Bank Loan Case Study Analysis

Project Description

- Our role is as a data analyst at a finance company that specializes in lending various types of loans to urban customers. our company faces a challenge: some customers who don't have a sufficient credit history take advantage of this and default on their loans. Your task is to use Exploratory Data Analysis (EDA) to analyze patterns in the data and ensure that capable applicants are not rejected.
- When a customer applies for a loan, your company faces two risks:
- If the applicant can repay the loan but is not approved, the company loses business.
- If the applicant cannot repay the loan and is approved, the company faces a financial loss.
- When a customer applies for a loan, there are four possible outcomes:
- Approved: The company has approved the loan application.
- Cancelled: The customer cancelled the application during the approval process.
- Refused: The company rejected the loan.
- Unused Offer: The loan was approved but the customer did not use it.
- Our goal in this project is to use EDA to understand how customer attributes and loan attributes influence the likelihood of default.

- We have been given 4 files:
- Application Data: Provides details about the current loan applications.
- Previous Application Data: Contains information about previous loan applications.
- Column Description: It defines the description of each column
- Notes for projects: It contains important notes related to project

Tools Used

- Microsoft Excel 2021
- Microsoft Power Point

Approach

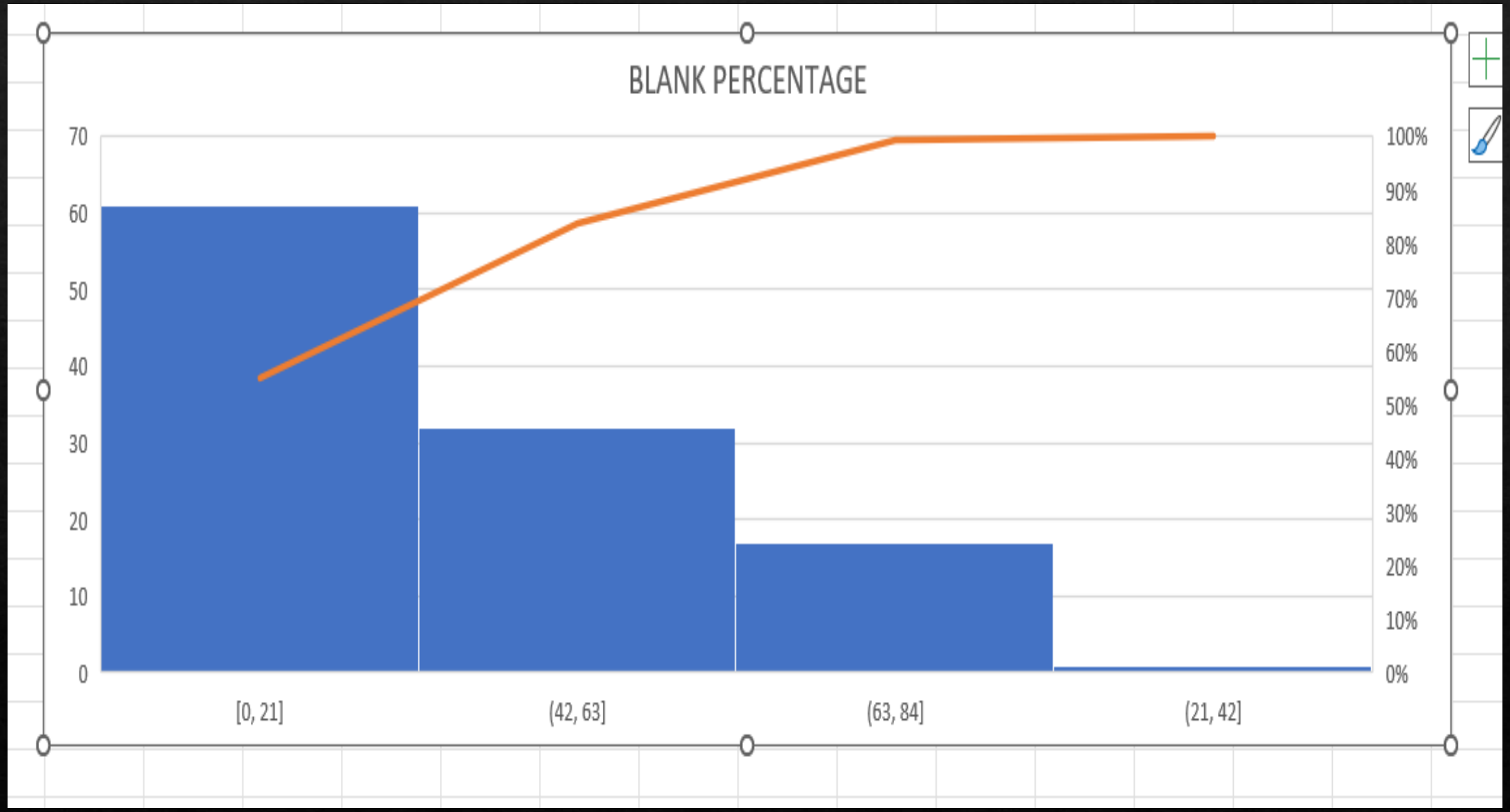
- Understanding Dataset
- Merging Dataset
- Data Preprocessing
- Data Analysis
- Data visualization
- Result

Task A:Data Analytics Tasks

- A. 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
- So for dealing with missing data ,we calculated the null value percentage i.e blank cells, the formula we used =(COUNTBLANK()/COUNT())*100
- Then we deleted the columns which had more than 30% null values.

DELETED COLUMNS

- OWN_CAR_AGE , OCCUPATION_TYPE , EXT_SOURCE_1 , APARTMENTS_AVG , BASEMENTARE_AVG
- YEARS_BEGINEXPLUATATION_AVG, YEARS_BUILD_AVG , COMMONAREA_AVG , ELEVATORS_AVG
- ENTRACES_AVG ,FLOORSMAX_AVG, FLOORSMIN_AVG , LANDAREA_AVG , LIVINGAPARTMENTS_AVG
- LIVINGAREA_AVG , NONLIVINGAPARTMENTS_AVG ,NONLIVINGAREA_AVG,APARTMENTS_MODE
- BASEMENTAREA_AVG , YEARS_BEGINEXPLUATATION_AVG , YEARS_BUILD_AVG,COMMONAREA_AVG
- ELEVATORS_AVG , ENTRANCES_AVG , FLOORSMAX_AVG , FLOORSMIN_AVG , LANDAREA_AVG
- LIVINGAREA_AVG ,LIVINGAPARTMENTS_AVG ,.....

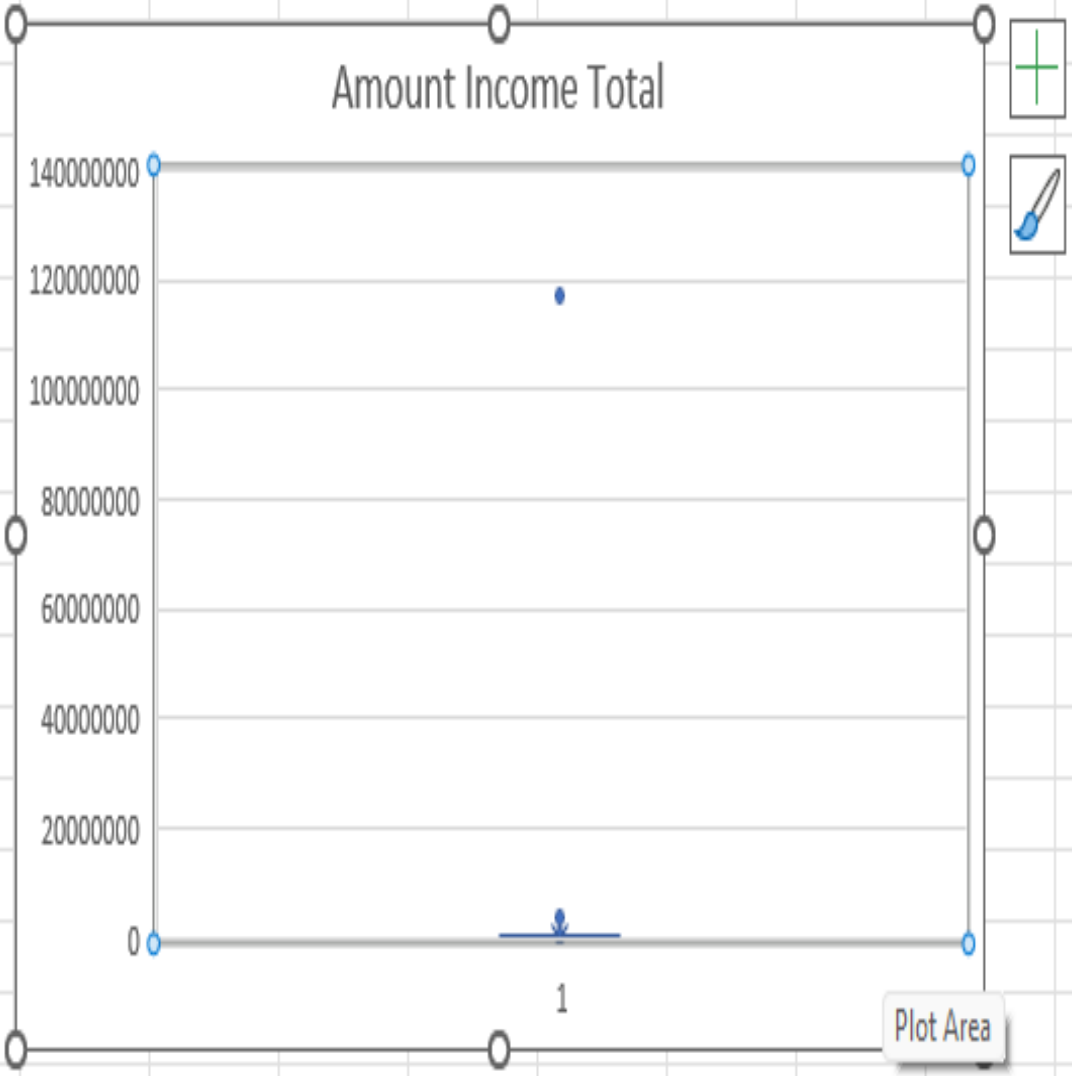

BLANK PERCENTAGE
70
60
50
40
30
20
10
0
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
[0, 21]
(42, 63]
(63, 84]
(21, 42]

