BANK LOAN CASE STUDY

PROJECT DESCRIPTION:

Overview:

The project aims to analyze various factors influencing bank loan approvals. The study's primary focus is to identify key variables that significantly impact loan decisions, understand the relationships between these variables, and provide actionable insights to optimize the loan approval process.

Objectives:

1. Identify Key Factors: Determine the most influential variables that affect loan approval decisions.
2. Analyze Relationships: Explore the correlations between different variables to understand their interplay.
3. Improve Decision-Making: Provide insights and recommendations to improve the efficiency and fairness of the loan approval process.

Approach:

1. Data Collection:
   - Gathered a comprehensive dataset of loan applicants, including demographic details, financial information, and loan specifics.
   - Ensured data quality by cleaning and preprocessing the dataset, handling missing values, and addressing any inconsistencies.

2. Exploratory Data Analysis (EDA):
   - Conducted EDA to gain initial insights into the dataset.
   - Visualized the distribution of key variables and identified patterns and trends.
   - Utilized heatmaps to visualize correlation coefficients between variables, helping to pinpoint significant relationships.

3. Insights and Recommendations:
   - Interpreted the results to draw actionable insights.
   - Provided recommendations to the bank on how to streamline the loan approval process, reduce default risks, and improve customer satisfaction.

By following this systematic approach, the project aims to enhance the bank's ability to make informed loan approval decisions, ultimately leading to better financial outcomes for both the bank and its customers.

APPROACH :

Data Collection and Preparation:
- Data Sources: Collected historical loan data, including applicant demographics, financial information, and loan details.
- Data Cleaning: Used Microsoft Excel 365 to address missing values, remove duplicates, and handle outliers, ensuring a high-quality dataset.

Exploratory Data Analysis (EDA):

Objective: To gain initial insights, understand data distribution, and identify patterns.

Techniques:
- Descriptive Statistics: Summarized key metrics using Excel functions like `AVERAGE`, `MEDIAN`, `SUM`, and `COUNT`.
- Visualizations: Created histograms, box plots, and scatter plots using Excel's built-in chart tools to explore relationships and distributions.
- Heatmaps: Utilized Excel's conditional formatting to create heatmaps, visualizing correlation coefficients between variables.

Feature Engineering:
Techniques:
- Creating New Variables: Based on domain knowledge, such as debt-to-income ratio.
- Binning: Grouped continuous variables into categories.
- Encoding Categorical Variables: Used Excel functions to convert categorical data into numerical format.
-Normalization and Scaling: Applied Excel functions to standardize and normalize numerical features for consistency.

Insights and Recommendations:
Objective: To draw actionable insights and provide recommendations.
Techniques:
- Feature Importance: Analyzed feature importance to understand key factors influencing loan approvals.
- Visualization: Created charts and graphs in Excel to effectively communicate findings and insights.
- Recommendations: Provided suggestions to optimize the loan approval process, reduce default risks, and improve customer satisfaction.

Tools Summary:
- Microsoft Excel 365: Primary tool for data cleaning, EDA, feature engineering, model development, evaluation, and visualization.

By leveraging the capabilities of Microsoft Excel 365, this approach allowed for a comprehensive analysis of the factors influencing bank loan approvals, ultimately leading to data-driven insights and recommendations.

TECH-STACK USED:

1. Microsoft Excel 365:

Version: Microsoft Excel 365

Purpose:
- Data Collection and Cleaning: Used for importing, cleaning, and organizing the dataset, handling missing values, and preparing the data for analysis.
- Exploratory Data Analysis (EDA): Performed initial data exploration, summary statistics, and visualizations like histograms, box plots, scatter plots, and heatmaps using Excel's built-in chart tools and conditional formatting.
- Feature Engineering: Created new variables, binned continuous variables, encoded categorical variables
- Visualization and Presentation: Created charts and graphs to effectively communicate findings and insights, and prepared final presentations.

By utilizing Microsoft Excel 365, we were able to efficiently manage, analyze, and visualize the data, leading to actionable insights and recommendations for the bank loan case study.

INSIGHTS:

Summary of Insights and Knowledge Gained:
During the Bank Loan Case Study project, several important insights and key findings were discovered. These insights provided a deeper understanding of the factors influencing loan approval decisions and highlighted meaningful trends and patterns within the data.

Key Findings:

1. Income and Loan Amount Correlation:
Higher-income applicants tend to receive larger loan amounts, suggesting that income is a crucial factor in determining loan eligibility.

2. Age and Employment Duration:

Older applicants tend to have shorter employment durations. This trend may impact loan approval decisions, as employment stability is often a key consideration.

3. Debt-to-Income Ratio:

Applicants with higher levels of debt relative to their income are less likely to be approved for loans, highlighting the importance of financial stability in the approval process.

4. Credit History and Loan Approval:

Applicants with a history of regular credit inquiries and responsible borrowing behavior are more likely to be approved for loans.

5. Region Population and Loan Approval:

This trend may be due to better economic opportunities and higher financial activity in densely populated areas.

6. Dependents and Loan Approval:

Applicants with more dependents are slightly less likely to be approved for loans, possibly due to increased financial responsibilities.

TECHNIQUES AND TOOLS USED:

Microsoft Excel 365:

- Data Cleaning: Utilized Excel functions to handle missing values and ensure data quality.
- Exploratory Data Analysis (EDA): Conducted EDA using descriptive statistics, histograms, scatter plots, and heatmaps to identify key patterns and relationships.
- Correlation Analysis: Applied Excel's `CORREL` function and conditional formatting to create a correlation matrix and visualize relationships between variables.
- Visualization: Created charts and graphs to effectively communicate findings and insights.

These insights provide valuable information to optimize the bank's loan approval process, reduce default risks, and enhance customer satisfaction. By leveraging Microsoft Excel 365, the project successfully identified critical factors influencing loan approvals and provided actionable recommendations.

RESULT:

Achievements:

1. Identification of Key Variables:
   - Successfully identified crucial factors influencing loan approval decisions, such as applicant income, credit history, debt-to-income ratio, and number of dependents.
   - Pinpointed correlations between these variables and their impact on loan outcomes.

2. Comprehensive Data Analysis:
   - Conducted thorough exploratory data analysis (EDA) to uncover patterns and trends within the data.
   - Created visualizations such as histograms, scatter plots, and heatmaps to present insights effectively.

3. Actionable Insights and Recommendations:
   - Derived actionable insights from the analysis, leading to practical recommendations for the bank.
   - Suggested strategies to improve the loan approval process, reduce default risks, and enhance customer satisfaction.

4. Enhanced Decision-Making:
   - Provided the bank with data-driven insights to make informed decisions on loan approvals.
   - Enabled the bank to better understand the characteristics of applicants who are likely to be approved or denied loans.

Contribution to Understanding:

- Deeper Insight into Loan Approval Factors:
  - Gained a comprehensive understanding of the factors that significantly impact loan approval decisions.
  - Recognized the importance of financial stability, credit history, and demographic factors in the loan approval process.

- Improved Analytical Skills:
  - Enhanced proficiency in using Microsoft Excel 365 for data analysis and visualization.
  - Developed skills in cleaning, organizing, and analyzing large datasets, as well as interpreting complex statistical relationships.

- Data-Driven Decision Making:
  - Emphasized the value of data-driven approaches in financial decision-making.
  - Demonstrated how analytical techniques can optimize business processes and improve outcomes.

- Practical Applications:
  - Provided practical recommendations that can be directly applied by the bank to refine their loan approval criteria and processes.
  - Highlighted the importance of continuous data analysis to adapt to changing economic conditions and applicant profiles.

Through this project, the understanding of the bank loan approval process was significantly enhanced, leading to the identification of key factors .The insights gained from this study are instrumental in guiding the bank towards more informed and effective decision-making.

DRIVE LINK :

BANK LOAN CASE STUDY REPORT PDF LINK :

https://drive.google.com/file/d/12oyPuOVzpnzI4MTTRjDNmiylSpIPSi5B/view?usp=sharing

BANK LOAN CASE STUDY POWER-POINT LINK :

https://docs.google.com/presentation/d/1MP6SEf2__rH2GIKnrDY3N4vuqrOixOVZ/edit?usp=sharing&ouid=101204343036685814262&rtpof=true&sd=true

DEMOCREATOR BANK LOAN CASE STUDY VIDEO LINK :

https://drive.google.com/file/d/1VoFdf4W7lgmrOGaCca5DgUy-m3ry5a8i/view?usp=sharing

The Given Datasets for Analysis link :

File 1:

https://docs.google.com/spreadsheets/d/1SqA29nV0MNZK-gNbi0JCR7zNlTNOlg_w/edit?usp=sharing&ouid=101204343036685814262&rtpof=true&sd=true

File 2 :
https://docs.google.com/spreadsheets/d/18dCjncxQ5Tst27YOO6nKTDiWj6UhmjFb/edit?usp=sharing&ouid=101204343036685814262&rtpof=true&sd=true

File 3 :

https://docs.google.com/spreadsheets/d/1_oEJTXjHwLOAU7iNdTCA4vpYrZ-aUQbV/edit?usp=sharing&ouid=101204343036685814262&rtpof=true&sd=true

The Excel sheet Columns taken for all Questionaire Link :

The columns used to find missing values link :

https://docs.google.com/spreadsheets/d/11NXivWjahu5285-i-83vVTD8g9CE9Oxu/edit?usp=sharing&ouid=101204343036685814262&rtpof=true&sd=true

The columns used to find outliers link :

https://docs.google.com/spreadsheets/d/11NXivWjahu5285-i-83vVTD8g9CE9Oxu/edit?usp=sharing&ouid=101204343036685814262&rtpof=true&sd=true

The columns taken to find data imbalance link :
https://docs.google.com/spreadsheets/d/1oVteOqRRD-49Eq0_ChBve4iZUBNufCLV/edit?usp=sharing&ouid=101204343036685814262&rtpof=true&sd=true

the columns chosen to perform univariate analysis link :

https://docs.google.com/spreadsheets/d/1N8ho6bKiiMOMwOGUt60p_Hj9SGPq88_L/edit?usp=sharing&ouid=101204343036685814262&rtpof=true&sd=true

the columns chosen to perform segmented univariate analysis link :

https://docs.google.com/spreadsheets/d/16sDbOypfawu5Guv0hyRxp8crEOJFisYP/edit?usp=sharing&ouid=101204343036685814262&rtpof=true&sd=true

the columns chosen to perform bivariate analysis link :

https://docs.google.com/spreadsheets/d/1cTCFZvgNtDc8gfoOWMKdS_ehlnsNCWKh/edit?usp=sharing&ouid=101204343036685814262&rtpof=true&sd=true

The columns taken to perform correlation coefficient including target variable (1) are given in the link below

https://docs.google.com/spreadsheets/d/15-QGp86lugTP0eNVSf_9fqLi1r9YWuQC/edit?usp=sharing&ouid=101204343036685814262&rtpof=true&sd=true

The columns taken to perform correlation coefficient including target variable (0) are given in the link below

https://docs.google.com/spreadsheets/d/1pilMPVgwnAurs6V9GU8HMX67tpJCjaQy/edit?usp=sharing&ouid=101204343036685814262&rtpof=true&sd=true

The Output for all Questionaire Link :

Outlier output link :

https://docs.google.com/spreadsheets/d/1b9cbpOIEyhBRMTom_q0YRpNZjvd5Zu1p/edit?usp=sharing&ouid=101204343036685814262&rtpof=true&sd=true

valid/invalid data point output link :

https://docs.google.com/spreadsheets/d/1TrCWbTN1DJhCJflxaRfpLZ9-ypGvzli3/edit?usp=sharing&ouid=101204343036685814262&rtpof=true&sd=true

counts comparing target output link :

https://docs.google.com/spreadsheets/d/10QDDrkD32XXK70ZV8w6Vbq-emor5eCMB/edit?usp=sharing&ouid=101204343036685814262&rtpof=true&sd=true

The descriptive analysis for univariate analysis output link :

https://docs.google.com/spreadsheets/d/1-cJcCWAY_A7eAVPrI7gKQY7S3k-6ho1m/edit?usp=sharing&ouid=101204343036685814262&rtpof=true&sd=true

segmented analysis segments found using pivot table output link :

https://docs.google.com/spreadsheets/d/19_q79GtofixKYSdwJPeJdAQ4VDfVO1jn/edit?usp=sharing&ouid=101204343036685814262&rtpof=true&sd=true

segmented analysis segment's descriptive analysis output link :

https://docs.google.com/spreadsheets/d/1ERFXED5g68gr3Vd_fwofd2Rog3rlfiHE/edit?usp=sharing&ouid=101204343036685814262&rtpof=true&sd=true

bivariate analysis segment's done using pivot tables output link :

https://docs.google.com/spreadsheets/d/1F0Fem3IY7msm2a3nPVK7IsPB7VBMjRG4/edit?usp=sharing&ouid=101204343036685814262&rtpof=true&sd=true

the descriptive analysis for bivariate analysis output link :

https://docs.google.com/spreadsheets/d/1QXr5Pp7rNYbdqPP3zGXIkr_lf9_Ddb0P/edit?usp=sharing&ouid=101204343036685814262&rtpof=true&sd=true

The coefficients found for target variable 1 output link :
https://docs.google.com/spreadsheets/d/1mkxx0A5Vz-E0xnF9Yu6W-utq0qtGXUGj/edit?usp=sharing&ouid=101204343036685814262&rtpof=true&sd=true

The Correlation Matrices for Target variable (1) output link :
https://docs.google.com/spreadsheets/d/1vayEb52rByP9-xnlNrS1Qzc8fsZ-YD6C/edit?usp=sharing&ouid=101204343036685814262&rtpof=true&sd=true

The coefficients found for target variable 0 output link :

https://docs.google.com/spreadsheets/d/1YmOyXEuEbEVg40XHfrVaRl2wqXoAc6V9/edit?usp=sharing&ouid=101204343036685814262&rtpof=true&sd=true

The Correlation Matrices for Target variable 0 output link :

https://docs.google.com/spreadsheets/d/1fT7Ao_s8Oj6tvAdj3JQhGi4pxx_Ebz7q/edit?usp=sharing&ouid=101204343036685814262&rtpof=true&sd=true

Data Analytics Tasks:

A.Identify Missing Data and Deal with it Appropriately:

1.Identify the missing data in the dataset and decide on an appropriate method to deal with it using Excel built-in functions and features. Create a bar chart or column chart to visualize the proportion of missing values for each variable.

OUTPUT :

	A	B	C	D	E	F
1	SK_ID_CURR	TARGET	NAME_CONTRACT_TYPE	CODE_GENDER	FLAG_OWN_CAR	FLAG_OWN_REALTY
2	100002	1	Cash loans	M	N	Y
3	100003	0	Cash loans	F	N	N
4	100004	0	Revolving loans	M	Y	Y
5	100006	0	Cash loans	F	N	Y
6	100007	0	Cash loans	M	N	Y
7	100008	0	Cash loans	M	N	Y
8	100009	0	Cash loans	F	Y	Y
9	100010	0	Cash loans	M	Y	Y
10	100011	0	Cash loans	F	N	Y
11	100012	0	Revolving loans	M	N	Y
12	100014	0	Cash loans	F	N	Y
13	100015	0	Cash loans	F	N	Y
14	100016	0	Cash loans	F	N	Y
15	100017	0	Cash loans	M	Y	N
16	100018	0	Cash loans	F	N	Y
17	100019	0	Cash loans	M	Y	Y

The Rest of the columns which I have used to find missing values is given in the below link :

https://docs.google.com/spreadsheets/d/11NXivWjahu5285-i-83vVTD8g9CE9Oxu/edit?usp=sharing&ouid=101204343036685814262&rtpof=true&sd=true

TO FIND BLANKS :

Used conditional formatting – new Rule – Format only cells that contain – format only that contain section –blanks –format -ok

To Find empty missing cells :

COUNT BLANK used to count the empty missing cells

COUNT BLANK FUNCTION :

=COUNTBLANK(A:A)

To find the total no of missing cells
COUNT A is used to count the total no of missing cells
=COUNTA (A:A)

So I have found the total no of missing values and the count of missing cells

As the the percentages of missing values is found to quantify the extent of missing data relative to the total no of entries for each variable

I have found percentages of all values

MISSING VALUES & PERCENTAGES

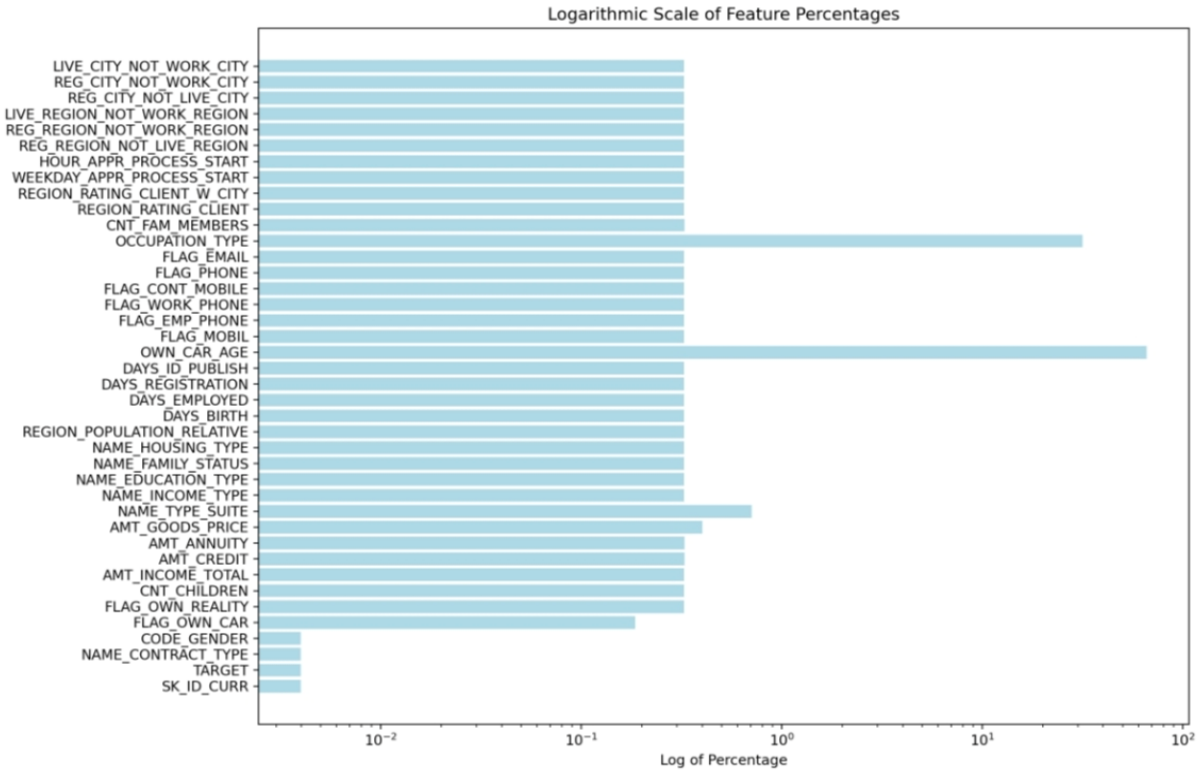
	A	B	C
1		MISSING VALUES	PERCENTAGE
2	SK_ID_CURR	2	0.003987082
3	TARGET	2	0.003987082
4	NAME_CONTRACT_TYPE	2	0.003987082
5	CODE_GENDER	2	0.003987082
6	FLAG_OWN_CAR	93	0.185399306
7	FLAG_OWN_REALITY	163	0.324947171
8	CNT_CHILDREN	163	0.324947171
9	AMT_INCOME_TOTAL	163	0.324947171
10	AMT_CREDIT	163	0.326940717
11	AMT_ANNUITY	164	0.326940712
12	AMT_GOODS_PRICE	201	0.400701726
13	NAME_TYPE_SUITE	355	0.707707029
14	NAME_INCOME_TYPE	163	0.324947171
15	NAME_EDUCATION_TYPE	163	0.324947171
16	NAME_FAMILY_STATUS	163	0.324947171
17	NAME_HOUSING_TYPE	163	0.324947171
18	REGION_POPULATION_RELATIVE	163	0.324947171
19	DAYS_BIRTH	163	0.324947171
20	DAYS_EMPLOYED	163	0.324947171
21	DAYS_REGISTRATION	163	0.324947171
22	DAYS_ID_PUBLISH	163	0.324947171
23	OWN_CAR_AGE	33113	66.01212073

24	FLAG_MOBIL	163	0.324947171
25	FLAG_EMP_PHONE	163	0.324947171
26	FLAG_WORK_PHONE	163	0.324947171
27	FLAG_CONT_MOBILE	163	0.324947171
28	FLAG_PHONE	163	0.324947171
29	FLAG_EMAIL	163	0.324947171
30	OCCUPATION_TYPE	15817	31.53183685
31	CNT_FAM_MEMBERS	164	0.326940712
32	REGION_RATING_CLIENT	163	0.324947171
33	REGION_RATING_CLIENT_W_CITY	163	0.324947171
34	WEEKDAY_APPR_PROCESS_START	163	0.324947171
35	HOUR_APPR_PROCESS_START	163	0.324947171
36	REG_REGION_NOT_LIVE_REGION	163	0.324947171
37	REG_REGION_NOT_WORK_REGIO	163	0.324947171
38	LIVE_REGION_NOT_WORK_REGIO	163	0.324947171
39	REG_CITY_NOT_LIVE_CITY	163	0.324947171
40	REG_CITY_NOT_WORK_CITY	163	0.324947171
41	LIVE_CITY_NOT_WORK_CITY	163	0.324947171
42	ORGANIZATION_TYPE	163	0.324947171
43	EXT_SOURCE_1	28335	56.48698218
44	EXT_SOURCE_2	289	0.576133328

45	EXT_SORCE_3	10107	20.14871815
46	APARTMENTS_AVG	25548	50.93098361
47	BASEMENTAREA_AVG	29362	58.53434871
48	YEARS_BEGINEXPLUATATION_AVG	24557	48.95538455
49	YEARS_BUILD_AVG	33402	66.58825406
50	COMMONAREA_AVG	35123	70.01913799
51	ELEVATORS_AVG	26814	53.45480643

So As by finding percentage of missing values I have found the proportion of the missing values

And chart used to visualize the proportion of missing values for each variable is given below



Logarithmic scale is used for the percentage differences in the chart to be seen more clearly

B. Identify Outliers in the Dataset:

2. Detect and identify outliers in the dataset using Excel statistical functions and features, focusing on numerical variables. Create box plots or scatter plots to visualize the distribution of numerical variables and highlight the outliers

OUTPUT :

	A	B	C	D	E	F
1	SK_ID_CURR	TARGET	NAME_CONTRACT_TYPE	CODE_GENDER	FLAG_OWN_CAR	FLAG_OWN_REALTY
2	100002	1	Cash loans	M	N	Y
3	100003	0	Cash loans	F	N	N
4	100004	0	Revolving loans	M	Y	Y
5	100006	0	Cash loans	F	N	Y
6	100007	0	Cash loans	M	N	Y
7	100008	0	Cash loans	M	N	Y
8	100009	0	Cash loans	F	Y	Y
9	100010	0	Cash loans	M	Y	Y
10	100011	0	Cash loans	F	N	Y
11	100012	0	Revolving loans	M	N	Y
12	100014	0	Cash loans	F	N	Y
13	100015	0	Cash loans	F	N	Y
14	100016	0	Cash loans	F	N	Y
15	100017	0	Cash loans	M	Y	N
16	100018	0	Cash loans	F	N	Y

The Rest of the columns which I have used to find outliers is given in the below link :

https://docs.google.com/spreadsheets/d/11NXivWjahu5285-i-83vVTD8g9CE9Oxu/edit?usp=sharing&ouid=101204343036685814262&rtpof=true&sd=true

I have found outliers using Z-score method & to calculate z score I have found the mean and standard deviation

To calculate Z-score I have subtracted mean from value and divide by standard deviation

To calculate Mean :

=AVERAGE(A2: A100)

To calculate standard deviation :

=STDEV.P (A2:A100)

z- score formula :

=(A2 - B1) / B2

I have found outliers using Z-score where values less than -3 and values greater than -3 are considered outliers

TheAMT_INCOME_TOTAL,AMT_CREDIT,AMT_ANNUITY,EXT_SOURCE_1,CNT_CHILDREN,REGION_RATING_CLIENT,EXT_SOURCE_3 are the columns from the dataset where outliers were found.

