BANK LOAN CASE STUDY

CONTENT

01 PROJECT DESCRIPTION

02 APPROACH

03 TECH-STACK USED

04 INSIGHTS

05 RESULT

PROJECT DESCRIPTION

The "Bank Loan Case Study" project utilizes data analytics to gain valuable insights in the lending industry. As a Data Analyst, the objective is to explore patterns and trends related to bank loan applications, identify factors influencing loan defaults, and create predictive models for assessing applicant risk. By segmenting customers based on their risk profiles and interpreting model results, actionable recommendations will be provided to improve the loan approval process and optimize lending strategies, enabling stakeholders to make data-driven decisions and gain a competitive edge in the financial services industry.

APPROACH

The "Bank Loan Case Study" project utilizes Excel for data analytics tasks to gain insights into the lending industry and loan approval process. Objectives include handling missing data effectively with COUNT, ISBLANK, and IF functions, detecting outliers in numerical variables using QUARTILE and IQR, and analyzing data imbalance with COUNTIF and SUM functions, visualized through pie charts or bar charts. The project aims to provide valuable insights for loan decision-making and risk management in a concise manner.

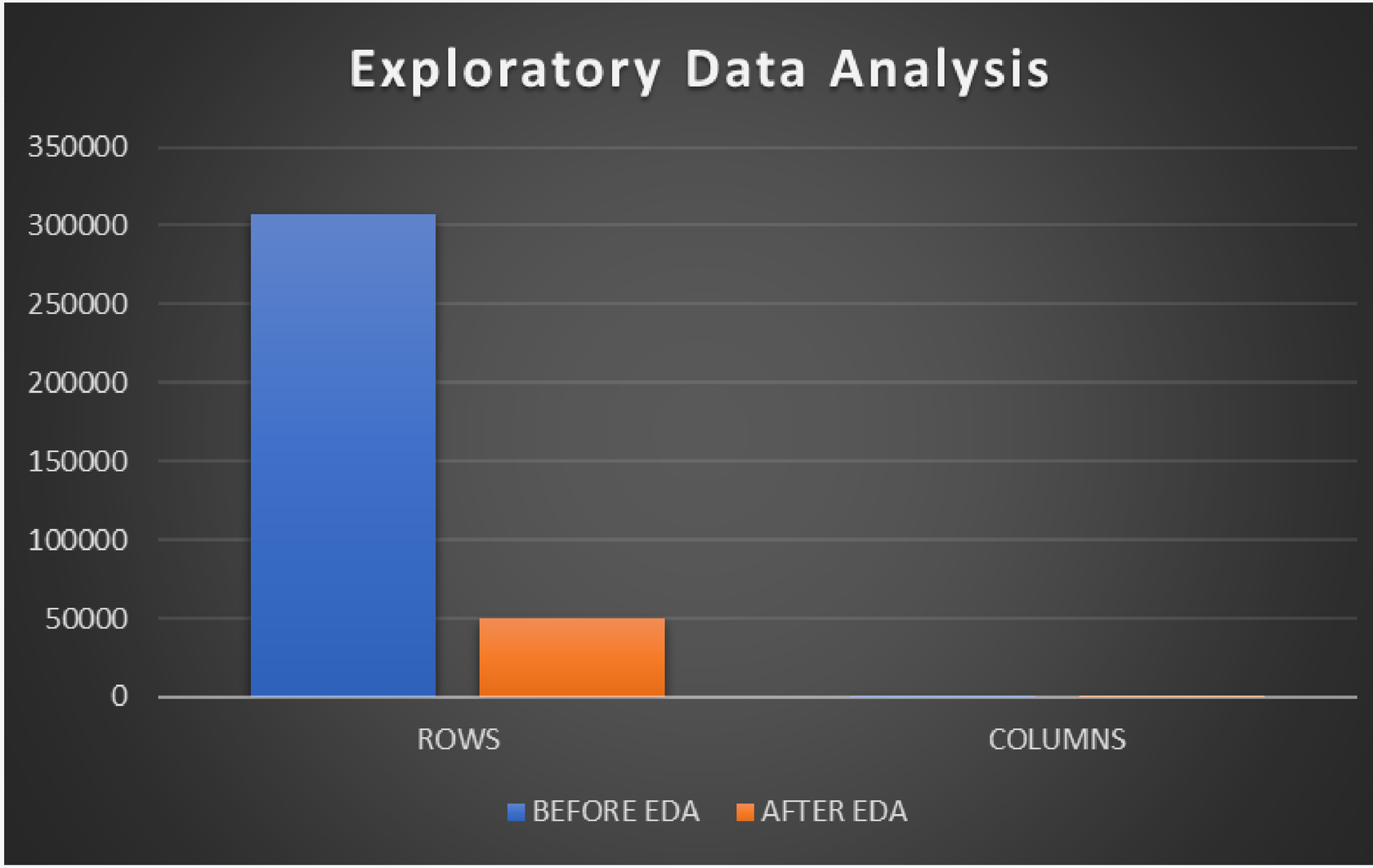
The main focus of the analysis will be to identify factors influencing loan defaults through exploratory data analysis (EDA). Using Excel functions like COUNT, AVERAGE, MEDIAN, and various statistical tools, we will perform univariate, segmented univariate, and bivariate analysis to gain insights into customer attributes, loan characteristics, and payment histories. By visualizing the relationships between variables and loan defaults using Excel's features like filters and pivot tables, we aim to uncover valuable patterns and correlations within different loan scenarios.

The project's deliverables will include actionable recommendations based on the findings from the Excel analysis, empowering the finance company to make informed decisions about loan approvals, loan amounts, and interest rates. The insights gained will support the company's goal of reducing loan defaults and optimizing its lending strategies, ensuring it remains competitive and successful in the dynamic world of banking and finance.

TECH-STACK USED

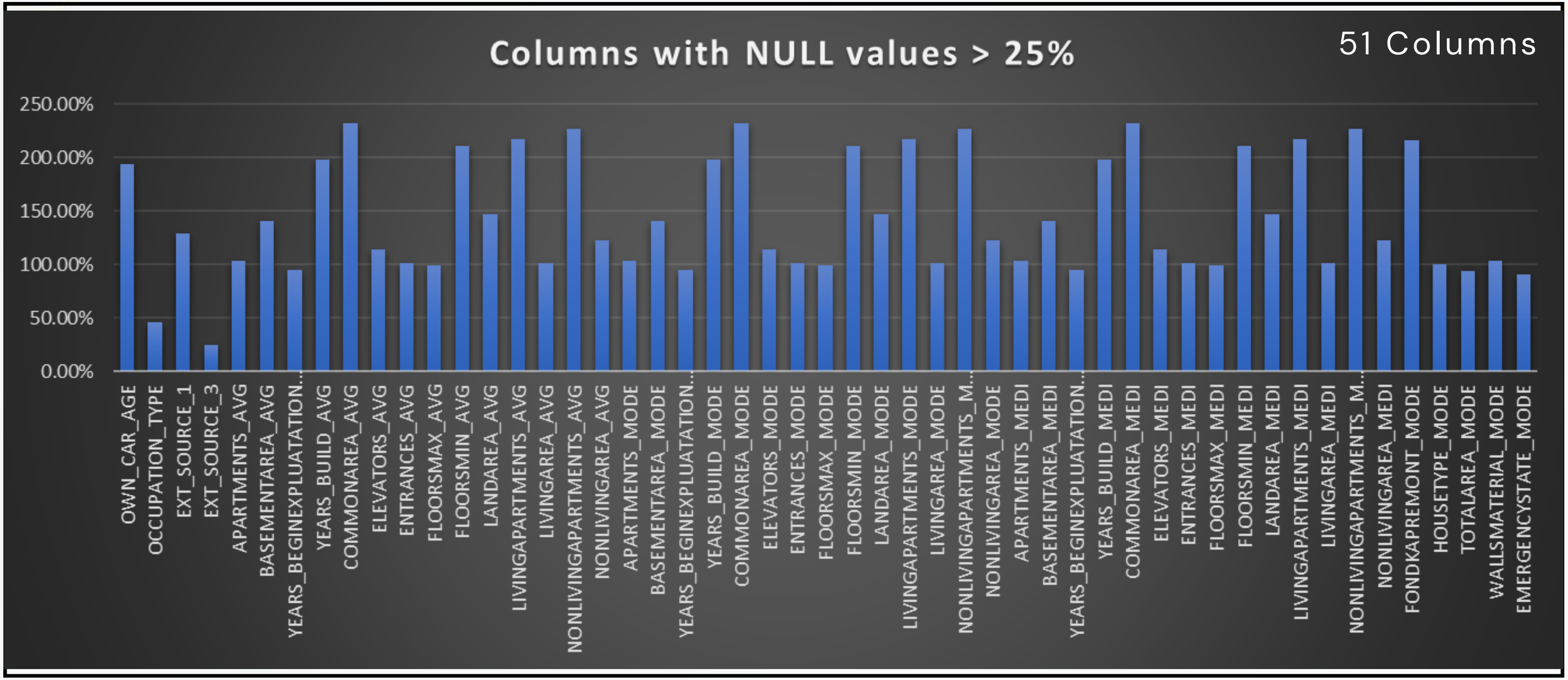
In the "Bank Loan Case Study" project, I will utilize Excel as the primary tool for data extraction and analysis. Leveraging Excel's powerful features and functions, I aim to efficiently process and analyze the bank loan dataset. By applying various data analysis techniques, I intend to gain valuable insights into the lending industry, identify factors influencing loan defaults, and provide actionable recommendations. This data-driven approach in Excel will facilitate informed decision-making and optimization of the loan approval process, contributing to a deeper understanding of the industry and improved decision-making for stakeholders. The use of Excel's functionalities will prove crucial in conducting thorough analysis and extracting valuable insights for the project's objectives.