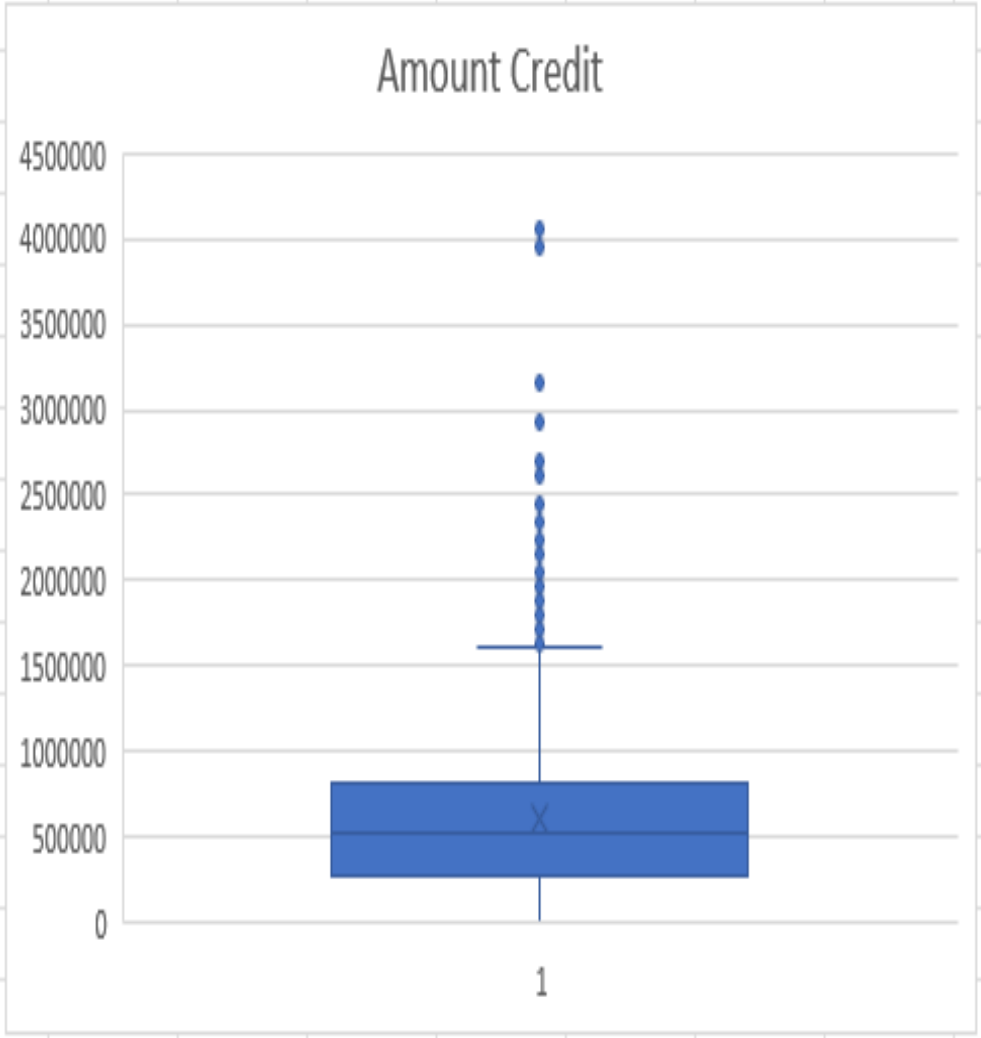
INSIGHTS

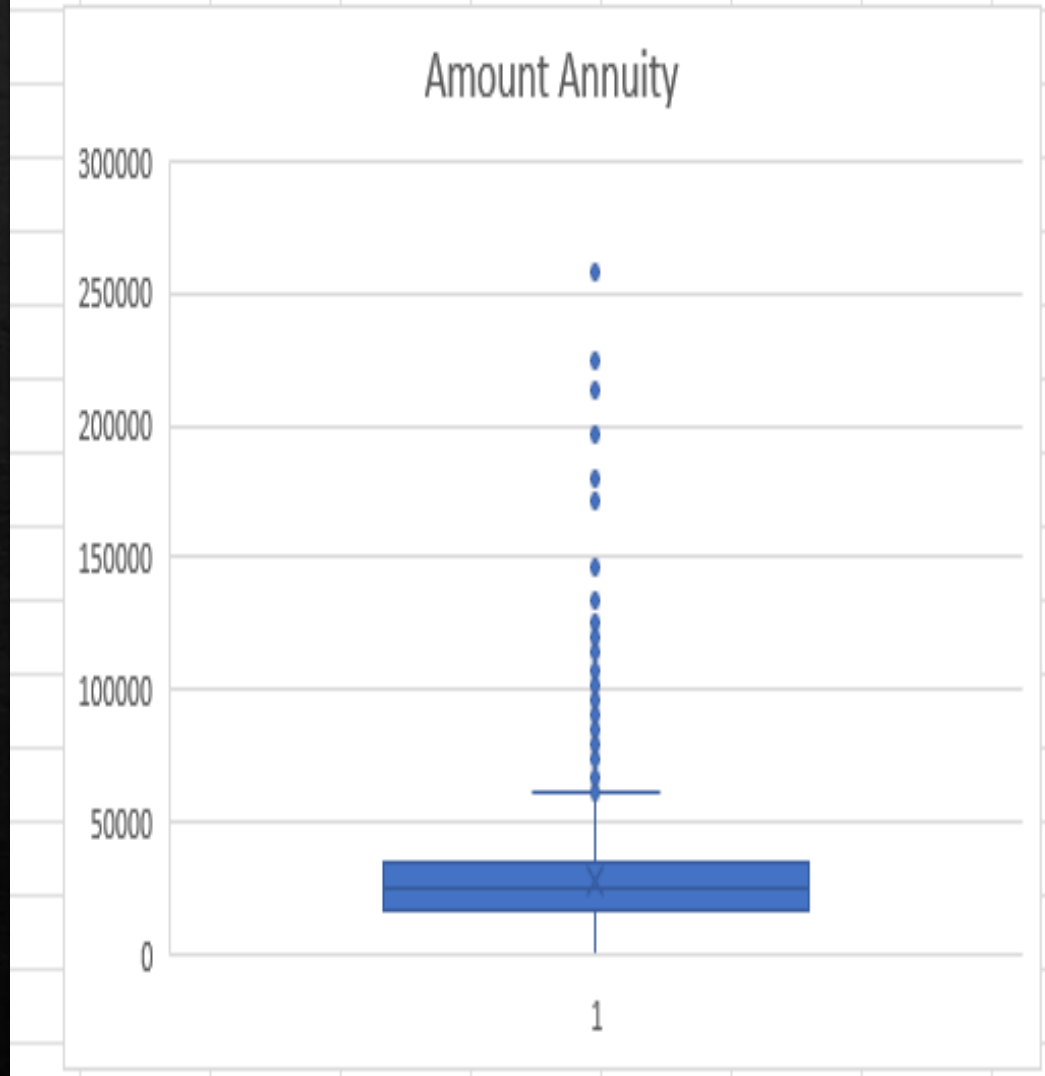
- Unnecessary columns had been deleted
- 50 Columns have been deleted from total number of columns which is 122 columns
- Columns whose blank percentage is more than 30% is deleted
- 72 columns are left out after deletion of unnecessary columns.

TASK B: Identify Outliers in the Dataset:

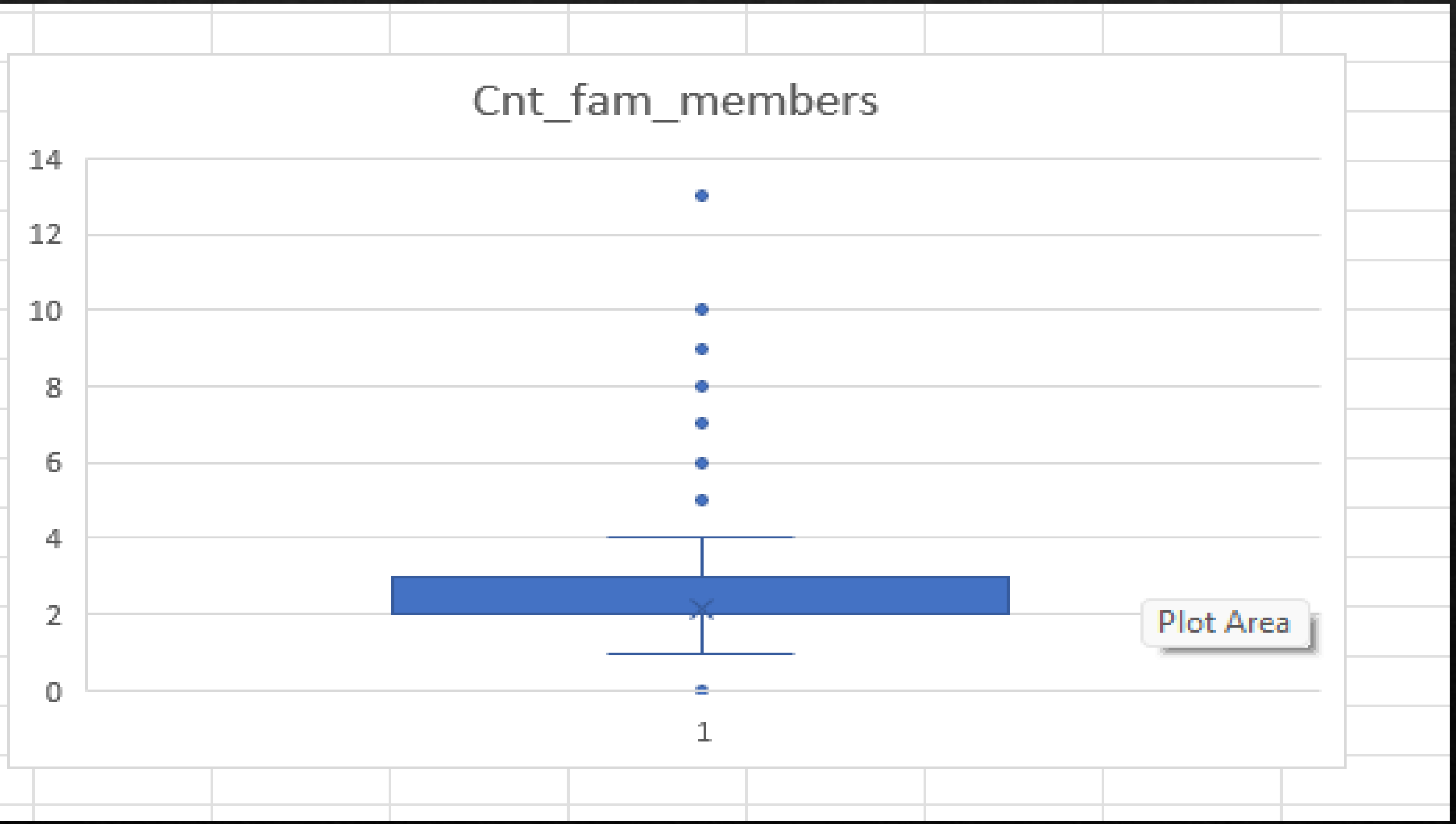
- Outliers can significantly impact the analysis and distort the results. You need to identify outliers in the loan application dataset.
- Detect and identify outliers in the dataset using Excel statistical functions and features, focusing on numerical variables.
- An outlier is an observation or data point that significantly differs from the rest of the data in a dataset. In simpler terms, it's a value that stands out from the general pattern of the data


