Bank Loan Case Study

Project Description: In this project I was working as a data analyst at a finance company that specializes in lending various types of loans to urban customers. My company faces a challenge: some customers who don't have a sufficient credit history take advantage of this and default on their loans. My task is to use Exploratory Data Analysis (EDA) to analyze patterns in the data and ensure that capable applicants are not rejected.

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

1. **Identify Missing Data and Deal with it Appropriately:** As a data analyst, you come across missing data in the loan application dataset. It is essential to handle missing data effectively to ensure the accuracy of the analysis.
   My Task: Identify the missing data in the dataset and decide on an appropriate method to deal with it using Excel built-in functions and features.

2. **Identify Outliers in the Dataset:** Outliers can significantly impact the analysis and distort the results. You need to identify outliers in the loan application dataset.
   My Task: Detect and identify outliers in the dataset using Excel statistical functions and features, focusing on numerical variables.

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

4. **Perform Univariate, Segmented Univariate, and Bivariate Analysis:** To gain insights into the driving factors of loan default, it is important to conduct various analyses on consumer and loan attributes.
   My Task: Perform univariate analysis to understand the distribution of individual variables, segmented univariate analysis to compare variable distributions for different scenarios, and bivariate analysis to explore relationships between variables and the target variable using Excel functions and features.

5. **Identify Top Correlations for Different Scenarios:** Understanding the correlation between variables and the target variable can provide insights into strong indicators of loan default.
   My Task: Segment the dataset based on different scenarios (e.g., clients with payment difficulties and all other cases) and identify the top correlations for each segmented data using Excel functions.

Approach: First I gone through dataset to know all the columns present in the table. Then I saw all the questions and thought of functions which could be used to answer each question. After that I applied those functions and found the answer to each question and plotted the graph wherever was required.

Tech-Stack Used: The software used for the project is Microsoft Excel 365. It is used to run the functions and get answers of each question. It is also used to plot the graphs.

Insights:

1. Identify Missing Data and Deal with it Appropriately:

   Function:-

   First, I found percentage of blank data in each column. I deleted all columns in which percentage of blank data was more than 25%.
   Then I filled blank cells as follow:
   =IF(ISBLANK(application_data!K2),MEDIAN(application_data!K$2:K$50000),application_data!K2)