INSIGHTS



- Initial dataset: 3,07,514 rows, 126 columns.
- Cleaned dataset: 49513 rows, 72 columns.
- By removing irrelevant columns and handling null values, the dataset was streamlined for analysis. This optimization improved efficiency and facilitated the extraction of meaningful insights.

INSIGHTS

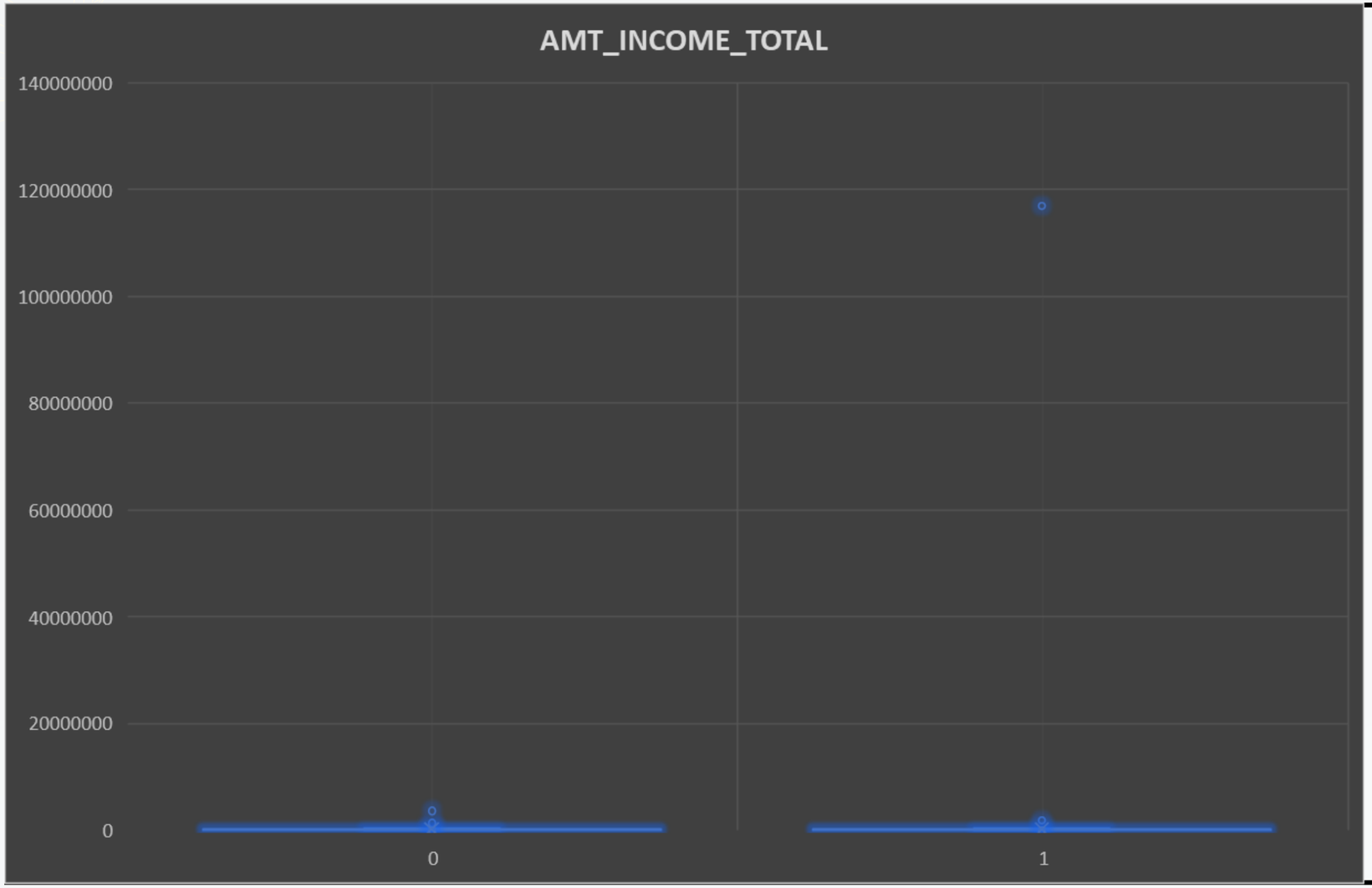

Columns with NULL values > 25%
51 Columns
250.00%
200.00%
150.00%
100.00%
50.00%
0.00%
OWN_CAR_AGE
OCCUPATION_TYPE
EXT_SOURCE_1
EXT_SOURCE_3
APARTMENTS_AVG
BASEMENTAREA_AVG
YEARS_BEGINEXPLUATATION...
YEARS_BUILD_AVG
COMMONAREA_AVG
ELEVATORS_AVG
ENTRANCES_AVG
FLOORSMAX_AVG
FLOORSMIN_AVG
LANDAREA_AVG
LIVINGAPARTMENTS_AVG
LIVINGAREA_AVG
NONLIVINGAPARTMENTS_AVG
NONLIVINGAREA_AVG
APARTMENTS_MODE
BASEMENTAREA_MODE
YEARS_BEGINEXPLUATATION...
YEARS_BUILD_MODE
COMMONAREA_MODE
ELEVATORS_MODE
ENTRANCES_MODE
FLOORSMAX_MODE
FLOORSMIN_MODE
LANDAREA_MODE
LIVINGAPARTMENTS_MODE
LIVINGAREA_MODE
NONLIVINGAPARTMENTS_M...
NONLIVINGAREA_MODE
APARTMENTS_MEDI
BASEMENTAREA_MEDI
YEARS_BEGINEXPLUATATION...
YEARS_BUILD_MEDI
COMMONAREA_MEDI
ELEVATORS_MEDI
ENTRANCES_MEDI
FLOORSMAX_MEDI
FLOORSMIN_MEDI
LANDAREA_MEDI
LIVINGAPARTMENTS_MEDI
LIVINGAREA_MEDI
NONLIVINGAPARTMENTS_M...
NONLIVINGAREA_MEDI
FONDKAPREMONT_MODE
HOUSETYPE_MODE
TOTALAREA_MODE
WALLSMATERIAL_MODE
EMERGENCYSTATE_MODE

