Bank Loan Case Study

Ritika Chaudhary

Agenda

Project Overview

Business Objectives

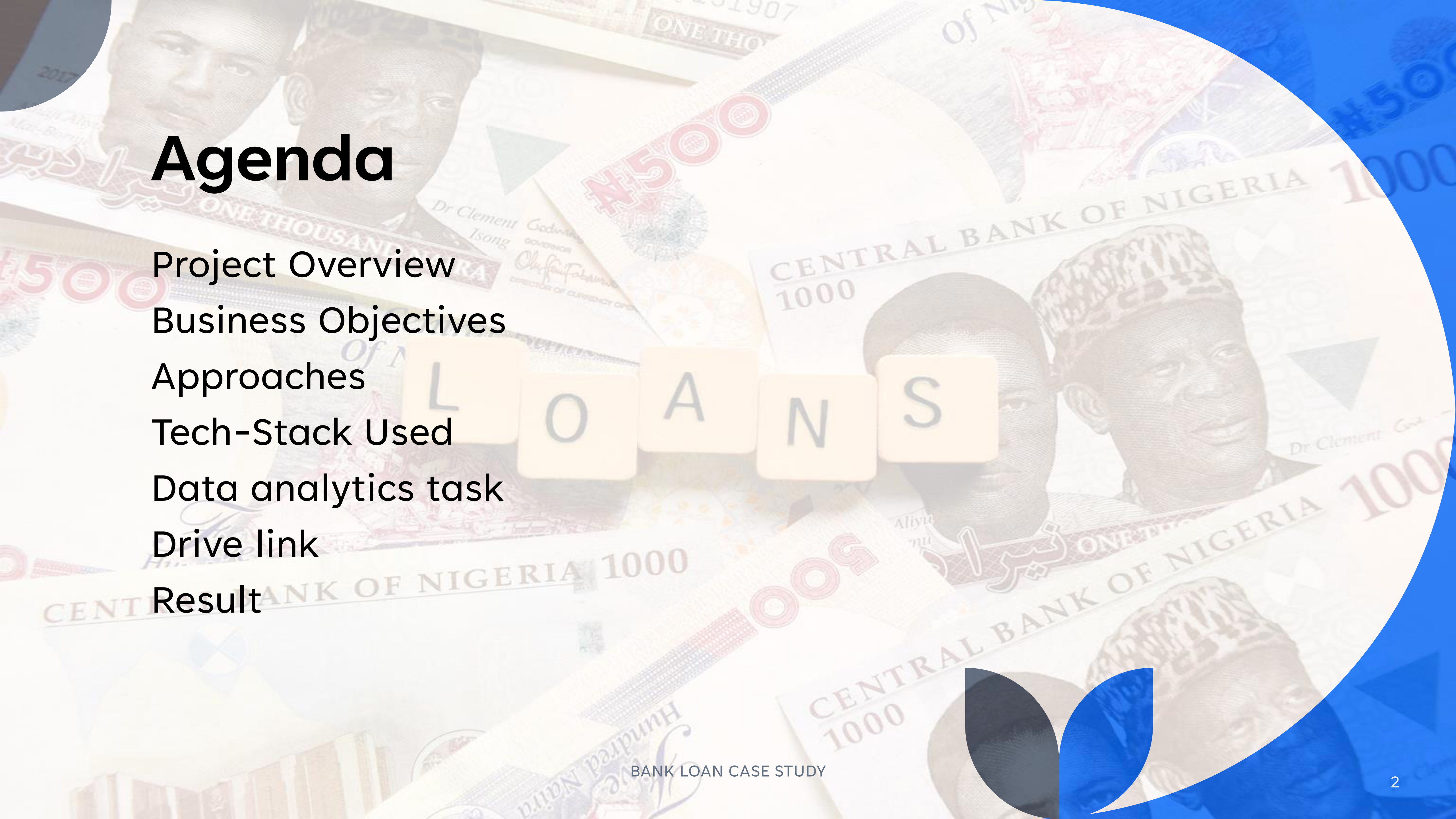
Approaches

Tech-Stack Used

Data analytics task

Drive link

Result



Project Overview

This project revolves around utilising exploratory data analysis (EDA) techniques to address a critical challenge faced by finance company which specialise in offering various types of loans to urban customers but the issue of loan defaults, especially among applicants with insufficient credit histories has been a recurring problem. The objective is to make more informed lending decisions ensuring that capital applicants are not rejected while minimising the financial losses due to defaults.

Business Objectives

The main aim of this project is to identify patterns that indicate if a customer will have difficulty paying their installments. This information can be used to make decisions such as denying the loan, reducing the amount of loan or lending at a higher rate to risk applicants. The company wants to understand the key factors behind the loan default so it can make better decisions about loan approval.

Approaches

- **Data collection:** Download the data set.
- **Data cleaning:** Handling duplicates by removing rows. Deleting the columns with more than 50% blank cells.
- **Outlier detection:** Identify outliers in the data set that might skew analysis.
- **Data summary:** Calculates statistical measures such as averages, medians, quartiles and range. Create visualisation like box and whisker, column charts, heatmaps and histograms for better understanding of data.
- Documented the findings and insights in a report format.

Tech-Stack used:

Microsoft Excel 2021

We selected Microsoft Excel for its robust data analysis capabilities, statistical functions, and versatile visualisation tools.

Microsoft PowerPoint 2021

Utilised for creating a visually engaging and informative project report presentation.

Data Analytics Task

Task A: Identify Missing data

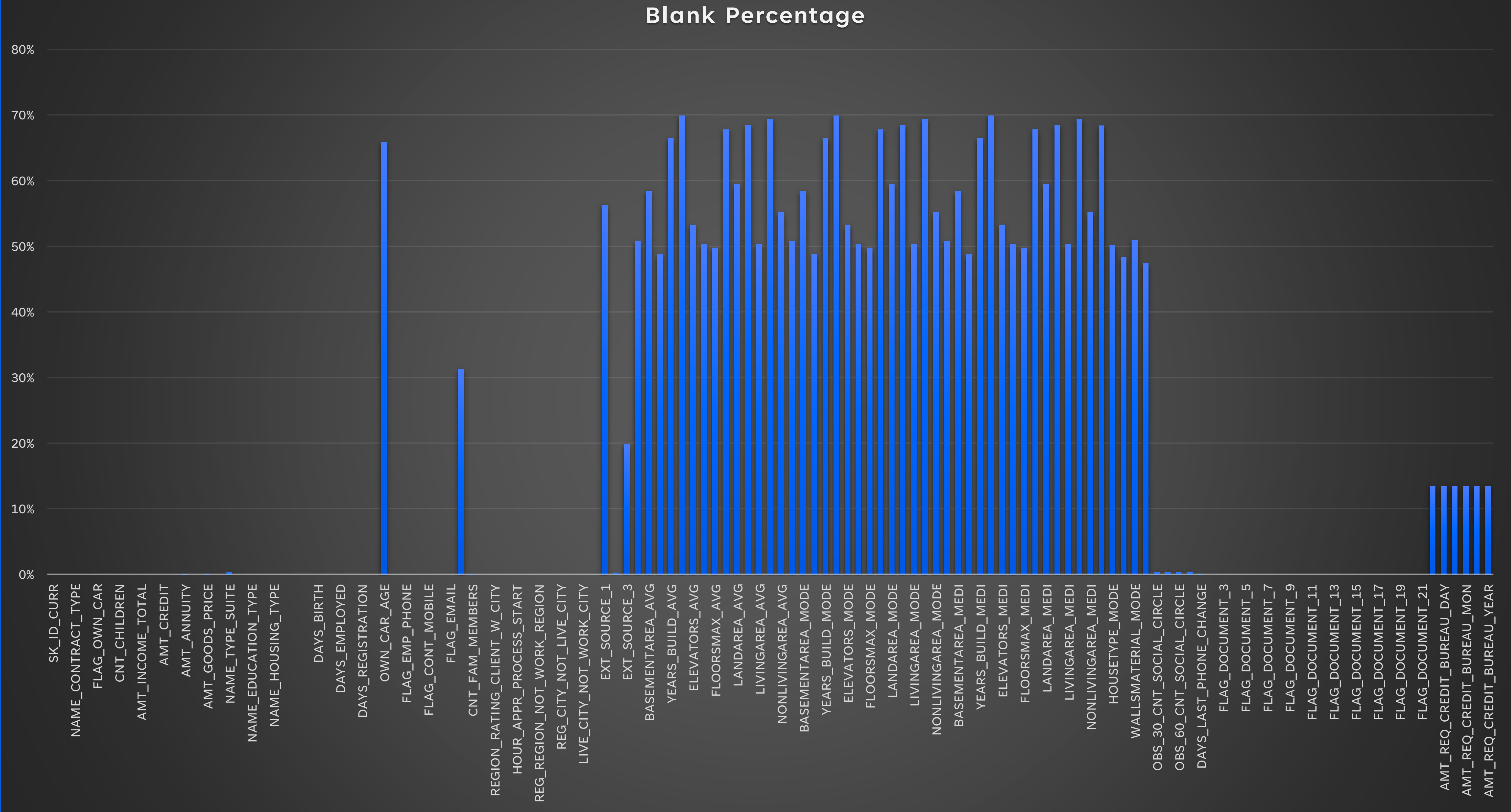
Identify the missing data in the dataset and decide on an appropriate method to deal with it using Excel built-in functions and features.