Amount Income Total
140000000
120000000
100000000
80000000
60000000
40000000
20000000
0
1
Plot Area


Amount Credit
4500000
4000000
3500000
3000000
2500000
2000000
1500000
1000000
500000
0
1


Amount Annuity
300000
250000
200000
150000
100000
50000
0
1


Amount Good Price
4500000
4000000
3500000
3000000
2500000
2000000
1500000
1000000
500000
0
1


Cnt_fam_members
14
12
10
8
6
4
2
0
1
Plot Area

Insights

- Amount credit , Amount goods price, Amount Annuity has the highest number of ouliers
- Count of family numbers and amount income total has less outliers when compared to other columns
- For understanding outliers we have created box plot for the various numerical columns
- It is an outlier if it is located outside the whisker of the box plot
- The whiskers are basically vertical lines which are located typically 1.5 times the interquartile range above the third quartile or below the first quartile

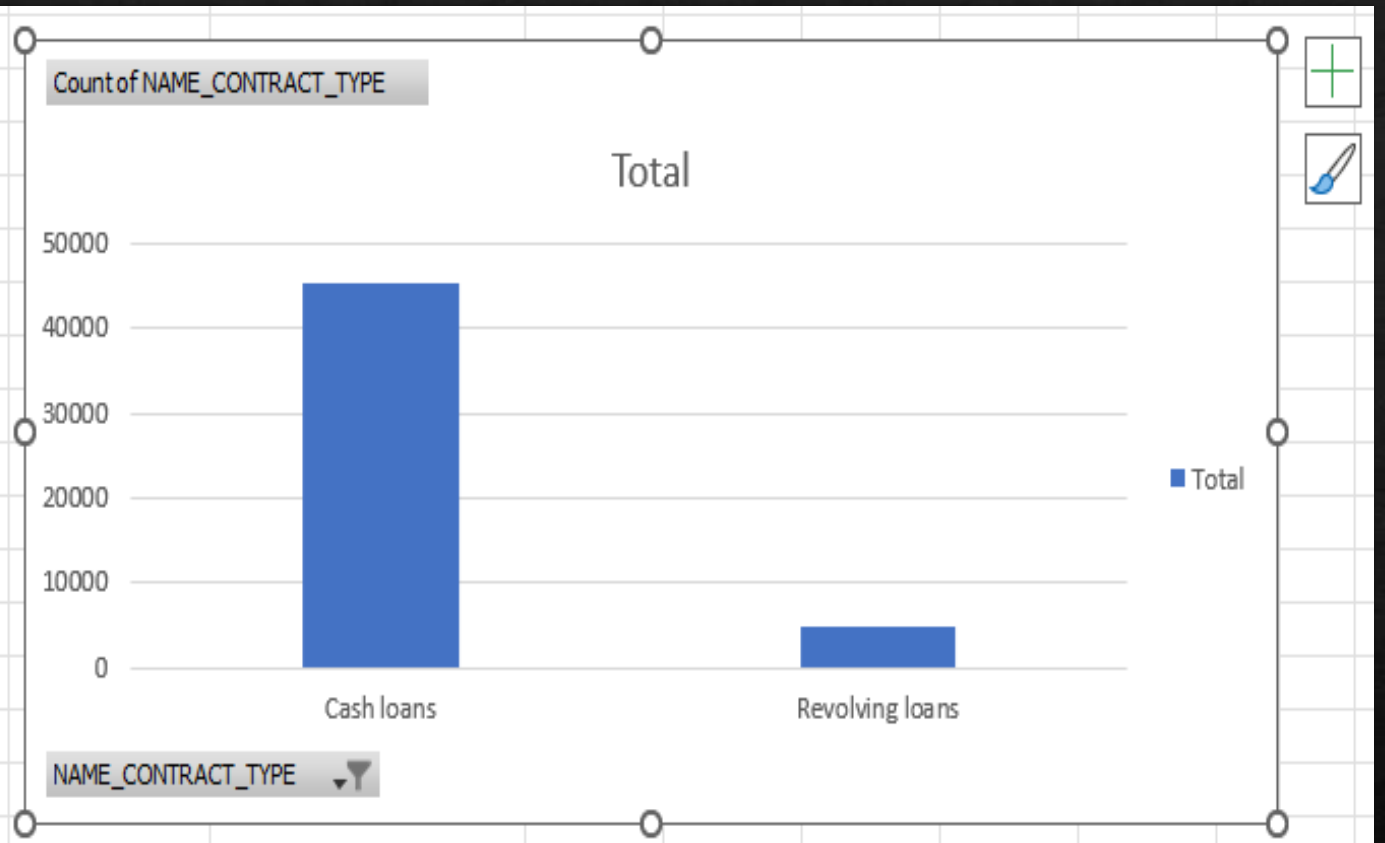
TASK C: Analyze Data Imbalance

- Data imbalance can affect the accuracy of the analysis, especially for binary classification problems. Understanding the data distribution is crucial for building reliable models.
- Determine if there is data imbalance in the loan application dataset and calculate the ratio of data imbalance using Excel functions.

DATA IMBALANCE

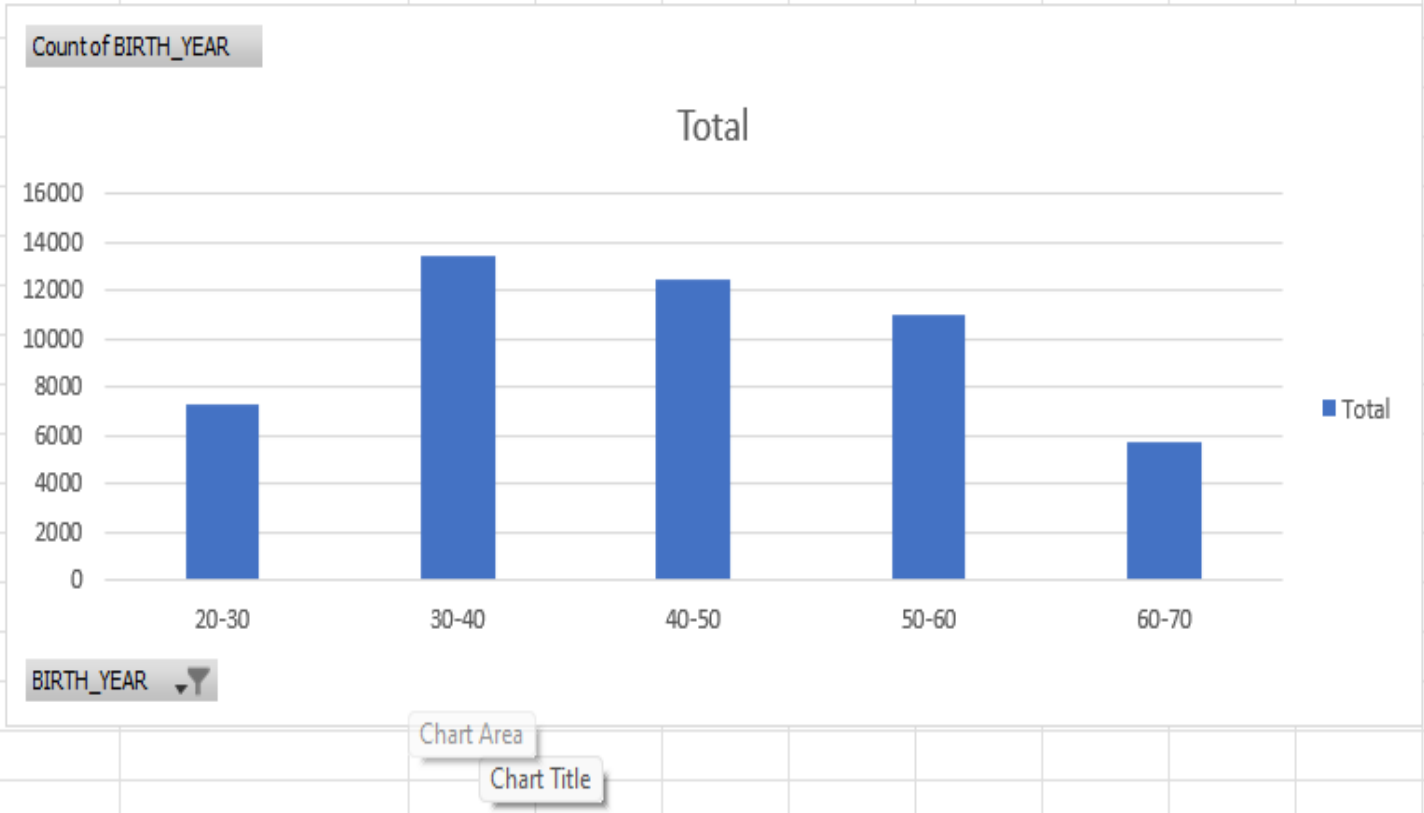
- Data Imbalance refers to an unequal distribution of classes or outcomes within the dataset related to loan approval or denial. In other words, it means that one class (e.g : approved loans) significantly outnumbers the other class (e.g: denied loans) in the dataset. This imbalance can pose challenges when building predictive models, as the model may become biased towards the majority class and may not perform well on the minority class.