INSIGHTS

OWN_CAR_AGE	ENTRANCES_AVG	YEARS_BEGINEXPLUATATION_MODE	NONLIVINGAPARTMENTS_MODE	FLOORSMIN_MEDI
OCCUPATION_TYPE	FLOORSMAX_AVG	YEARS_BUILD_MODE	NONLIVINGAREA_MODE	LANDAREA_MEDI
EXT_SOURCE_1	FLOORSMIN_AVG	COMMONAREA_MODE	APARTMENTS_MEDI	LIVINGAPARTMENTS_MEDI
EXT_SOURCE_3	LANDAREA_AVG	ELEVATORS_MODE	BASEMENTAREA_MEDI	LIVINGAREA_MEDI
APARTMENTS_AVG	LIVINGAPARTMENTS_AVG	ENTRANCES_MODE	YEARS_BEGINEXPLUATATION_MEDI	NONLIVINGAPARTMENTS_MEDI
BASEMENTAREA_AVG	LIVINGAREA_AVG	FLOORSMAX_MODE	YEARS_BUILD_MEDI	NONLIVINGAREA_MEDI
YEARS_BEGINEXPLUATATION_AVG	NONLIVINGAPARTMENTS_AVG	FLOORSMIN_MODE	COMMONAREA_MEDI	FONDKAPREMONT_MODE
YEARS_BUILD_AVG	NONLIVINGAREA_AVG	LANDAREA_MODE	ELEVATORS_MEDI	HOUSETYPE_MODE
COMMONAREA_AVG	APARTMENTS_MODE	LIVINGAPARTMENTS_MODE	ENTRANCES_MEDI	TOTALAREA_MODE
ELEVATORS_AVG	BASEMENTAREA_MODE	LIVINGAREA_MODE	FLOORSMAX_MEDI	WALLSMATERIAL_MODE
				EMERGENCYSTATE_MODE

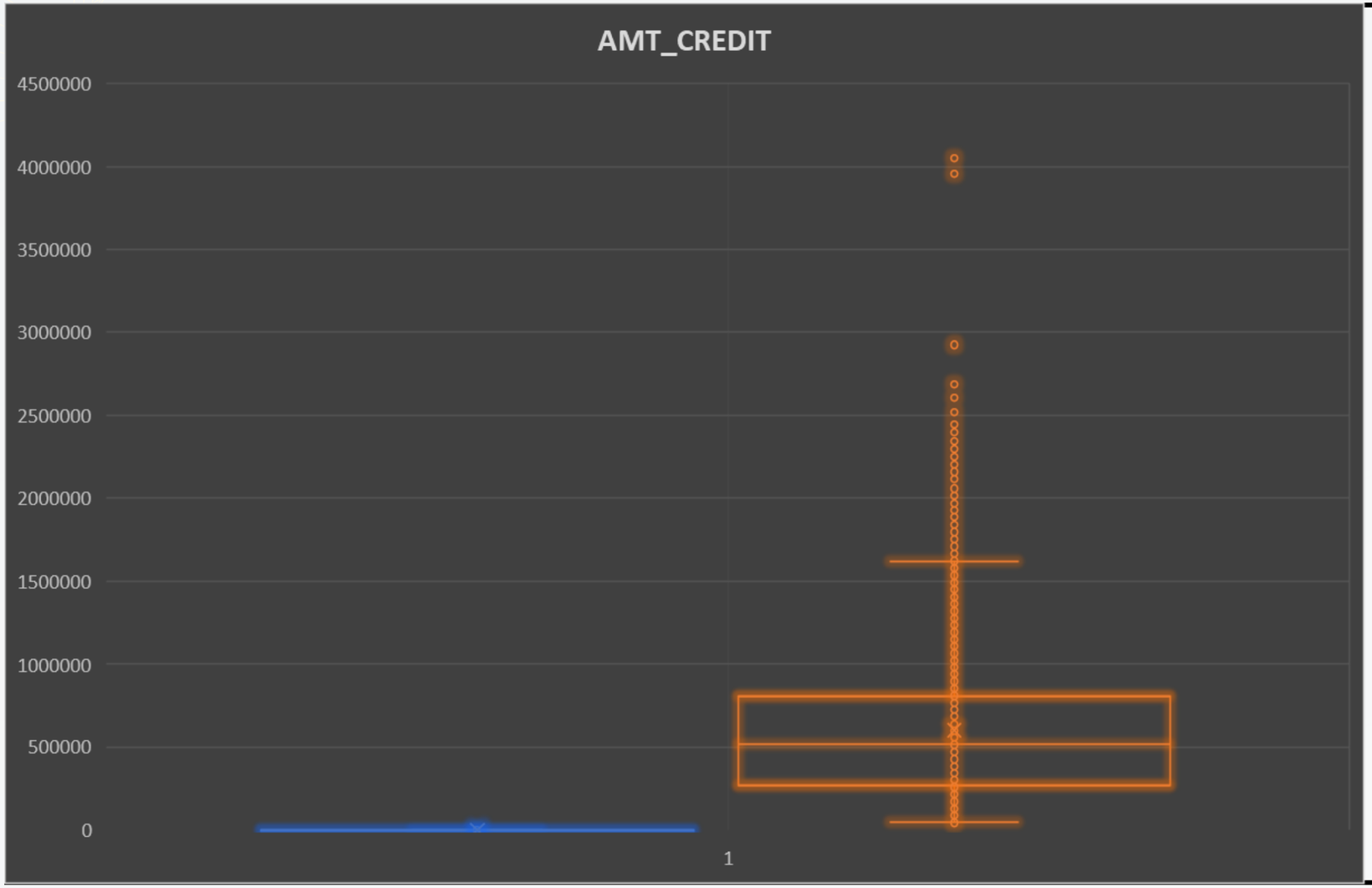
- There are 51 columns with null values > 25%.
- Most of these columns contains residential details of the applicants.
- Columns with null values < 25% are imputed using Mean, Median, or Mode as appropriate. This ensures data completeness for further analysis.