The percentage of missing values varies across columns, with some columns having no missing data while others exhibit significant levels of missingness, up to 70%. This insight highlights the importance of addressing missing data effectively to ensure the accuracy of our analysis. In subsequent tasks, we will employ appropriate methods such as imputation to handle these missing values, ensuring that our analysis is based on complete and reliable task.

TOTAL COLUMNS	ROWS	BLANKS
• 122	• 50000	• 1488212

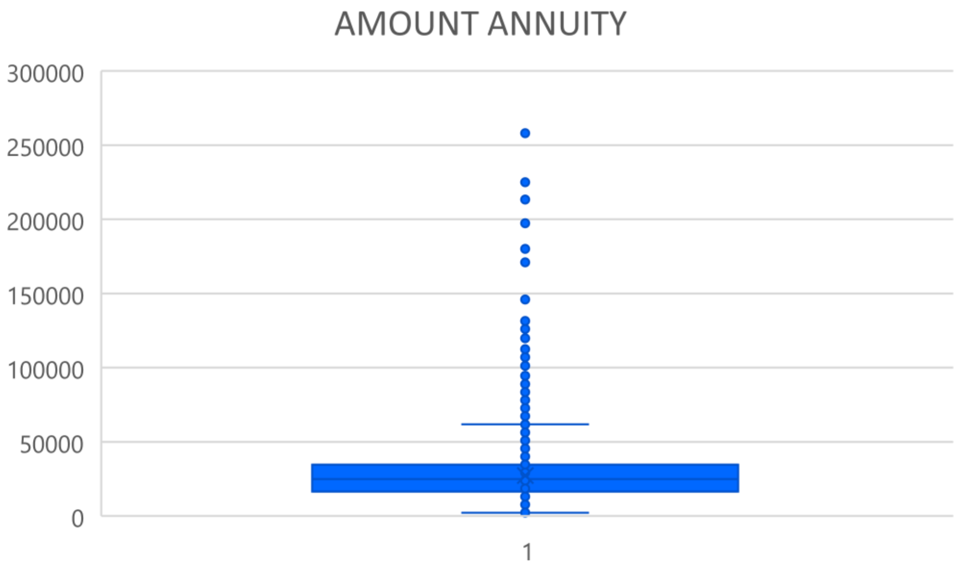
Columns having more than 50 % blanks were deleted. Following is the list of deleted columns:

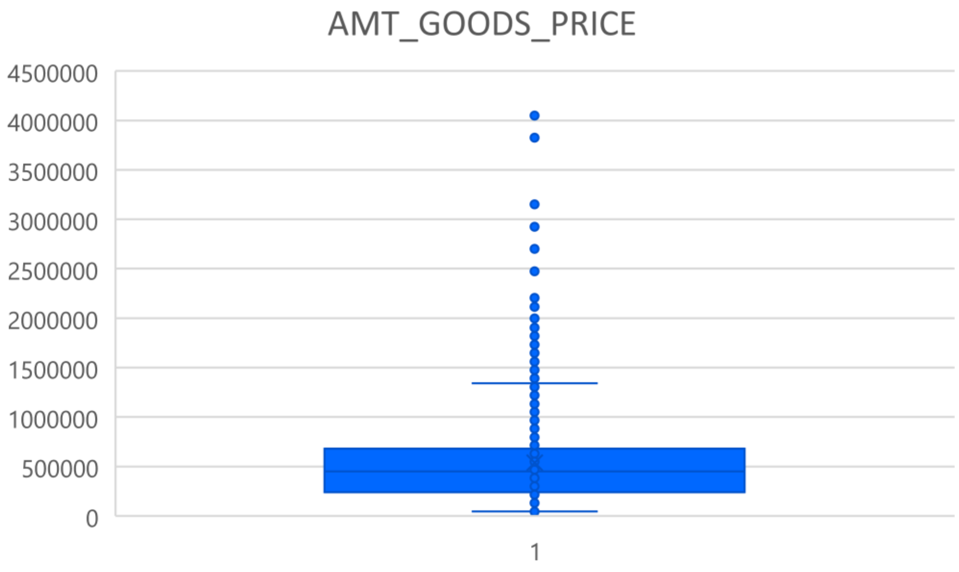
Deleted Columns			
OWN_CAR_AGE	BASEMENTAREA_AVG	NONLIVINGAPARTMENTS_AVG	LANDAREA_MEDI
EXT_SOURCE_1	YEARS_BUILD_AVG	NONLIVINGAREA_AVG	LIVINGAPARTMENTS_MEDI
APARTMENTS_AVG	COMMONAREA_AVG	APARTMENTS_MODE	LIVINGAREA_MEDI
ELEVATORS_AVG	ENTRANCES_AVG	BASEMENTAREA_MODE	NONLIVINGAPARTMENTS_MEDI
FLOORSMIN_AVG	LANDAREA_AVG	FLOORSMIN_MODE	NONLIVINGAREA_MEDI
LIVINGAPARTMENTS_AVG	LIVINGAREA_AVG	LANDAREA_MODE	FONDKAPREMONT_MODE
NONLIVINGAPARTMENTS_AVG	NONLIVINGAREA_AVG	LIVINGAPARTMENTS_MODE	HOUSETYPE_MODE
APARTMENTS_MODE	BASEMENTAREA_MODE	LIVINGAREA_MODE	WALLSMATERIAL_MODE
NONLIVINGAPARTMENTS_MODE	NONLIVINGAREA_MODE	APARTMENTS_MEDI	
BASEMENTAREA_MEDI	YEARS_BUILD_MEDI	COMMONAREA_MEDI	
ELEVATORS_MEDI	ENTRANCES_MEDI	FLOORSMIN_MEDI	


Blank Percentage
80%
70%
60%
50%
40%
30%
20%
10%
0%
SK_ID_CURR
NAME_CONTRACT_TYPE
FLAG_OWN_CAR
CNT_CHILDREN
AMT_INCOME_TOTAL
AMT_CREDIT
AMT_ANNUITY
AMT_GOODS_PRICE
NAME_TYPE_SUITE
NAME_EDUCATION_TYPE
NAME_HOUSING_TYPE
DAYS_BIRTH
DAYS_EMPLOYED
DAYS_REGISTRATION
OWN_CAR_AGE
FLAG_EMP_PHONE
FLAG_CONT_MOBILE
FLAG_EMAIL
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_1
EXT_SOURCE_3
BASEMENTAREA_AVG
YEARS_BUILD_AVG
ELEVATORS_AVG
FLOORSMAX_AVG
LANDAREA_AVG
LIVINGAREA_AVG
NONLIVINGAREA_AVG
BASEMENTAREA_MODE
YEARS_BUILD_MODE
ELEVATORS_MODE
FLOORSMAX_MODE
LANDAREA_MODE
LIVINGAREA_MODE
NONLIVINGAREA_MODE
BASEMENTAREA_MEDI
YEARS_BUILD_MEDI
ELEVATORS_MEDI
FLOORSMAX_MEDI
LANDAREA_MEDI
LIVINGAREA_MEDI
NONLIVINGAREA_MEDI
HOUSETYPE_MODE
WALLSMATERIAL_MODE
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

