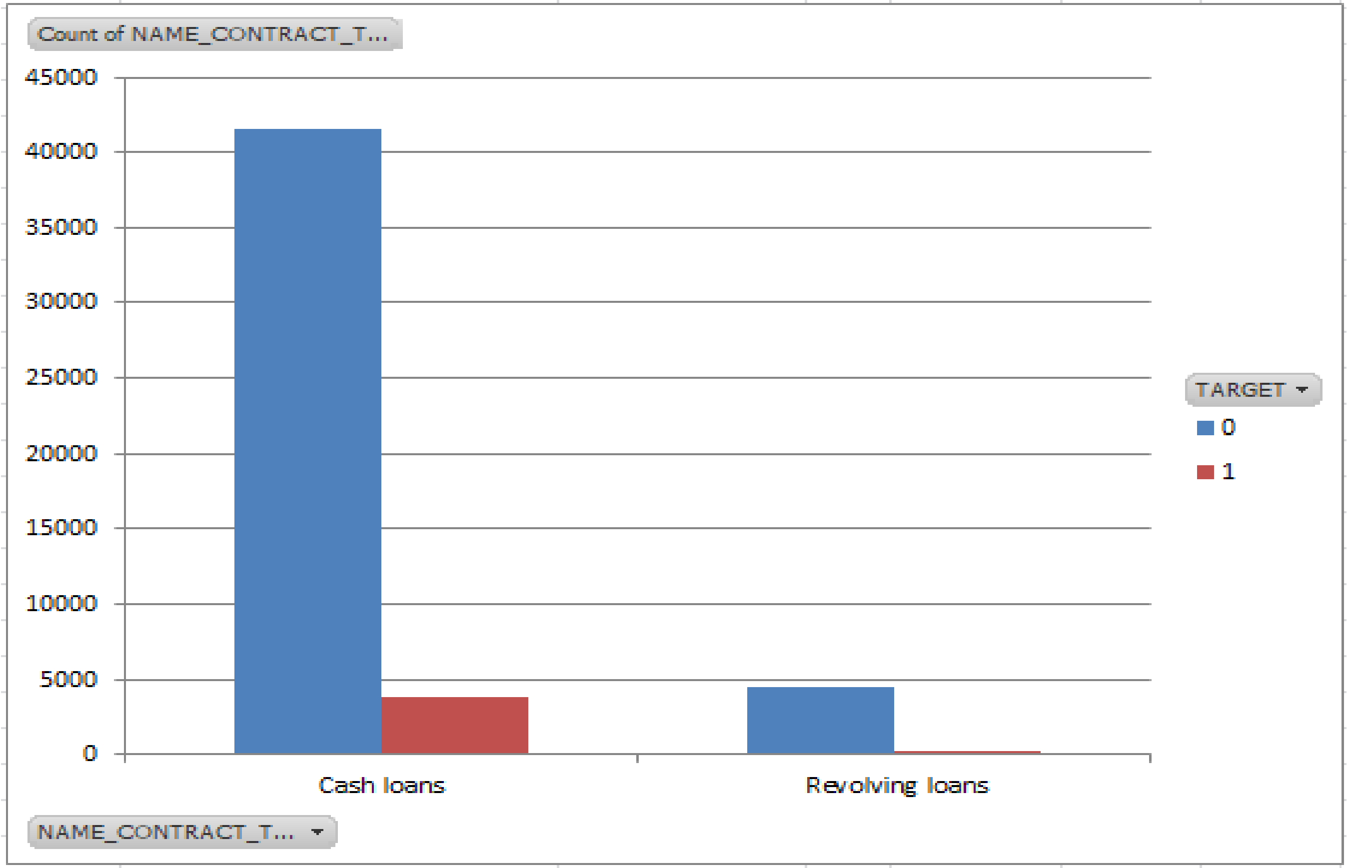
B. NAME_FAMILY_STATUS VS TARGET

Count of NAME_FAMILY_STATUS Row Labels	Column Labels 0	1	Grand Total
Civil marriage	4377	482	4859
Married	29699	2395	32094
Separated	2870	272	3142
Single / not married	6577	729	7306
Unknown	1		1
Widow	2449	148	2597
Grand Total	**45973**	**4026**	**49999**