	A	B	C	D	E	F	G
1	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	EXT_SOURCE_1	CNT_CHILDREN	REGION_RATING_CLIENT	EXT_SOURCE_3
2	360000	1663987.5	86989.5	0.083036967	3	1	0.062103038
3	450000	1755000	64107	0.311267311	3	3	0.070108844
4	360000	2250000	73611	0.774761413	3	3	0.077498546
5	360000	1710000	83515.5	0.587334047	3	1	0.077498546
6	540000	1800000	77494.5	0.319760172	3	3	0.071055125
7	360000	1971072	66262.5	0.72204445	3	3	0.077498546
8	540000	2250000	65956.5	0.464831117	3	3	0.051681767
9	360000	1724220	62698.5	0.721939769	3	1	0.051681767
10	360000	1971072	67500	0.115634337	3	3	0.000527265
11	450000	2286211.5	67500	0.565654882	3	3	0.065109614
12	450000	1800000	67500	0.43770902	3	1	0.079059696
13	382500	1971072	67500	0.561948409	3	1	0.053478087
14	360000	1971072	62019	0.600395905	3	3	0.072496677
15	765000	1885500	116266.5	0.297913509	3	3	0.0588259
16	450000	1649646	72778.5	0.274422372	3	1	0.052036407
17	360000	2085120	72778.5	0.842763466	4	3	0.000527265
18	360000	2125953	79065	0.804586121	3	3	0.076473519
19	765000	1661418	62019	0.468208057	3	1	0.000527265

The rest of the outliers output is in the link below :

https://docs.google.com/spreadsheets/d/1b9cbpOIEyhBRMTom_q0YRpNZjvd5Zu1p/edit?usp=sharing&ouid=101204343036685814262&rtpof=true&sd=true

To find if all the datapoints are valid or not I have used these below functions

FOR AMT_INCOME_TOTAL

=IF(OR(A208<10000,A208>1000000),"INVALID","VALID")

FOR AMT_CREDIT

=IF(OR(B2<5000,B2>500000),"INVALID","VALID")

IN AMT_CREDIT column all the values are invalid

FOR AMT_ANNUITY

=IF(OR(C2<1000,C2>50000),"INVALID","VALID")

In AMT_ANNUITY column all the values are invalid

FOR EXT_SORCE_1

=IF(OR(D2<0,D2>1),"INVALID","VALID")

FOR CNT_CHILDREN

=IF(OR(E2<0,E2>10),"INVALID","VALID")

FOR REGION_RATING_CLIENT

=IF(OR(F2<1,F2>3),"INVALID","VALID")

FOR EXT_SOURCE_3

=IF(OR(G2<0,G2>1),"INVALID","VALID")

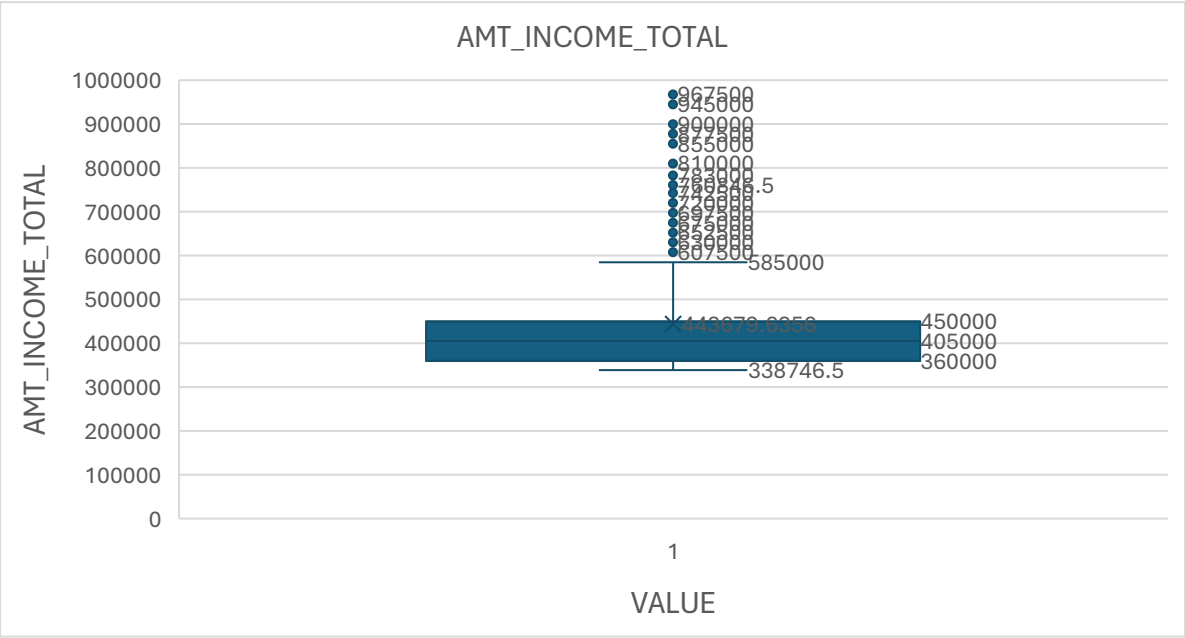
K	L	M	N	O	P	Q
AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	EXT_SOURCE_1	CNT_CHIDREN	REGION_RATING_CL	EXT_SOURCE_3
VALID	INVALID	INVALID	VALID	VALID	VALID	VALID
VALID	INVALID	INVALID	VALID	VALID	VALID	VALID
VALID	INVALID	INVALID	VALID	VALID	VALID	VALID
VALID	INVALID	INVALID	VALID	VALID	VALID	VALID
VALID	INVALID	INVALID	VALID	VALID	VALID	VALID
VALID	INVALID	INVALID	VALID	VALID	VALID	VALID
VALID	INVALID	INVALID	VALID	VALID	VALID	VALID
VALID	INVALID	INVALID	VALID	VALID	VALID	VALID
VALID	INVALID	INVALID	VALID	VALID	VALID	VALID
VALID	INVALID	INVALID	VALID	VALID	VALID	VALID
VALID	INVALID	INVALID	VALID	VALID	VALID	VALID
VALID	INVALID	INVALID	VALID	VALID	VALID	VALID
VALID	INVALID	INVALID	VALID	VALID	VALID	VALID
VALID	INVALID	INVALID	VALID	VALID	VALID	VALID
VALID	INVALID	INVALID	VALID	VALID	VALID	VALID
VALID	INVALID	INVALID	VALID	VALID	VALID	VALID
VALID	INVALID	INVALID	VALID	VALID	VALID	VALID
VALID	INVALID	INVALID	VALID	VALID	VALID	VALID
VALID	INVALID	INVALID	VALID	VALID	VALID	VALID
VALID	INVALID	INVALID	VALID	VALID	VALID	VALID
VALID	INVALID	INVALID	VALID	VALID	VALID	VALID
VALID	INVALID	INVALID	VALID	VALID	VALID	VALID
VALID	INVALID	INVALID	VALID	VALID	VALID	VALID
VALID	INVALID	INVALID	VALID	VALID	VALID	VALID

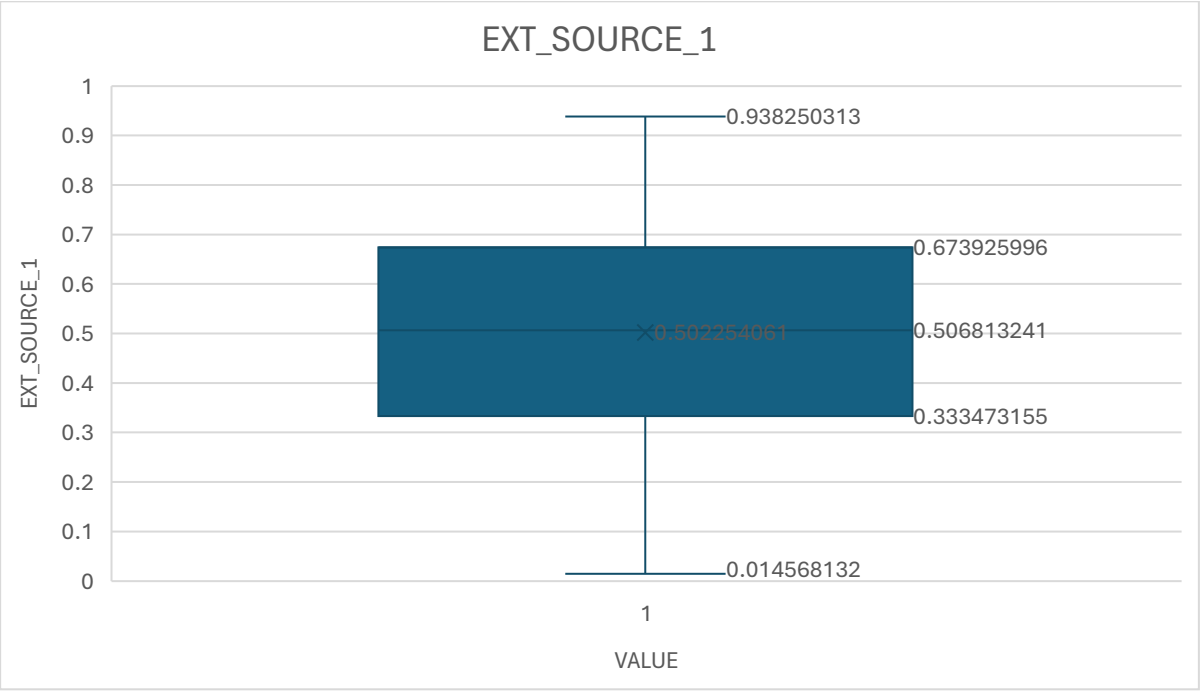
The valid/invalid data points of the data are there in file link below

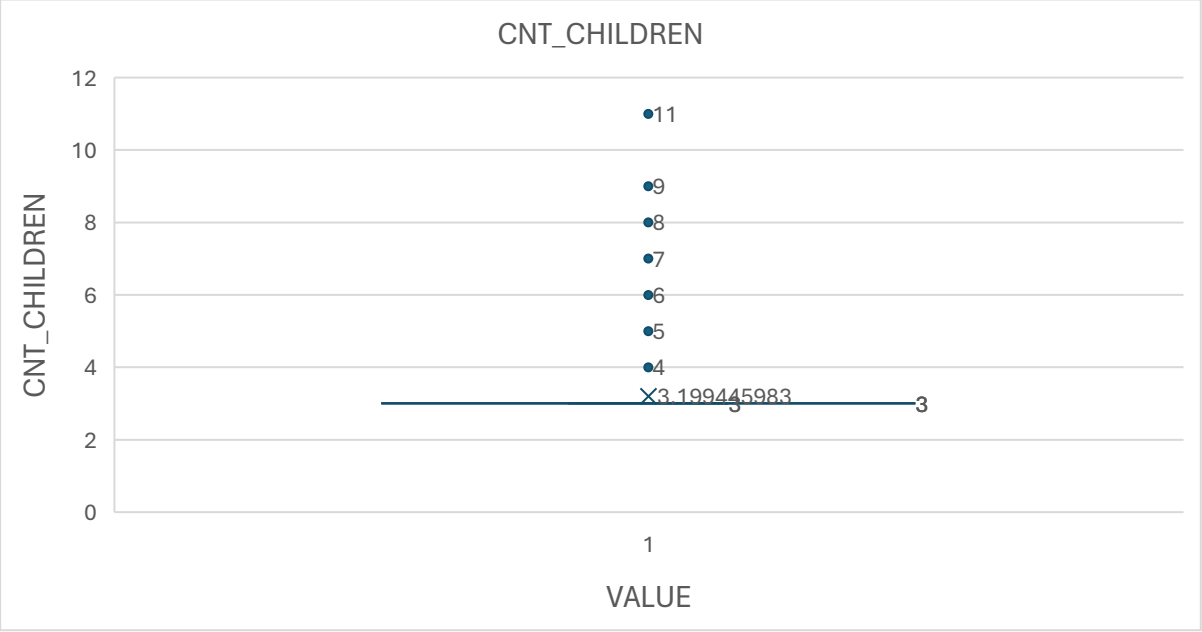
https://docs.google.com/spreadsheets/d/1TrCWbTN1DJhCJflxaRfpLZ9-ypGvzli3/edit?usp=sharing&ouid=101204343036685814262&rtpof=true&sd=true

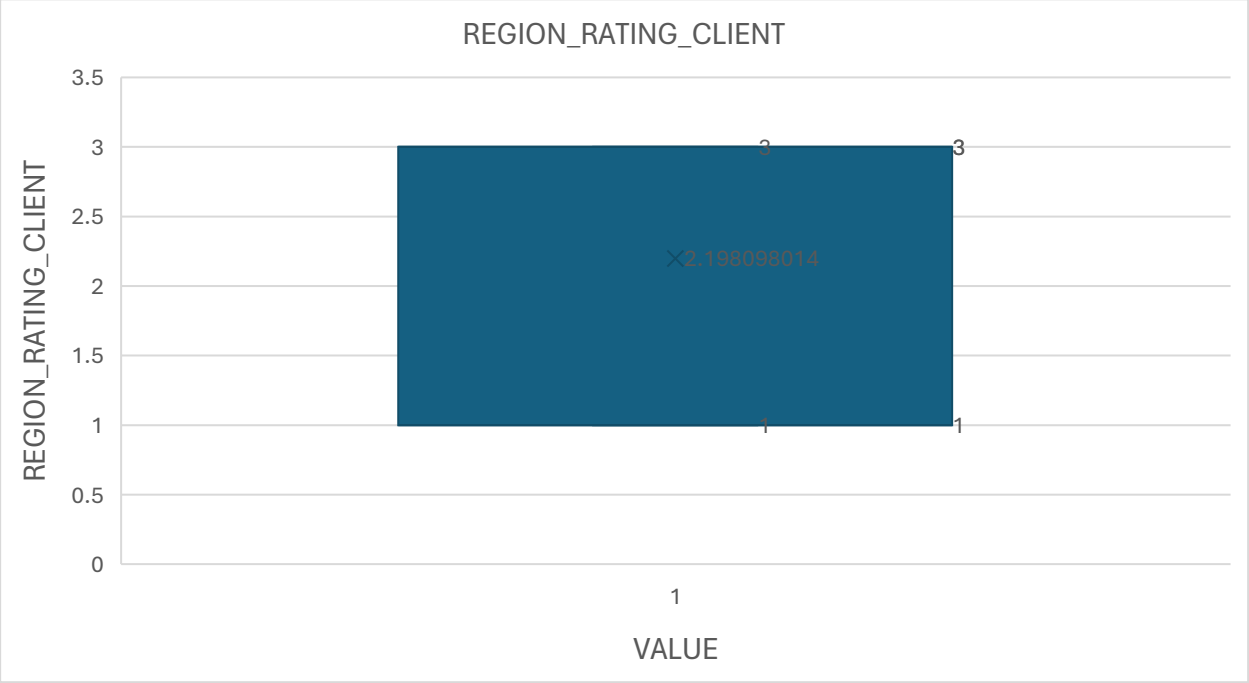
The Data points are used to plot these below box plots

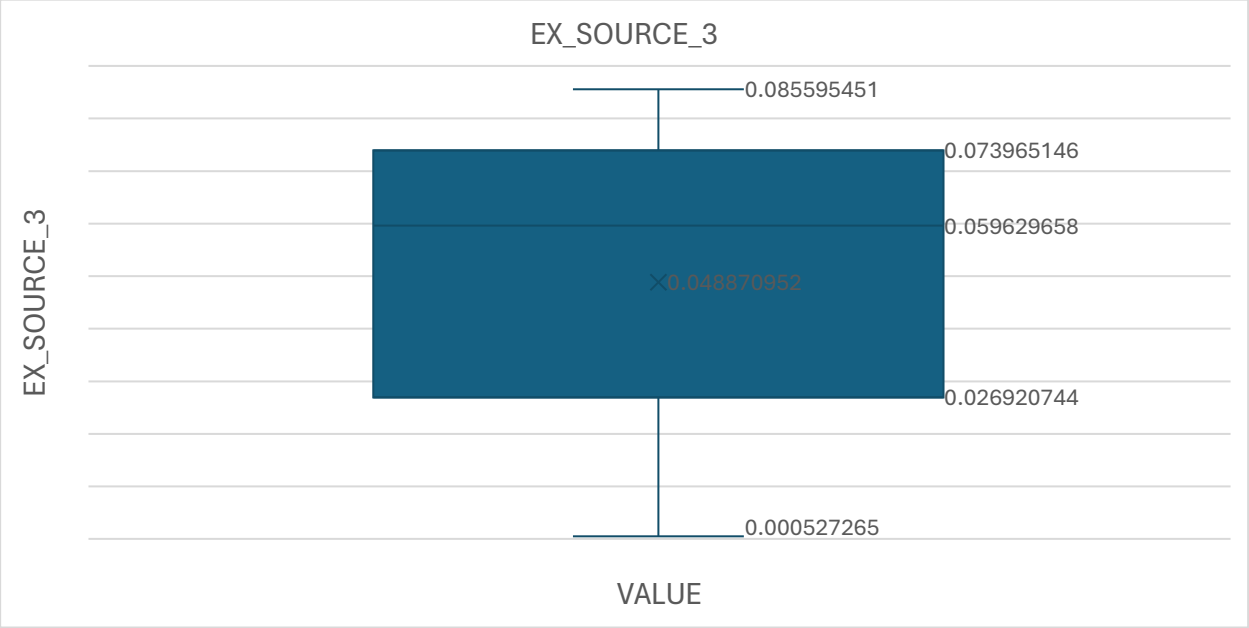
BOX PLOTS :




EXT_SOURCE_1
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
EXT_SOURCE_1
0.938250313
0.673925996
0.502254061
0.506813241
0.333473155
0.014568132
1
VALUE


CNT_CHILDREN
12
10
8
6
4
2
0
CNT_CHILDREN
11
9
8
7
6
5
4
3.199445983
3
3
1
VALUE


REGION_RATING_CLIENT
REGION_RATING_CLIENT
3.5
3
2.5
2
1.5
1
0.5
0
3
3
2.198098014
1
1
1
VALUE


EX_SOURCE_3
EX_SOURCE_3
0.085595451
0.073965146
0.059629658
0.048870952
0.026920744
0.000527265
VALUE

C. Analyze Data Imbalance:

3.Determine if there is data imbalance in the loan application dataset and calculate the ratio of data imbalance using Excel functions. Create a pie chart or bar chart to visualize the distribution of the target variable and highlight the class imbalance.

OUTPUT :

The columns taken are :

code_gender,name_type_suite,name_income_type,name_education_type, name_family_status,name_housing_type,occupation_type,organization_type, housetype_mode,wallsmateal_mode,emergency state_mode

	A	B	C	D	E	F	G	H	I	J
1	CODE_GENDER	NAME_TYPE_SUITE	NAME_INCOME_TYPE	NAME_EDUCATION_TYPE	NAME_FAMILY_STATUS	NAME_HOUSING_TYPE	ORGANIZATION_	HOUSETYPE_MODE	WALLSMATERIAL_MODE	EMERGENCYSTATE_MODE
2	M	Unaccompanied	Working	Secondary / secondary special	Single / not married	House / apartment	Business Entity T	block of flats	Stone, brick	No
3	F	Family	State servant	Higher education	Married	House / apartment	School	block of flats	Block	No
4	M	Unaccompanied	Working	Secondary / secondary special	Single / not married	House / apartment	Government	block of flats	Panel	No
5	F	Unaccompanied	Working	Secondary / secondary special	Civil marriage	House / apartment	Business Entity T	block of flats	Panel	No
6	M	Unaccompanied	Working	Secondary / secondary special	Single / not married	House / apartment	Religion	block of flats	Stone, brick	No
7	M	Spouse, partner	State servant	Secondary / secondary special	Married	House / apartment	Other	block of flats	Stone, brick	No
8	F	Unaccompanied	Commercial associate	Higher education	Married	House / apartment	Business Entity T	block of flats	Panel	No
9	M	Unaccompanied	State servant	Higher education	Married	House / apartment	Other	block of flats	Mixed	No
10	F	Children	Pensioner	Secondary / secondary special	Married	House / apartment	XNA	block of flats	Panel	No
11	M	Unaccompanied	Working	Secondary / secondary special	Single / not married	House / apartment	Electricity	block of flats	Stone, brick	No
12	F	Unaccompanied	Working	Higher education	Married	House / apartment	Medicine	block of flats	Wooden	No
13	F	Children	Pensioner	Secondary / secondary special	Married	House / apartment	XNA	block of flats	Panel	Yes
14	F	Unaccompanied	Working	Secondary / secondary special	Married	House / apartment	Business Entity T	block of flats	Others	No
15	M	Unaccompanied	Working	Secondary / secondary special	Married	House / apartment	Self-employed	block of flats	Block	No
16	F	Unaccompanied	Working	Secondary / secondary special	Married	House / apartment	Transport: type 2	block of flats	Panel	No
17	M	Family	Working	Secondary / secondary special	Single / not married	Rented apartment	Business Entity T	block of flats	Stone, brick	No
18	M	Unaccompanied	Working	Secondary / secondary special	Married	House / apartment	Government	block of flats	Stone, brick	No
19	F	Unaccompanied	Working	Secondary / secondary special	Married	House / apartment	Construction	block of flats	Stone, brick	No
20	F	Other_A	Working	Secondary / secondary special	Widow	House / apartment	Housing	block of flats	Stone, brick	No
21	F	Unaccompanied	State servant	Higher education	Single / not married	House / apartment	Kindergarten	block of flats	Stone, brick	No
22	M	Unaccompanied	Working	Secondary / secondary special	Married	House / apartment	Self-employed	block of flats	Panel	No
23	F	Unaccompanied	Commercial associate	Secondary / secondary special	Married	House / apartment	Trade: type 7	block of flats	Panel	No
24	F	Unaccompanied	Working	Secondary / secondary special	Married	Rented apartment	Self-employed	block of flats	Panel	No

The columns taken to find data imbalance are given in the link below :
https://docs.google.com/spreadsheets/d/1oVteOqRRD-49Eq0_ChBve4iZUBNufCLV/edit?usp=sharing&ouid=101204343036685814262&rtpof=true&sd=true

=COUNTIFS(A2:A50000,"1",B2:B50000,"M")

I have used countif function to calculate count of each category comparing to Target column

	A	B	C	D	E	F	G	H	I	J
1		CODE_GENDER			NAME_TYPE_SUITE					
2		F	M	XNA	Children	Family	Group of people	Other_A	Other_B	Spouse, partner
3	TARGET									
4	0	30559	15412	2	502	6037	36	129	246	1704
5	1	2264	1762	0	40	512	0	8	13	145

The rest of the output of counts where I have calculated the count of each category comparing to Target column is in the link below

https://docs.google.com/spreadsheets/d/10QDDrkD32XXK70ZV8w6Vbq-emor5eCMB/edit?usp=sharing&ouid=101204343036685814262&rtpof=true&sd=true

I have calculated the sum of target 0 and target 1 rows using sum function

=SUM(A1:A10)

I have calculated total of (sum of target 0 and target 1 rows) using sum function

=SUM(A1:A10,B1:B10)

I have calculated proportions of target variable using proportion formula

=B6/B8

I have calculated the proportion of target 0 and target 1

6	**SUM(0)**	422931
7	**SUM (1)**	36980
8	**TOTAL SUM**	459911
9	**PROPORTION(0)**	0.919593139
10	**PROPORTION(1)**	0.080406861

I have calculated the ratio of data imbalance through the formula

Ratio of data imbalance = proportion of Majority value / proportion of Minority value

Majority value(target variable 0) = 0.919593139

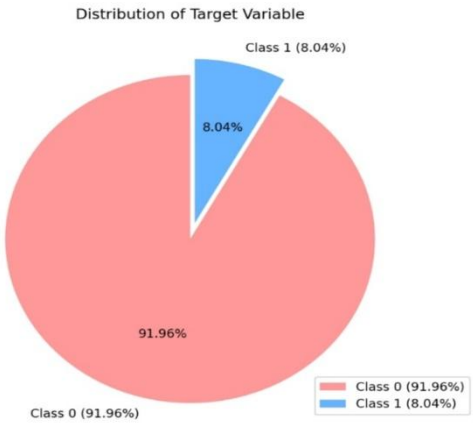
Minority value (target variable 1) = 0.080406861

Ratio of data imbalance = 0.919593139 / 0.080406861 = 11.44

Then I have calculated the percentage of both target 0 & target 1 proportions

12	**TARGET**	
13	**0**	91.96%
14	**1**	8.04%

I have inserted a pie chart to display the distribution of target variable and highlighted the class imbalance



DISTRIBUTION OF TARGET VARIABLE

NOTE: The chart indicates a ratio imbalance of 11.44,highlighting a significant disparity among the target variable

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

4. 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. Create histograms, bar charts, or box plots to visualize the distributions of variables. Create stacked bar charts or grouped bar charts to compare variable distributions across different scenarios. Create scatter plots or heatmaps to visualize the relationships between variables and the target variable.