Task B: Identify Outliers In The Dataset

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



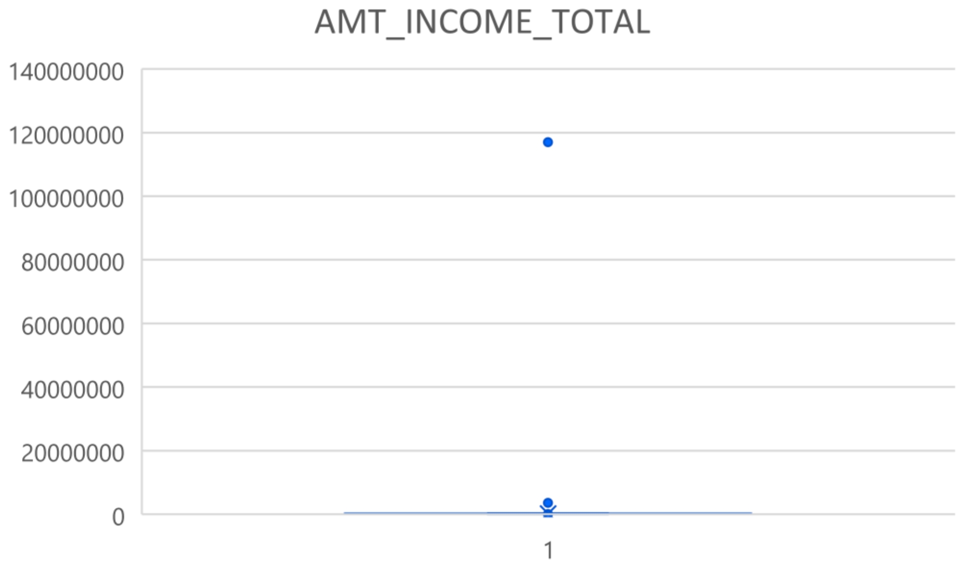


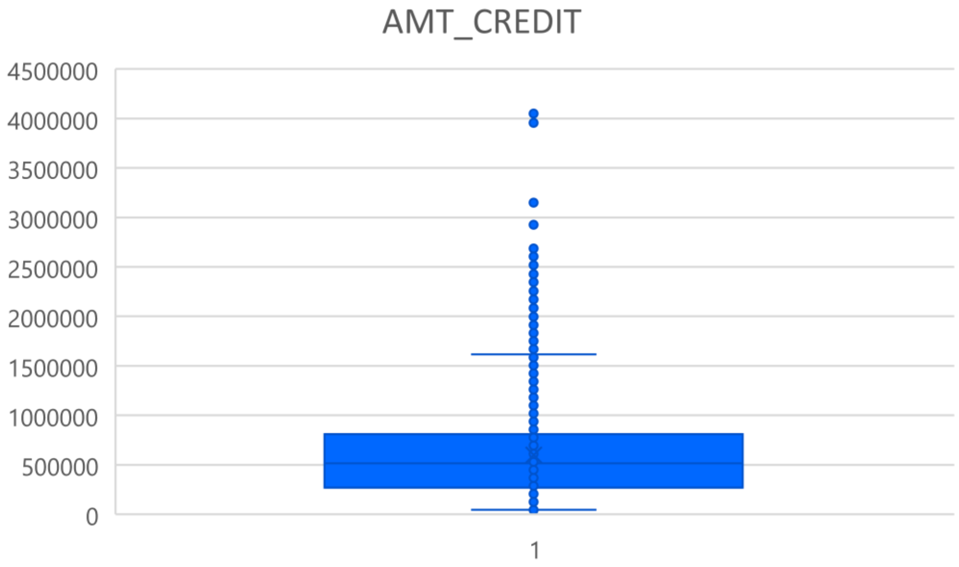
Amt annuity has outliers ranging from 50,000 to 1,32,000

Amount goods price has outliers ranging from 3L to 23L.

Task B: Identify Outliers In The Dataset

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



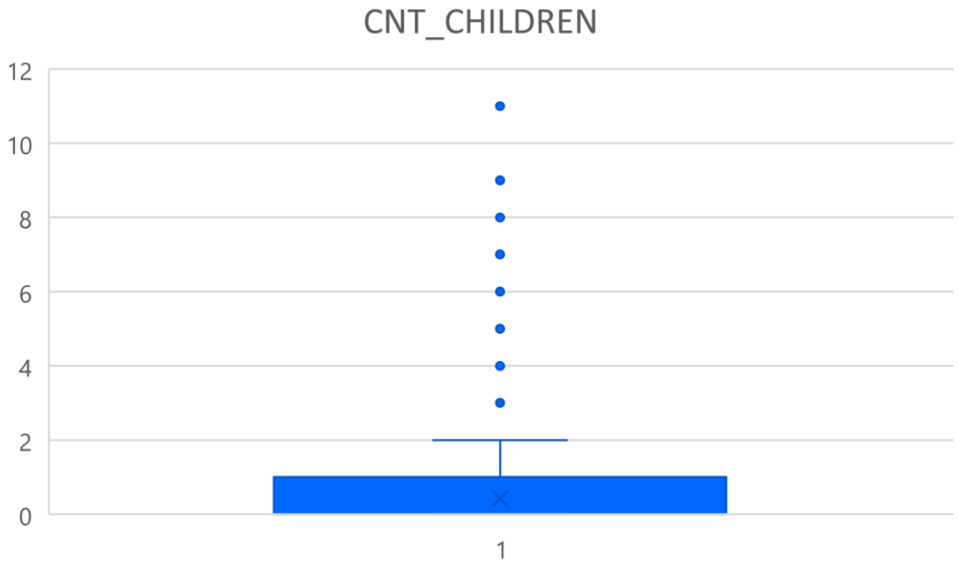


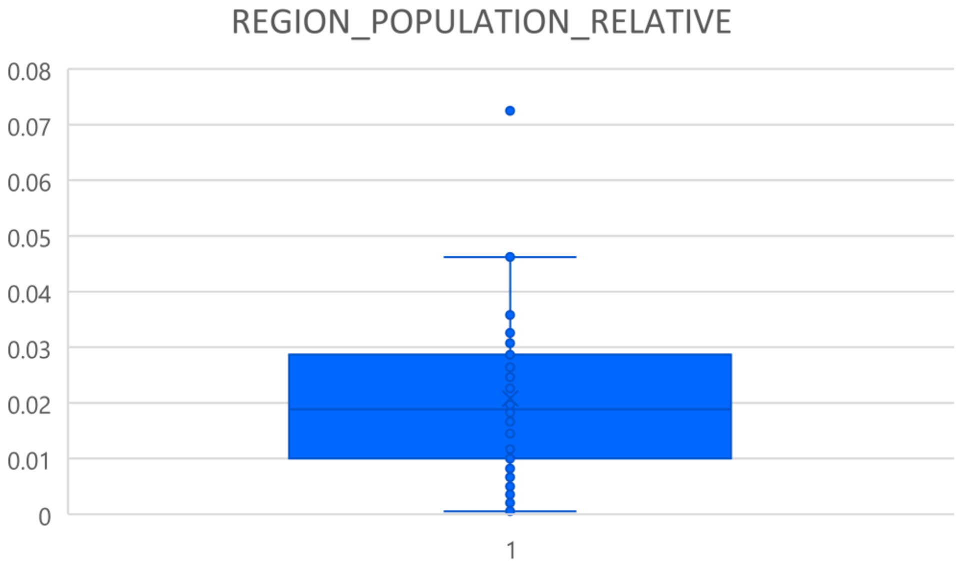
Amount income has outliers ranging from 0 to 36,00,000.

Amount credit has most outliers ranging from 1.3L to 27L.

Task B: Identify Outliers In The Dataset

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





Count of children has most outliers ranging from 0 to 2.

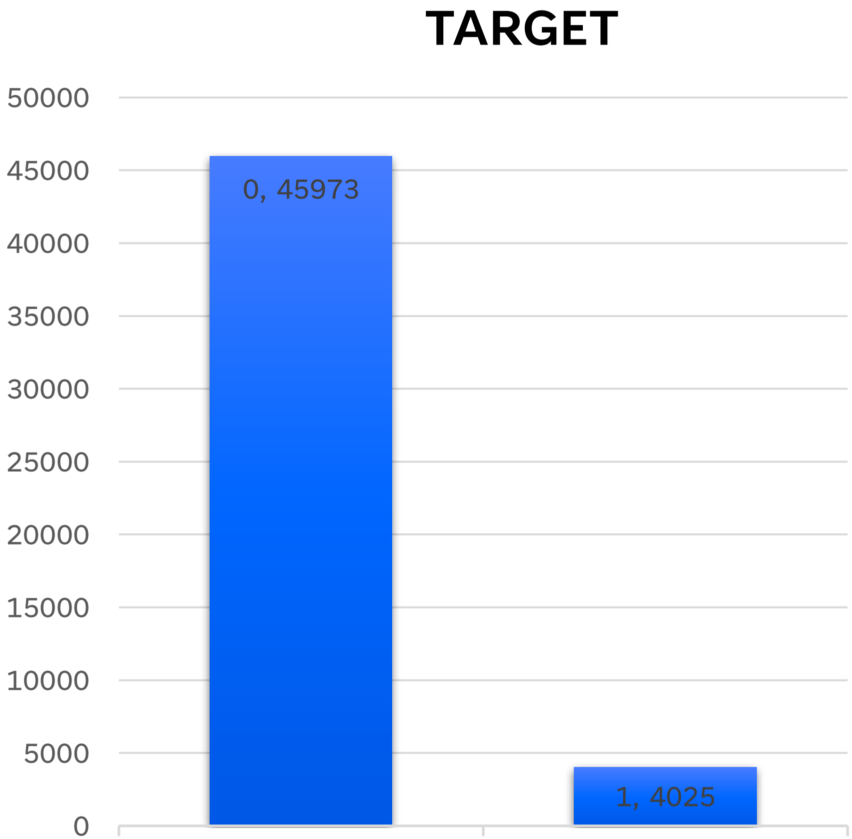
Region population relative has most outliers ranging from 0 to 0.046.