INSIGHTS



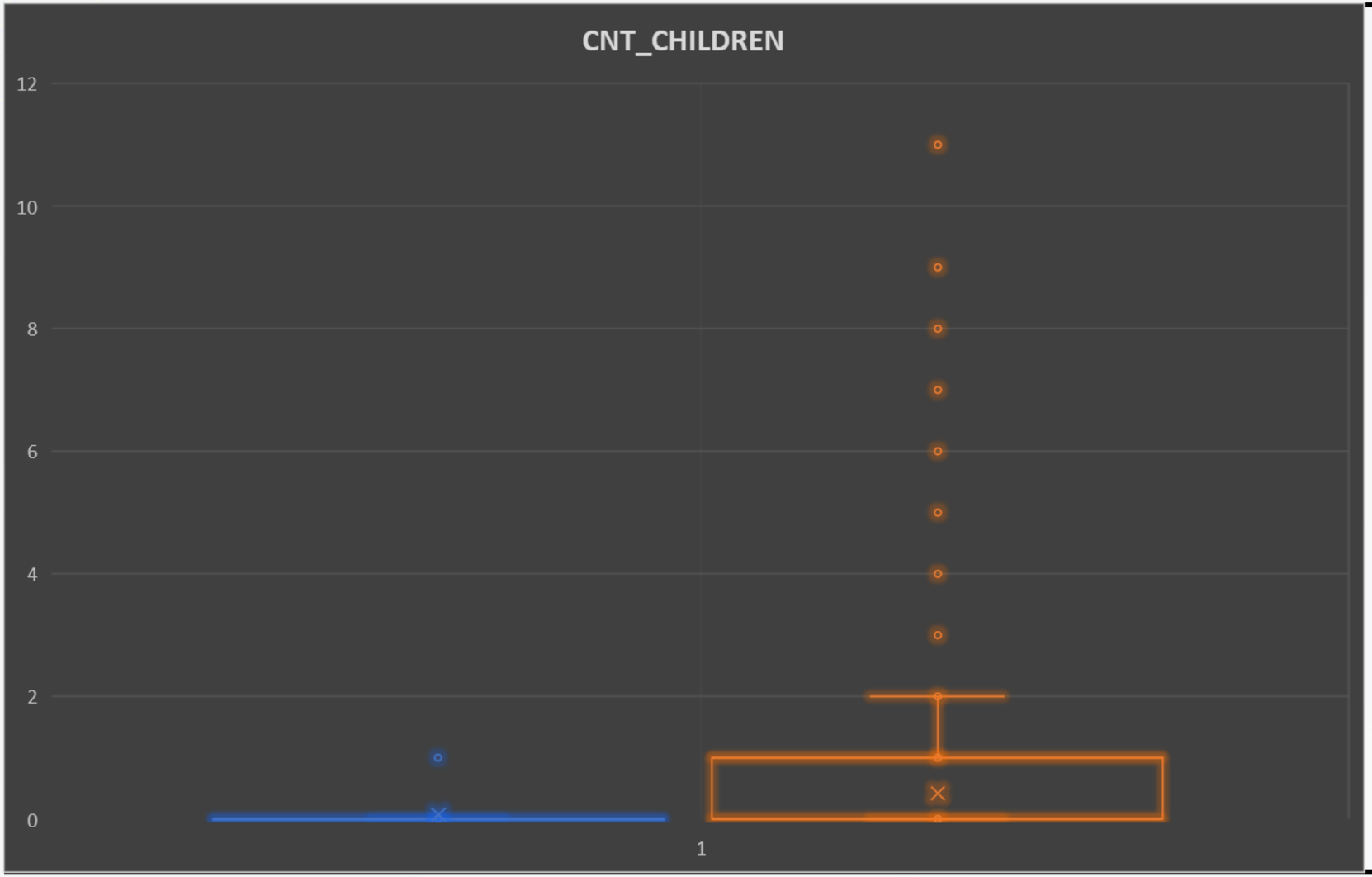
- In the AMT_INCOME_TOTAL column, we have identified the presence of outliers, with one particular data point showing an exceptionally high salary of 117,000,000. Such a value is considered an extreme outlier, as it significantly deviates from the typical income values in the dataset.

INSIGHTS



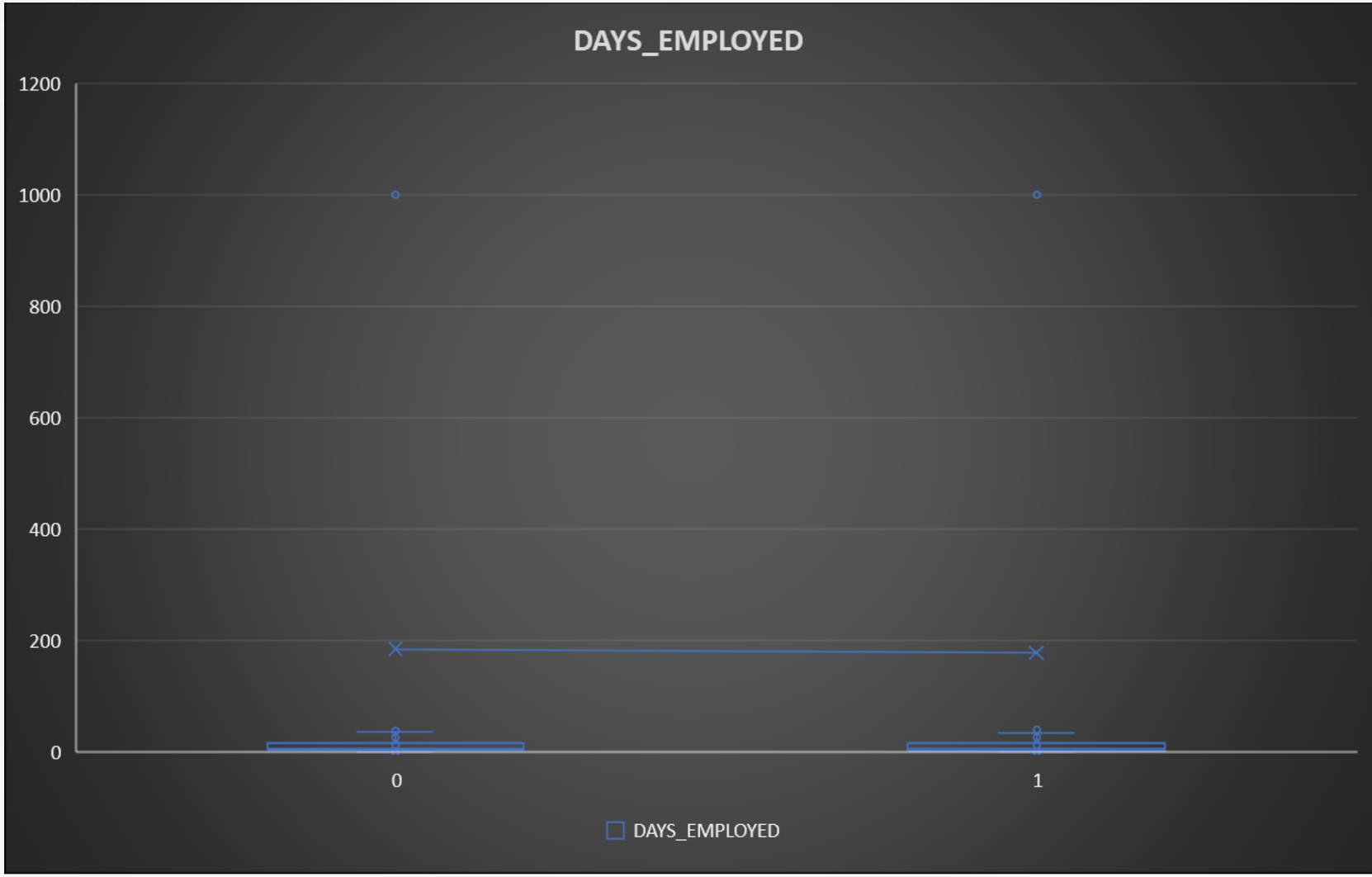
- The presence of outliers in the AMT_CREDIT column indicates that some loan amounts i.e. ~40,50,000 are unusually higher than the majority of loan values. These outliers may represent special cases or exceptional loan requests that significantly deviate from the typical loan amounts.