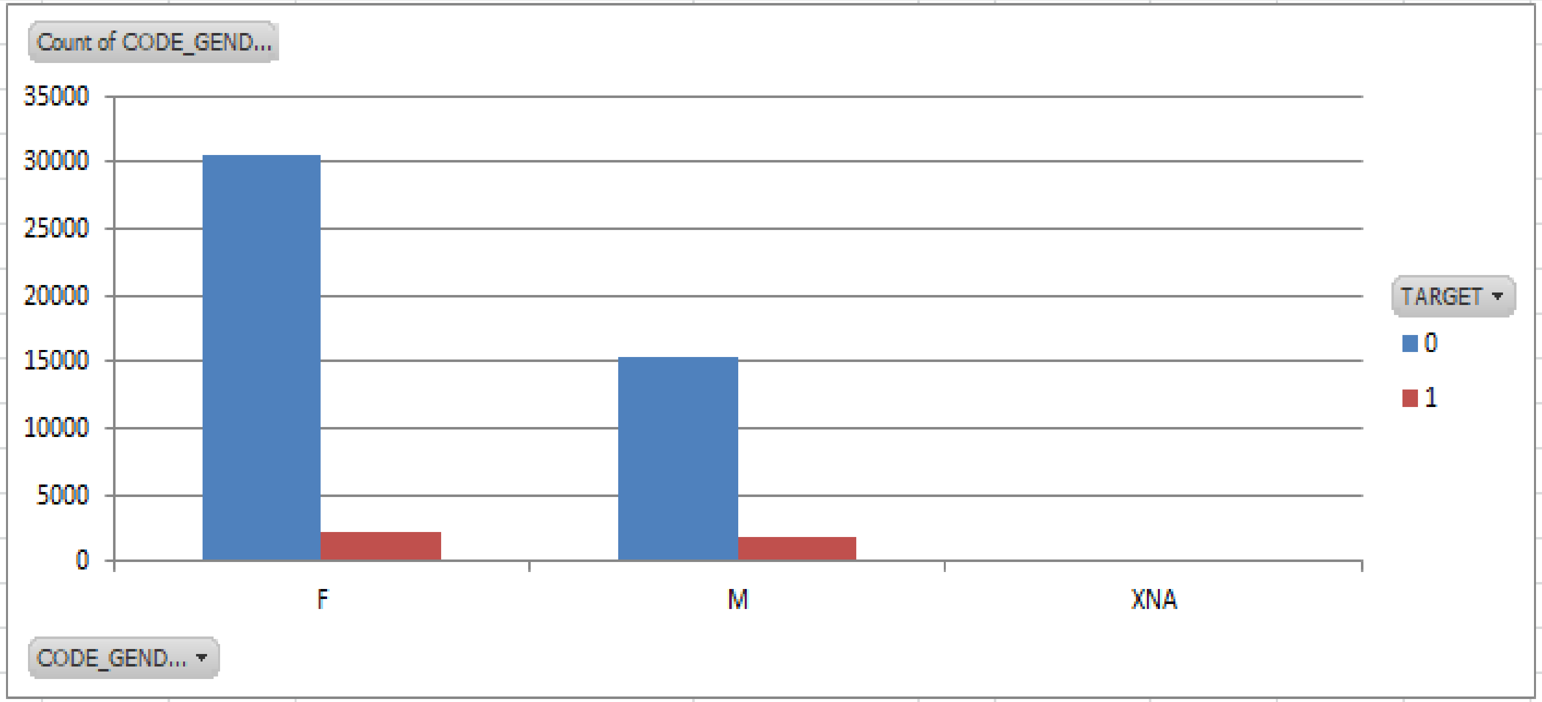
D. NAME_CONTRACT_TYPE VS TARGET

Count of NAME_CONTRACT_TYPE Row Labels	Column Labels 0	1	Grand Total
Cash loans	41484	3792	45276
Revolving loans	4489	234	4723
Grand Total	**45973**	**4026**	**49999**



C. CODE_GENDER VS TARGET

Count of CODE_GENDER	Column Labels		
Row Labels	0	1	Grand Total
F	30559	2264	32823
M	15412	1762	17174
XNA	2		2
Grand Total	45973	4026	49999



- **Insights:** there are very few targets 1 applicant who draw an income of more than 50 Lacs and above which can be the reason for the difficulties in the payments. Also, maximum applicants (0,1) draw an income between 1.25 Lacs to 1.5 Lacs but there are applicants which are having payment difficulties despite belonging to the same income range.

5) Identify Top Correlations for Different Scenarios:

Understanding the correlation between variables and the target variable can provide insights into strong indicators of loan default.

Task: Segment the dataset based on different scenarios (e.g., clients with payment difficulties and all other cases) and identify the top correlations for each segmented data using Excel functions.

Results:

5. Identify Top Correlations for Different Scenarios:

NOTE THAT THESE ARE THE CORRELATIONS FOR TARGET 1

Column	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	DAYS_BIRTH	DAYS_EMPLOYED	DAYS_ID_PUBLISH	REGION_RATING_CLIENT
CNT_CHILDREN	1						
AMT_INCOME_TOTAL	0.036319722	1					
AMT_CREDIT	0.005705458	0.377965752	1				
DAYS_BIRTH	0.335876269	0.073769425	-0.05108418	1			
DAYS_EMPLOYED	-0.243591518	-0.162702675	-0.07736722	-0.61528998	1		
DAYS_ID_PUBLISH	-0.032537221	0.032286356	-0.00829019	0.270073313	-0.27222439	1	
REGION_RATING_CLIENT	0.021288992	-0.205031899	-0.10255648	0.00902485	0.040505636	-0.008097427	1

NOTE THAT THESE ARE THE CORRELATIONS FOR TARGET 1							
Column	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	DAYS_BIRTH	DAYS_EMPLOYED	DAYS_ID_PUBLISH	REGION_RATING_CLIENT
CNT_CHILDREN	1						
AMT_INCOME_TOTAL	0.010110177	1					
AMT_CREDIT	0.007601905	0.015271444	1				
DAYS_BIRTH	0.2496732	0.009033662	-0.142506035	1			
DAYS_EMPLOYED	-0.189324184	-0.011555963	0.016039571	-0.58147904	1		
DAYS_ID_PUBLISH	-0.042360717	-0.009122006	-0.043771901	0.247896571	-0.230063668	1	
REGION_RATING_CLIENT	0.055515557	-0.012846697	-0.045024534	0.045027112	-0.009145883	-0.008097427	1

- **Insights:** There are many correlations between the columns and the highest correlated column is DAYS_BIRTH. Dark pink is the weakest correlation.

- **Conclusion:** This project helps in handling the large datasets. How exploratory data analysis can be applied to large datasets. When dealing with the large datasets it is also important to select only those columns which are extremely useful to our analysis. Finding correlations columns can become very convenient while dealing with large datasets as it saves time selecting which columns should be considered for analysis. The project also helps in understanding the various terminologies used in the banking domain.

- **Link for Resultant Dataset:** **https://docs.google.com/spreadsheets/d/1hY9jMqhbbmnWij0kBmjzCBfGj_gK5I7_/edit?usp=sharing&ouid=101431809048624548912&rtpof=true&sd=true**

Thank You