Task C: Analyze Data Imbalance

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

The target variable analysis reveals a class in balance issue with significantly high number of instances of “0” class compared to the “1” class

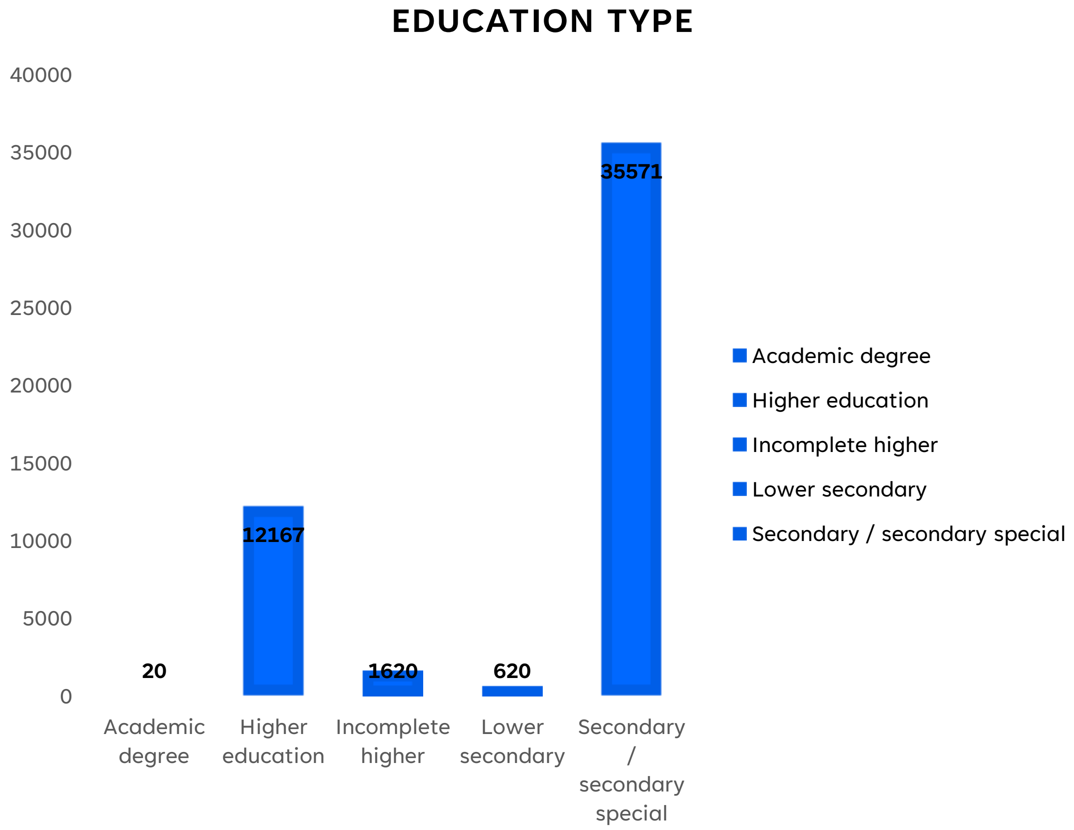
TARGET	Count
0	45973
1	4025
Grand Total	49998



Task C: Analyze Data Imbalance

The majority of loan applicants have completed Secondary/ secondary special education, followed by higher education, indicating a diverse educational background among applicants.

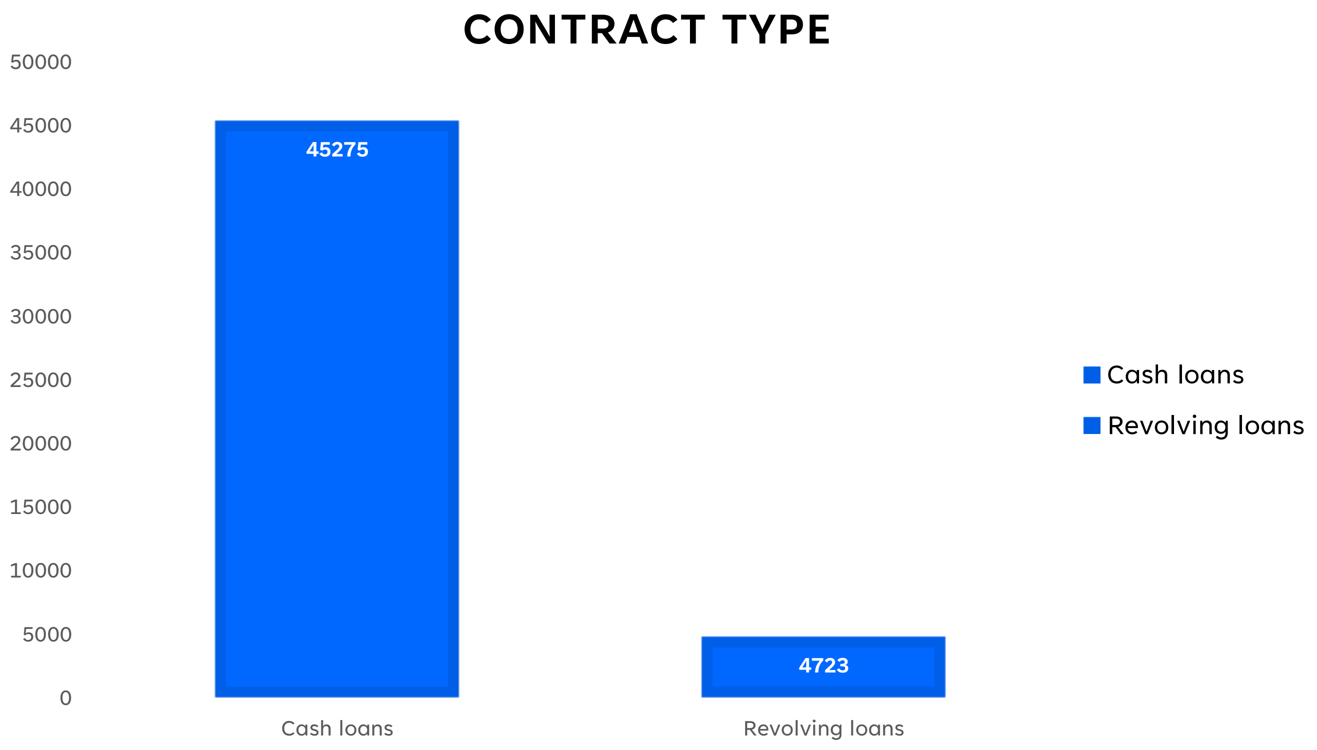
Education type	Count of Secondary / secondary special
Academic degree	20
Higher education	12167
Incomplete higher	1620
Lower secondary	620
Secondary / secondary special	35571
Grand Total	**49998**



Task C: Analyze Data Imbalance

The majority of loans are Cash loans (45,275), while a smaller proportion are Revolving loans (4723).