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



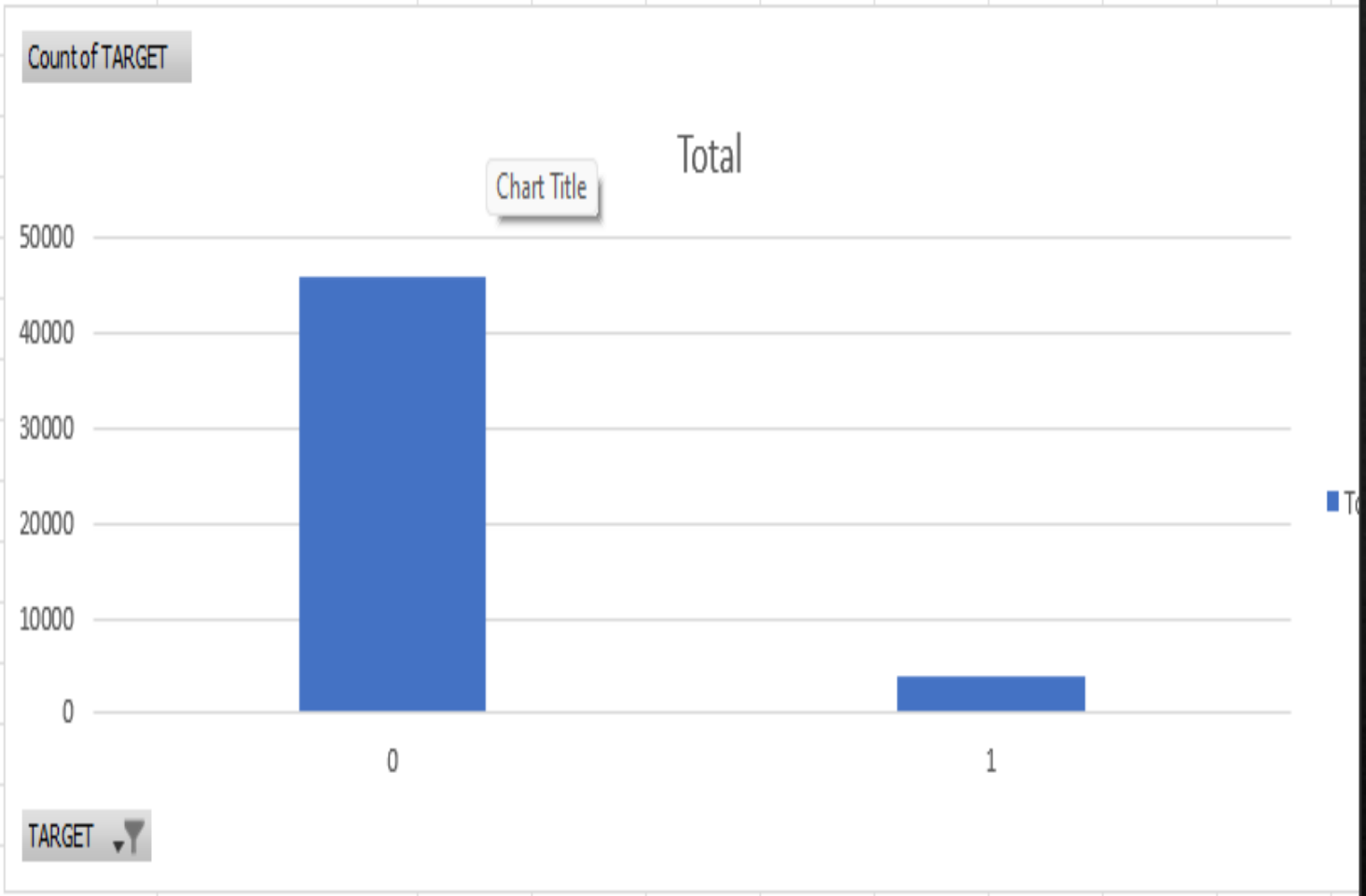
- Insight:
- Cash Loans are very high when compare to revolving loans.
- Cash loan is about 45276 which is around 40000 to 50000 range where as revolving loan is nothing when compared it is just 4723.

Row Labels	Count of BIRTH_YEAR
20-30	7296
30-40	13423
40-50	12491
50-60	11021
60-70	5768
Grand Total	**49999**



- Insights:
- Count of birth year between 30-40 year range is the highest after that the second highest is 40-50 year range

Row Labels	Count of TARGET
0	45973
1	4026
Grand Total	**49999**



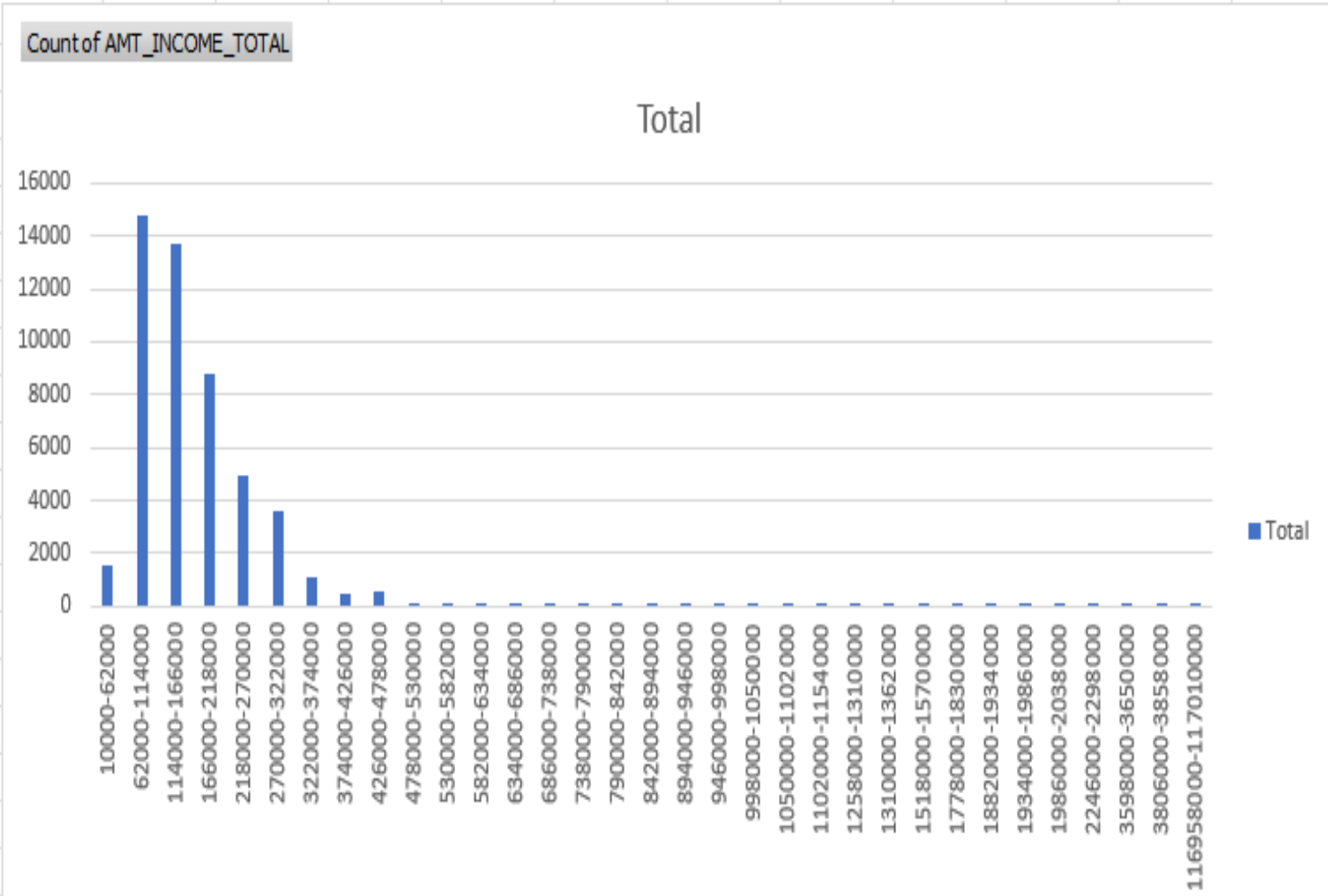
Insights:

Target 1 has the least number of count when compared to target 0.

TASK D: Perform Univariate, Segmented Univariate, and Bivariate Analysis:

- Univariate: "univariate" refers to the analysis or consideration of a single variable at a time. When you're looking at one thing.
- Segmented Univariate:"Segmented univariate" analysis typically refers to the examination of a single variable within different segments or subgroups of a population
- Bivariate analysis involves the simultaneous analysis of two variables to understand the relationship between them.

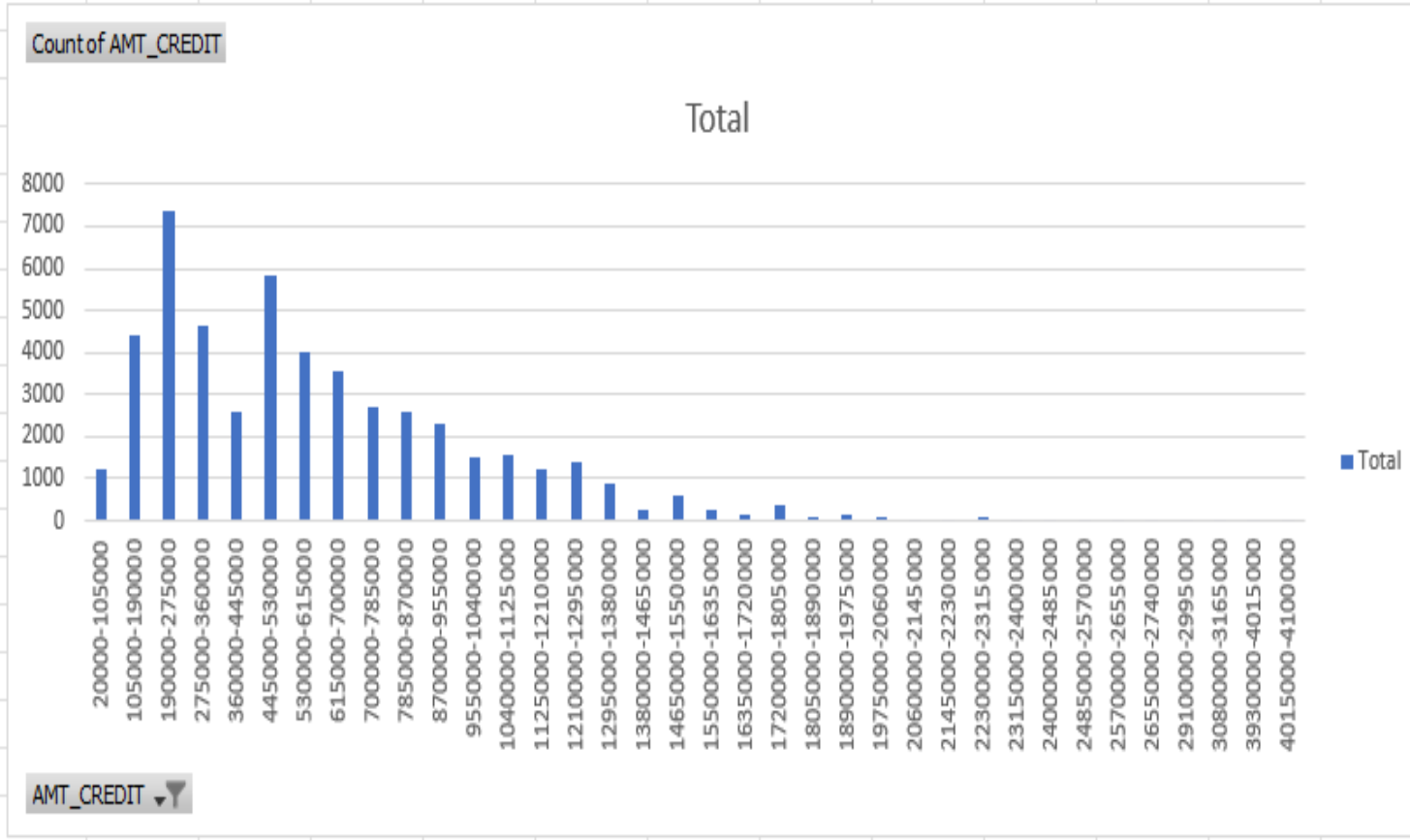
Row Labels	Count of AMT_INCOME_TOTAL
10000-62000	1498
62000-114000	14803
114000-166000	13702
166000-218000	8815
218000-270000	4943
270000-322000	3612
322000-374000	1122
374000-426000	490
426000-478000	514
478000-530000	55
530000-582000	127
582000-634000	71
634000-686000	113
686000-738000	24
738000-790000	13
790000-842000	21
842000-894000	5
894000-946000	30
946000-998000	1



Insights:

Amount income total between 62000-114000 has the highest number of count then 114000-166000 has the second highest number of count and 166000-218000 is the third highest rest of all has very less when compared.