OUTPUT :

1) UNIVARIATE ANALYSIS :

For univariate analysis the columns AMT_INCOME_TOTAL, AMT_CREDIT, AMT_ANNUITY, CNT_CHILDREN, DAYS_BIRTH, DAYS_EMPLOYED, AMT_ANNUITY, AMT_APPLICATION, AMT_CREDIT, AMT_DOWN_PAYMENT, AMT_GOODS_PRICE, NAME_CONTRACT_TYPE, NAME_CONTRACT_STATUS are taken

	A	B	C	D	E	F	G
1	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	CNT_CHILDREN	DAYS_BIRTH	DAYS_EMPLOYED	AMT_ANNUITY
2	202500	406597.5	24700.5	0	-9461	-637	1730.43
3	270000	1293502.5	35698.5	0	-16765	-1188	25188.615
4	67500	135000	6750	0	-19046	-225	15060.735
5	135000	312682.5	29686.5	0	-19005	-3039	47041.335
6	121500	513000	21865.5	0	-19932	-3038	31924.395
7	99000	490495.5	27517.5	0	-16941	-1588	23703.93
8	171000	1560726	41301	1	-13778	-3130	11368.62
9	360000	1530000	42075	0	-18850	-449	13832.775

The rest of the columns chosen to perform univariate analysis are in the link below :

https://docs.google.com/spreadsheets/d/1N8ho6bKiiMOMwOGUt60p_Hj9SGPq88_L/edit?usp=sharing&ouid=101204343036685814262&rtpof=true&sd=true

For Descriptive analysis -

TofindCount,Mean,Std,Minimum,Maximum,Median,Range,Mode,Percentage,Variance,quartile,percentile values These following functions are used :

1. COUNT FUNCTION:

=COUNT(A1:A50000)

2.MEAN FUNCTION:

=AVERAGE(A1: A50000)

3.STANDARD DEVIATION FUNCTION :

=STDEV.S(A1:A50000)

4.MINIMUM FUNCTION :

=MIN(A1:A50000)

5.MAXIMUM FUNCTION :

=MAX(A1:A50000)

6.RANGE FUNCTION :

=MAX(A1:A50000) – MIN(A1:A50000)

7.MODE FUNCTION :

=MODE(A1:A50000)

8.VARIANCE FUNCTION:

=VAR.S(A1: A50000)

9.QUARTILE FUNCTION:

Q1 = QUARTILE.INC(A1:A50000,1)

Q2 = QUARTILE.INC(A1:A50000,2)

Q3 = QUARTILE.INC(A1:A50000,3)

10.PERCENTILE FUNCTION:

1st Percentile:

=PERCENTILE.INC(A1:A50000, 1/100)

5th Percentile :

=PERCENTILE.INC(A1:A50000, 5/100)

10th Percentile:

=PERCENTILE.INC(A1:A50000, 10/100)

25th Percentile:

=PERCENTILE.INC(A1:A50000, 25/100)

50th Percentile :

=PERCENTILE.INC(A1:A50000, 50/100)

75th Percentile:

=PERCENTILE.INC(A1:A50000, 75/100)

90th Percentile :

=PERCENTILE.INC(A1:A50000, 90/100)

95th Percentile:

=PERCENTILE.INC(A1:A50000, 95/100)

99th Percentile:

=PERCENTILE.INC(A1:A50000, 99/100)

Percentage was also found for all columns

	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	CNT_CHILDREN	DAYS_BIRTH	DAYS_EMPLOYED	AMT_ANNUITY.1
count	49999	49999	49998	49999	49999	49999	39408
mean	170767.5905	599700.5815	27107.37736	0.419848397	-16022.04208	63219.42449	15482.20397
std	531819.0951	402415.4339	14562.94444	0.724038548	4361.40027	140794.6057	14530.99679
min	25650	45000	2052	0	-25184	-17531	0
max	117000000	4050000	258025.5	11	-7680	365243	234478.395
Median	145800	514777.5	24939	0	-15731	-1221	10879.8075
Range	116974350	4005000	255973.5	11	17504	382774	234478.395
mode	135000	450000	9000	0	-13429	365243	2250
percentage	12.315789	43.2505127	1.95494978	41.98	-1.155125	4.55939615	0.88006075
variance	2.82832E+11	1.61938E+11	212079350.6	0.524231818	19021812.32	19823120985	211149867.6
quantiles(0.25)	112500	270000	16456.5	0	-19644	-2786	6122.835
quantiles(0.5)	145800	514777.5	24939	0	-15731	-1221	10879.8075
quantiles(0.75)	202500	808650	34596	1	-12378.5	-292	19668.915
Percentile(1%)	45000	76410	6174	0	-24426.02	-10942.18	2117.33505
Percentile(5%)	67500	135000	9000	0	-23197	-6824.2	2666.58075
percentile(10%)	81000	180000	11002.5	0	-22180.2	-4942.4	3708.3105
percentile(25%)	112500	270000	16456.5	0	-19644	-2786	6122.835
percentile(50%)	145800	514777.5	24939	0	-15731	-1221	10879.8075
percentile(75%)	202500	808650	34596	1	-12378.5	-292	19668.915
percentile(90%)	270000	1132573.5	45954	2	-10291.8	365243	33643.845
percentile(95%)	337500	1350000	53248.5	2	-9381.9	365243	45000
percentile(99%)	477045	1847703.6	70007.31	3	-8244	365243	69105.61395

COUNTS CALCULATED FOR CATEGORICAL COLUMN :

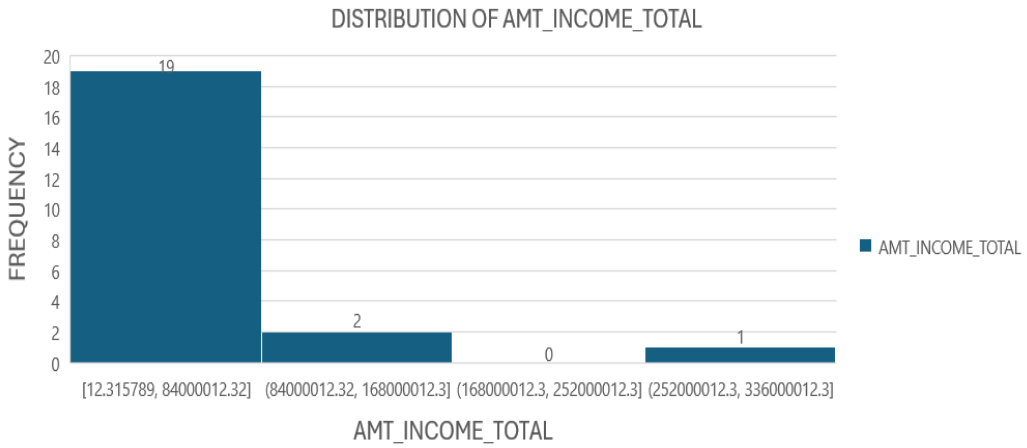
	NAME_CONTRACT_TYPE COUNTS	NAME_CONTRACT_STATUS COUNTS
Approved	0	31885
canceled	1	8594
cash loans	20855	0
consumer loans	23510	0
refused	0	8660
revolving loans	5625	0
unused offer	0	859
XNA	8	0

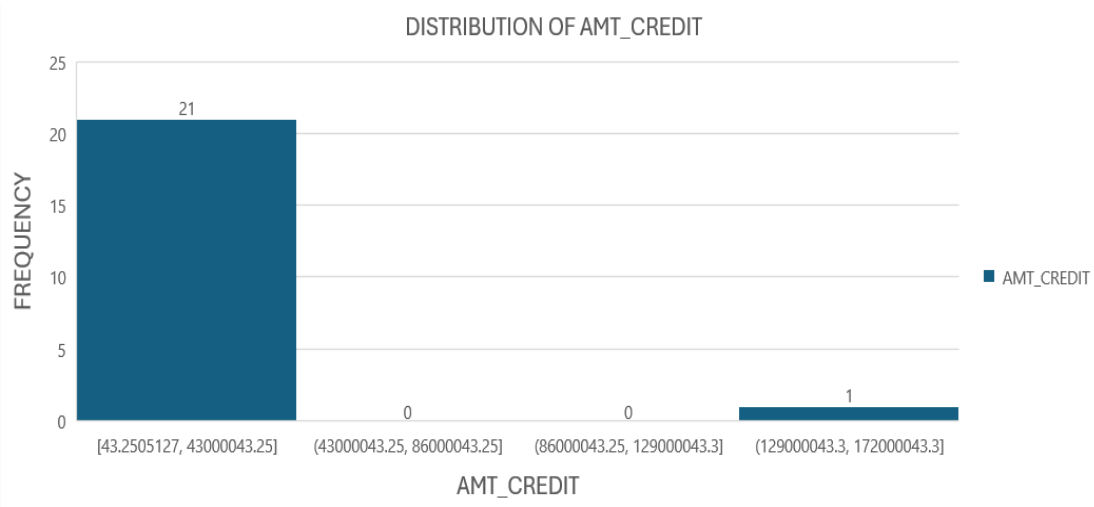
The rest of the descriptive analysis columns output link is given below :

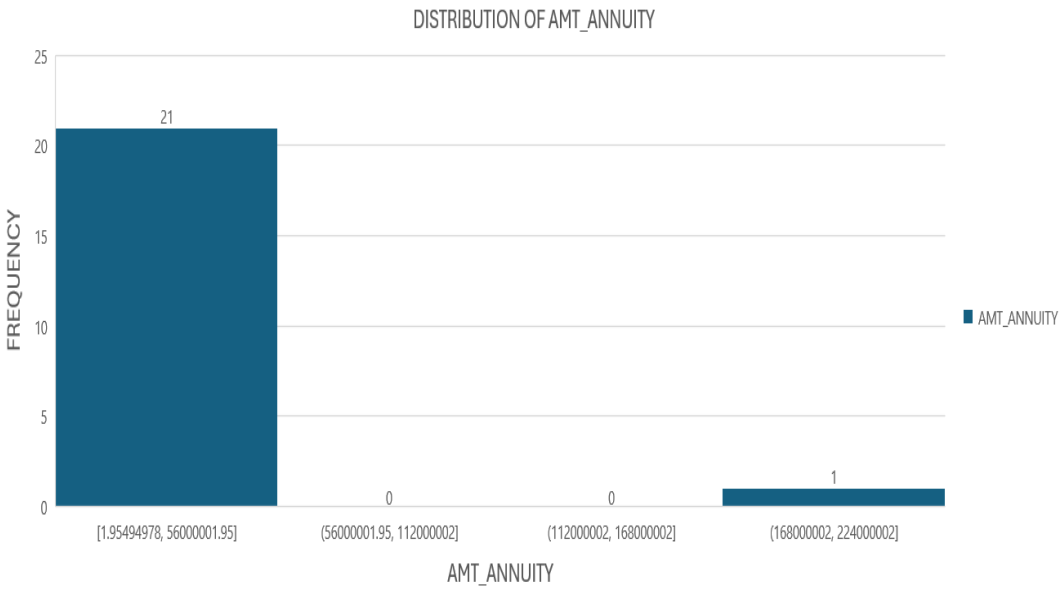
https://docs.google.com/spreadsheets/d/1-cJcCWAY_A7eAVPrI7gKQY7S3k-6ho1m/edit?usp=sharing&ouid=101204343036685814262&rtpof=true&sd=true

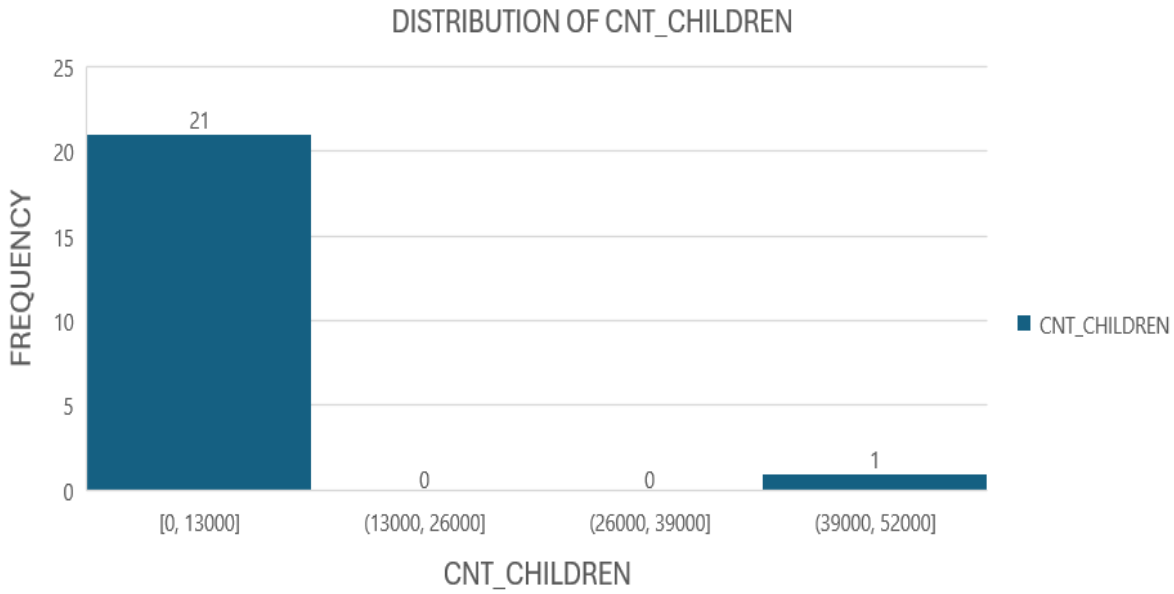
I have used histograms and bar chart to show distribution of individual variable below

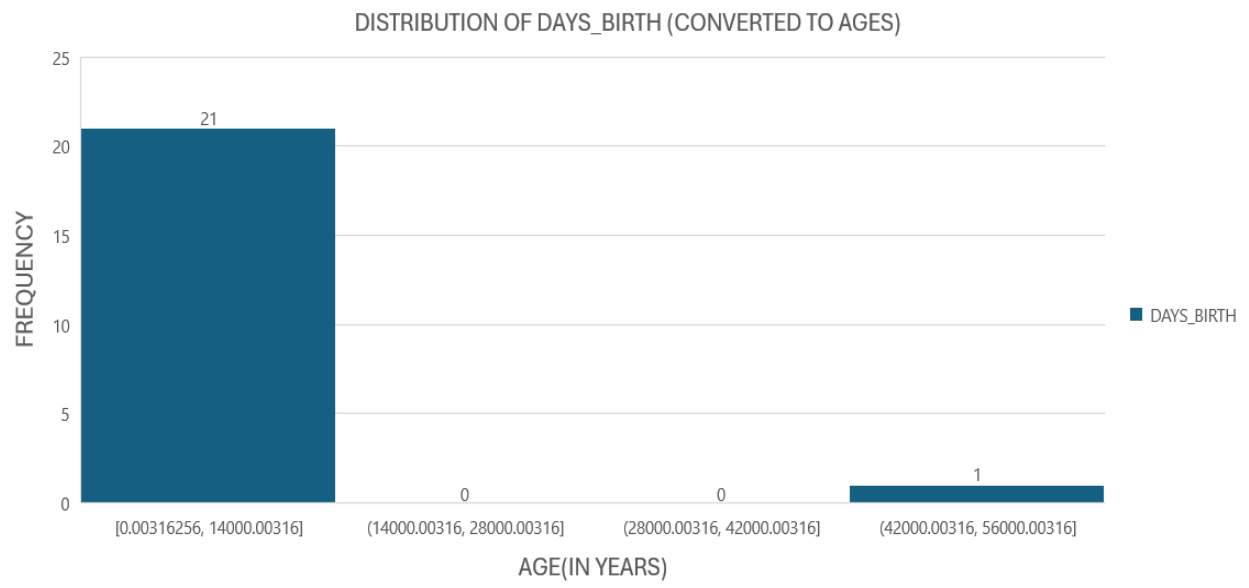
HISTOGRAMS :


DISTRIBUTION OF AMT_INCOME_TOTAL
FREQUENCY
20
18
16
14
12
10
8
6
4
2
0
19
2
0
1
[12.315789, 84000012.32]
(84000012.32, 168000012.3]
(168000012.3, 252000012.3]
(252000012.3, 336000012.3]
AMT_INCOME_TOTAL
AMT_INCOME_TOTAL


DISTRIBUTION OF AMT_CREDIT
FREQUENCY
25
20
15
10
5
0
21
0
0
1
[43.2505127, 43000043.25]
(43000043.25, 86000043.25]
(86000043.25, 129000043.3]
(129000043.3, 172000043.3]
AMT_CREDIT
AMT_CREDIT


DISTRIBUTION OF AMT_ANNUITY
FREQUENCY
25
20
15
10
5
0
21
0
0
1
[1.95494978, 56000001.95]
(56000001.95, 112000002]
(112000002, 168000002]
(168000002, 224000002]
AMT_ANNUITY
AMT_ANNUITY

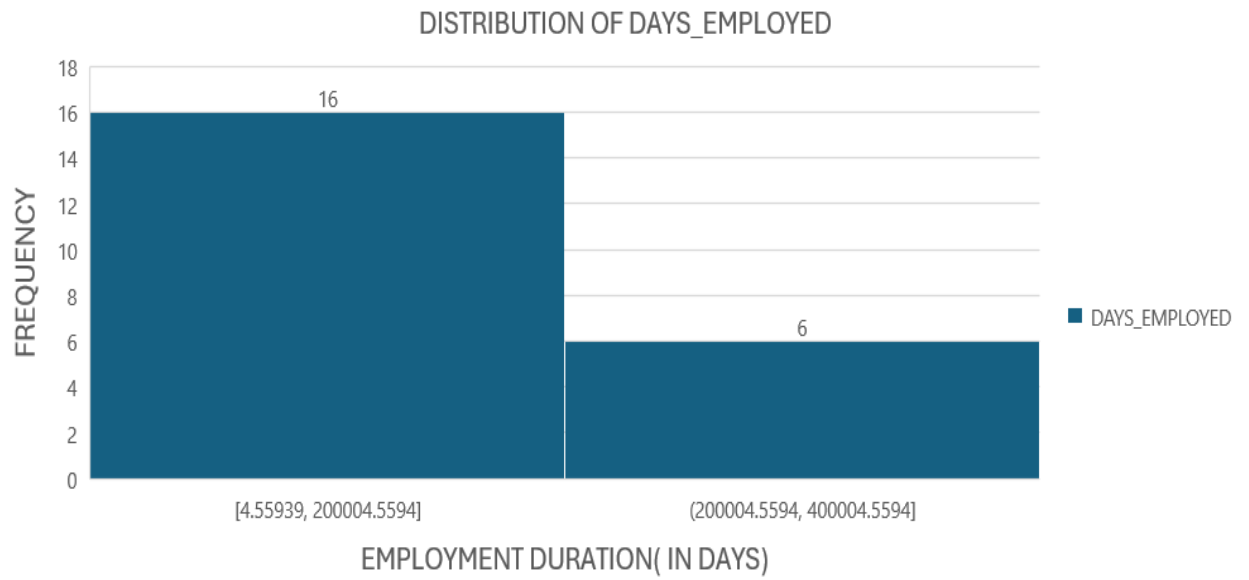

DISTRIBUTION OF CNT_CHILDREN
FREQUENCY
25
20
15
10
5
0
21
0
0
1
[0, 13000]
(13000, 26000]
(26000, 39000]
(39000, 52000]
CNT_CHILDREN
CNT_CHILDREN

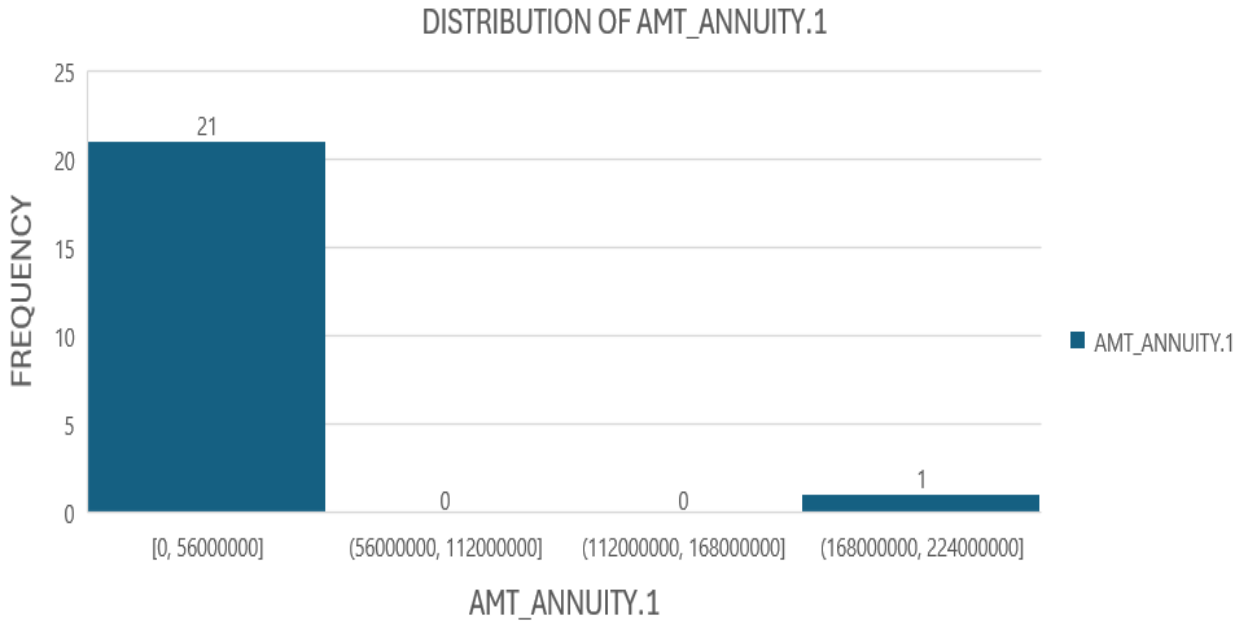


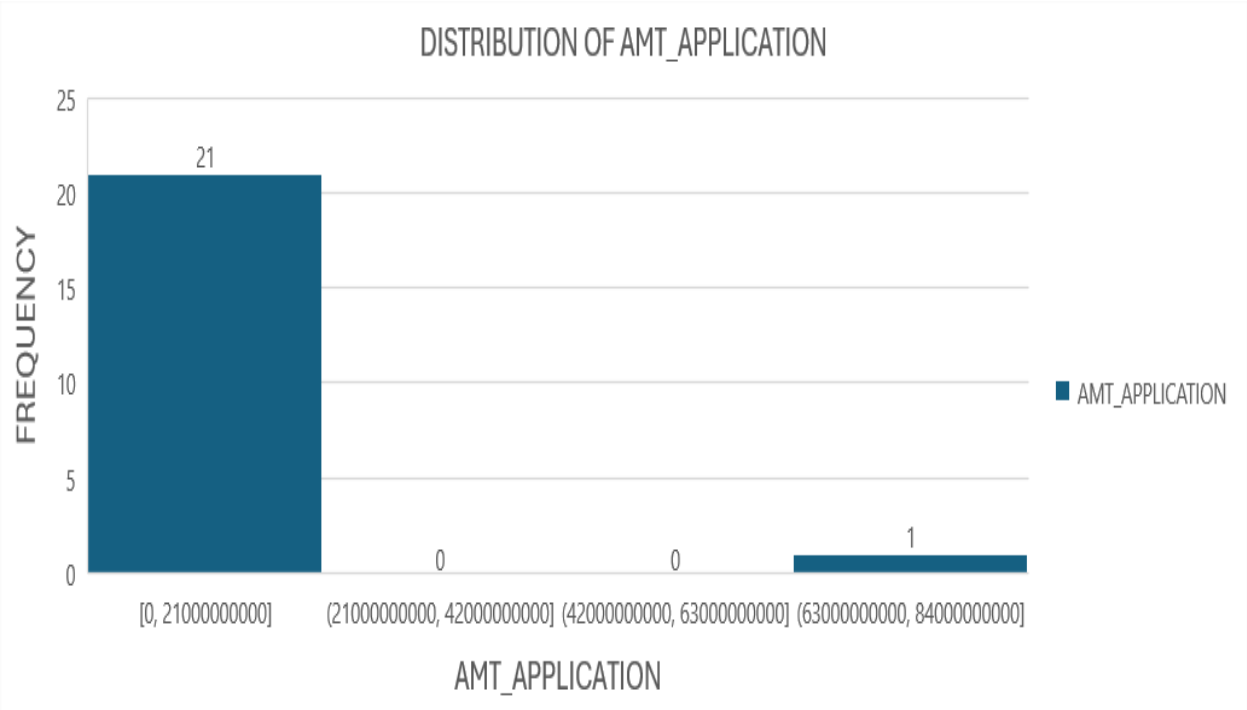
DAYS_BIRTH column is converted to years for positive values on X – Axis