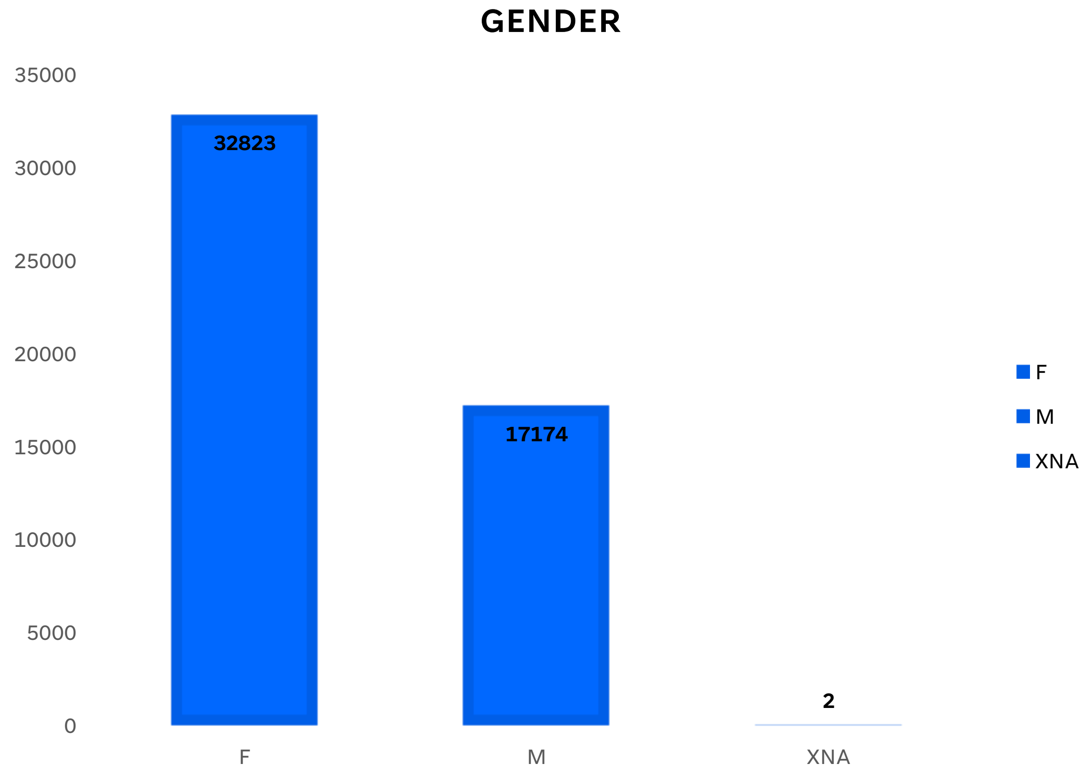
CONTRACT TYPE	Count
Cash loans	45275
Revolving loans	4723
Grand Total	49998



Task C: Analyze Data Imbalance

The CODE_GENDER analysis reveals a gender distribution among loan applicants with a higher number of females(32,823) compared to males(17,174).

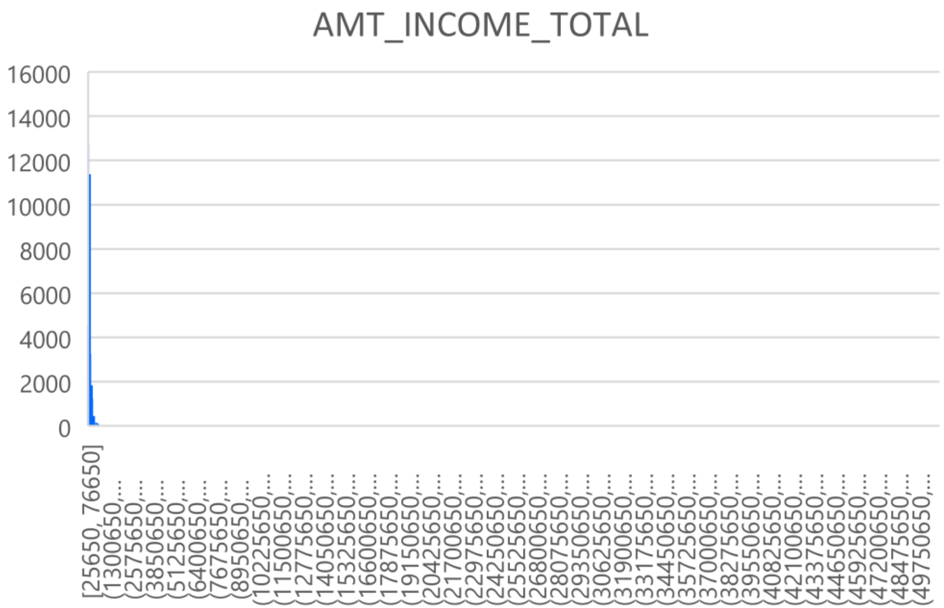
GENDER	Count
F	32823
M	17174
XNA	2
Grand Total	49999

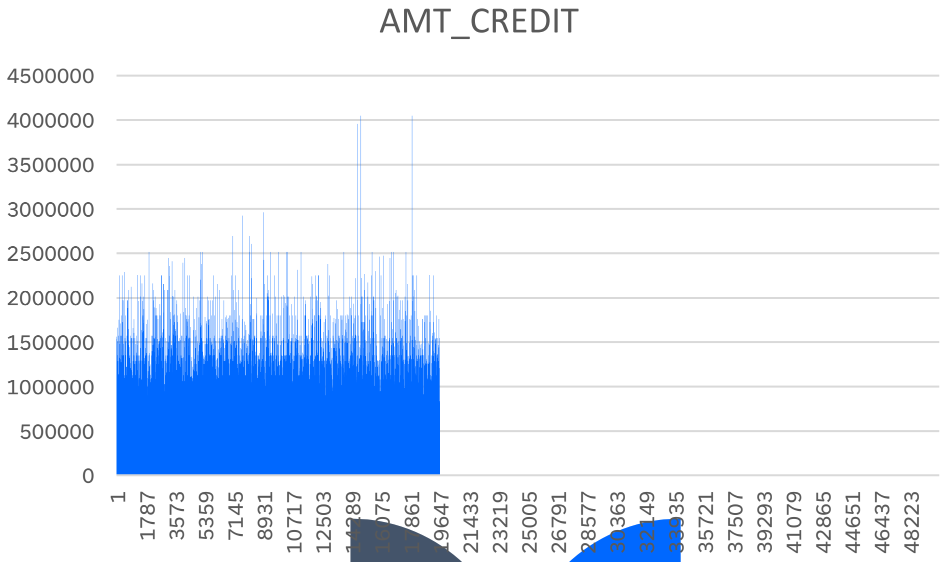


Task D: Perform Univariate, Segmented Univariate And Bivariate Analysis

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 value using Excel functions and features.

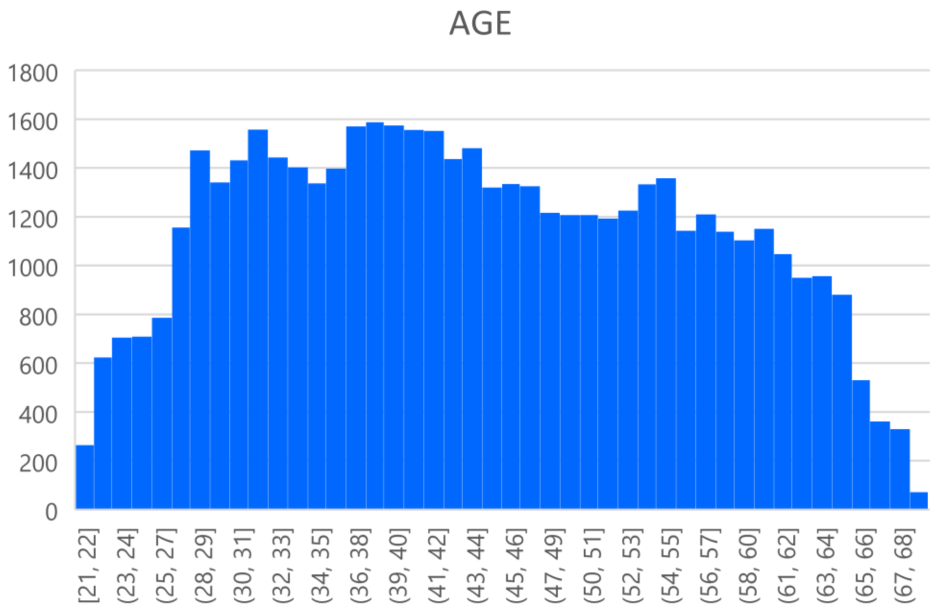
Univariate/ Segmented Univariate Analysis