Row Labels	Count of AMT_CREDIT
20000-105000	1210
105000-190000	4427
190000-275000	7346
275000-360000	4659
360000-445000	2613
445000-530000	5858
530000-615000	4028
615000-700000	3555
700000-785000	2701
785000-870000	2570
870000-955000	2309
955000-104000	1496
1040000-11250	1547
1125000-12100	1222
1210000-12950	1375
1295000-13800	904
1380000-14650	268

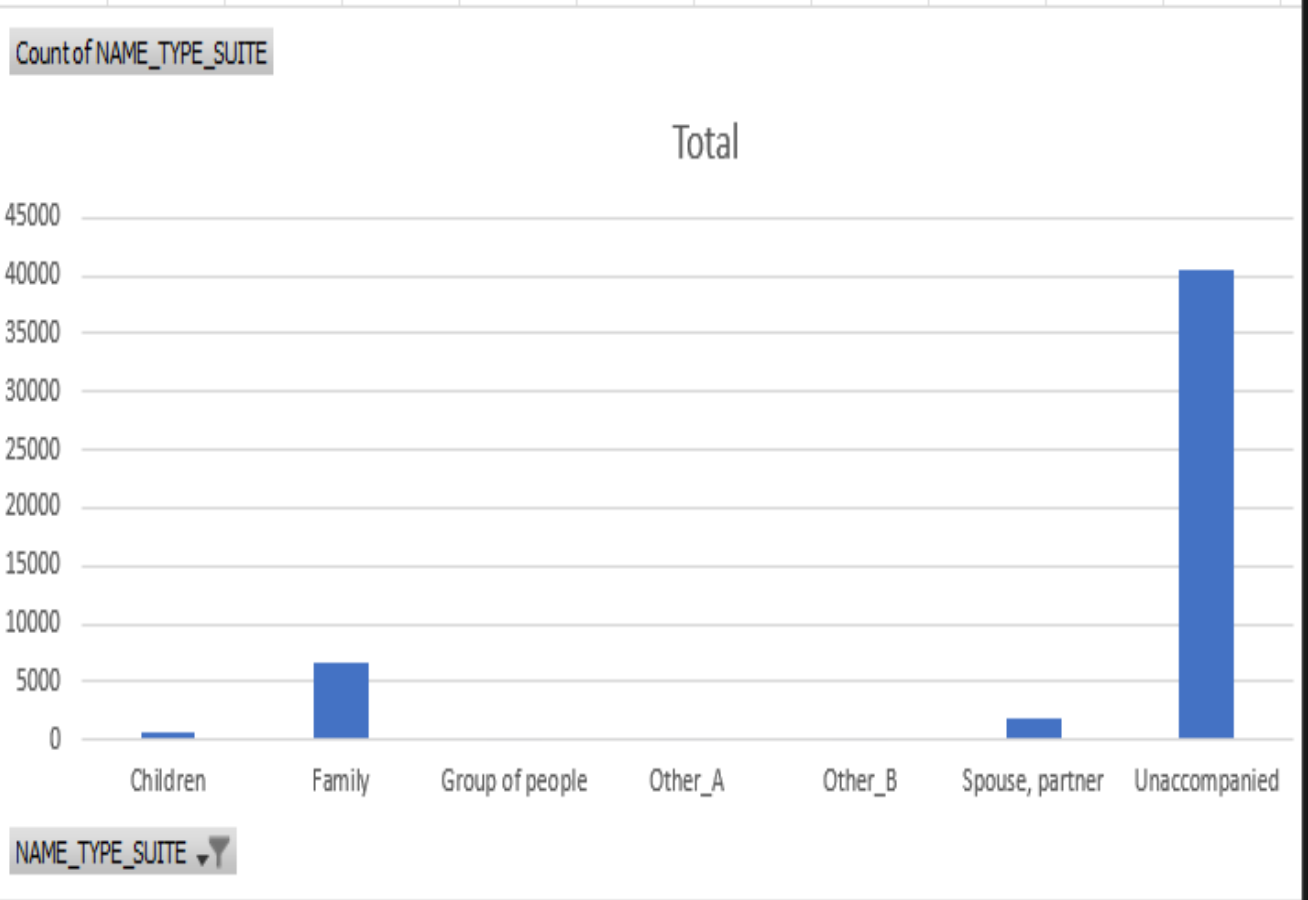


Insights:

Amount credit with highest number of count is between the range 190000-275000

445000-530000 has the second highest all of the rest have low count when compared to these two.

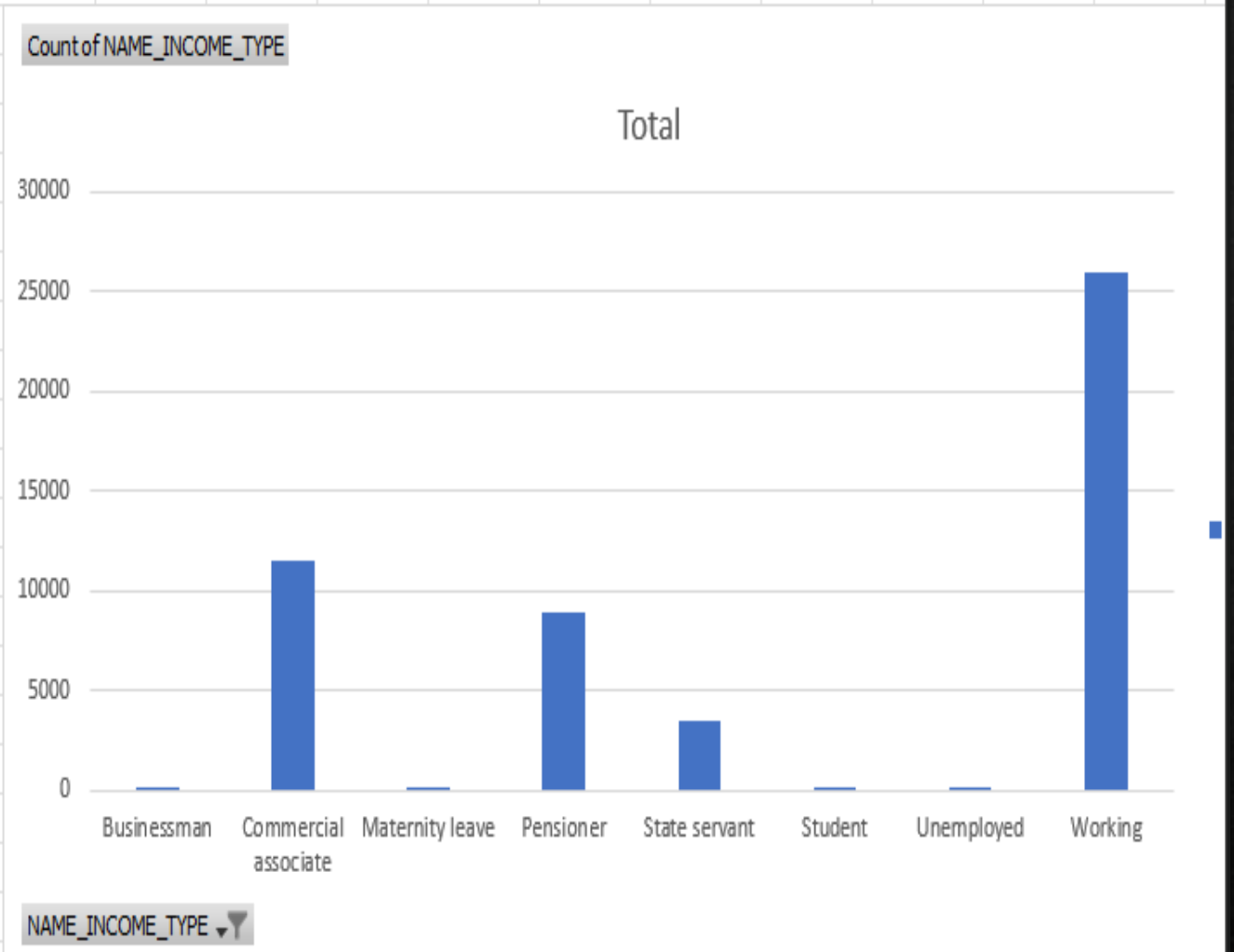
Row Labels	Count of NAME_TYPE_SUITE
Children	542
Family	6549
Group of people	36
Other_A	137
Other_B	259
Spouse, partner	1849
Unaccompanied	40435
Grand Total	**49807**



Insights:

Unaccompanied has the highest count when compared to rest of them.