INSIGHTS



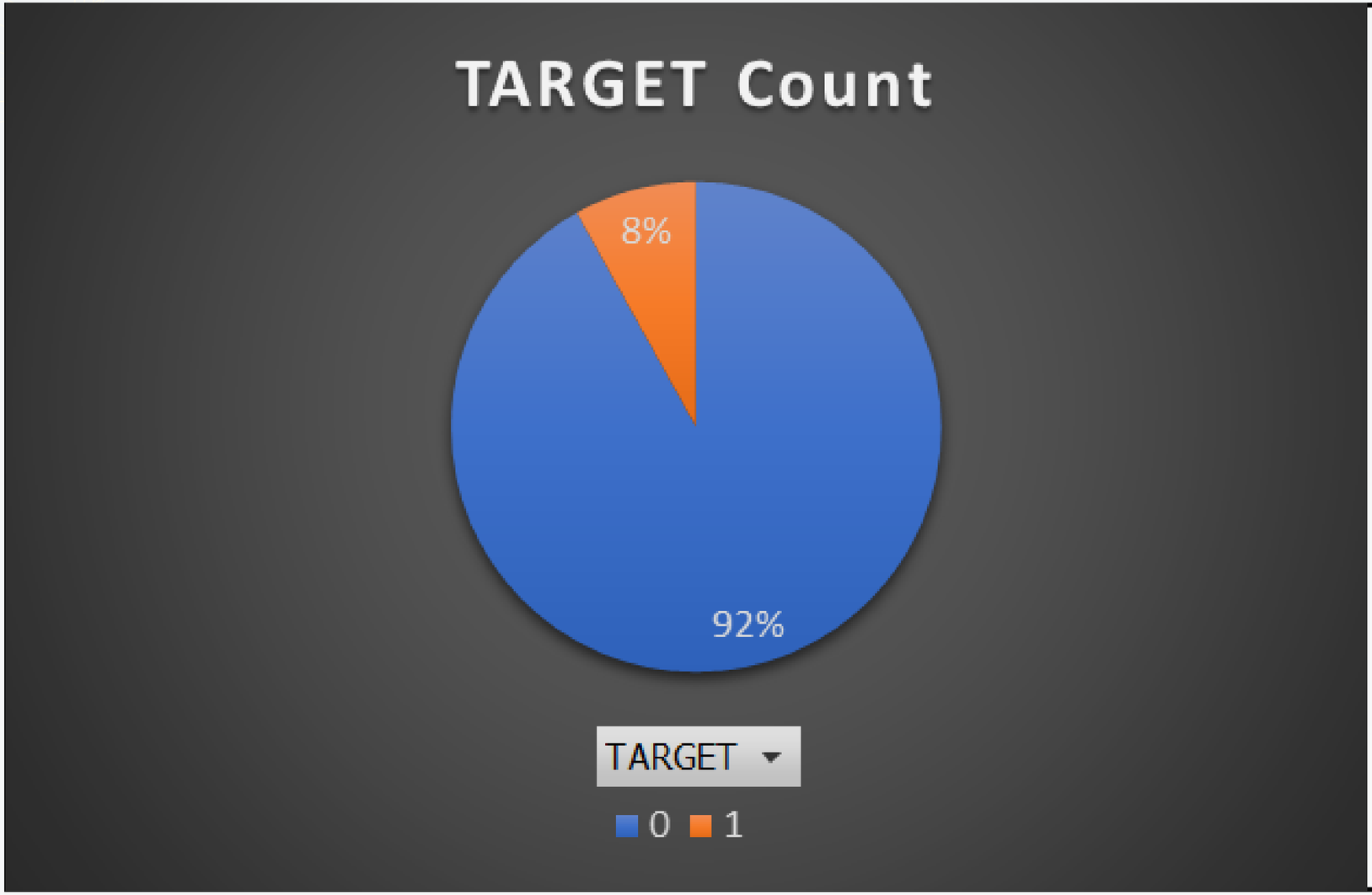
- The CNT_CHILDREN column displays outliers, indicating that there are individuals with an unusually high number of children, such as 19 children, which might not be realistic in today's context.
- The distribution of the CNT_CHILDREN column reveals that the majority of loan applicants have no children or a small number of children, as indicated by the mode and median values being 0.

INSIGHTS



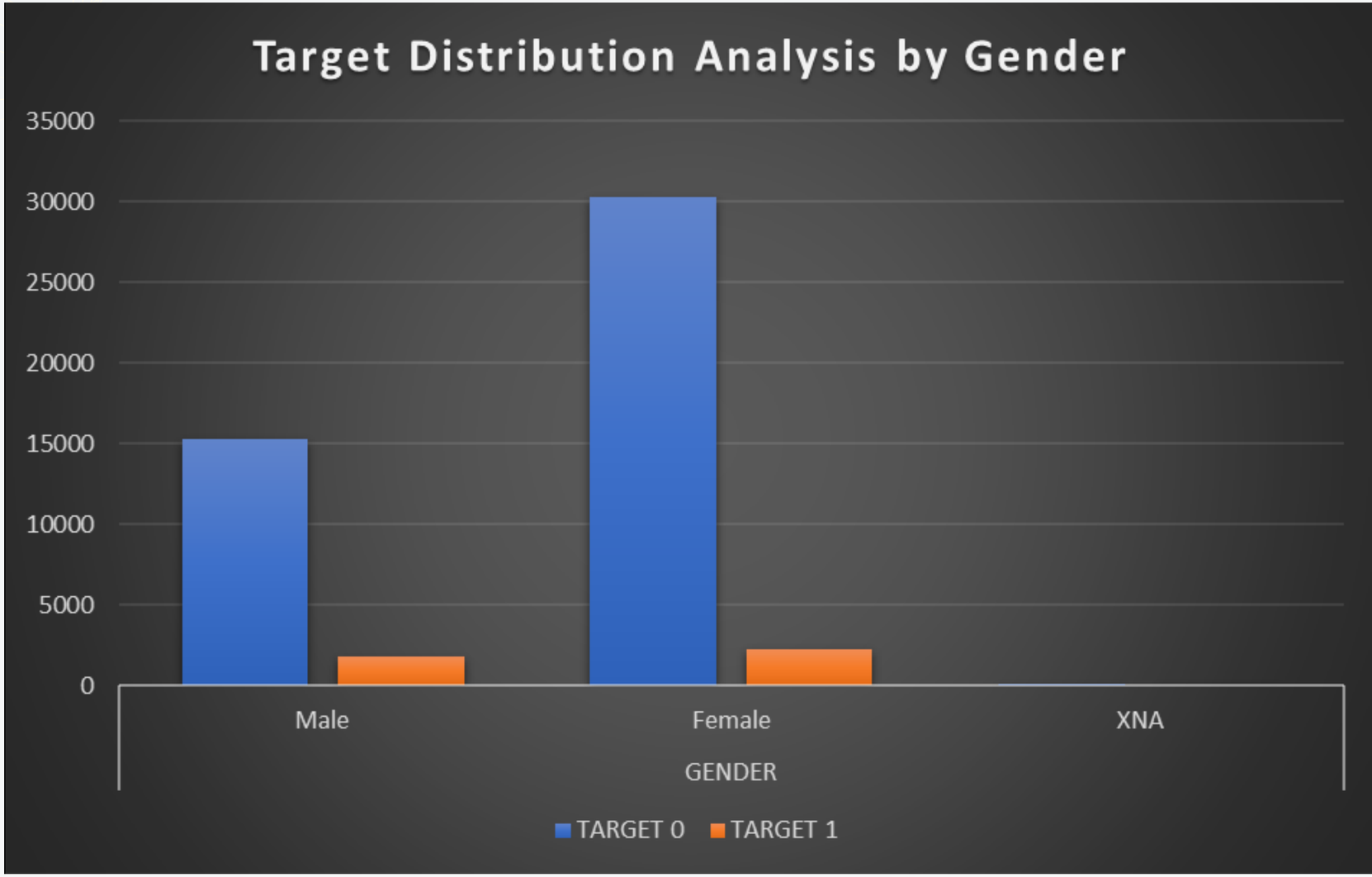
- The DAYS_EMPLOYED column displays significant outliers, indicating that some individuals have been employed for over 1001 years, which is clearly impossible and indicates a data issue.
- The mean employment duration of 183.92 suggests an unusually high value due to erroneous entries of 1001 years, impacting the analysis.

INSIGHTS



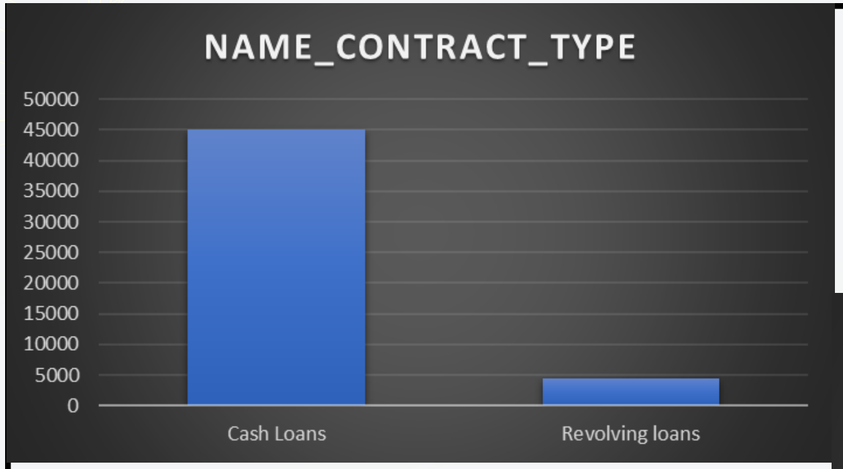
- The dataset shows a significant class imbalance, with approximately 92% of loan applicants categorized as loan re-payers and only 8% as defaulters.
- Ratio (1):(0) = 11.36
- This imbalance indicates that the number of defaulters is substantially lower than the number of re-payers in the dataset.