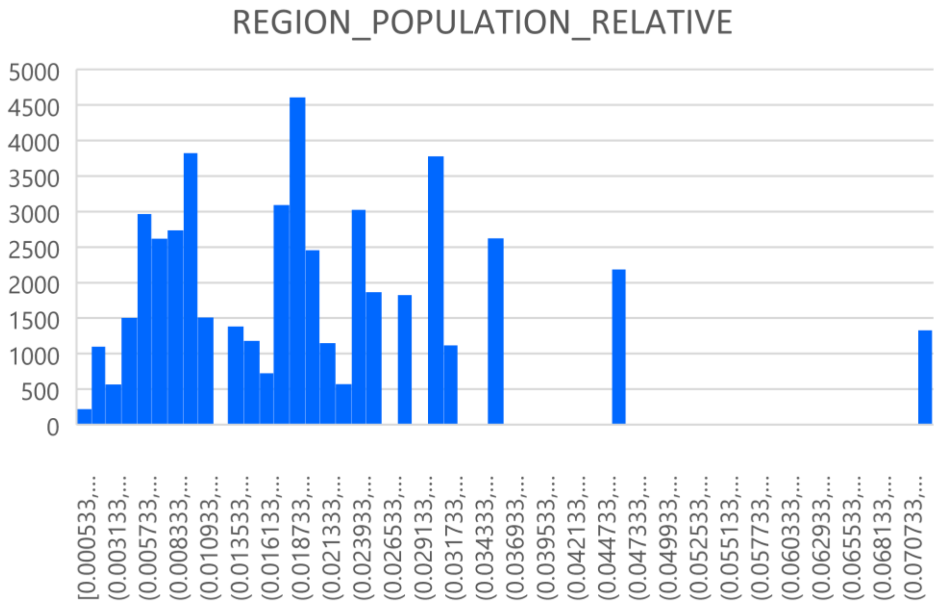




Task D: Perform Univariate, Segmented Univariate And Bivariate Analysis

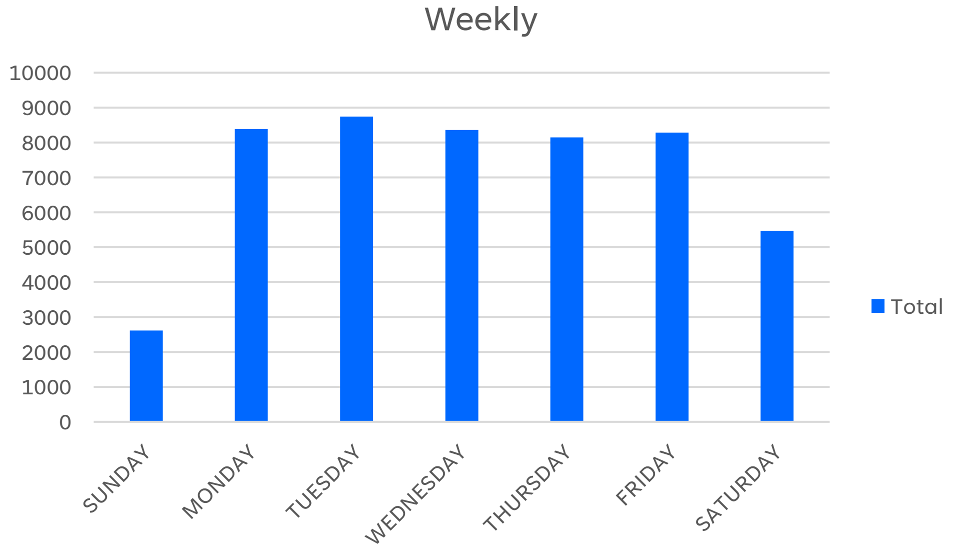
Univariate / Segmented Univariate Analysis

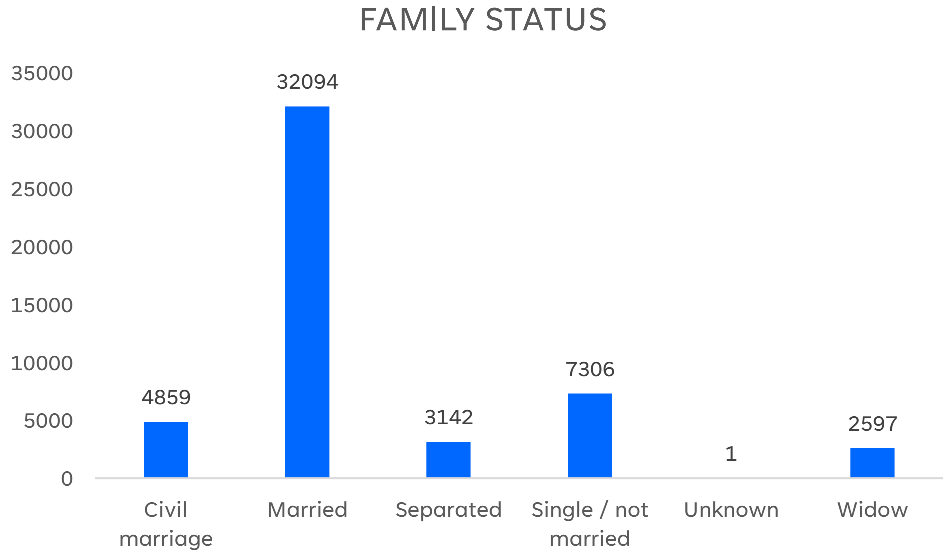




Task D: Perform Univariate, Segmented Univariate And Bivariate Analysis

Univariate / Segmented Univariate Analysis

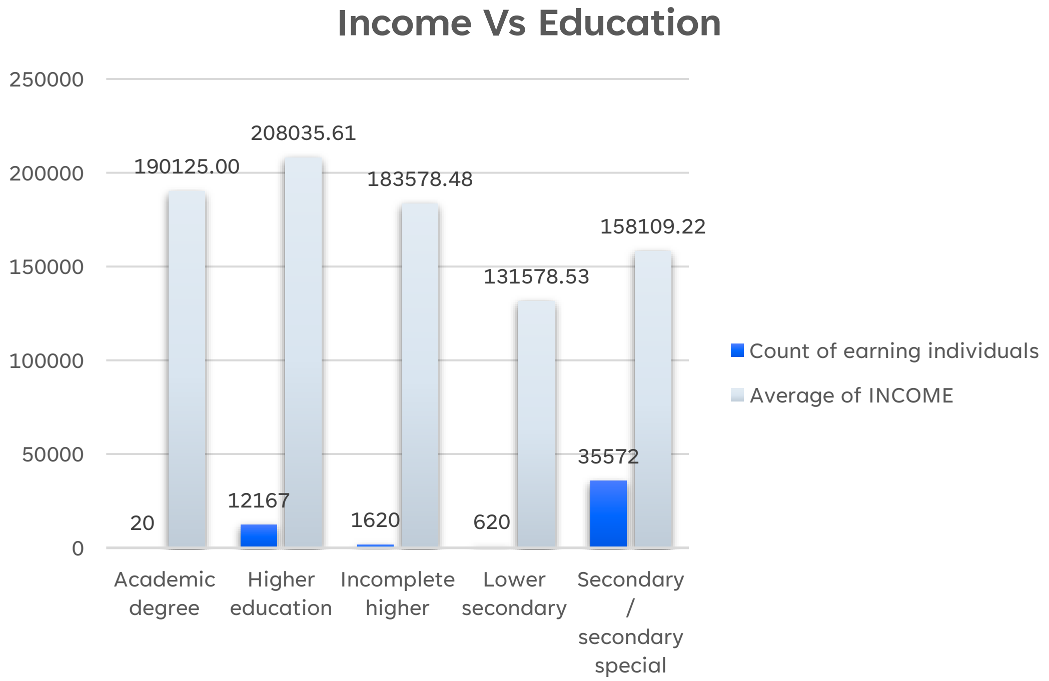


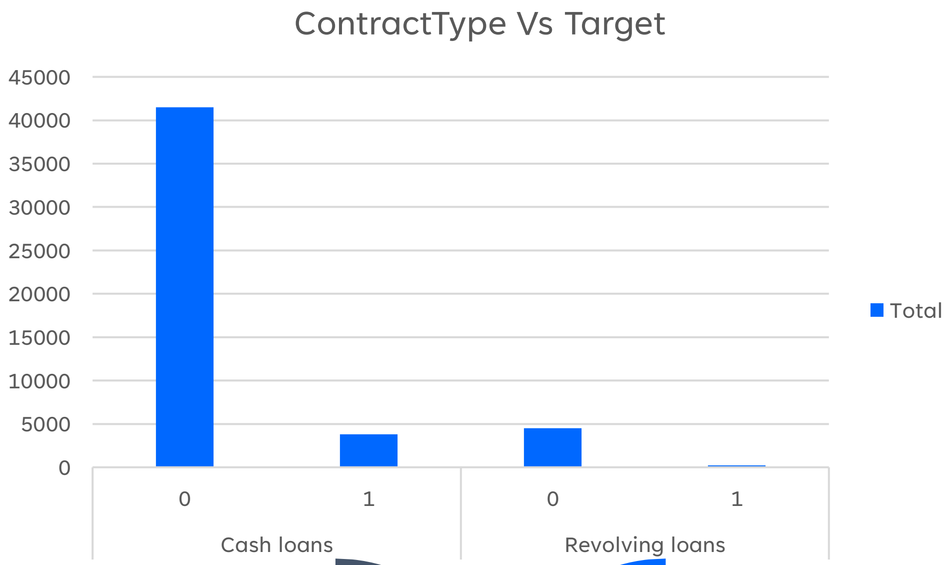


Task D: Perform Univariate, Segmented Univariate And Bivariate Analysis

.