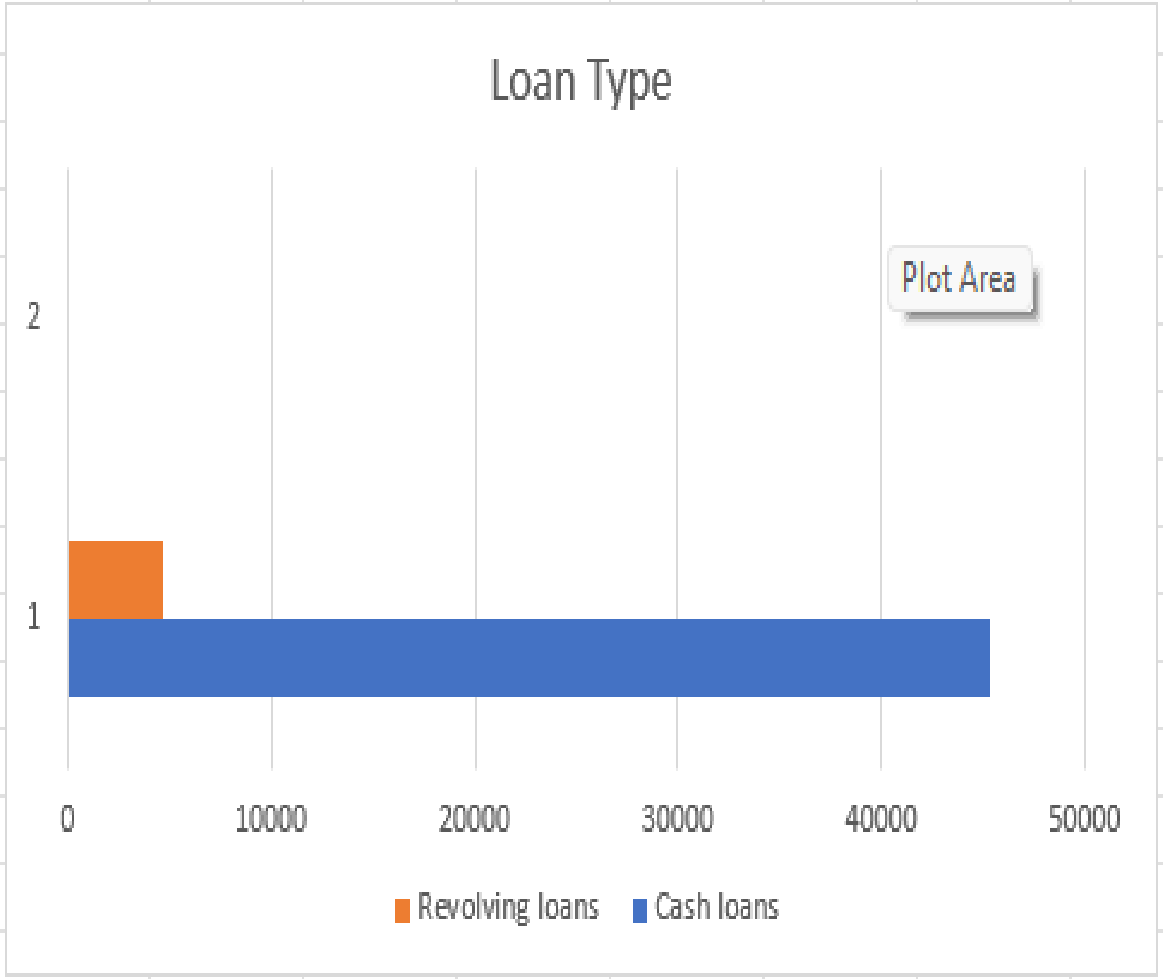
Row Labels	Count of NAME_INCOME_TYPE
Businessman	2
Commercial associate	11543
Maternity leave	1
Pensioner	8920
State servant	3512
Student	5
Unemployed	6
Working	26010
Grand Total	**49999**



Insights:
working income type has the highest count when compared to other types.

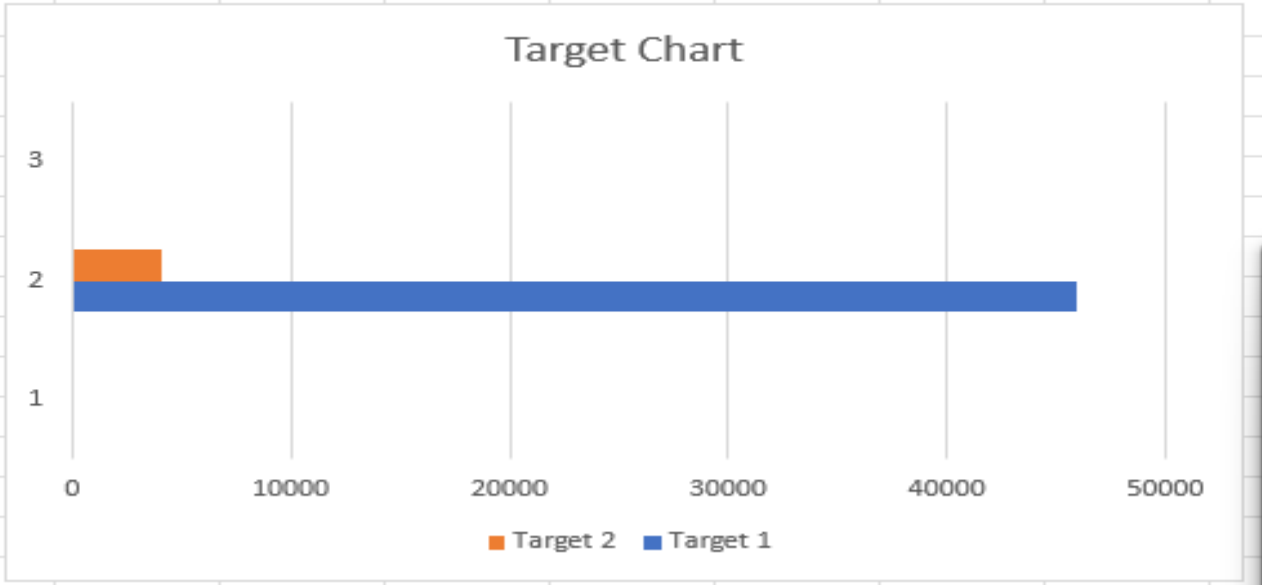
Segmented Univariate Graphs

Cash loans	45276	90.55381108
Revolving loans	4723	9.446188924

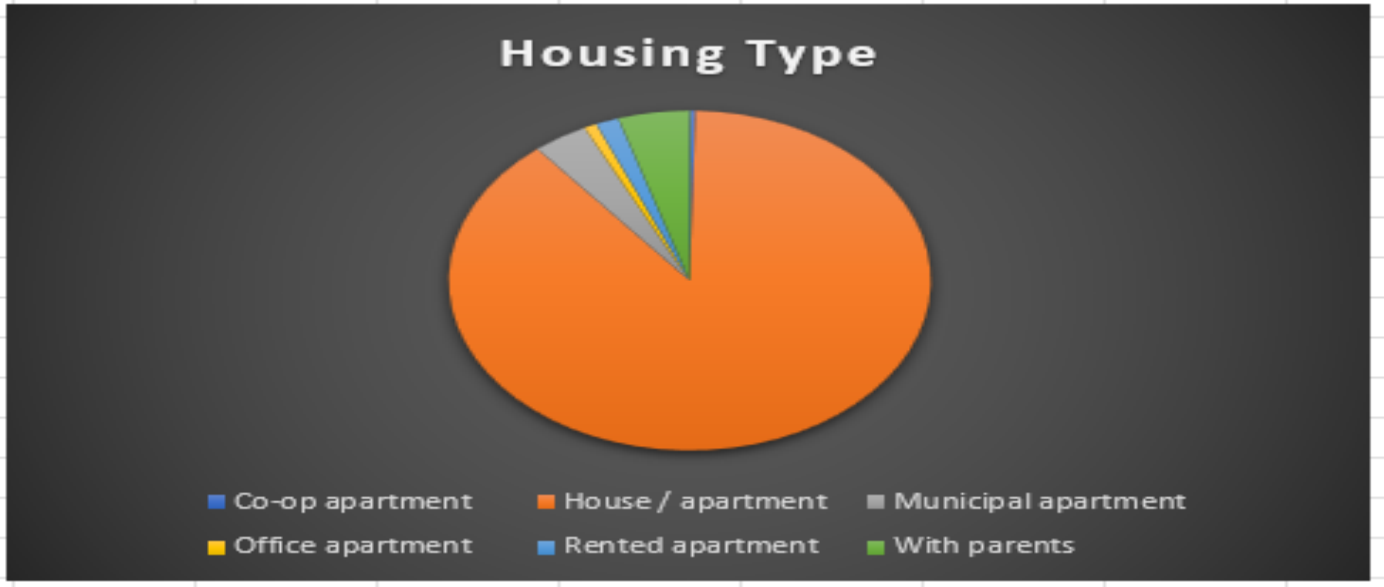


Target 1	0	45973	91.94784
Target 2	1	4026	8.052161

Chart Area

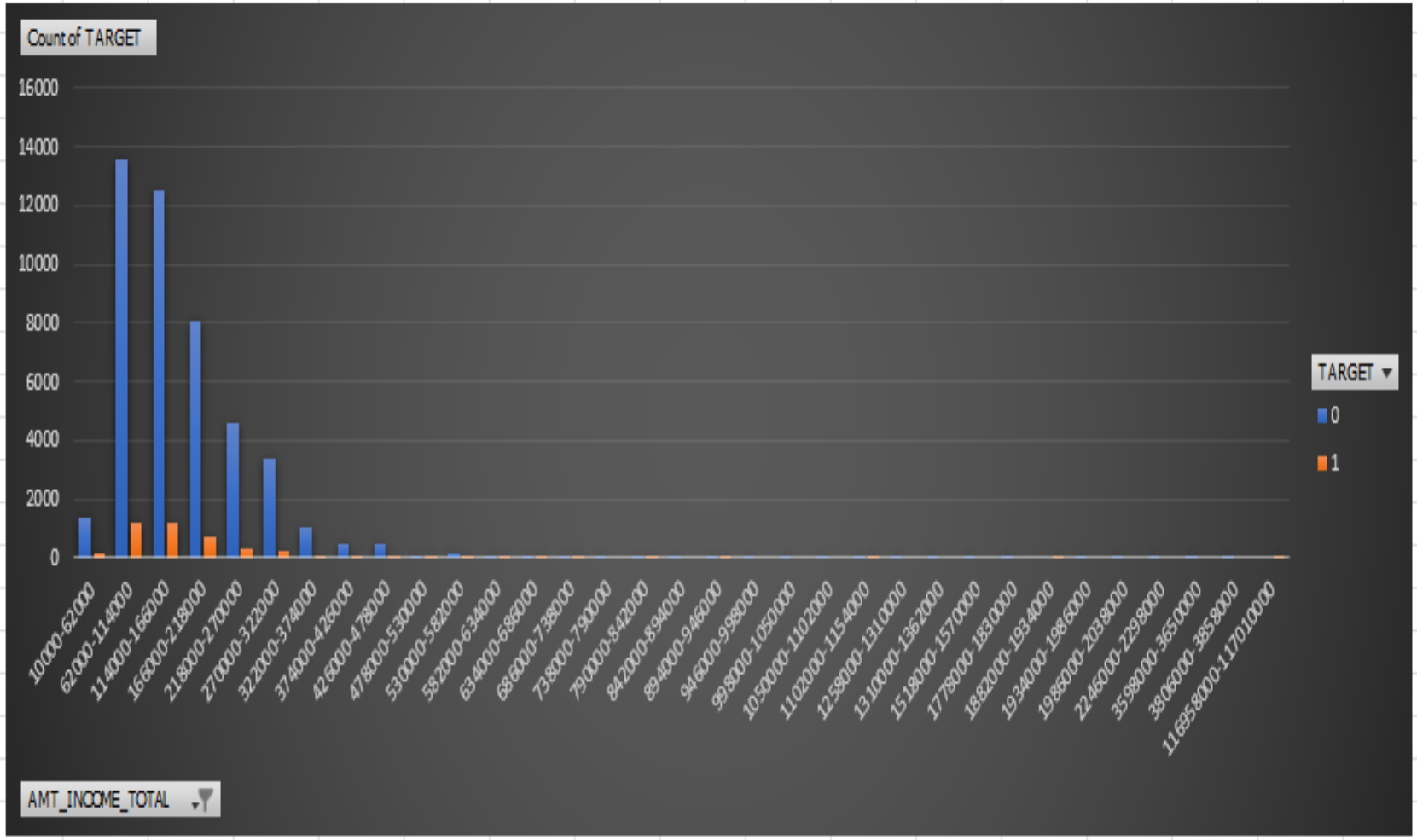


Co-op apartment	191
House / apartment	44368
Municipal apartment	1845
Office apartment	427
Rented apartment	769
With parents	2399