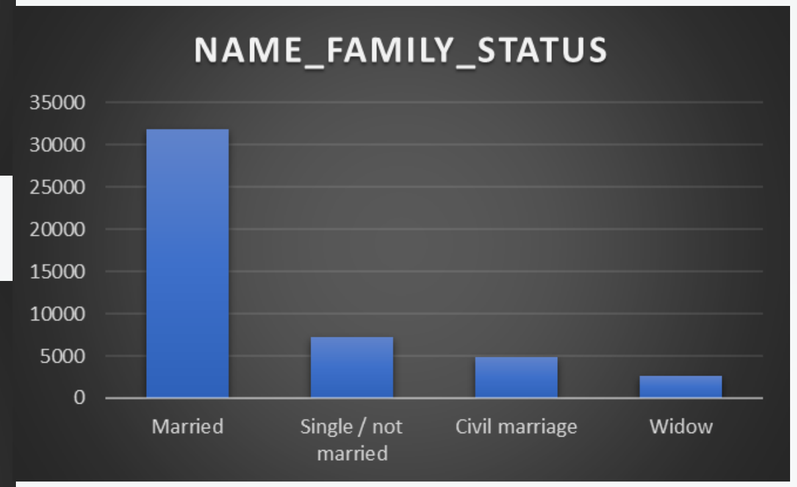
INSIGHTS

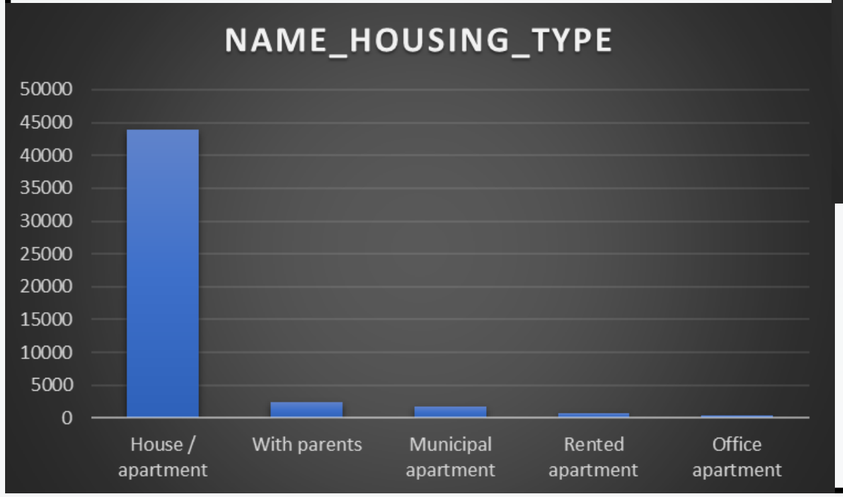


- The analysis reveals that a higher proportion of female loan applicants are loan repayers compared to male applicants.
- Percentage of Female Loan Repayers: 94.9%
- Percentage of Male Loan Repayers: 93.2%

INSIGHTS



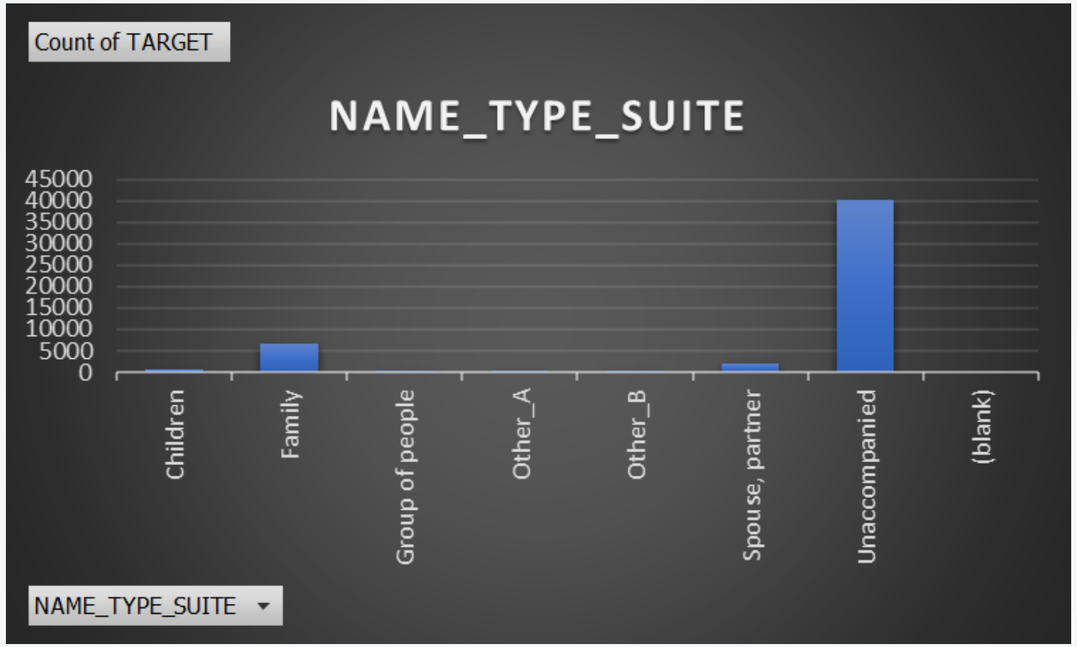


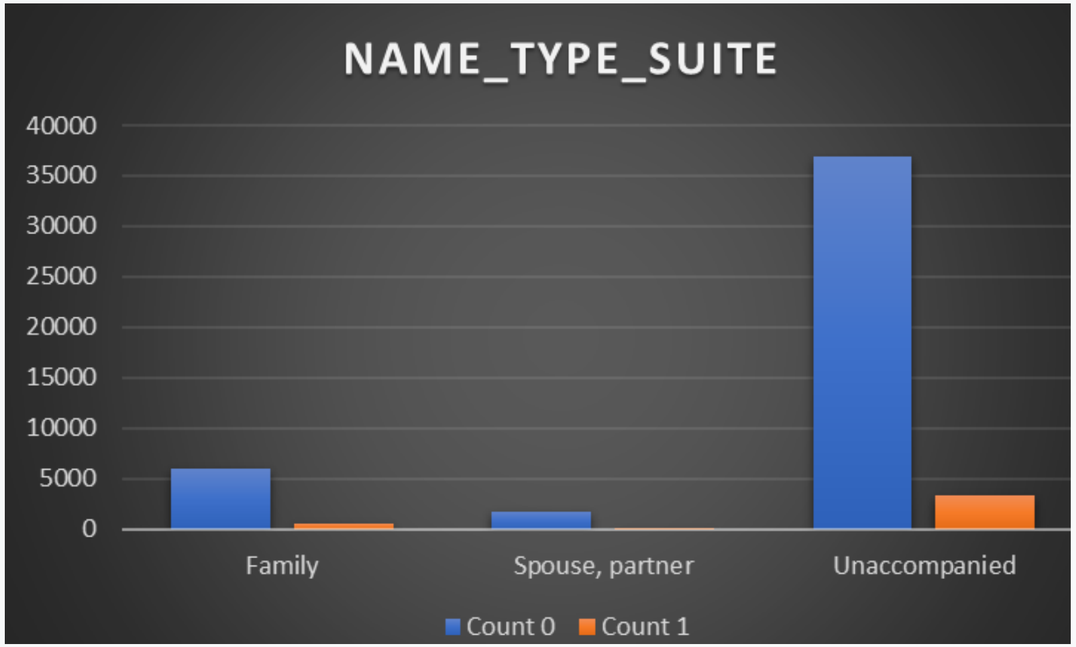


The dataset exhibits significant imbalance in both the TARGET variable, favoring applicants likely to repay loans, and the NAME_CONTRACT_TYPE variable, with more Cash loans than Revolving loans. Similarly, there's an imbalance in the NAME_FAMILY_STATUS variable, primarily representing married individuals, and in NAME_HOUSING_TYPE, where the majority reside in House/Apartments.