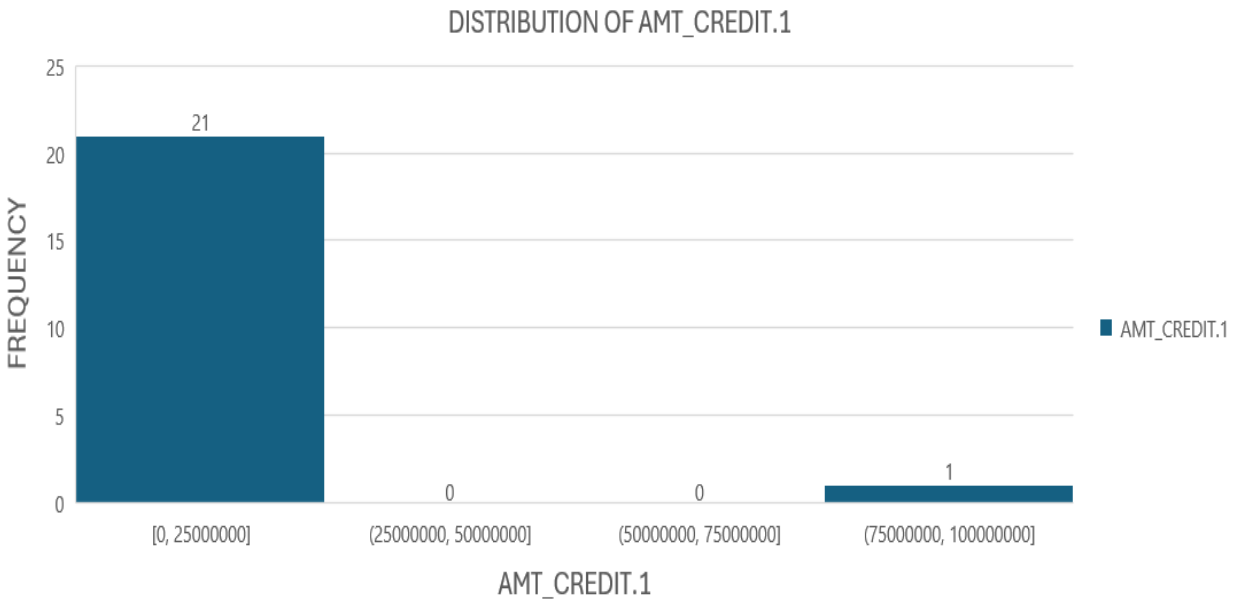
The formula used to convert age in days into age in years :

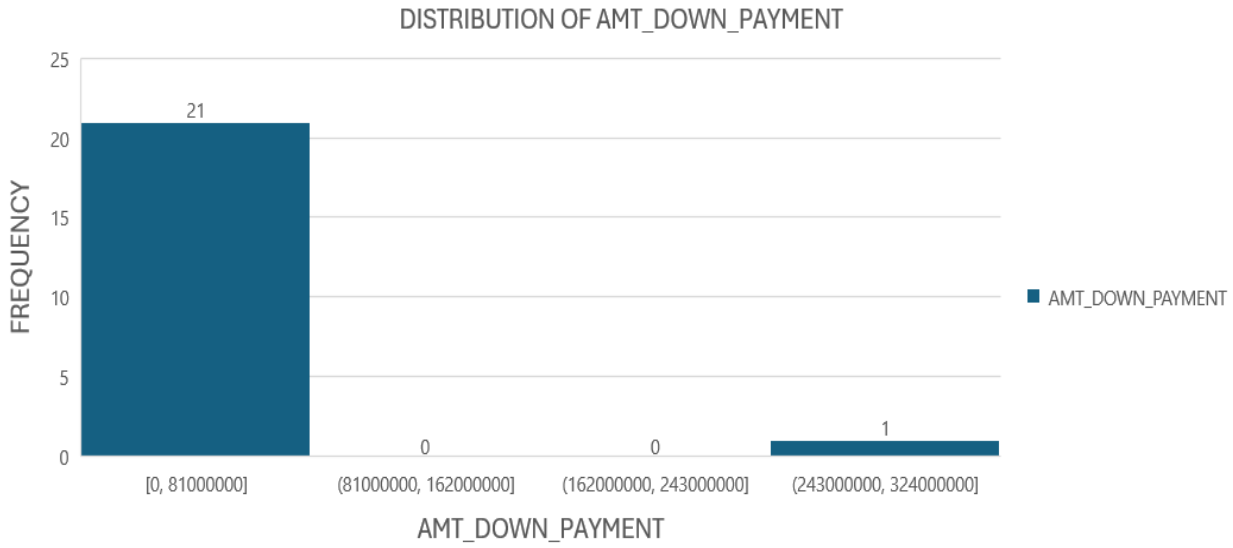
= the age value / 365.25



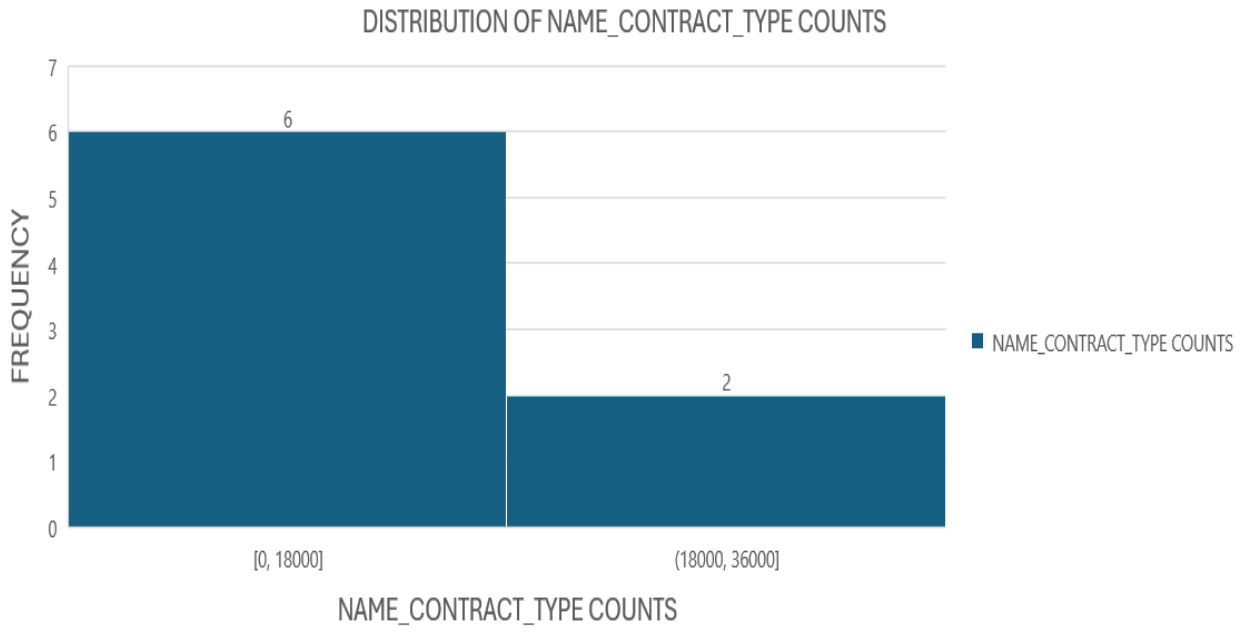

DISTRIBUTION OF AMT_ANNUITY.1
FREQUENCY
25
20
15
10
5
0
21
0
0
1
[0, 56000000]
(56000000, 112000000]
(112000000, 168000000]
(168000000, 224000000]
AMT_ANNUITY.1
AMT_ANNUITY.1

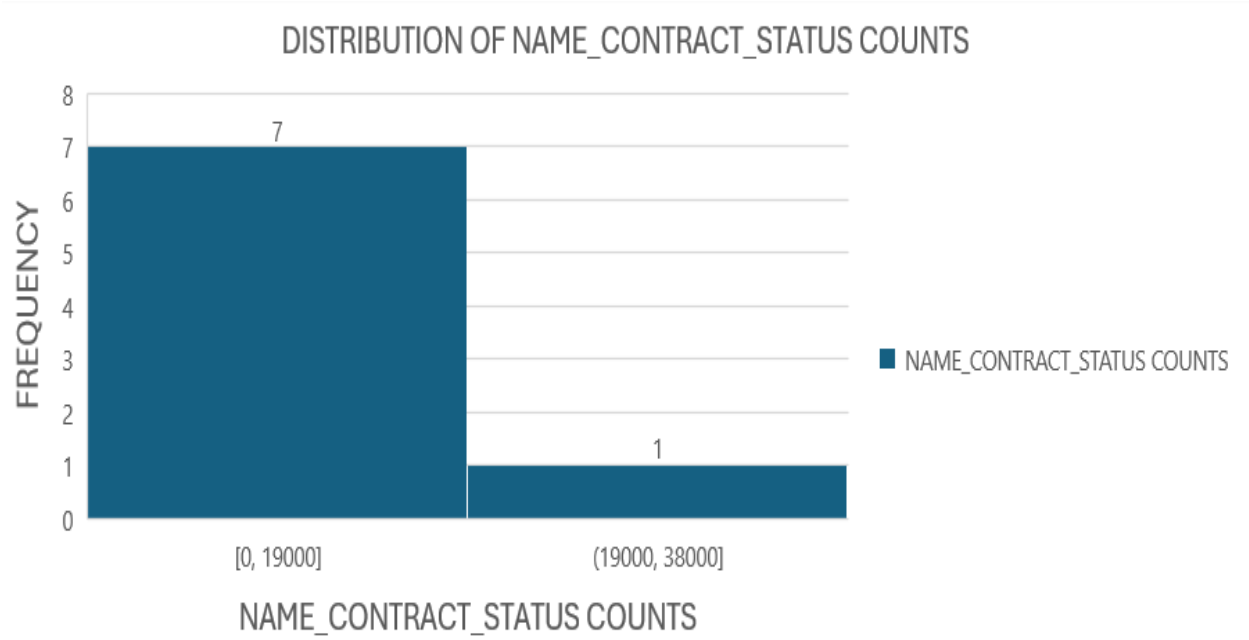

DISTRIBUTION OF AMT_APPLICATION
FREQUENCY
25
20
15
10
5
0
21
0
0
1
[0, 21000000000]
(21000000000, 42000000000]
(42000000000, 63000000000]
(63000000000, 84000000000]
AMT_APPLICATION
AMT_APPLICATION


DISTRIBUTION OF AMT_CREDIT.1
FREQUENCY
25
20
15
10
5
0
21
0
0
1
[0, 25000000]
(25000000, 50000000]
(50000000, 75000000]
(75000000, 100000000]
AMT_CREDIT.1
AMT_CREDIT.1

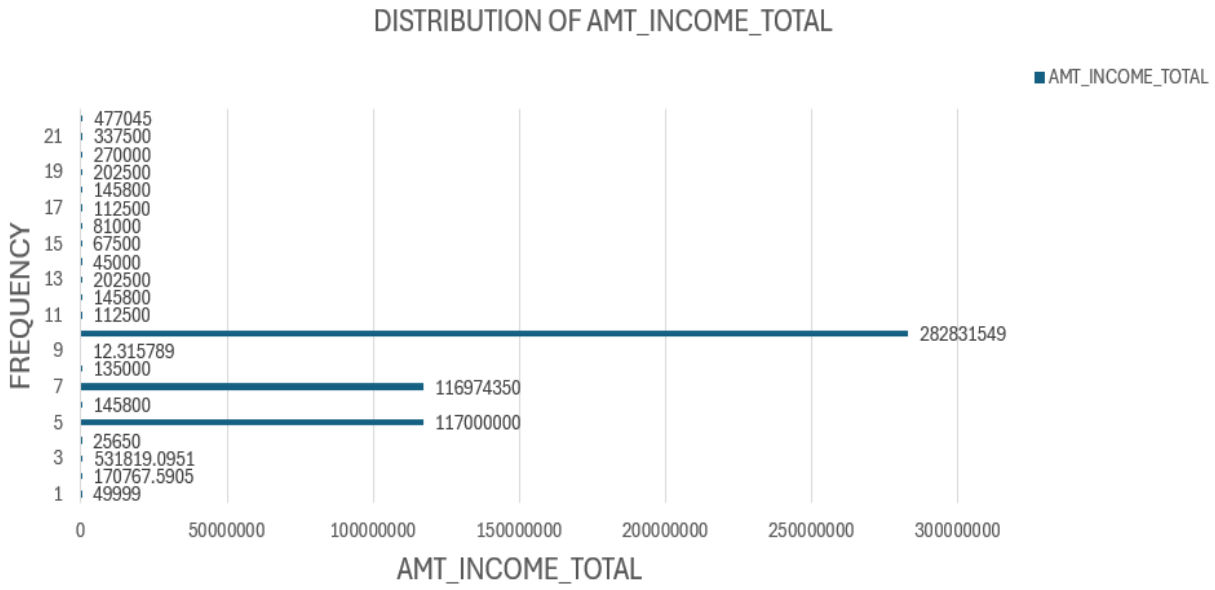

DISTRIBUTION OF AMT_DOWN_PAYMENT
FREQUENCY
25
20
15
10
5
0
21
0
0
1
[0, 81000000]
(81000000, 162000000]
(162000000, 243000000]
(243000000, 324000000]
AMT_DOWN_PAYMENT
AMT_DOWN_PAYMENT

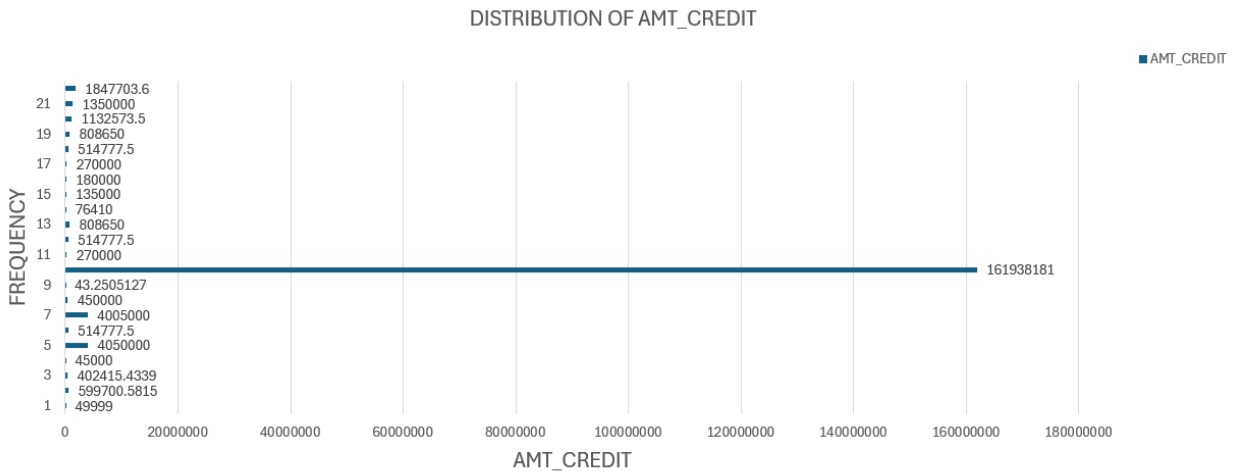
DISTRIBUTION OF VARIABLE COUNTS

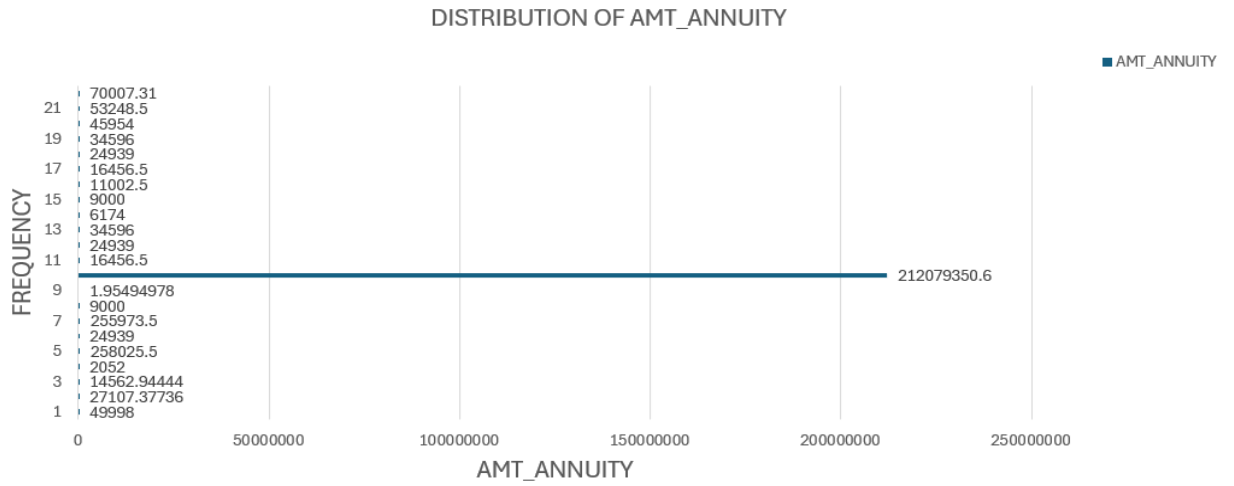


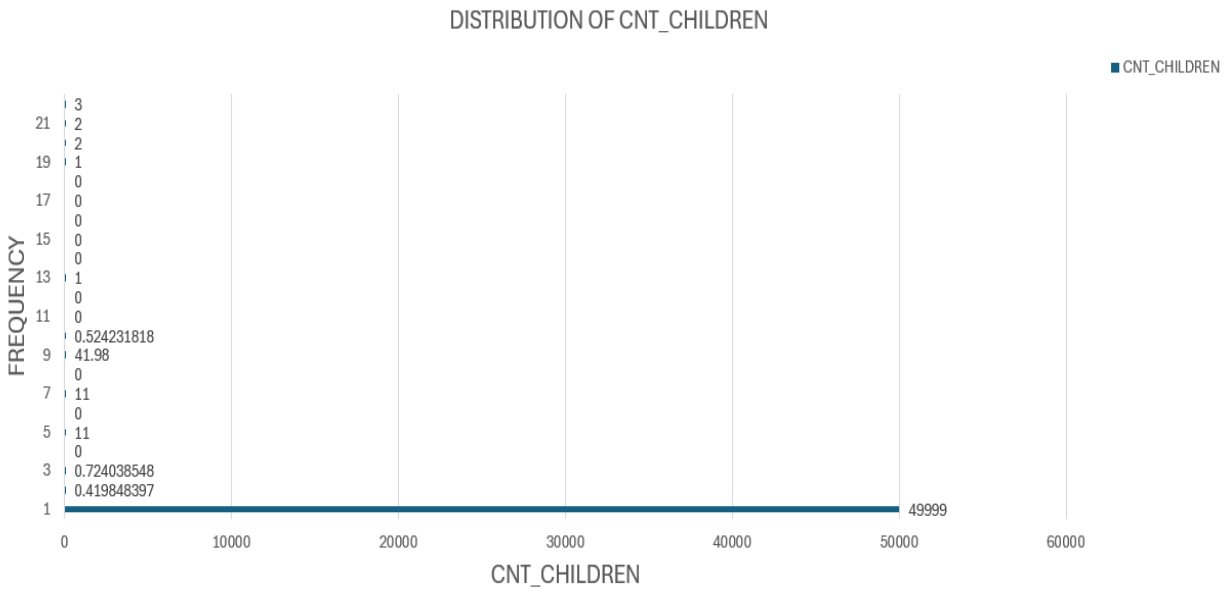

DISTRIBUTION OF NAME_CONTRACT_STATUS COUNTS
FREQUENCY
8
7
6
5
4
3
2
1
0
7
1
[0, 19000]
(19000, 38000]
NAME_CONTRACT_STATUS COUNTS
NAME_CONTRACT_STATUS COUNTS

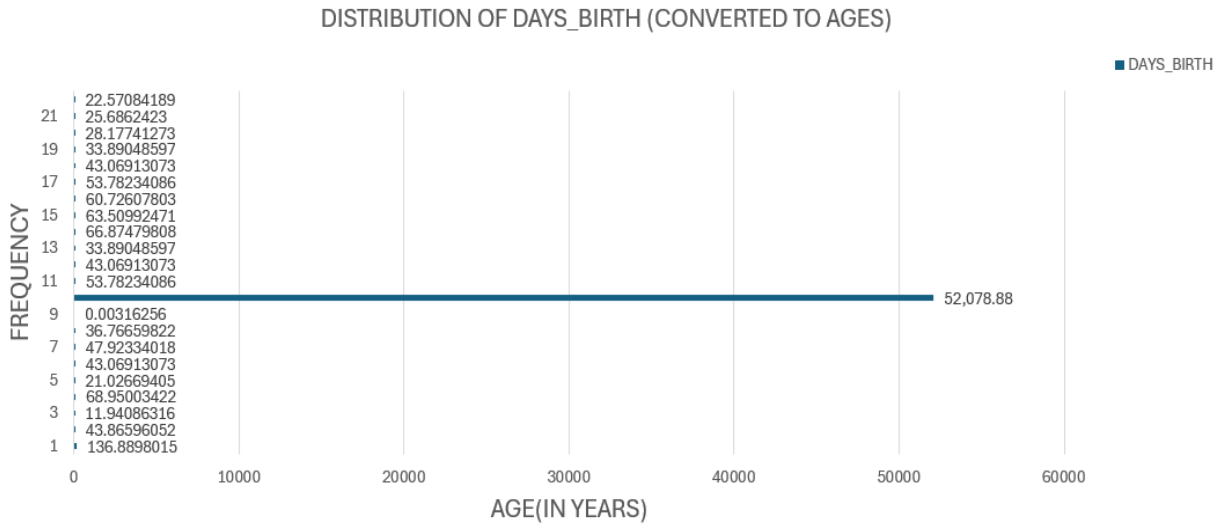
BAR CHARTS :


DISTRIBUTION OF AMT_INCOME_TOTAL
AMT_INCOME_TOTAL
FREQUENCY
21
19
17
15
13
11
9
7
5
3
1
477045
337500
270000
202500
145800
112500
81000
67500
45000
202500
145800
112500
12.315789
135000
145800
25650
531819.0951
170767.5905
49999
282831549
116974350
117000000
0
50000000
100000000
150000000
200000000
250000000
300000000
AMT_INCOME_TOTAL


DISTRIBUTION OF AMT_CREDIT
AMT_CREDIT
FREQUENCY
1847703.6
1350000
1132573.5
808650
514777.5
270000
180000
135000
76410
808650
514777.5
270000
161938181
43.2505127
450000
4005000
514777.5
4050000
45000
402415.4339
599700.5815
49999
0
20000000
40000000
60000000
80000000
100000000
120000000
140000000
160000000
180000000
AMT_CREDIT


DISTRIBUTION OF AMT_ANNUITY
AMT_ANNUITY
FREQUENCY
70007.31
53248.5
45954
34596
24939
16456.5
11002.5
9000
6174
34596
24939
16456.5
212079350.6
1.95494978
9000
255973.5
24939
258025.5
2052
14562.94444
27107.37736
49998
0
50000000
100000000
150000000
200000000
250000000
AMT_ANNUITY