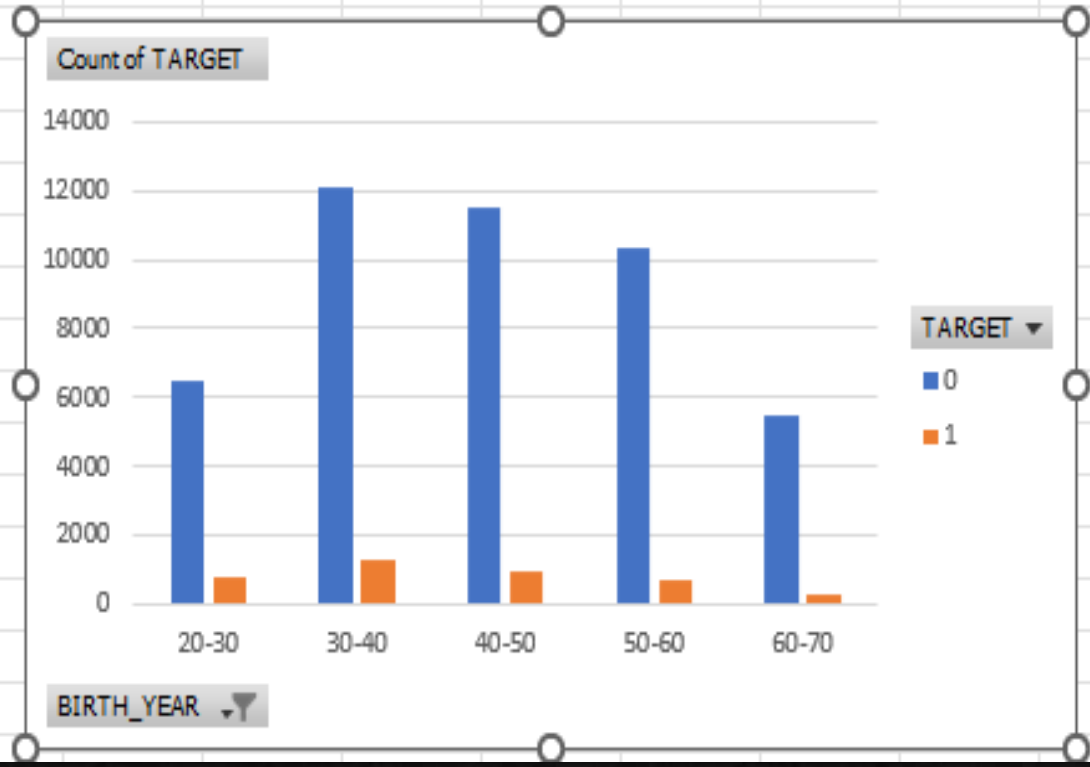


Bivariate Analysis Graphs/Charts

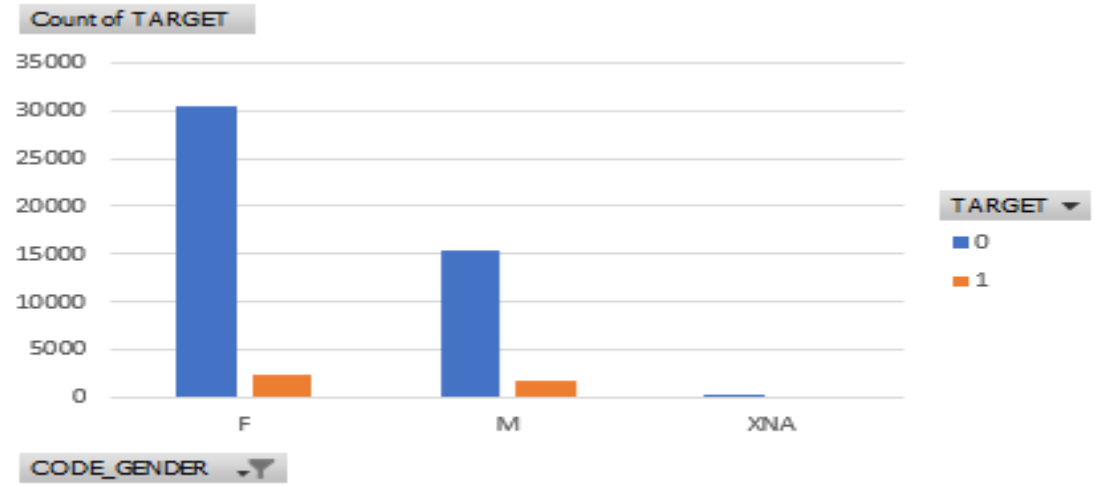
Count of TARGET	Column Labels		
Row Labels	0	1	Grand Total
10000-62000	1382	116	1498
62000-114000	13558	1245	14803
114000-166000	12482	1220	13702
166000-218000	8094	721	8815
218000-270000	4617	326	4943
270000-322000	3384	228	3612
322000-374000	1064	58	1122
374000-426000	450	40	490
426000-478000	476	38	514
478000-530000	52	3	55
530000-582000	117	10	127
582000-634000	66	5	71
634000-686000	106	7	113
686000-738000	22	2	24
738000-790000	13		13
790000-842000	19	2	21
842000-894000	5		5
894000-946000	28	2	30
946000-998000	1		1



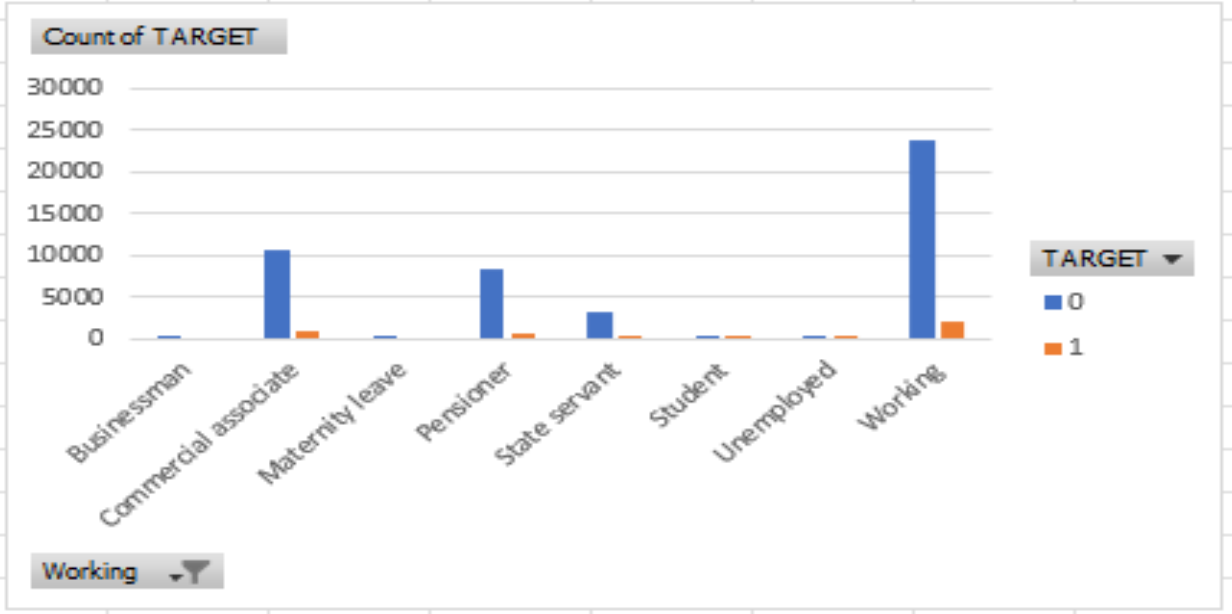
Count of TARGET	Column Labels		
Row Labels	0	1	Grand Total
20-30	6478	818	7296
30-40	12112	1311	13423
40-50	11551	940	12491
50-60	10353	668	11021
60-70	5479	289	5768
Grand Total	45973	4026	49999



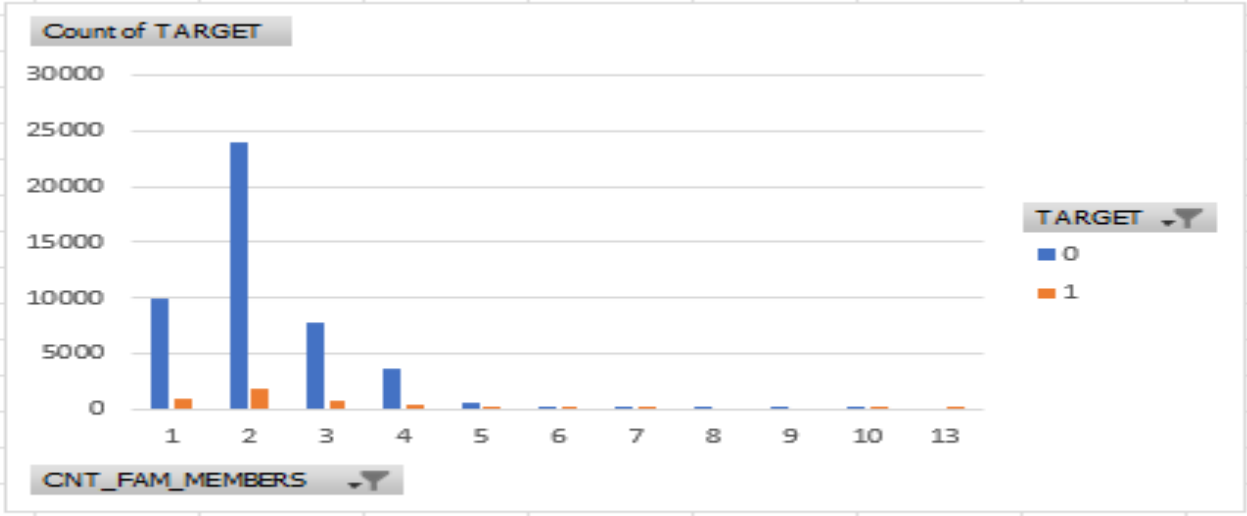
Count of TARGET	Column Labels		
Row Labels	0	1	Grand Total
F	30559	2264	32823
M	15412	1762	17174
XNA	2		2
Grand Total	45973	4026	49999