INSIGHTS

UNIVARIATE/SEGMENTED UNIVARIATE ANALYSIS

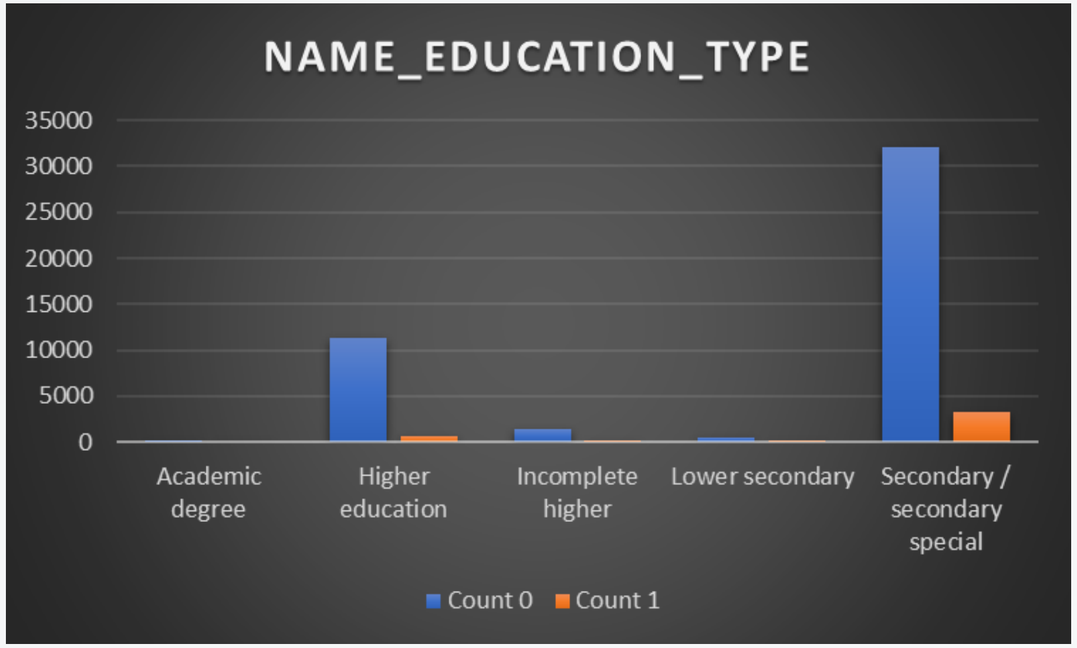




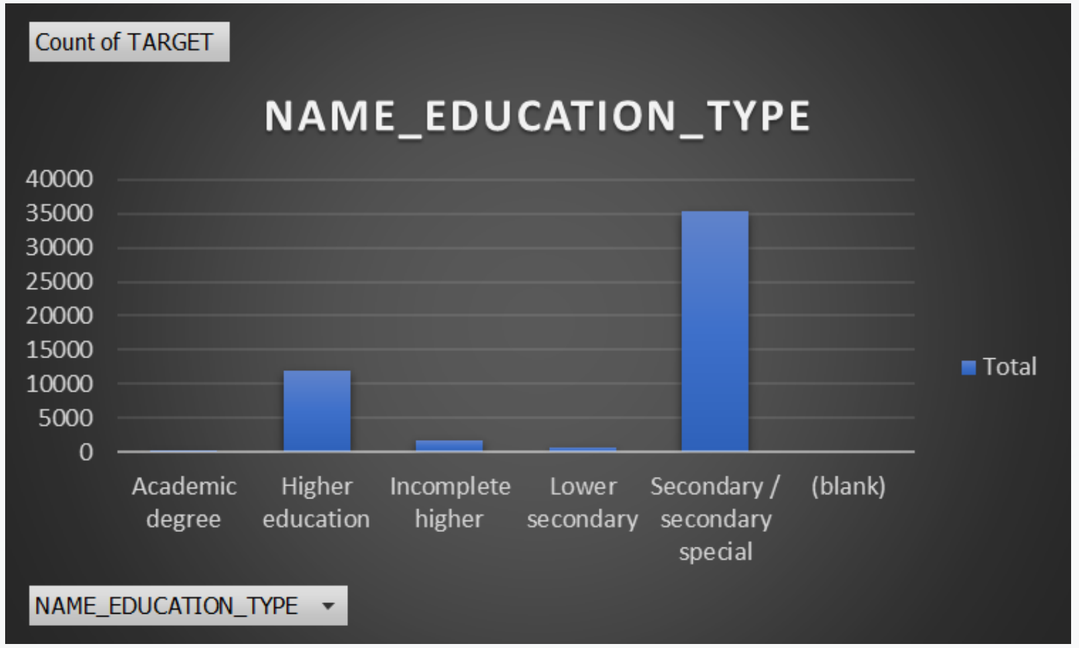
The majority of account holders are individuals applying for loans on an individual basis, followed by families seeking financial assistance.

INSIGHTS

UNIVARIATE/SEGMENTED UNIVARIATE ANALYSIS



Clients with a secondary education have taken the highest number of loans, while those with an academic degree have taken the lowest.