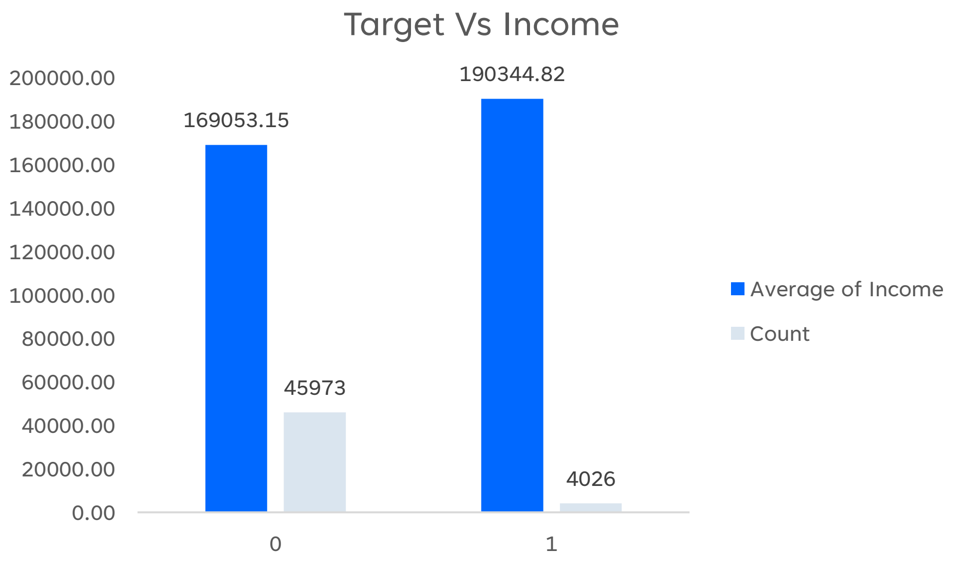
Bivariate Analysis

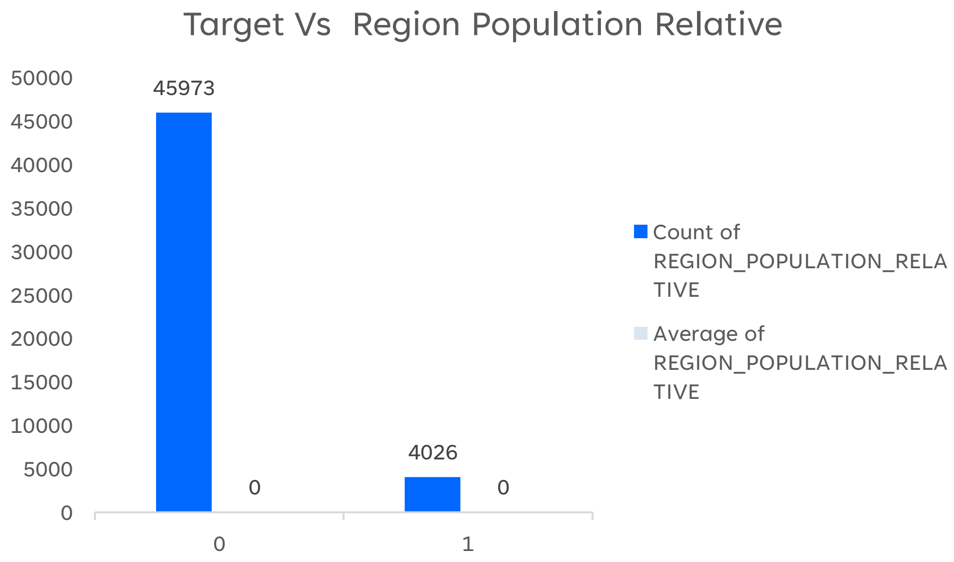




Task D: Perform Univariate, Segmented Univariate And Bivariate Analysis

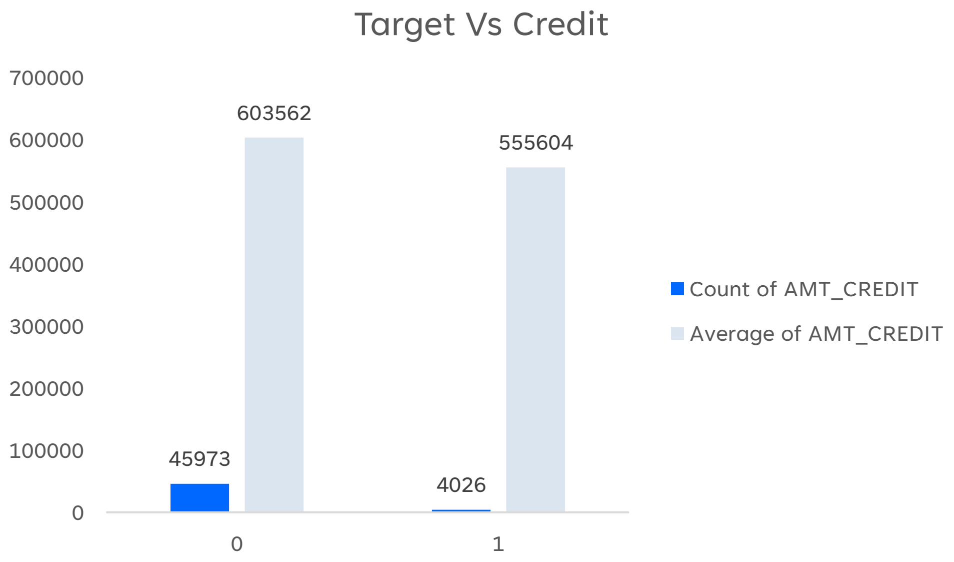
Bivariate Analysis

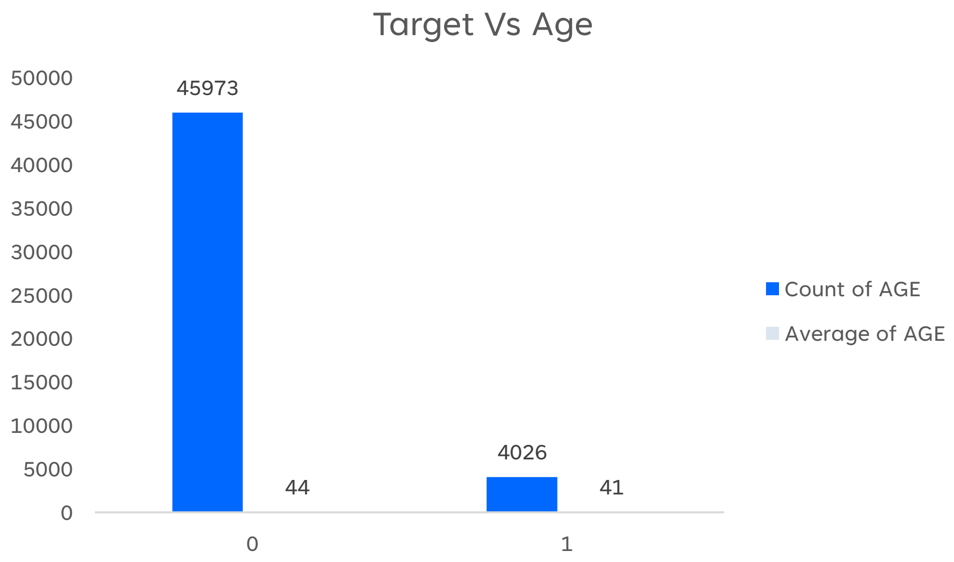




Task D: Perform Univariate, Segmented Univariate And Bivariate Analysis

Bivariate Analysis

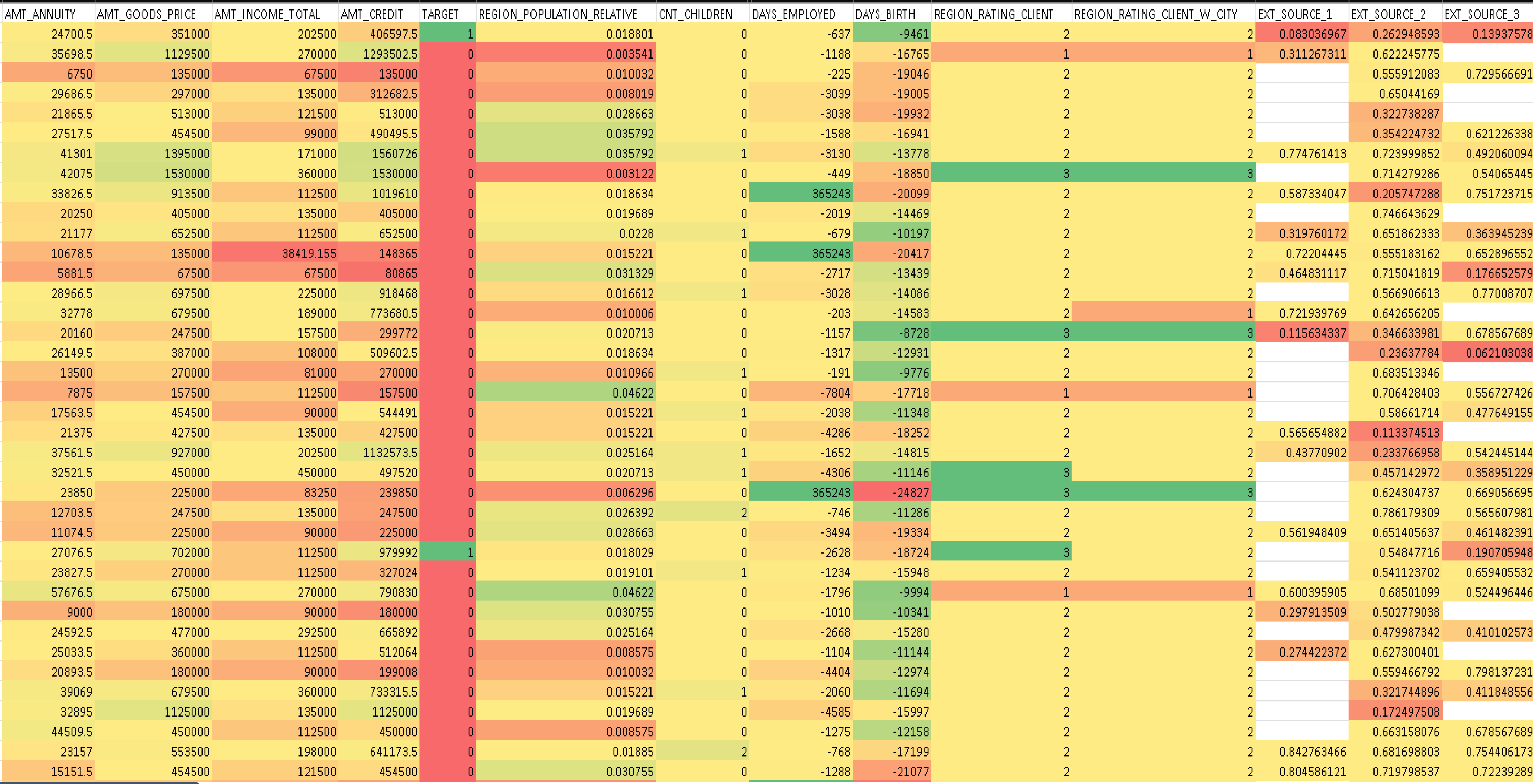




Task E: Identify top correlations for different scenarios

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.

Columns	Corrrelation Coefficient
AMT_ANNUITY	-0.012399094
AMT_CREDIT	-0.032428347
AMT_GOODS_PRICE	-0.041306523
AMT_INCOME_TOTAL	0.010893745
CNT_CHILDREN	0.010893745
DAYS_BIRTH	0.076787685
DAYS_EMPLOYED	-0.040294905
EXT_SOURCE_1	-0.156806293
EXT_SOURCE_2	-0.158424274
EXT_SOURCE_3	-0.181275965
REGION_POPULATION_RELATIVE	-0.040799172
REGION_RATING_CLIENT	0.066130148
REGION_RATING_CLIENT_W_CITY	0.067079294



AMT_ANNUITY	AMT_GOODS_PRICE	AMT_INCOME_TOTAL	AMT_CREDIT	TARGET	REGION_POPULATION_RELATIVE	CNT_CHILDREN	DAYS_EMPLOYED	DAYS_BIRTH	REGION_RATING_CLIENT	REGION_RATING_CLIENT_W_CITY	EXT_SOURCE_1	EXT_SOURCE_2	EXT_SOURCE_3
24700.5	351000	202500	406597.5	1	0.018801	0	-637	-9461	2	2	0.083036967	0.262948593	0.13937578
35698.5	1129500	270000	1293502.5	0	0.003541	0	-1188	-16765	1	1	0.311267311	0.622245775	
6750	135000	67500	135000	0	0.010032	0	-225	-19046	2	2		0.555912083	0.729566691
29686.5	297000	135000	312682.5	0	0.008019	0	-3039	-19005	2	2		0.65044169	
21865.5	513000	121500	513000	0	0.028663	0	-3038	-19932	2	2		0.322738287	
27517.5	454500	99000	490495.5	0	0.035792	0	-1588	-16941	2	2		0.354224732	0.621226338
41301	1395000	171000	1560726	0	0.035792	1	-3130	-13778	2	2	0.774761413	0.723999852	0.492060094
42075	1530000	360000	1530000	0	0.003122	0	-449	-18850	3	3		0.714279286	0.54065445
33826.5	913500	112500	1019610	0	0.018634	0	365243	-20099	2	2	0.587334047	0.205747288	0.751723715
20250	405000	135000	405000	0	0.019689	0	-2019	-14469	2	2		0.746643629	
21177	652500	112500	652500	0	0.0228	1	-679	-10197	2	2	0.319760172	0.651862333	0.363945239
10678.5	135000	38419.155	148365	0	0.015221	0	365243	-20417	2	2	0.72204445	0.555183162	0.652896552
5881.5	67500	67500	80865	0	0.031329	0	-2717	-13439	2	2	0.464831117	0.715041819	0.176652579
28966.5	697500	225000	918468	0	0.016612	1	-3028	-14086	2	2		0.566906613	0.77008707
32778	679500	189000	773680.5	0	0.010006	0	-203	-14583	2	1	0.721939769	0.642656205	
20160	247500	157500	299772	0	0.020713	0	-1157	-8728	3	3	0.115634337	0.346633981	0.678567689
26149.5	387000	108000	509602.5	0	0.018634	0	-1317	-12931	2	2		0.23637784	0.062103038