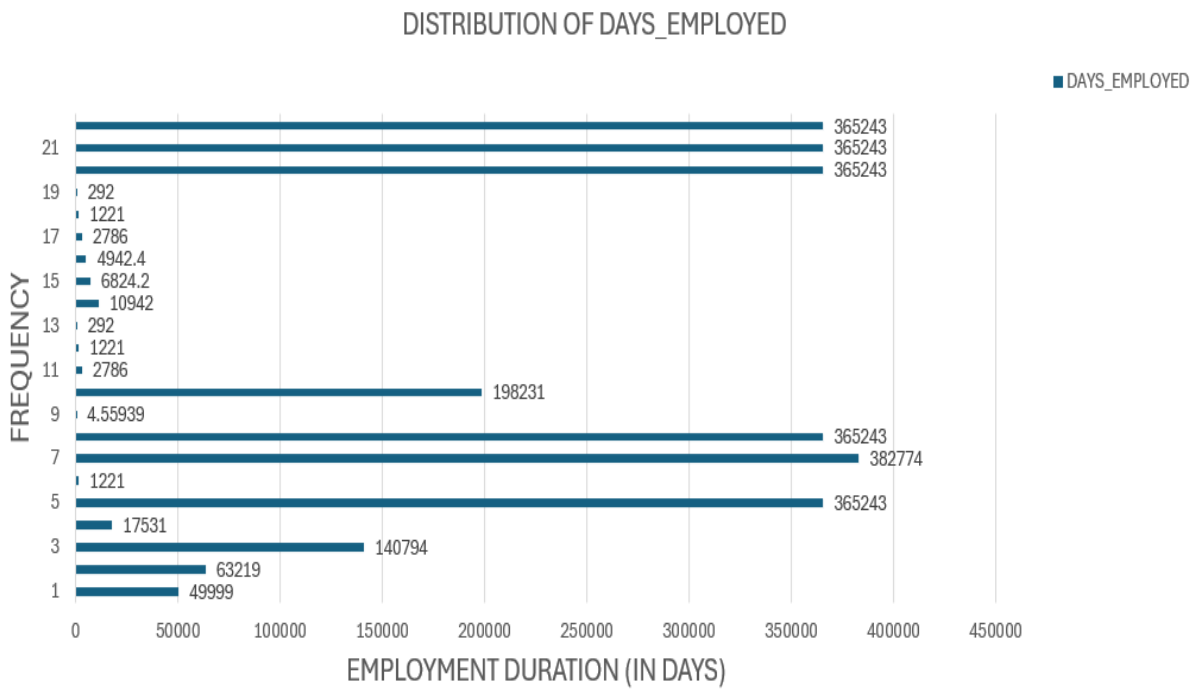


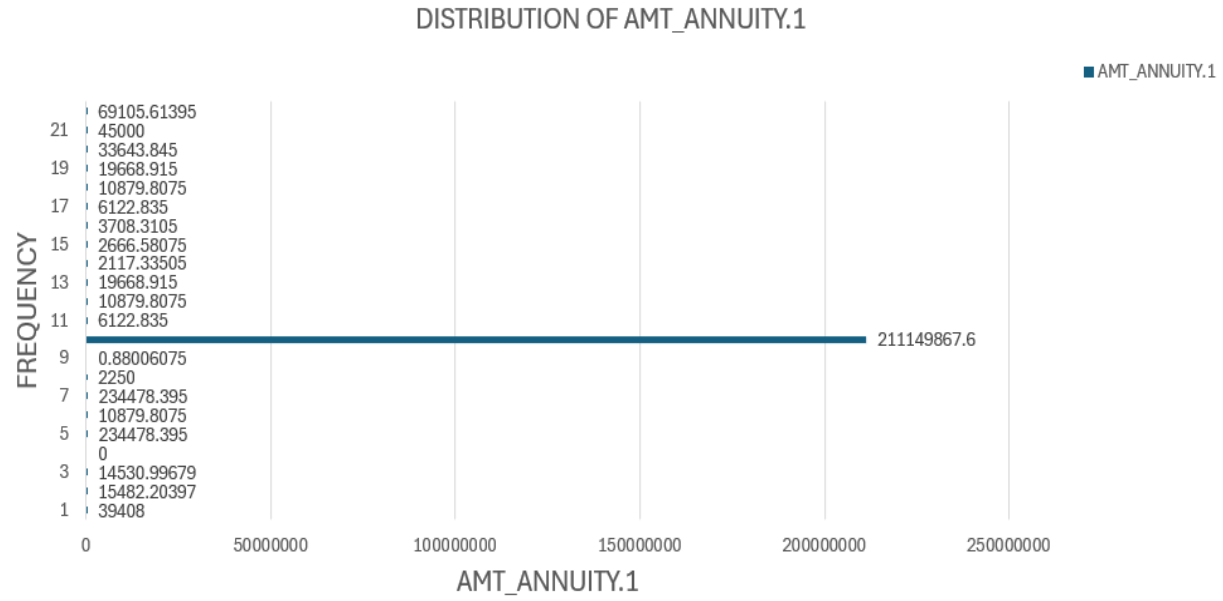


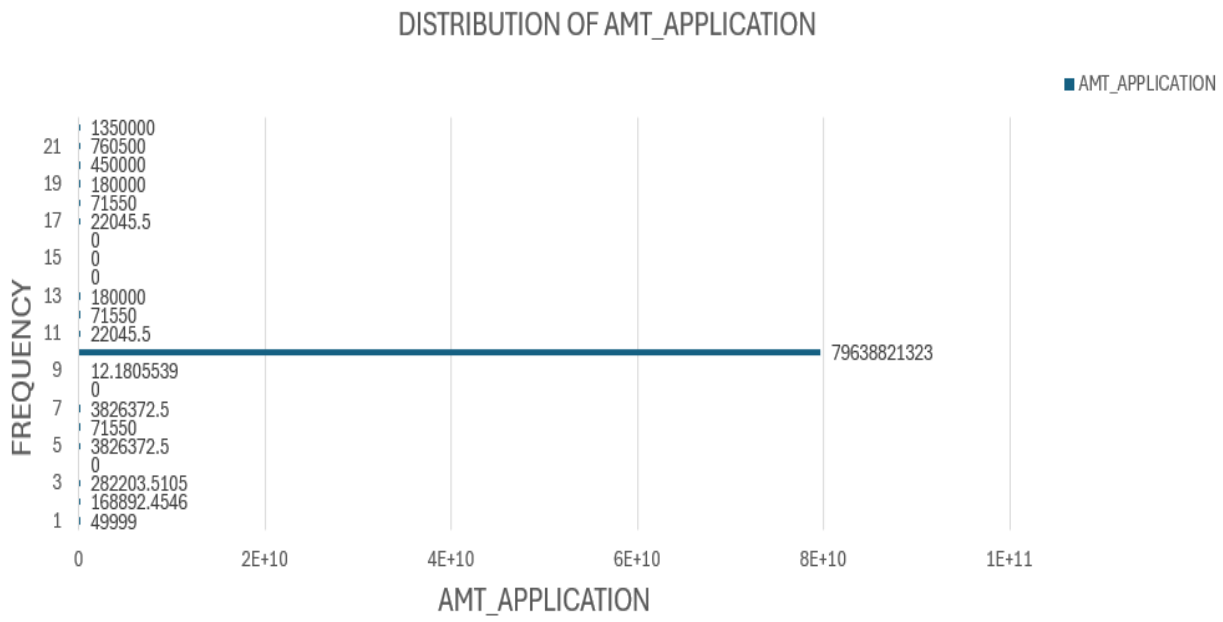
DAYS_BIRTH column is converted to years for positive values on X – Axis

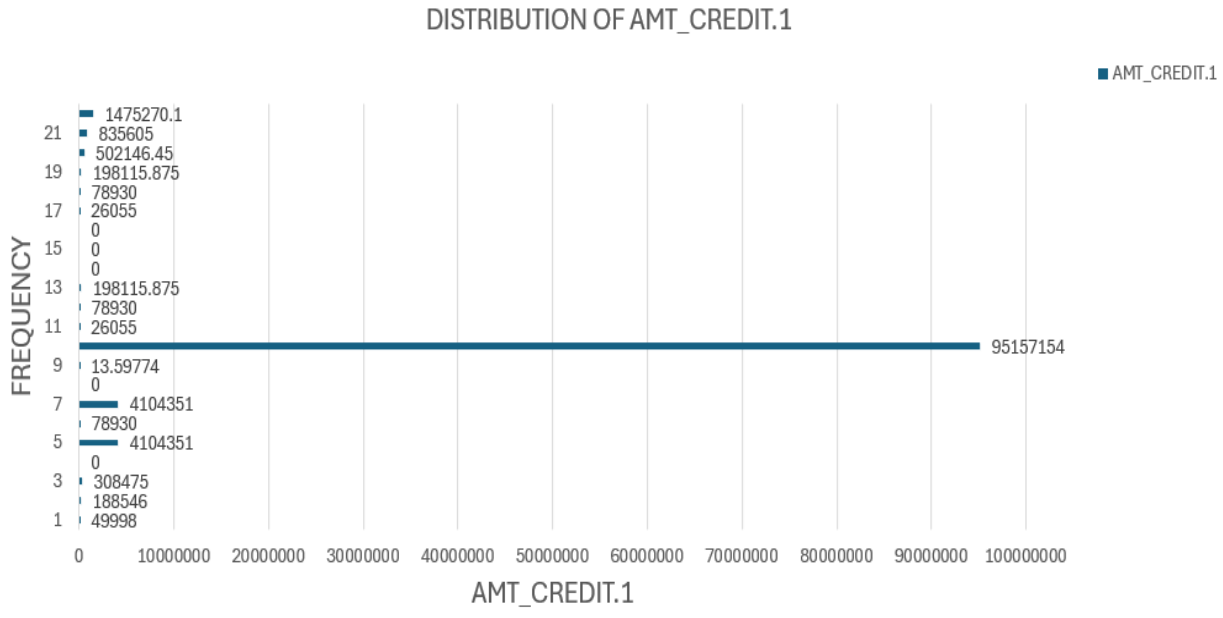
The formula used to convert age in days into age in years :

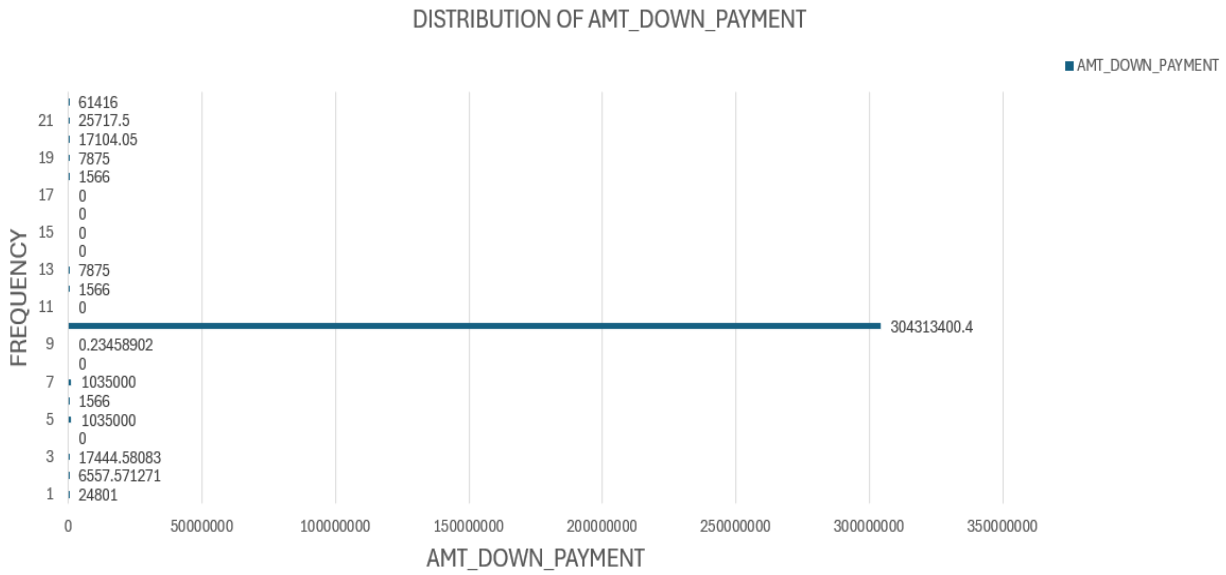
= the age value / 365.25

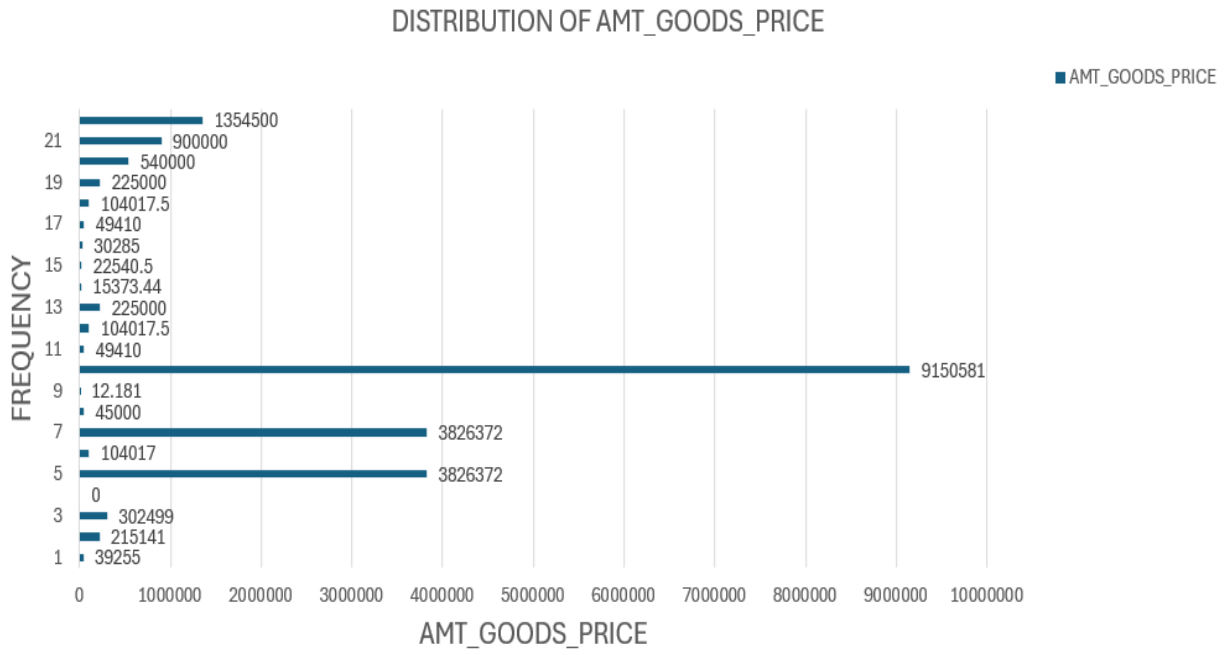

DISTRIBUTION OF DAYS_EMPLOYED
DAYS_EMPLOYED
FREQUENCY
EMPLOYMENT DURATION (IN DAYS)
365243
365243
365243
292
1221
2786
4942.4
6824.2
10942
292
1221
2786
198231
4.55939
365243
382774
1221
365243
17531
140794
63219
49999
0
50000
100000
150000
200000
250000
300000
350000
400000
450000
1
3
5
7
9
11
13
15
17
19
21


DISTRIBUTION OF AMT_ANNUITY.1
AMT_ANNUITY.1
FREQUENCY
AMT_ANNUITY.1
69105.61395
45000
33643.845
19668.915
10879.8075
6122.835
3708.3105
2666.58075
2117.33505
19668.915
10879.8075
6122.835
211149867.6
0.88006075
2250
234478.395
10879.8075
234478.395
0
14530.99679
15482.20397
39408
0
50000000
100000000
150000000
200000000
250000000
1
3
5
7
9
11
13
15
17
19
21

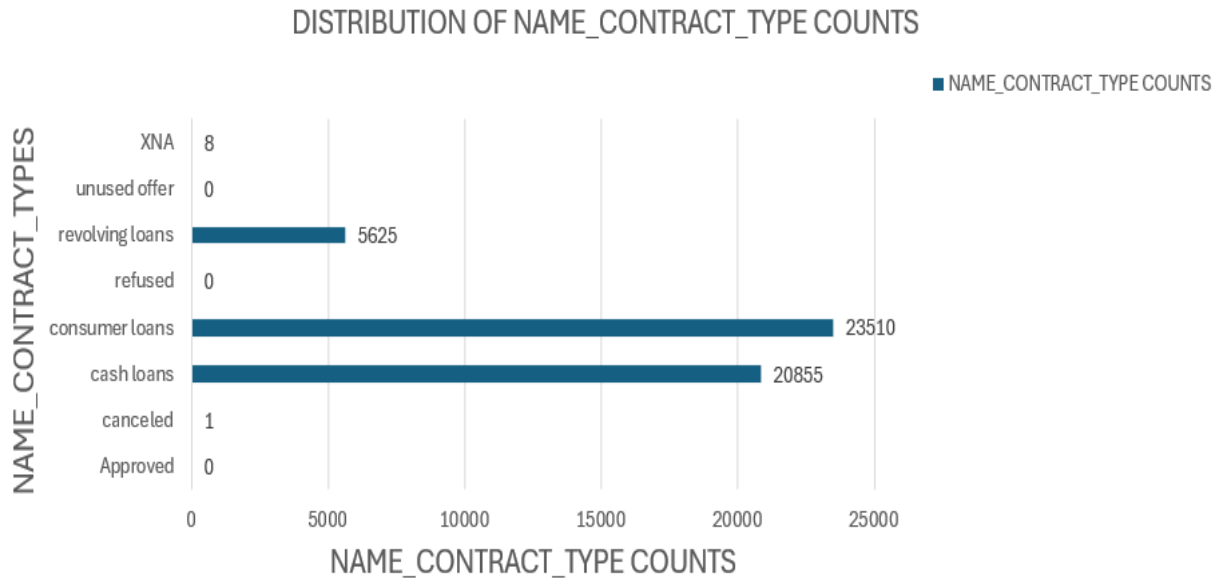

DISTRIBUTION OF AMT_APPLICATION
AMT_APPLICATION
FREQUENCY
1350000
760500
450000
180000
71550
22045.5
0
0
0
180000
71550
22045.5
79638821323
12.1805539
0
3826372.5
71550
3826372.5
0
282203.5105
168892.4546
49999
0
2E+10
4E+10
6E+10
8E+10
1E+11
AMT_APPLICATION

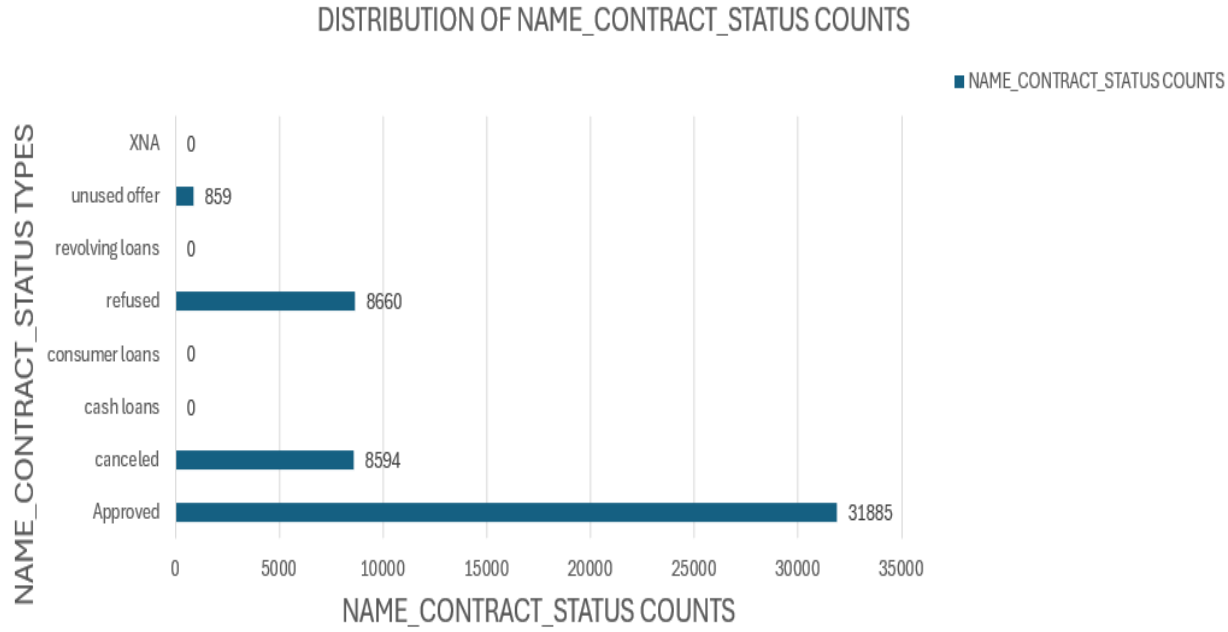

DISTRIBUTION OF AMT_CREDIT.1
AMT_CREDIT.1
FREQUENCY
1475270.1
835605
502146.45
198115.875
78930
26055
0
0
0
198115.875
78930
26055
95157154
13.59774
0
4104351
78930
4104351
0
308475
188546
49998
0
10000000
20000000
30000000
40000000
50000000
60000000
70000000
80000000
90000000
100000000
AMT_CREDIT.1


DISTRIBUTION OF AMT_DOWN_PAYMENT
AMT_DOWN_PAYMENT
FREQUENCY
AMT_DOWN_PAYMENT
61416
25717.5
17104.05
7875
1566
0
0
0
0
7875
1566
0
304313400.4
0.23458902
0
1035000
1566
1035000
0
17444.58083
6557.571271
24801
0
50000000
100000000
150000000
200000000
250000000
300000000
350000000


DISTRIBUTION OF AMT_GOODS_PRICE
AMT_GOODS_PRICE
FREQUENCY
AMT_GOODS_PRICE
1354500
900000
540000
225000
104017.5
49410
30285
22540.5
15373.44
225000
104017.5
49410
9150581
12.181
45000
3826372
104017
3826372
0
302499
215141
39255
0
1000000
2000000
3000000
4000000
5000000
6000000
7000000
8000000
9000000
10000000

DISTRIBUTION OF VARIABLE COUNTS





2)SEGMENTED UNIVARIATE ANALYSIS:

For Segmented univariate Analysis the columns CODE_GENDER, NAME_INCOME_TYPE, NAME_EDUCATION_TYPE, NAME_FAMILY_STATUS, NAME_CLIENT_TYPE, NAME_PAYMENT_TYPE, CHANNEL_TYPE, NAME_CONTRACT_TYPE, NAME_CONTRACT_TYPE.1, OWN_CAR_AGE are taken

	A	B	C	D	E	F
1	CODE_GENDER	NAME_INCOME_TYPE	NAME_EDUCATION_TYPE	NAME_FAMILY_STATUS	NAME_CLIENT_TYPE	NAME_PAYMENT_TYPE
2	F	Working	Secondary / secondary special	Single / not married	Repeater	Cash through the bank
3	F	State servant	Higher education	Married	Repeater	XNA
4	F	Working	Secondary / secondary special	Single / not married	Repeater	Cash through the bank
5	F	Working	Secondary / secondary special	Civil marriage	Repeater	Cash through the bank
6	F	Working	Secondary / secondary special	Single / not married	Repeater	Cash through the bank
7	F	State servant	Secondary / secondary special	Married	Repeater	Cash through the bank
8	F	Commercial associate	Higher education	Married	Repeater	XNA
9	F	State servant	Higher education	Married	Repeater	XNA
10	F	Pensioner	Secondary / secondary special	Married	Repeater	XNA
11	F	Working	Secondary / secondary special	Single / not married	Repeater	XNA
12	F	Working	Higher education	Married	Repeater	Cash through the bank
13	F	Pensioner	Secondary / secondary special	Married	Repeater	Cash through the bank
14	F	Working	Secondary / secondary special	Married	Repeater	Cash through the bank
15	F	Working	Secondary / secondary special	Married	New	Cash through the bank
16	F	Working	Secondary / secondary special	Married	New	Cash through the bank
17	F	Working	Secondary / secondary special	Single / not married	New	Cash through the bank
18	M	Working	Secondary / secondary special	Married	Repeater	Cash through the bank
19	M	Working	Secondary / secondary special	Married	Repeater	XNA
20	M	Working	Secondary / secondary special	Widow	Repeater	Cash through the bank

The rest of the columns chosen to perform segmented univariate analysis are in the given link below :

https://docs.google.com/spreadsheets/d/16sDbOypfawu5Guv0hyRxp8crEOJFisYP/edit?usp=sharing&ouid=101204343036685814262&rtpof=true&sd=true

The Segments chosen are given below :

1) CODE_GENDER segmented by NAME_INCOME_TYPE

2) NAME_INCOME_TYPE segmented by NAME_EDUCATION_TYPE

3) NAME_EDUCATION_ TYPE segmented by NAME_FAMILY_STATUS

4) NAME_FAMILY_STATUS segmented by NAME_CLIENT_TYPE

5) NAME_CLIENT_TYPE segmented by NAME_PAYMENT_TYPE

6) NAME_PAYMENT_TYPE segmented by CHANNEL_TYPE

7) NAME_CONTRACT_TYPE segmented by CHANNEL_TYPE

8) NAME_CONTRACT_TYPE segmented by CODE_GENDER

9) NAME_CONTRACT_TYPE.1 segmented by OWN_CAR_AGE

10) OWN_CAR_AGE segmented by CODE_GENDER

These segments are done using pivot table

CODE_GENDER segmented by NAME_INCOME_TYPE

CODE_GENDER	Businessman	Commercial associate	Maternity leave	Pensioner	State servant
F	0	7221	0	7151	
M	2	4499	1	1678	
XNA	0	0	0	0	

NAME_INCOME_TYPE segmented by NAME_EDUCATION_TYPE

NAME_INCOME_TYPE	Academic degree	Higher education	Incomplete higher	Lower secondary	Secondary / secondary special
Businessman	0	2	0	0	
Commercial associate	4	3964	551	71	
Maternity leave	0	1	0	0	
Pensioner	1	1323	76	239	
State servant	6	1451	113	12	
Student	0	2	0	0	
Unemployed	0	5	1	0	
Working	5	5493	879	236	

The Rest of the segments done using pivot tables are in the given link below :

https://docs.google.com/spreadsheets/d/19_q79GtofixKYSdwJPeJdAQ4VDfVO1jn/edit?usp=sharing&ouid=101204343036685814262&rtpof=true&sd=true

The Descriptive Analysis is done to each of the segment's output found using pivot table

For Descriptive analysis -

TofindCount,Mean,Std,Minimum,Maximum,Median,Range,Mode,Percentage,Variance,quartile,percentile values These following functions are used :

1. COUNT FUNCTION:

=COUNT(A1:A50000)

2.MEAN FUNCTION:

=AVERAGE(A1: A50000)

3.STANDARD DEVIATION FUNCTION :

=STDEV.S(A1:A50000)

4.MINIMUM FUNCTION :

=MIN(A1:A50000)