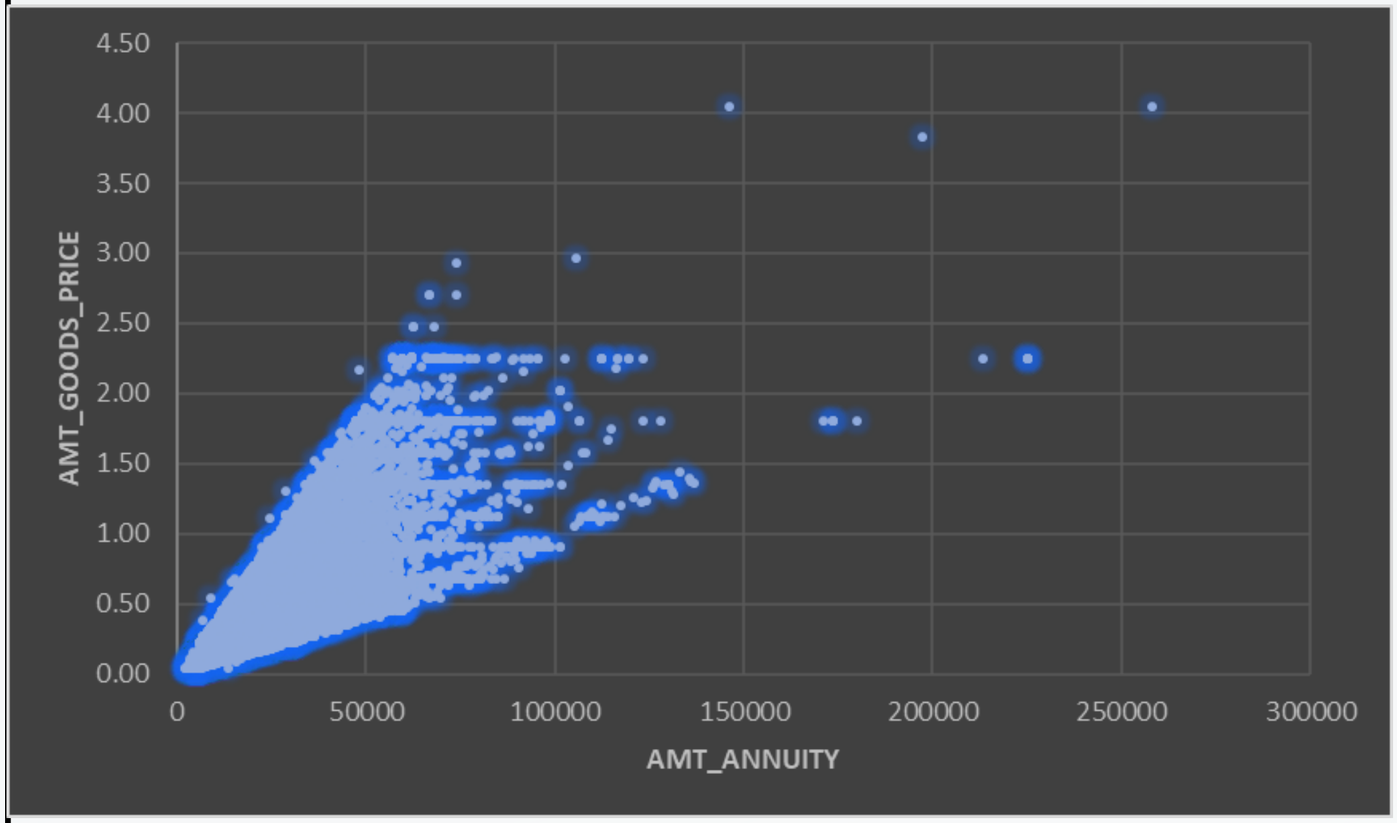


The lowest and highest percentages of defaults were recorded as follows:

- Academic degree: 1.87%
- Lower secondary: 10.72%

INSIGHTS

BIVARIATE ANALYSIS



when the cost of the goods rises, there is a corresponding increase in the loan amount, leading to a higher annuity payment. The correlation between AMT_GOODS_PRICE and AMT_ANNUITY is also evident from the plotted data.

INSIGHTS

CORRELATION MATRIX FOR APPLICANTS WHO CONSISTENTLY MEET PAYMENT DEADLINES

	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	REGION_POPULATION_RELATIVE	DAYS_BIRTH	DAYS_EMPLOYED	DAYS_REGISTRATION	DAYS_ID_PUBLISH	REGION_RATING_CLIENT
CNT_CHILDREN	1	0.036915228	0.005643033	0.026447123	-0.024704049	-0.336852199	-0.245480583	-0.183465558	0.032879976	0.021670909
AMT_INCOME_TOTAL	0.036915228	1	0.376945591	0.45010547	0.180686936	-0.07380461	-0.162817482	-0.068586685	-0.032684165	-0.203744585
AMT_CREDIT	0.005643033	0.376945591	1	0.769984817	0.096402036	0.052183834	-0.074455576	-0.007143557	0.007620887	-0.103131926
AMT_ANNUITY	0.026447123	0.45010547	0.769984817	1	0.117923175	-0.010100274	-0.111608403	-0.034206668	-0.010360142	-0.130520387
REGION_POPULATION_RELATIVE	-0.024704049	0.180686936	0.096402036	0.117923175	1	0.030949039	-0.00654157	0.059648498	0.002879137	-0.538280443
DAYS_BIRTH	-0.336852199	-0.07380461	0.052183834	-0.010100274	0.030949039	1	0.623363973	0.335907632	0.269292385	-0.00960868
DAYS_EMPLOYED	-0.245480583	-0.162817482	-0.074455576	-0.111608403	-0.00654157	0.623363973	1	0.209748333	0.273956183	0.04055974
DAYS_REGISTRATION	-0.183465558	-0.068586685	-0.007143557	-0.034206668	0.059648498	0.335907632	0.209748333	1	0.103307692	-0.083227345
DAYS_ID_PUBLISH	0.032879976	-0.032684165	0.007620887	-0.010360142	0.002879137	0.269292385	0.273956183	0.103307692	1	0.007057714
REGION_RATING_CLIENT	0.021670909	-0.203744585	-0.103131926	-0.130520387	-0.538280443	-0.00960868	0.04055974	-0.083227345	0.007057714	1

INSIGHTS

The heat map presented earlier displays the correlation between various variables for the target (0), representing applicants who consistently meet payment deadlines.

The color scheme used in the heat map ranges from blue to white, where blue represents the strongest correlations, and white indicates the weakest correlations.

Noteworthy correlations observed between the variables are:

1. AMT_ANNUITY TO AMT_CREDIT
2. DAYS_EMPLOYED TO DAYS_BIRTH.
3. AMT_TOTAL_INCOME TO AMT_CREDIT

INSIGHTS

CORRELATION MATRIX FOR APPLICANTS EXPERIENCING PAYMENT DIFFICULTIES

	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	REGION_POPULATION_RELATIVE	DAYS_BIRTH	DAYS_EMPLOYED	DAYS_REGISTRATION	DAYS_ID_PUBLISH	REGION_RATING_CLIENT
CNT_CHILDREN	1	0.010114731	0.007426825	0.029282836	-0.020674428	-0.250131116	-0.190479825	-0.152465328	0.04463221	0.056070474
AMT_INCOME_TOTAL	0.010114731	1	0.015151694	0.017905212	-0.006231907	-0.008985259	-0.01172534	0.009624679	0.00917932	-0.01285733
AMT_CREDIT	0.007426825	0.015151694	1	0.749265697	0.065625879	0.145232913	0.019919471	0.043819007	0.044126096	-0.043179483
AMT_ANNUITY	0.029282836	0.017905212	0.749265697	1	0.070104804	0.010483725	-0.077415246	-0.020671313	0.021483667	-0.060120518
REGION_POPULATION_RELATIVE	-0.020674428	-0.006231907	0.065625879	0.070104804	1	0.018199573	0.009291934	0.045901786	0.005950488	-0.42864488
DAYS_BIRTH	-0.250131116	-0.008985259	0.145232913	0.010483725	0.018199573	1	0.588635224	0.288197045	0.24764622	-0.047128367
DAYS_EMPLOYED	-0.190479825	-0.01172534	0.019919471	-0.077415246	0.009291934	0.588635224	1	0.1940134	0.233683323	-0.010672728
DAYS_REGISTRATION	-0.152465328	0.009624679	0.043819007	-0.020671313	0.045901786	0.288197045	0.1940134	1	0.090250818	-0.114877337
DAYS_ID_PUBLISH	0.04463221	0.00917932	0.044126096	0.021483667	0.005950488	0.24764622	0.233683323	0.090250818	1	-0.026470214
REGION_RATING_CLIENT	0.056070474	-0.01285733	-0.043179483	-0.060120518	-0.42864488	-0.047128367	-0.010672728	-0.114877337	-0.026470214	1

INSIGHTS

The heat map presented earlier displays the correlation between various variables for the target (1), representing applicants who consistently meet payment deadlines.

The color scheme used in the heat map ranges from blue to white, where blue represents the strongest correlations, and white indicates the weakest correlations.

Noteworthy correlations observed between the variables are:

1. AMT_ANNUITY TO AMT_CREDIT
2. DAYS_EMPLOYED TO DAYS_BIRTH.
3. AMT_TOTAL_INCOME TO AMT_CREDIT

RESULT

During the course of working on the "Bank Loan Case Study" project, I achieved significant milestones and gained valuable expertise. Firstly, I conducted efficient data analysis using Excel, making the most of its functions and features to derive meaningful insights. Throughout this project, I honed my data analytics skills, especially in dealing with loan-related variables such as credit scores, loan amounts, and interest rates. This analysis deepened my understanding of credit risk assessment, loan approval processes, and customer behavior within the banking sector. Additionally, I strengthened my analytical thinking, problem-solving, and data visualization capabilities, equipping me with practical knowledge and skills applicable to future projects in the financial domain.

RESULT

In this project, I've included the Excel file containing both the dataset and data analysis, providing access for transparent and interactive exploration of movie data. Please download this file and use Microsoft Excel for the best experience.

https://docs.google.com/spreadsheets/d/1J_zbDUfmFeHSRF_Byepw9AdQ2ioOYuLH/edit?usp=sharing&ouid=118332653323880560357&rtpof=true&sd=true

THANK YOU!

PRESENTED BY: MIHIR PATEL