13500	270000	81000	270000	0	0.010966	1	-191	-9776	2	2		0.683513346	
7875	157500	112500	157500	0	0.04622	0	-7804	-17718	1	1		0.706428403	0.556727426
17563.5	454500	90000	544491	0	0.015221	1	-2038	-11348	2	2		0.58661714	0.477649155
21375	427500	135000	427500	0	0.015221	0	-4286	-18252	2	2	0.565654882	0.113374513	
37561.5	927000	202500	1132573.5	0	0.025164	1	-1652	-14815	2	2	0.43770902	0.233766958	0.542445144
32521.5	450000	450000	497520	0	0.020713	1	-4306	-11146	3	2		0.457142972	0.358951229
23850	225000	83250	239850	0	0.006296	0	365243	-24827	3	3		0.624304737	0.669056695
12703.5	247500	135000	247500	0	0.026392	2	-746	-11286	2	2		0.786179309	0.565607981
11074.5	225000	90000	225000	0	0.028663	0	-3494	-19334	2	2	0.561948409	0.651405637	0.461482391
27076.5	702000	112500	979992	1	0.018029	0	-2628	-18724	3	2		0.54847716	0.190705948
23827.5	270000	112500	327024	0	0.019101	1	-1234	-15948	2	2		0.541123702	0.659405532
57676.5	675000	270000	790830	0	0.04622	0	-1796	-9994	1	1	0.600395905	0.68501099	0.524496446
9000	180000	90000	180000	0	0.030755	0	-1010	-10341	2	2	0.297913509	0.502779038	
24592.5	477000	292500	665892	0	0.025164	0	-2668	-15280	2	2		0.479987342	0.410102573
25033.5	360000	112500	512064	0	0.008575	0	-1104	-11144	2	2	0.274422372	0.627300401	
20893.5	180000	90000	199008	0	0.010032	0	-4404	-12974	2	2		0.559466792	0.798137231
39069	679500	360000	733315.5	0	0.015221	1	-2060	-11694	2	2		0.321744896	0.411848556
32895	1125000	135000	1125000	0	0.019689	0	-4585	-15997	2	2		0.172497508	
44509.5	450000	112500	450000	0	0.008575	0	-1275	-12158	2	2		0.663158076	0.678567689
23157	553500	198000	641173.5	0	0.01885	2	-768	-17199	2	2	0.842763466	0.681698803	0.754406173
15151.5	454500	121500	454500	0	0.030755	0	-1288	-21077	2	2	0.804586121	0.719798537	0.72239289

DRIVE LINK

https://docs.google.com/spreadsheets/d/1yJ5ychdWfzMnnQNwJNuXsjyK0iHrqnDe/edit?usp=sharing&ouid=102610263604488941402&rtpof=true&sd=true

Result

This project on exploratory data analysis (EDA) of a loan application data set has significantly advanced my skills and knowledge it enabled me to master critical data handling and cleaning techniques, statistical analysis and data visualisation using Excel. Moreover, I developed valuable problem-solving skills, particularly in addressing data quality issues such as missing data, outliers, data imbalances. This project has enhanced my ability to think analytically and communicate complex data insights effectively. It has strengthened my portfolio as a budding data scientist and analyst.

Thank you

Ritika Chaudhary

ameliarc2105@gmail.com