5.MAXIMUM FUNCTION :

=MAX(A1:A50000)

6.RANGE FUNCTION :

=MAX(A1:A50000) – MIN(A1:A50000)

7.MODE FUNCTION :

=MODE(A1:A50000)

8.VARIANCE FUNCTION:

=VAR.S(A1: A50000)

9.QUARTILE FUNCTION:

Q1 = QUARTILE.INC(A1:A50000,1)

Q2 = QUARTILE.INC(A1:A50000,2)

Q3 = QUARTILE.INC(A1:A50000,3)

10.PERCENTILE FUNCTION:

1st Percentile:

=PERCENTILE.INC(A1:A50000, 1/100)

5th Percentile :

=PERCENTILE.INC(A1:A50000, 5/100)

10th Percentile:

=PERCENTILE.INC(A1:A50000, 10/100)

25th Percentile:

=PERCENTILE.INC(A1:A50000, 25/100)

50th Percentile :

=PERCENTILE.INC(A1:A50000, 50/100)

75th Percentile:

=PERCENTILE.INC(A1:A50000, 75/100)

90th Percentile :

=PERCENTILE.INC(A1:A50000, 90/100)

95th Percentile:

=PERCENTILE.INC(A1:A50000, 95/100)

99th Percentile:

=PERCENTILE.INC(A1:A50000, 99/100)

Percentage was also found for all columns

CODE_GENDER segmented by NAME_INCOME_TYPE

CODE_GENDER	Businessman	Commercial associate	Maternity leave	Pensioner	State servant	Student
F	0	7221	0	7151	2519	1
M	2	4499	1	1678	976	3
XNA	0	0	0	0	0	0
count	3	3	3	3	3	3
mean	675.9090909	2722.818182	1845.636364	763.7727273	2440.0625	346.375
std	1517.649684	5076.551357	4904.395342	1803.59432	5105.476073	803.4144878
min	0	0	0	0	0	0
max	6303	22586	22909	7151	19307	2311
Median	4.5	562.5	94.5	11	1.5	14.5
Range	6303	22586	22909	7151	19307	2311
mode	0	0.0,2.0	0	0	0	1
percentage	37.5	37.5	37.5	37.5	37.5	37.5
variance	2303260.563	25771373.68	24053093.67	3252952.47	26065885.93	645474.8393
quantiles(0.25)	0	5.75	0.25	0.25	0	1
quantiles(0.5)	4.5	562.5	94.5	11	1.5	14.5
quantiles(0.75)	659	3613.25	1750.5	238.25	2064.5	139.75
Percentile(1%)	0	0	0	0	0	0.07
Percentile(5%)	0	0.05	0	0	0	0.35
percentile(10%)	0	1.1	0	0	0	0.7
percentile(25%)	0	5.75	0.25	0.25	0	1
percentile(50%)	4.5	562.5	94.5	11	1.5	14.5
percentile(75%)	659	3613.25	1750.5	238.25	2064.5	139.75
percentile(90%)	1385.4	7048.2	3340.8	2077.6	7160	948.1

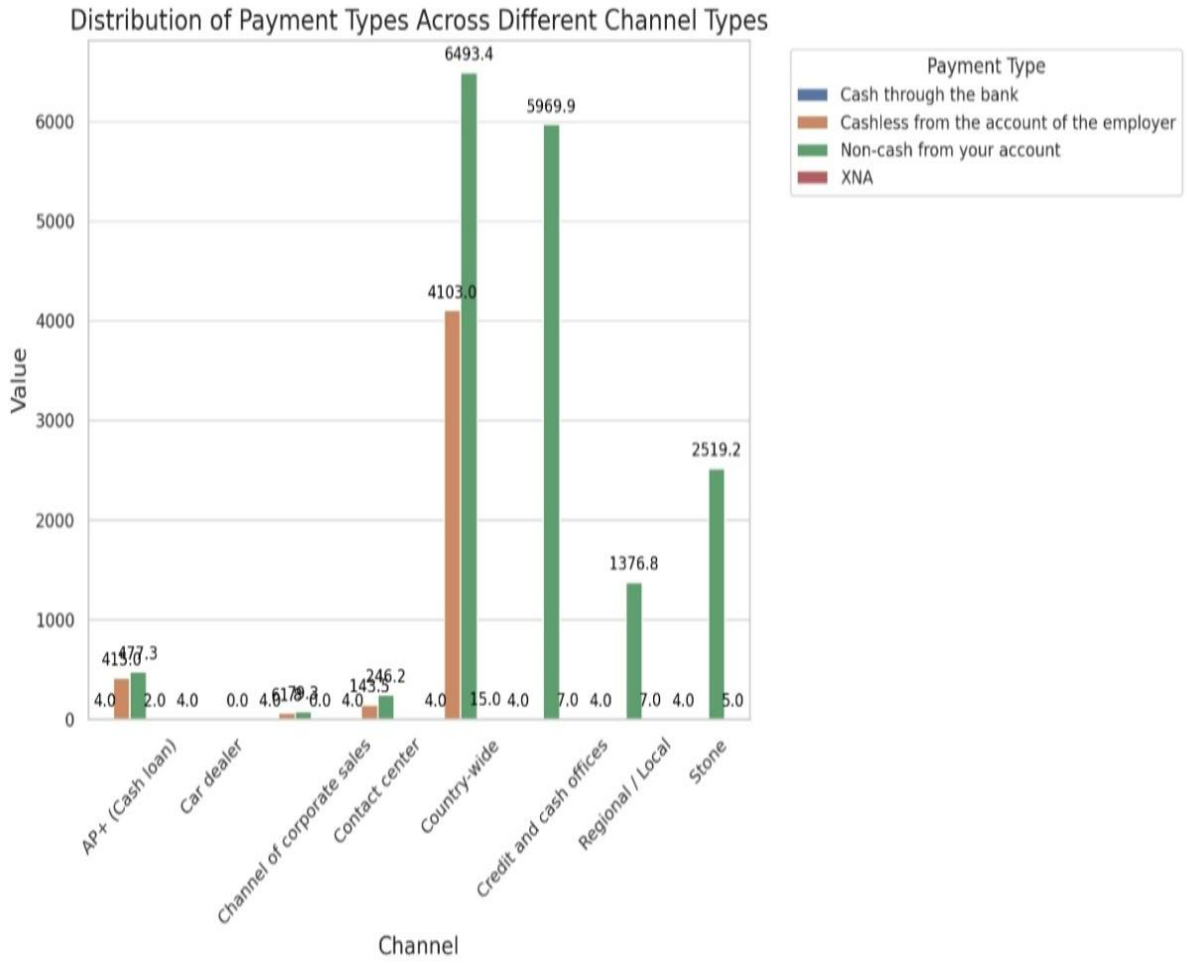
The Rest of the segment's descriptive analysis output is given in the below link :

https://docs.google.com/spreadsheets/d/1ERFXED5g68gr3Vd_fwofd2Rog3rlfiHE/edit?usp=sharing&ouid=101204343036685814262&rtpof=true&sd=true

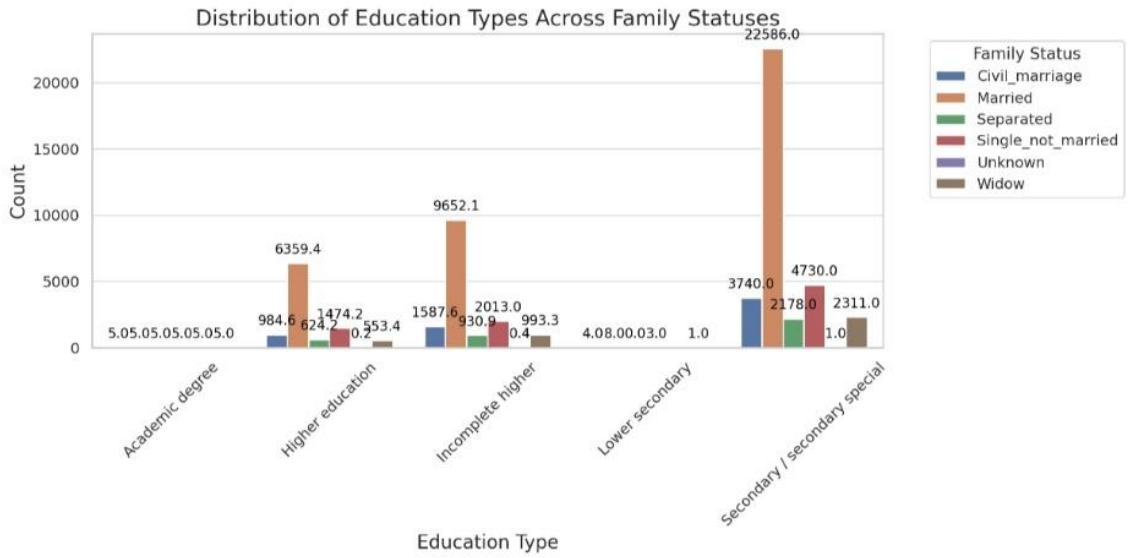
The bar charts plotted to compare variable distributions across different scenarios are given below

BAR CHARTS:

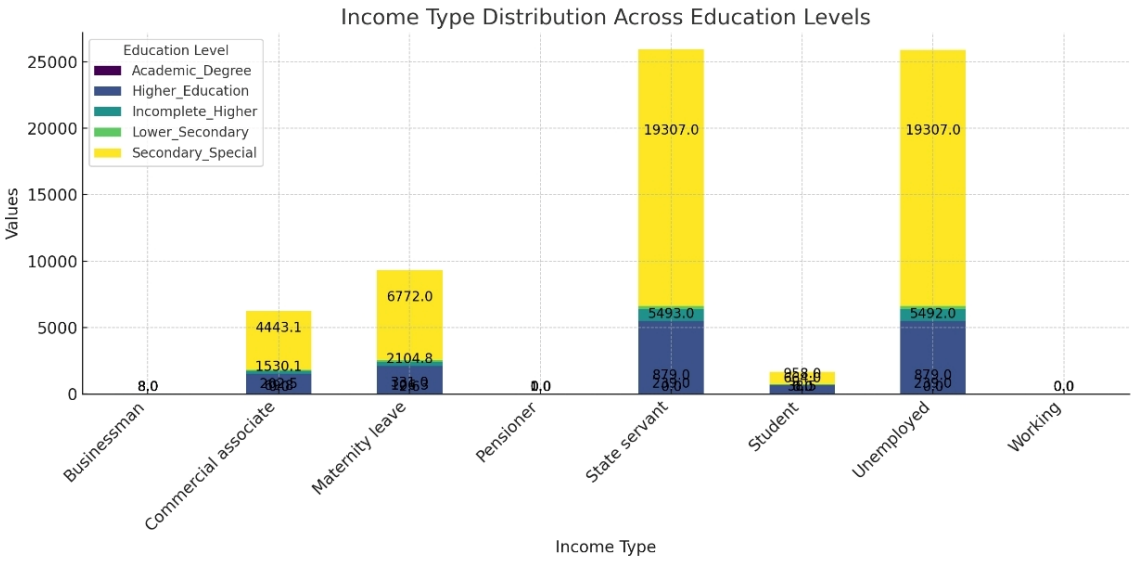
NAME_PAYMENT_TYPE segmented by CHANNEL_TYPE



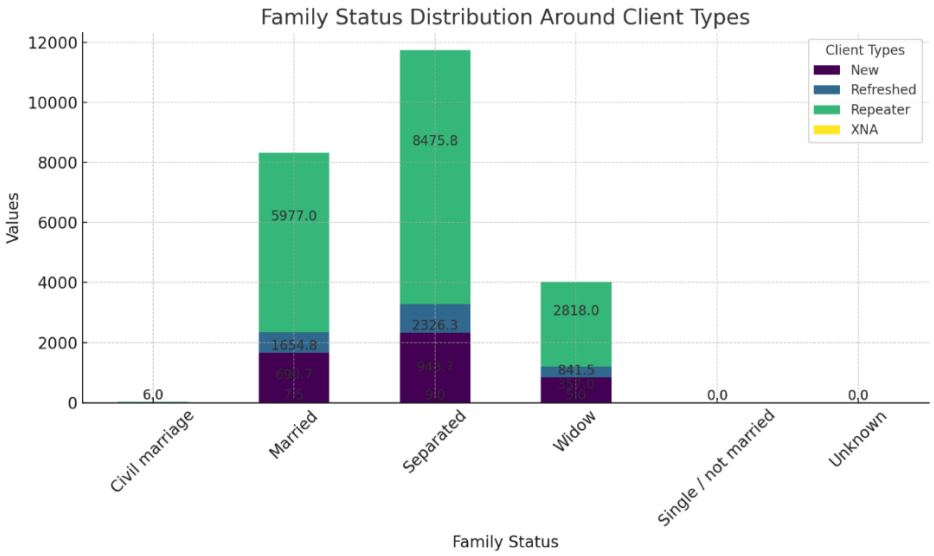
NAME_EDUCATION_ TYPE segmented by NAME_FAMILY_STATUS



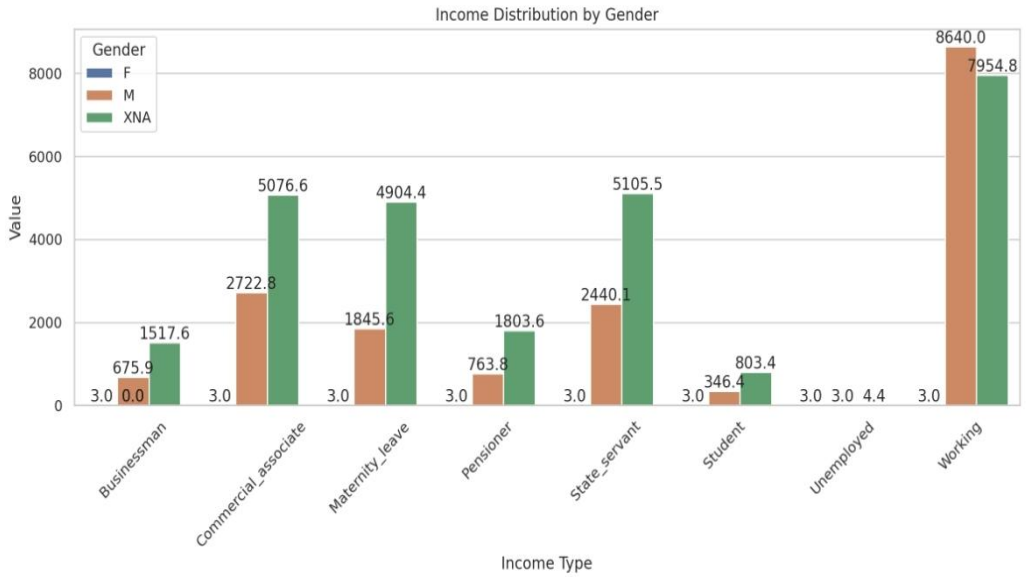
NAME_INCOME_TYPE segmented by NAME_EDUCATION_TYPE



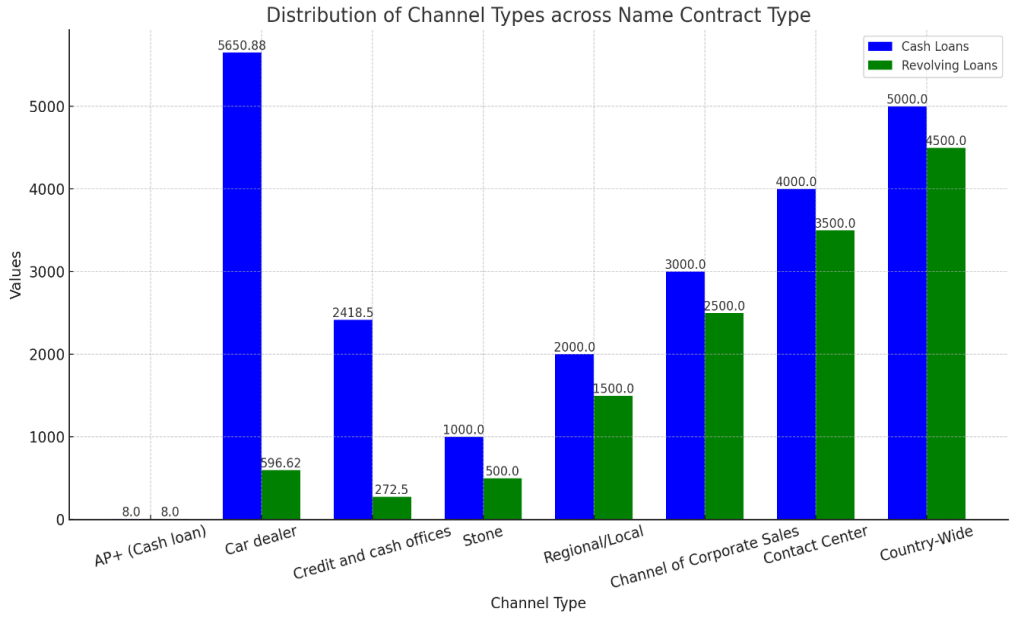
NAME_FAMILY_STATUS segmented by NAME_CLIENT_TYPE



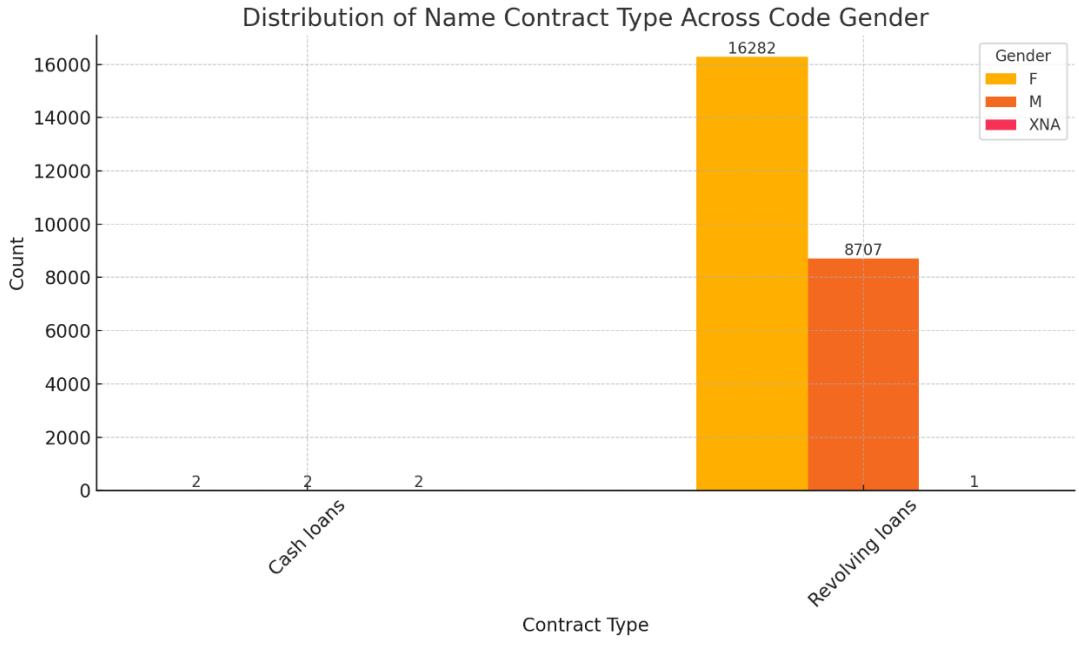
CODE_GENDER segmented by NAME_INCOME_TYPE



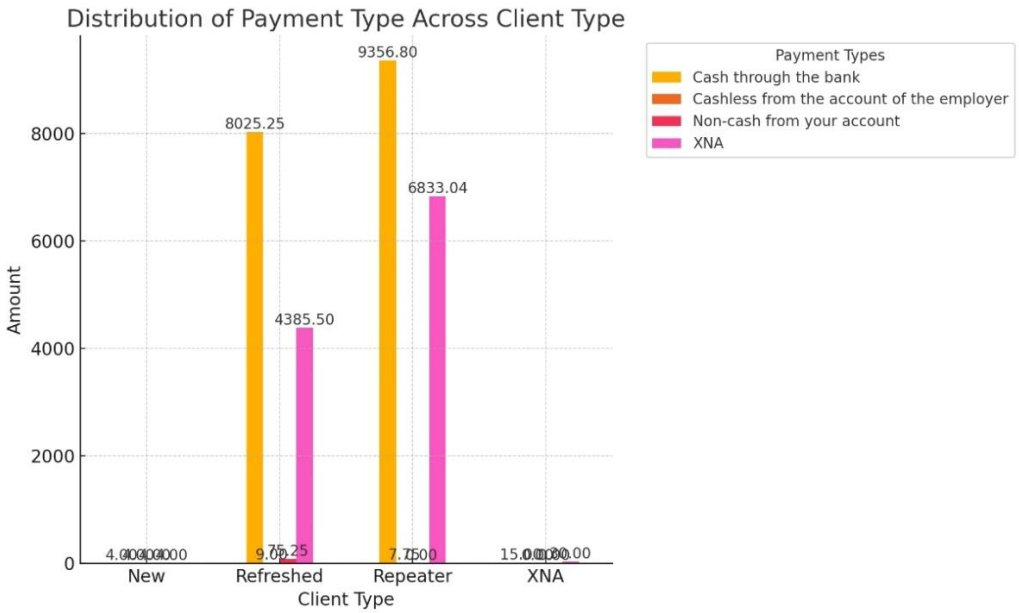
CHANNEL _TYPE segmented by NAME_CONTRACT_TYPE



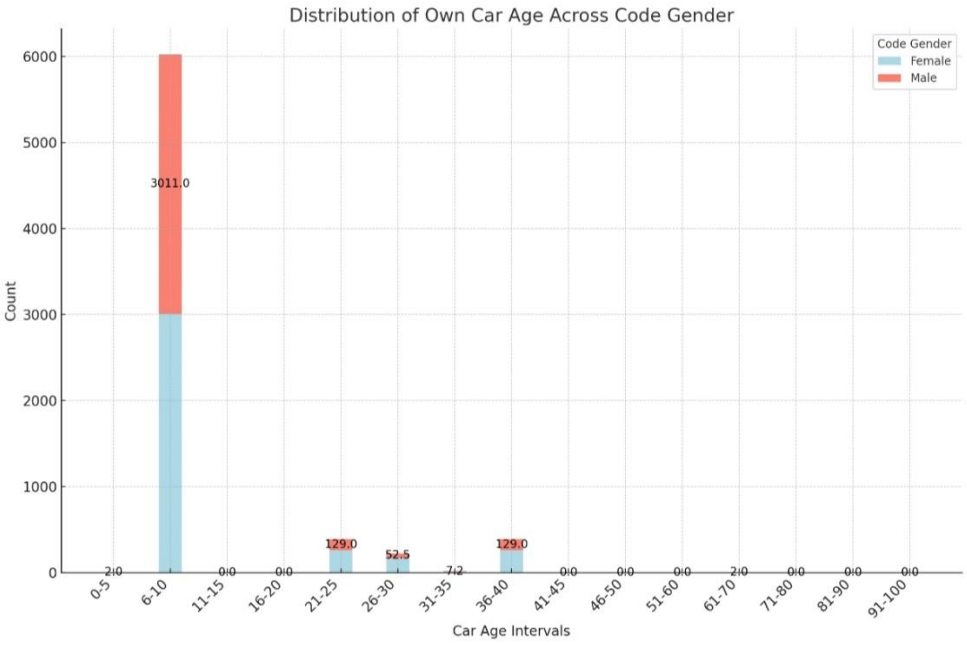
NAME_CONTRACT_TYPE segmented by CODE_GENDER



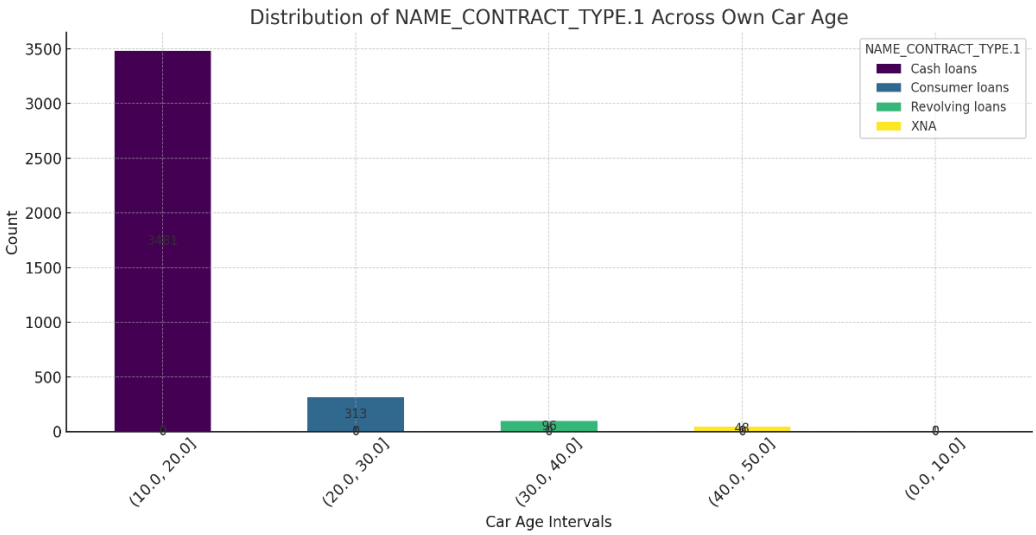
NAME_CLIENT_TYPE segmented by NAME_PAYMENT_TYPE