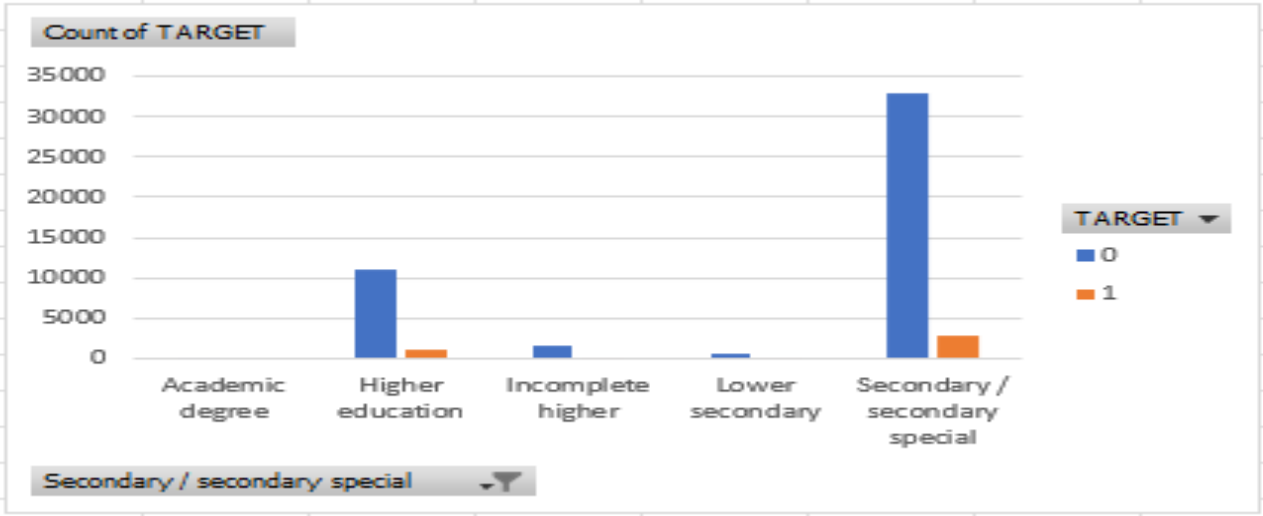
Count of TARGET	Column Labels		
Row Labels	0	1	Grand Total
Businessman	2		2
Commercial asso	10605	938	11543
Maternity leave	1		1
Pensioner	8249	671	8920
State servant	3229	283	3512
Student	4	1	5
Unemployed	5	1	6
Working	23877	2132	26009
Grand Total	45972	4026	49998



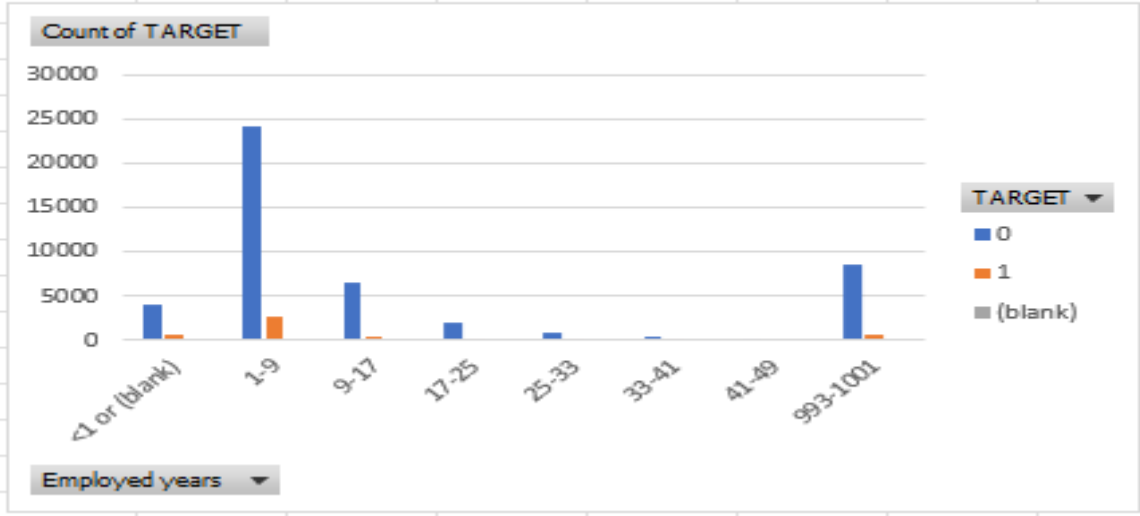
Count of TARGET	Column Labels		
Row Labels	0	1	Grand Total
1	9951	922	10873
2	23901	1906	25807
3	7858	777	8635
4	3651	349	4000
5	538	54	592
6	55	13	68
7	9	3	12
8	6		6
9	2		2
10	1	1	2
13		1	1
Grand Total	45972	4026	49998



Count of TARGET	Column Labels		
Row Labels	0	1	Grand Total
Academic degree	20		20
Higher education	11138	1029	12167
Incomplete highe	1500	120	1620
Lower secondary	574	46	620
Secondary / secor	32740	2831	35571
Grand Total	**45972**	**4026**	**49998**



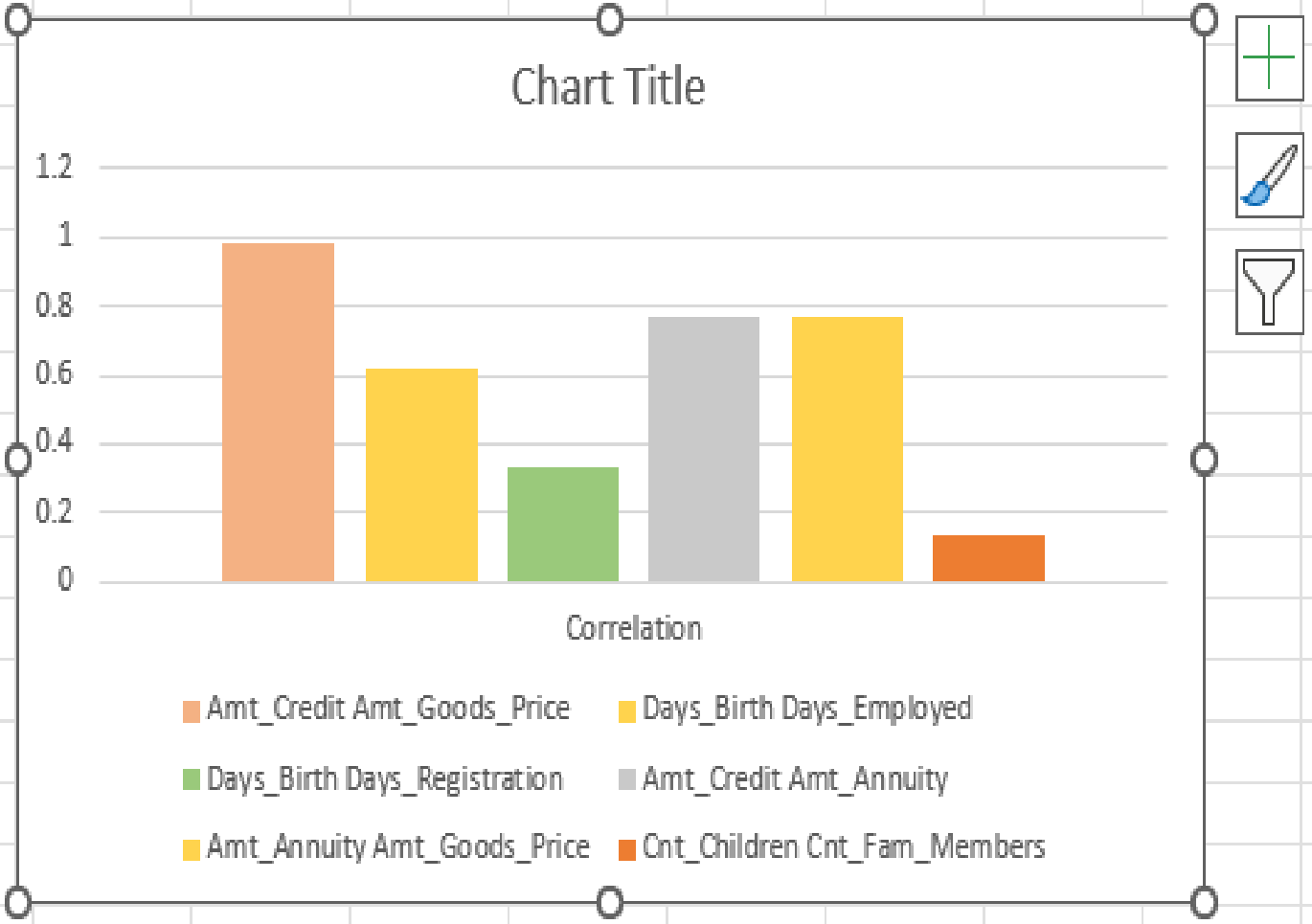
Count of TARGET	Column Labels			
Row Labels	0	1	(blank)	Grand Total
<1 or (blank)	4057	491		4548
1-9	24100	2521		26621
9-17	6447	383		6830
17-25	1947	96		2043
25-33	716	26		742
33-41	262	6		268
41-49	24			24
993-1001	8421	503		8924
Grand Total	**45974**	**4026**		**50000**



TASK E: Identify Top Correlations for Different Scenarios

- Understanding the correlation between variables and the target variable can provide insights into strong indicators of loan default.
- Correlation is a statistical measure that quantifies the degree to which two variables are related or move together. It assesses the strength and direction of a linear relationship between two numerical variable

Var1	Var2	Correlation
Amt_Credit	Amt_Goods_Price	0.986944513
Days_Birth	Days_Employed	0.621810207
Days_Birth	Days_Registration	0.333691949
Amt_Credit	Amt_Annuity	0.769467164
Amt_Annuity	Amt_Goods_Price	0.774416374
Cnt_Children	Cnt_Fam_Members	0.131156266



Result

- This project involved extensive use of excel . The major challenge was working with such huge data , this project helped me understand how to work with huge datasets.
- This helped me understand how datasets are merged to anlyze the data.
- The dataset involved a lot of missing data and outliers, handling them was a task and this project helped me understand outliers and missing data handling.
- We got to know about univariate , segmented univariate and bivariate analysis.
- We also got to know about Data Imbalance.

LINK FOR EXCEL SHEET

- mailto:https://docs.google.com/spreadsheets/d/1SAeCV19y5Q25hkoxy3RGIVkJF5_pIVLg/edit?usp=drivesdk&ouid=113826139200146158008&rtpof=true&sd=true

LINK FOR VIDEO

- mailto:https://drive.google.com/file/d/1SCE4NEqSzVH8y3h8WbIx0ZNsOTGusmyU/view?usp=drivesdk