OWN_CAR_AGE segmented by CODE_GENDER



NAME_CONTRACT_TYPE.1 segmented by OWN_CAR_AGE



3. bivariate analysis

For Bivariate analysis the columns AMT_CREDIT , NAME_CONTRACT_TYPE .1, AMT_APPLICATION,

AMT_CREDIT 1, AMT_INCOME_TOTAL, TARGET, AMT_DOWN_PAYMENT are taken

	A	B	C	D	E	F	G
1	AMT_CREDIT	NAME_CONTRACT_TYPE .1	AMT_APPLICATION	AMT_CREDIT 1	AMT_INCOME_TOTAL	TARGET	AMT_DOWN_PAYMENT
2	17145	Consumer loans	17145	17145	202500	1	0
3	679671	Cash loans	607500	679671	270000	0	0
4	136444.5	Cash loans	112500	136444.5	67500	0	0
5	470790	Cash loans	450000	470790	135000	0	12649.5
6	404055	Cash loans	337500	404055	121500	0	1350
7	340573.5	Cash loans	315000	340573.5	99000	0	0
8	0	Cash loans	0	0	171000	0	9000
9	0	Cash loans	0	0	360000	0	0
10	0	Cash loans	0	0	112500	0	0
11	0	Cash loans	0	0	135000	0	0
12	335754	Cash loans	270000	335754	112500	0	0
13	246397.5	Cash loans	211500	246397.5	38419.155	0	13500
14	174361.5	Cash loans	148500	174361.5	67500	0	0
15	57564	Consumer loans	53779.5	57564	225000	0	4500
16	27252	Consumer loans	26550	27252	189000	0	0
17	119853	Consumer loans	126490.5	119853	157500	0	0
18	27297	Consumer loans	26955	27297	108000	0	7573.5
19	180000	Revolving loans	180000	180000	81000	0	9000

The rest of the Full Data of the columns chosen to perform bivariate analysis are in the given link below :

https://docs.google.com/spreadsheets/d/1cTCFZvgNtDc8gfoOWMKdS_ehlnsNCWKh/edit?usp=sharing&ouid=101204343036685814262&rtpof=true&sd=true

The Segments chosen are given below :

1) AMT_CREDIT segmented by AMT_CREDIT_1

2) APPLICATION_AMT segmented by TARGET

3) INCOME_TOTAL segmented by DOWN_PAYMENT

4) INCOME_TOTAL segmented by TARGET

5) CREDIT_AMT segmented by INCOME_TOTAL

6) DOWN_PAYMENT segmented by APPLICATION_AMT

7) CREDIT_AMT_1 segmented by DOWN_PAYMENT

8) CONTRACT_TYPE_1 segmented by TARGET

These segments are done using pivot table

	A	B
1		
2	amt_credit segmented by amt_credit_1	
3		
4		AMT_CREDIT
5	0	9435
6	6948	1
7	7879.5	1
8	8281.08	1
9	8649	1
10		
11	Application_amt segmented by target	
12		
13		AMT_APPLICATION
14	0	45973
15	1	4026

The Rest of the segments done using pivot tables are in the given link below :

https://docs.google.com/spreadsheets/d/1F0Fem3IY7msm2a3nPVK7IsPB7VBMjRG4/edit?usp=sharing&ouid=101204343036685814262&rtpof=true&sd=true

The Correlation Analysis is done to each of the segment's output found using pivot table

To find Correlation Coefficient I have used the CORREL() Function :

= CORREL(A5 : A9 , B5 : B9)

AMT_CREDIT SEGMENTED BY AMT_CREDIT_1

2	amt_credit segmented by amt_credit_1	
3		
4		AMT_CREDIT
5	0	9435
6	6948	1
7	7879.5	1
8	8281.08	1
9	8649	1

The Correlation between amt_credit_1 & amt_credit : -0.984

Indicating a negative correlation between the two variables suggesting that as the amt_credit_1 increases the amt_credit tends to decrease

APPLICATION_AMT SEGMENTED BY TARGET

11	Application_amt segmented by target	
12		
13		AMT_APPLICATION
14	0	45973
15	1	4026

The Correlation between Application_amt & target : -0.2955118903717489

Indicating a negative correlation between the two variables suggesting that as the amt of application increases the target variable tends to decrease

INCOME_TOTAL SEGMENTED BY DOWN_PAYMENT

17	Income_Total segmented by Down_payment	
18		
19		AMT_INCOME_TOTAL
20	0	11912
21	0.045	1
22	0.09	5
23	0.135	1
24	0.18	2

The Correlation between AMT_Income_Total & Down_payment : -0.707
Indicating a strong negative correlation between the two variables suggesting that as the down payment increases the total income tends to decrease

INCOME_TOTAL SEGMENTED BY TARGET

26	INCOME_TOTAL segmented by TARGET	
27		
28		AMT_INCOME_TOTAL
29	0	45973
30	1	4026

The Correlation between Income_Total & Target : - 0.9999999999999999

Indicating a negative correlation between the two variables suggesting that as the income_total decreases, the target value increases or vice versa

CREDIT_AMT SEGMENTED BY INCOME_TOTAL

32	CREDIT_AMT segmented by INCOME_TOTAL	
33		
34		AMT_CREDIT
35	25650	2
36	27000	9
37	28350	1
38	28575	1
39	28800	1

The Correlation between Credit_amt & Income_Total : -0.3992454827590575

Indicating a negative correlation between the two variables suggesting that as the income increases the credit amount tends to decrease

DOWN_PAYMENT SEGMENTED BY APPLICATION_AMT

41	DOWN_PAYMENT segmented by APPLICATION_AMT	
42		
43		AMT_DOWN_PAYMENT
44	0	5364
45	6120	0
46	6916.5	1
47	8281.08	0
48	9450	1

The Correlation between Down_payment & Application_Amt : -0.9376

Indicating a negative correlation between the two variables suggesting that as the application_amt increases the amt_down_payment tends to decrease and vice versa

CREDIT_AMT_1 SEGMENTED BY DOWN_PAYMENT

50	CREDIT_AMT_1 segmented by DOWN_PAYMENT	
51		
52		AMT_CREDIT 1
53	0	11912
54	0.045	1
55	0.09	5
56	0.135	1
57	0.18	2

The Correlation between Down_payment & Credit_amt_1 is -0.7071

Indicating a negative correlation between the two variables suggesting that as the Down_payment increases the credit_amt_1 tends to decrease and vice versa

CONTRACT_TYPE_1 SEGMENTED BY TARGET

50	CREDIT_AMT_1 segmented by DOWN_PAYMENT	
51		
52		AMT_CREDIT 1
53	0	11912
54	0.045	1
55	0.09	5
56	0.135	1
57	0.18	2

The Correlation between Contract_type_1 & Target is -1.0000

Indicating a negative correlation between the two variables suggesting that as the target increases, the contract_type_1 value decreases

The Descriptive Analysis is done to each one of columns taken

For Descriptive analysis -

TofindCount,Mean,Std,Minimum,Maximum,Median,Range,Mode,Percentage,Variance,quartile,percentile values These following functions are used :

1. COUNT FUNCTION:

=COUNT(A1:A50000)

2.MEAN FUNCTION:

=AVERAGE(A1: A50000)

3.STANDARD DEVIATION FUNCTION :

=STDEV.S(A1:A50000)

4.MINIMUM FUNCTION :

=MIN(A1:A50000)

5.MAXIMUM FUNCTION :

=MAX(A1:A50000)

6.RANGE FUNCTION :

=MAX(A1:A50000) – MIN(A1:A50000)

7.MODE FUNCTION :

=MODE(A1:A50000)

8.VARIANCE FUNCTION:

=VAR.S(A1: A50000)

9.QUARTILE FUNCTION:

Q1 = QUARTILE.INC(A1:A50000,1)

Q2 = QUARTILE.INC(A1:A50000,2)

Q3 = QUARTILE.INC(A1:A50000,3)

10.PERCENTILE FUNCTION:

1st Percentile:

=PERCENTILE.INC(A1:A50000, 1/100)

5th Percentile :

=PERCENTILE.INC(A1:A50000, 5/100)

10th Percentile:

=PERCENTILE.INC(A1:A50000, 10/100)

25th Percentile:

=PERCENTILE.INC(A1:A50000, 25/100)

50th Percentile :

=PERCENTILE.INC(A1:A50000, 50/100)

75th Percentile:

=PERCENTILE.INC(A1:A50000, 75/100)

90th Percentile :

=PERCENTILE.INC(A1:A50000, 90/100)

95th Percentile:

=PERCENTILE.INC(A1:A50000, 95/100)

99th Percentile:

=PERCENTILE.INC(A1:A50000, 99/100)

Percentage was also found for all columns

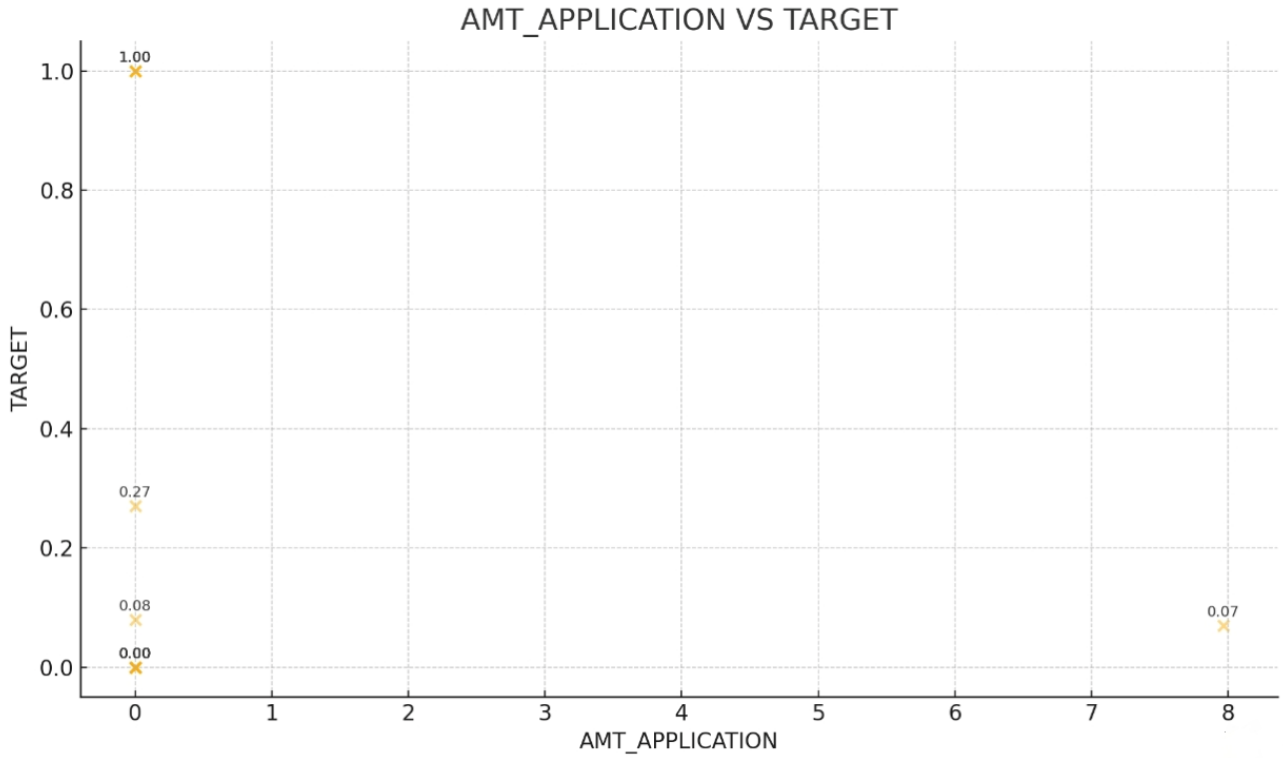
	A	B	C	D	E	F	G
1		AMT_CREDIT	AMT_APPLICATION	AMT_CREDIT 1	AMT_INCOME_TOTAL	TARGET	AMT_DOWN_PAYMENT
2	**count**	49,999	49,999	49,999	49,999	49,999	24,801
3	**mean**	188,542.89	168,892.45	188,542.89	170,767.59	0.08	6,557.57
4	**std**	308,470.52	282,200.69	308,473.60	531,819.10	0.27	17,444.58
5	**min**	0	0	0	25,650.00	0	0
6	**max**	4,104,351.00	3,826,372.50	4,104,351.00	117,000,000.00	1.00	1,035,000.00
7	**Median**	78,907.50	71,550	78,907.50	145,800.00	0.00	1,566.00
8	**Range**	4,104,351.00	3,826,372.50	4,104,351.00	116,974,350.00	1.00	1,035,000.00
9	**mode**	0.00	0.00	0.00	135,000.00	0.00	0.00
10	**percentage**	18.87%	78.12%	81.13%	100.00%	8.05%	25.78%
11	**variance**	95,154,059,586.69	79,637,228,515.08	95,155,962,744.02	282,831,549,942.21	0.07	304,313,400.40
12	**quantiles(0.25)**	26,055.00	22,045.50	26,055.00	112,500.00	0.00	0.00
13	**quantiles(0.5)**	78,907.50	71,550.00	78,907.50	145,800.00	0.00	1,566.00
14	**quantiles(0.75)**	198,126.00	180,000.00	198,105.75	202,500.00	0.00	7,875.00
15	**Percentile(1%)**	0.00	0.00	0.00	45,000.00	0.00	0.00
16	**Percentile(5%)**	0.00	0.00	0.00	67,500.00	0.00	0.00
17	**percentile(10%)**	0.00	0.00	0.00	81,000.00	0.00	0.00
18	***percentile(25%)***	26,055.00	22,045.50	26,055.00	112,500.00	0.00	0.00
19	**percentile(50%)**	78,907.50	71,550.00	78,907.50	145,800.00	0.00	1,566.00

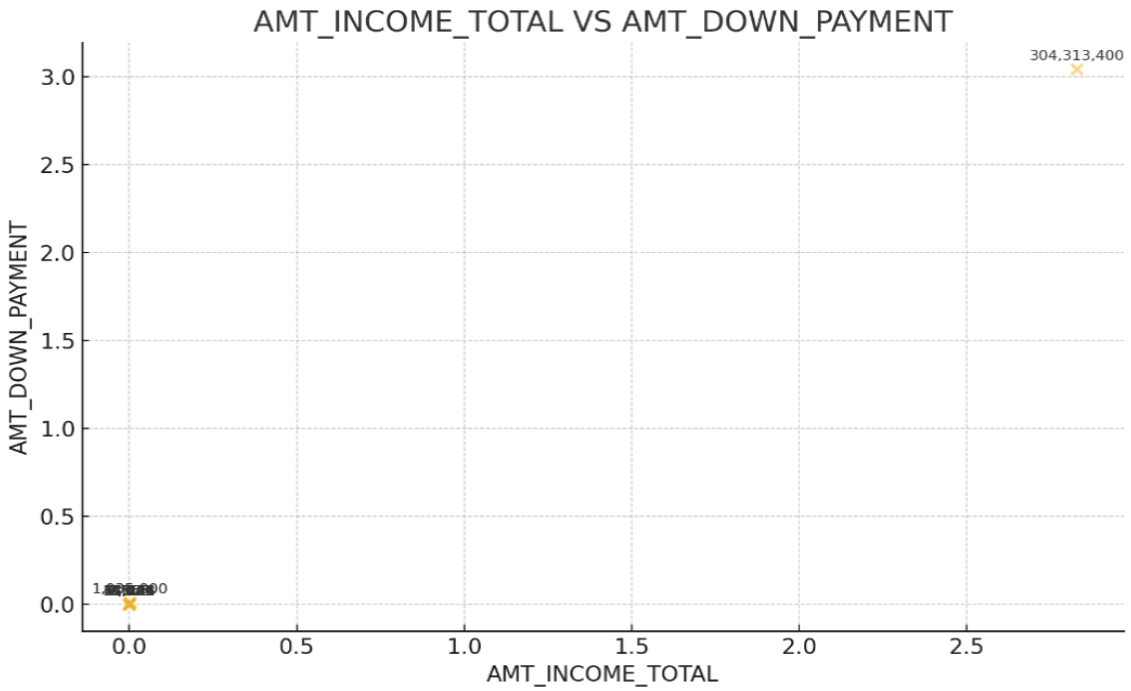
The descriptive analysis performed to the rest of the columns is given in the below link :

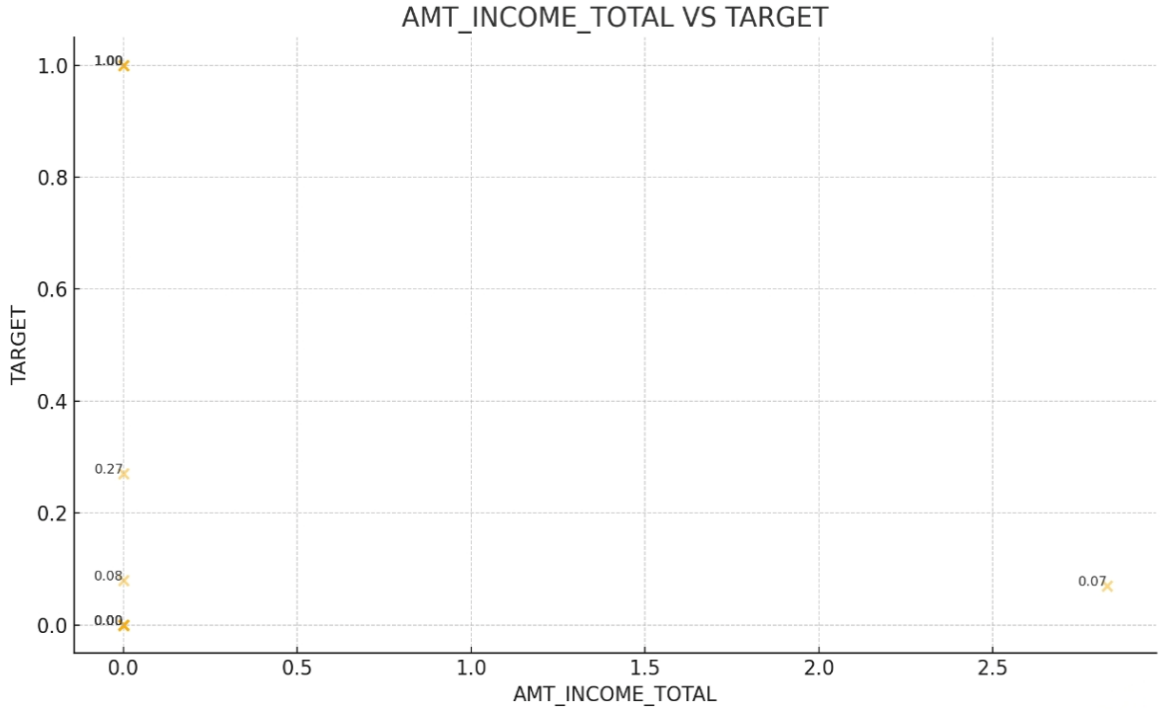
https://docs.google.com/spreadsheets/d/1QXr5Pp7rNYbdqPP3zGXIkr_lf9_Ddb0P/edit?usp=sharing&ouid=101204343036685814262&rtpof=true&sd=true

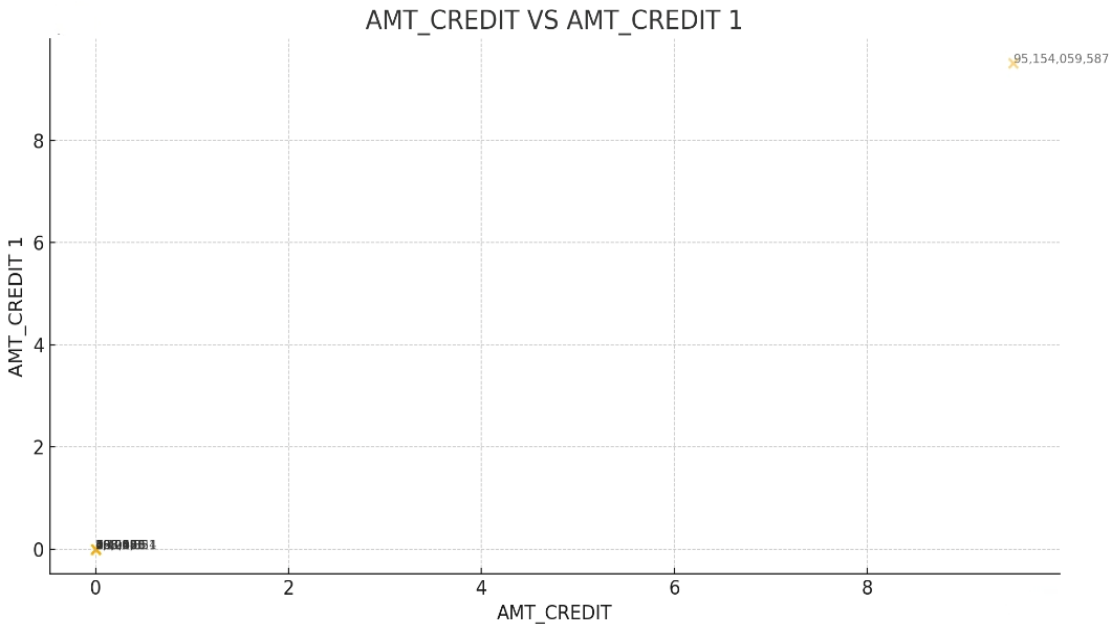
The scatter plots plotted to visualize the relationships between variables and target variable are given below

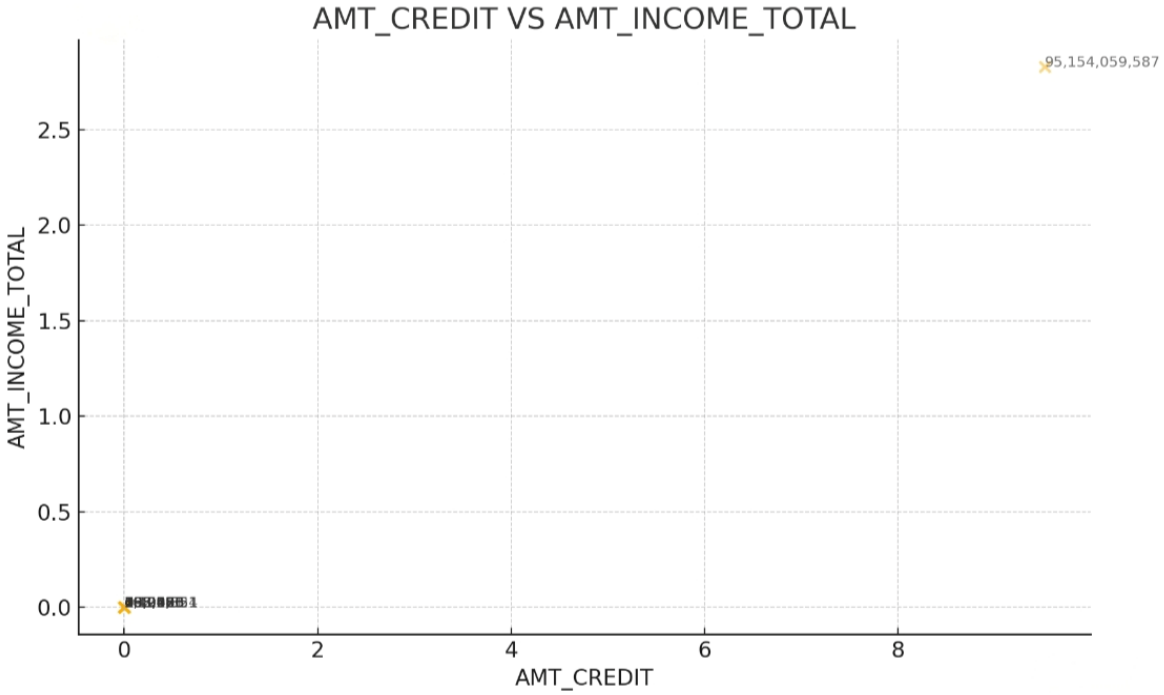
SCATTER PLOTS :

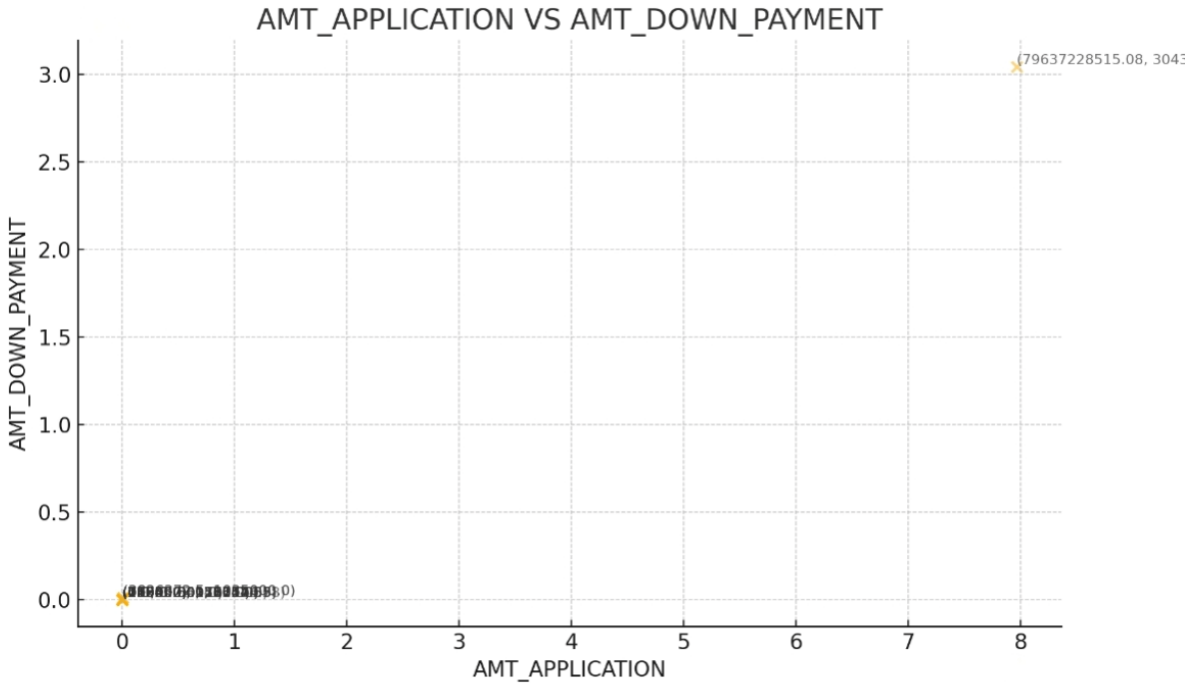


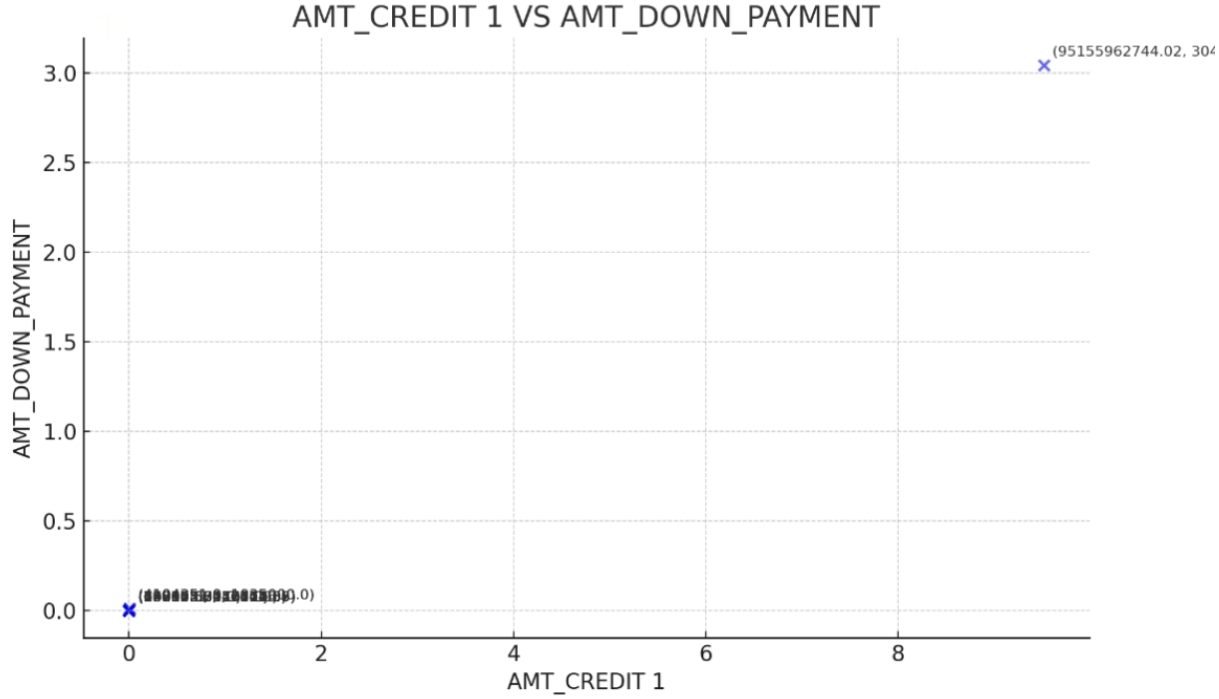

AMT_INCOME_TOTAL VS AMT_DOWN_PAYMENT
304,313,400
AMT_DOWN_PAYMENT
AMT_INCOME_TOTAL
0.0
0.5
1.0
1.5
2.0
2.5
3.0

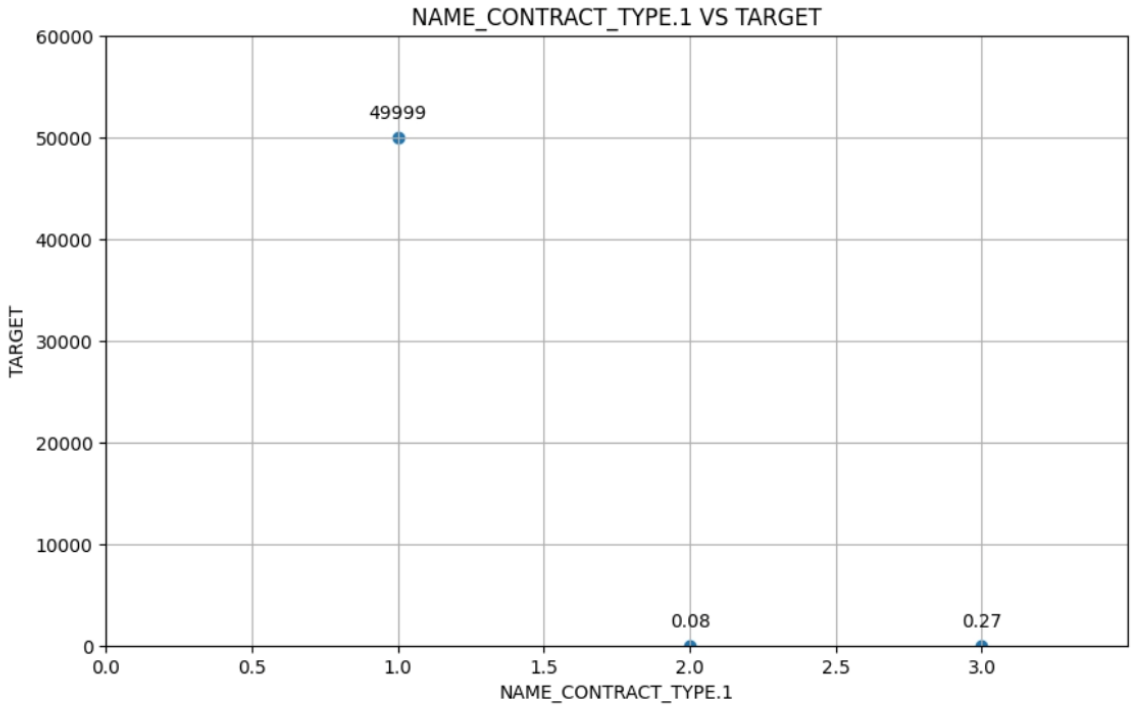

AMT_INCOME_TOTAL VS TARGET
1.00
0.27
0.08
0.00
0.07
TARGET
AMT_INCOME_TOTAL
0.0
0.2
0.4
0.6
0.8
1.0
0.5
1.5
2.0
2.5


AMT_CREDIT VS AMT_CREDIT 1
95,154,059,587
AMT_CREDIT 1
AMT_CREDIT
0
2
4
6
8


AMT_CREDIT VS AMT_INCOME_TOTAL
95,154,059,587
AMT_INCOME_TOTAL
AMT_CREDIT
0.0
0.5
1.0
1.5
2.0
2.5
0
2
4
6
8


AMT_APPLICATION VS AMT_DOWN_PAYMENT
AMT_DOWN_PAYMENT
AMT_APPLICATION
(79637228515.08, 3043
3.0
2.5
2.0
1.5
1.0
0.5
0.0
0
1
2
3
4
5
6
7
8


AMT_CREDIT 1 VS AMT_DOWN_PAYMENT
AMT_DOWN_PAYMENT
AMT_CREDIT 1
(95155962744.02, 304
3.0
2.5
2.0
1.5
1.0
0.5
0.0
0
2
4
6
8



E. **Identify Top Correlations for Different Scenarios:**

5. 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. Create correlation matrices or heatmaps to visualize the correlations between variables within each segment. Highlight the top correlated variables for each scenario using different colors or shading.

OUTPUT :

The segments found are :

1) CNT_CHILDREN

2) AMT_INCOME_TOTAL

3) AMT_CREDIT

4) REGION_POPULATION_RELATIVE

5) DAYS_BIRTH

6) DAYS_EMPLOYED

7) DAYS_ID_PUBLISH

8) REGION_RATING_CLIENT

9)AMT_GOODS_PRICE

10)AMT_REQ_CREDIT_BUREAU_YEAR

Including Target variable taken

	A	B	C	D	E	F	G	H
1	TARGET	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	REGION_POPULATION_RELATIVE	DAYS_BIRTH	DAYS_EMPLOYED	DAYS_ID_PUBLISH
2	1	0	202500	406597.5	0.018801	-9461	-637	-2120
3	1	0	270000	1293502.5	0.003541	-16765	-1188	-291
4	1	0	67500	135000	0.010032	-19046	-225	-2531
5	1	0	135000	312682.5	0.008019	-19005	-3039	-2437
6	1	0	121500	513000	0.028663	-19932	-3038	-3458
7	1	0	99000	490495.5	0.035792	-16941	-1588	-477
8	1	1	171000	1560726	0.035792	-13778	-3130	-619
9	1	0	360000	1530000	0.003122	-18850	-449	-2379
10	1	0	112500	1019610	0.018634	-20099	365243	-3514
11	1	0	135000	405000	0.019689	-14469	-2019	-3992
12	1	1	112500	652500	0.0228	-10197	-679	-738
13	1	0	38419.155	148365	0.015221	-20417	365243	-2512
14	1	0	67500	80865	0.031329	-13439	-2717	-3227
15	1	1	225000	918468	0.016612	-14086	-3028	-4911
16	1	0	189000	773680.5	0.010006	-14583	-203	-2056
17	1	0	157500	299772	0.020713	-8728	-1157	-1368
18	1	0	108000	509602.5	0.018634	-12931	-1317	-3866

The columns taken to find correlation coefficient including target variable (1) are given in the link below

https://docs.google.com/spreadsheets/d/15-QGp86lugTP0eNVSf_9fqLi1r9YWuQC/edit?usp=sharing&ouid=101204343036685814262&rtpof=true&sd=true

=correl(A2:A50000,B2:B50000)

The Correal Function is used between target variable (1) and the other variables to find the coefficients

TARGET (1)	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_GOODS_PRICE	REGION_POPULATION_RELATIVE	DAYS_BIRTH	DAYS_EMPLOYED	DAYS_ID_PUBLISH	REGION_RATING_CLIENT	AMT_REQ_CREDIT_BUREAU_YEAR
CNT_CHILDREN	1	0.009588558	0.00497156	0.002085777	-0.025555665	0.329263754	-0.239693041	-0.032115773	0.025913889	0.005882652
AMT_INCOME_TOTAL	0.009588558	1	0.069315897	-0.000190769	0.029841469	0.016002774	-0.031615555	0.003506646	-0.038188511	-0.005089418
AMT_CREDIT	0.00497156	0.069315897	1	0.013258331	0.095111221	-0.053942658	-0.070471393	-0.012228765	-0.100507425	0.003898937
AMT_GOODS_PRICE	0.002085777	-0.000190769	0.013258331	1	0.00576511	-0.003153225	-0.00424617	-0.00093887	-0.00141928	-0.004311374
REGION_POPULATION_RELATIVE	-0.025555665	0.029841469	0.095111221	0.00576511	1	-0.032513748	-0.004101686	-0.004345136	-0.532667302	0.002975614
DAYS_BIRTH	0.329263754	0.016002774	-0.053942658	-0.003153225	-0.032513748	1	-0.613553972	0.270825141	0.016779196	0.00771461
DAYS_EMPLOYED	-0.239693041	-0.031615555	-0.070471393	-0.00424617	-0.004101686	-0.613553972	1	-0.270382022	0.034321673	-0.006483373
DAYS_ID_PUBLISH	-0.032115773	0.003506646	-0.012228765	-0.00093887	-0.004345136	0.270825141	-0.270382022	1	-0.002307011	0.012634162
REGION_RATING_CLIENT	0.025913889	-0.038188511	-0.100507425	-0.00141928	-0.532667302	0.016779196	0.034321673	-0.002307011	1	0.004033946
AMT_REQ_CREDIT_BUREAU_YEAR	0.005882652	-0.005089418	0.003898937	-0.004311374	0.002975614	0.00771461	-0.006483373	0.012634162	0.004033946	1

The Full data of coefficients found are in the link below :
https://docs.google.com/spreadsheets/d/1mkxx0A5Vz-E0xnF9Yu6W-utq0qtGXUGj/edit?usp=sharing&ouid=101204343036685814262&rtpof=true&sd=true

The conditional formatting rules are used to highlight Correlations

Conditional Formatting Rules

Maximum = 1

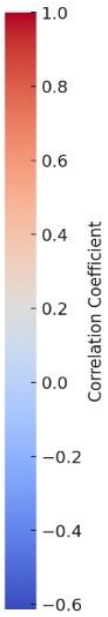
Minimum = - 1

Midpoint = 0

THE CORRELATION MATRICES FOR TARGET VARIABLE (1) done using conditional formatting

TARGET (1)	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_GOODS_PRICE	REGION_POPULATION_RELATIVE	DAYS_BIRTH	DAYS_EMPLOYED	DAYS_ID_PUBLISH	REGION_RATING_CLIENT	AMT_REQ_CREDIT_BUREAU_YEAR
CNT_CHILDREN	1	0.009588558	0.00497156	0.002085777	-0.025555665	0.329263754	-0.239693041	-0.032115773	0.025913889	0.005882852
AMT_INCOME_TOTAL	0.009588558	1	0.069315897	-0.000190769	0.029841469	0.016002774	-0.031615555	0.003506646	-0.038188511	-0.005089418
AMT_CREDIT	0.00497156	0.069315897	1	0.013258331	0.095111221	-0.059342658	-0.070471393	-0.012228765	-0.100507425	0.003898937
AMT_GOODS_PRICE	0.002085777	-0.000190769	0.013258331	1	0.00576511	-0.003153225	-0.00424617	-0.00093887	-0.00141928	-0.004311374
REGION_POPULATION_RELATIVE	-0.025555665	0.029841469	0.095111221	0.00576511	1	-0.032513748	-0.004101686	-0.004345136	-0.532667302	0.002975614
DAYS_BIRTH	0.329263754	0.016002774	-0.059342658	-0.003153225	-0.032513748	1	-0.613553972	0.270825141	0.016779196	0.00771461
DAYS_EMPLOYED	-0.239693041	-0.031615555	-0.070471393	-0.00424617	-0.004101686	-0.613553972	1	-0.270382022	0.034321673	-0.006483373
DAYS_ID_PUBLISH	-0.032115773	0.003506646	-0.012228765	-0.00093887	-0.004345136	0.270825141	-0.270382022	1	-0.002307011	0.012634162
REGION_RATING_CLIENT	0.025913889	-0.038188511	-0.100507425	-0.00141928	-0.532667302	0.016779196	0.034321673	-0.002307011	1	0.004039346
AMT_REQ_CREDIT_BUREAU_YEAR	0.005882852	-0.005089418	0.003898937	-0.004311374	0.002975614	0.00771461	-0.006483373	0.012634162	0.004039346	1

LEGEND :



Color	Explanation
RED	Negative correlation
BLUE	positive correlation
WHITE	No correlation

Explanation of Color Coding:

1.Blue:

- Explanation: positive correlation.
- Example: As one variable increases, the other also increases strongly.

2. White:
- Explanation: No correlation.
- Example: No linear relationship between the variables.

3. Red:
- Explanation: negative correlation.
- Example: As one variable increases, the other shows a minimal decrease.

The Clear Correlation Matrices for Target variable (1) is in the given link below :
https://docs.google.com/spreadsheets/d/1vayEb52rByP9-xnlNrS1Qzc8fsZ-YD6C/edit?usp=sharing&ouid=101204343036685814262&rtpof=true&sd=true

The segments found are :

1) CNT_CHILDREN

2) AMT_INCOME_TOTAL

3) AMT_CREDIT

4) REGION_POPULATION_RELATIVE

5) DAYS_BIRTH

6) DAYS_EMPLOYED

7) DAYS_ID_PUBLISH

8) REGION_RATING_CLIENT

9)AMT_GOODS_PRICE

10)AMT_REQ_CREDIT_BUREAU_YEAR

Including Target variable taken

	A	B	C	D	E	F	G	H
1	TARGET	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	REGION_POPULATION_RELATIVE	DAYS_BIRTH	DAYS_EMPLOYED	DAYS_ID_PUBLISH
2	0	0	202500	406597.5	0.018801	-9461	-637	-2120
3	0	0	270000	1293502.5	0.003541	-16765	-1188	-291
4	0	0	67500	135000	0.010032	-19046	-225	-2531
5	0	0	135000	312682.5	0.008019	-19005	-3039	-2437
6	0	0	121500	513000	0.028663	-19932	-3038	-3458
7	0	0	99000	490495.5	0.035792	-16941	-1588	-477
8	0	1	171000	1560726	0.035792	-13778	-3130	-619
9	0	0	360000	1530000	0.003122	-18850	-449	-2379
10	0	0	112500	1019610	0.018634	-20099	365243	-3514
11	0	0	135000	405000	0.019689	-14469	-2019	-3992
12	0	1	112500	652500	0.0228	-10197	-679	-738
13	0	0	38419.155	148365	0.015221	-20417	365243	-2512
14	0	0	67500	80865	0.031329	-13439	-2717	-3227
15	0	1	225000	918468	0.016612	-14086	-3028	-4911
16	0	0	189000	773680.5	0.010006	-14583	-203	-2056
17	0	0	157500	299772	0.020713	-8728	-1157	-1368

The columns taken to find correlation coefficient including target variable (0) are given in the link below

https://docs.google.com/spreadsheets/d/1pilMPVgwnAurs6V9GU8HMX67tpJCjaQy/edit?usp=sharing&ouid=101204343036685814262&rtpof=true&sd=true

=correl(A2:A50000,B2:B50000)

The Correal Function is used between target variable (0) and the other variables to find the coefficients

61	TARGET VARIABLE (0)	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	ION_POPULATION_RELA	DAYS_BIRTH	DAYS_EMPLOYED	DAYS_ID_PUBLISH	EGION_RATING_CLIE	AMT_GOODS_PRICE	AMT_REQ_CREDIT_BUREAU_YEAR
62	CNT_CHILDREN	1	0.009389816	0.006515123	-0.026689704	0.328293814	-0.239627071	-0.032268389	0.025908972	0.004130779	-0.001437756
63	AMT_INCOME_TOTAL	0.009389816	1	0.066433368	0.028425341	0.01586795	-0.03052779	0.003309333	-0.03657852	-0.000172738	0.000185516
64	AMT_CREDIT	0.006515123	0.066433368	1	0.095127525	-0.058641286	-0.070718891	-0.011346152	-0.099112917	0.011234511	-0.001744445
65	REGION_POPULATION_RELATIVE	-0.026689704	0.028425341	0.095127525	1	-0.030369054	-0.005920011	-0.002298217	-0.534389649	0.003609725	0.003684573
66	DAYS_BIRTH	0.328293814	0.01586795	-0.058641286	-0.030369054	1	-0.61451668	0.271792027	0.017541173	-0.002299615	-0.000614776
67	DAYS_EMPLOYED	-0.239627071	-0.03052779	-0.070718891	-0.005920011	-0.61451668	1	-0.270558903	0.034195347	-0.006196234	-0.00744605
68	DAYS_ID_PUBLISH	-0.032268389	0.003309333	-0.011346152	-0.002298217	0.271792027	-0.270558903	1	-0.003836517	0.001301719	0.004114932
69	REGION_RATING_CLIENT	0.025908972	-0.03657852	-0.099112917	-0.534389649	0.017541173	0.034195347	-0.003836517	1	0.003321731	-0.002785354
70	AMT_GOODS_PRICE	0.004130779	-0.000172738	0.011234511	0.003609725	-0.002299615	-0.006196234	0.001301719	0.003321731	1	0.00371245
71	AMT_REQ_CREDIT_BUREAU_YEAR	-0.001437756	0.000185516	-0.001744445	0.003684573	-0.000614776	-0.00744605	0.004114932	-0.002785354	0.00371245	1

The Full data of coefficients found are in the link below :

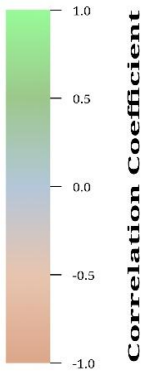
https://docs.google.com/spreadsheets/d/1YmOyXEuEbEVg40XHfrVaRl2wqXoAc6V9/edit?usp=sharing&ouid=101204343036685814262&rtpof=true&sd=true

The conditional formatting rules are used to highlight Correlations
Conditional Formatting Rules
Maximum = 1
Minimum = - 1
Midpoint = 0

THE CORRELATION MATRICES FOR TARGET VARIABLE (0) done using conditional formatting

TARGET VARIABLE (0)	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	REGION_POPULATION_RELATIVE	DAYS_BIRTH	DAYS_EMPLOYED	DAYS_ID_PUBLISH	REGION_RATING_CLIENT	AMT_GOODS_PRICE	AMT_REQ_CREDIT_BUREAU_YEAR
CNT_CHILDREN	1	0.009389816	0.006515123	-0.026689704	0.328293814	-0.239627071	-0.032268389	0.025908972	0.004130779	-0.001437756
AMT_INCOME_TOTAL	0.009389816	1	0.066433368	0.028425341	0.01586795	-0.03052779	0.003309333	-0.03657852	-0.000172738	0.000185516
AMT_CREDIT	0.006515123	0.066433368	1	0.095127525	-0.058641286	-0.070718891	-0.011346152	-0.099112917	0.011234511	-0.001744445
REGION_POPULATION_RELATIVE	-0.026689704	0.028425341	0.095127525	1	-0.030369054	-0.005920011	-0.002298217	-0.534389649	0.003609725	0.003684573
DAYS_BIRTH	0.328293814	0.01586795	-0.058641286	-0.030369054	1	-0.61451668	0.271792027	0.017541173	-0.002299615	-0.000614776
DAYS_EMPLOYED	-0.239627071	-0.03052779	-0.070718891	-0.005920011	-0.61451668	1	-0.270558903	0.034195347	-0.006196234	-0.00744605
DAYS_ID_PUBLISH	-0.032268389	0.003309333	-0.011346152	-0.002298217	0.271792027	-0.270558903	1	-0.003836517	0.001301719	0.004114932
REGION_RATING_CLIENT	0.025908972	-0.03657852	-0.099112917	-0.534389649	0.017541173	0.034195347	-0.003836517	1	0.003321731	-0.002785354
AMT_GOODS_PRICE	0.004130779	-0.000172738	0.011234511	0.003609725	-0.002299615	-0.006196234	0.001301719	0.003321731	1	0.00371245
AMT_REQ_CREDIT_BUREAU_YEAR	-0.001437756	0.000185516	-0.001744445	0.003684573	-0.000614776	-0.00744605	0.004114932	-0.002785354	0.00371245	1

LEGEND :



Color	Explanation
Brown	No correlation
Green	Positive correlation
Blue	negative correlation

Explanation of Color Coding:

1.Green :
- Explanation: positive correlation.
- Example: As one variable increases, the other also increases strongly.

2.Brown :
- Explanation: No correlation.
- Example: No linear relationship between the variables.

3. Blue :
- Explanation: negative correlation.
- Example: As one variable increases, the other shows a minimal decrease.

The Clear Correlation Matrices for Target variable (0) is in the given link below :

https://docs.google.com/spreadsheets/d/1fT7Ao_s8Oj6tvAdj3JQhGi4pxx_Ebz7q/edit?usp=sharing&ouid=101204343036685814262&rtpof=true&sd=true