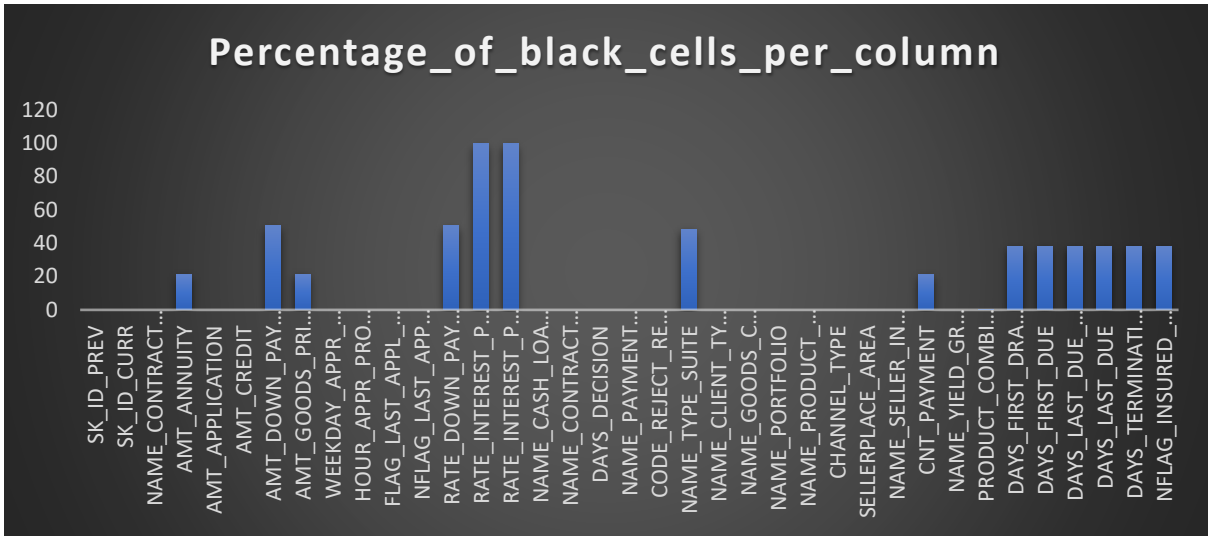
   Output:-

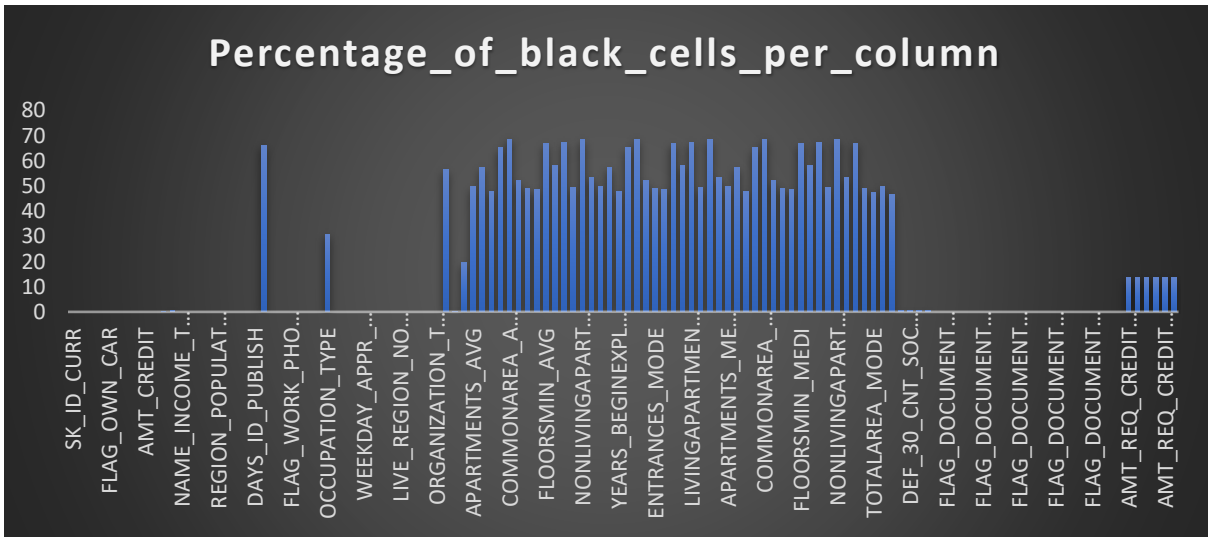
   https://docs.google.com/spreadsheets/d/1OZwBS2eC2zznkfd8_TQyGbaM46sVeKNZ/edit?usp=sharing&ouid=106942457558004201317&rtpof=true&sd=true

Column	Percentage_of_black_cells
SK_ID_PREV	0
SK_ID_CURR	0
NAME_CONTRACT_TYPE	0
AMT_ANNUITY	21.18442369
AMT_APPLICATION	0
AMT_CREDIT	0
AMT_DOWN_PAYMENT	50.39700794
AMT_GOODS_PRICE	21.48842977
WEEKDAY_APPR_PROCESS_START	0
HOUR_APPR_PROCESS_START	0
FLAG_LAST_APPL_PER_CONTRACT	0
NFLAG_LAST_APPL_IN_DAY	0
RATE_DOWN_PAYMENT	50.39700794
RATE_INTEREST_PRIMARY	99.6699934
RATE_INTEREST_PRIVILEGED	99.6699934
NAME_CASH_LOAN_PURPOSE	0
NAME_CONTRACT_STATUS	0
DAYS_DECISION	0
NAME_PAYMENT_TYPE	0
CODE_REJECT_REASON	0
NAME_TYPE_SUITE	48.48696974
NAME_CLIENT_TYPE	0
NAME_GOODS_CATEGORY	0
NAME_PORTFOLIO	0
NAME_PRODUCT_TYPE	0

CHANNEL_TYPE	0
SELLERPLACE_AREA	0
NAME_SELLER_INDUSTRY	0
CNT_PAYMENT	21.18442369
NAME_YIELD_GROUP	0
PRODUCT_COMBINATION	0.01600032
DAYS_FIRST_DRAWING	38.32076642
DAYS_FIRST_DUE	38.32076642
DAYS_LAST_DUE_1ST_VERSION	38.32076642
DAYS_LAST_DUE	38.32076642
DAYS_TERMINATION	38.32076642
NFLAG_INSURED_ON_APPROVAL	38.32076642

Graph:





2. Identify Outliers in the Dataset:

   Function:-

   =QUARTILE.EXC(H2:H50000,1)
   =QUARTILE.EXC(H2:H50000,3)

   Then I used conditional formatting to highlight the values which are less than lower bound and higher than upper bound.

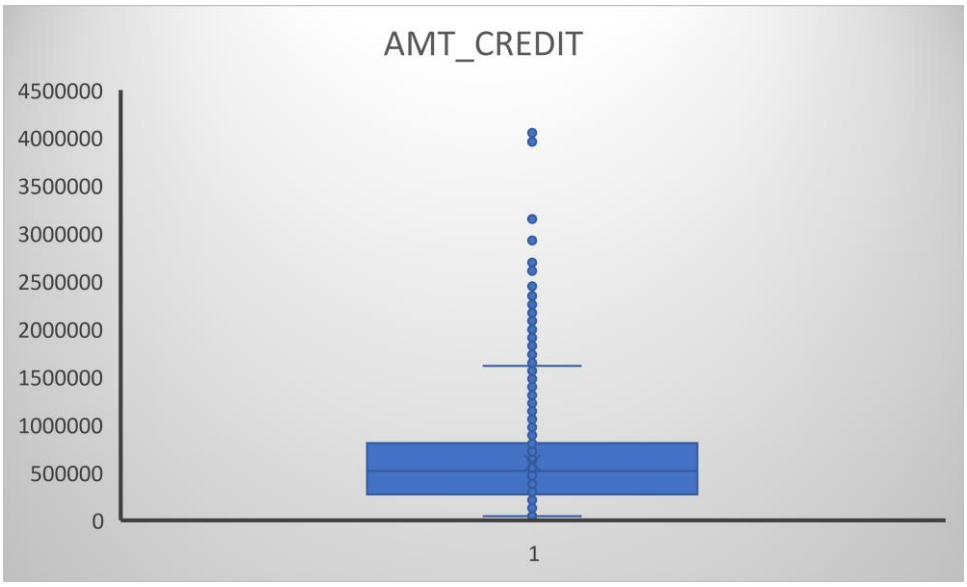
   Output:-

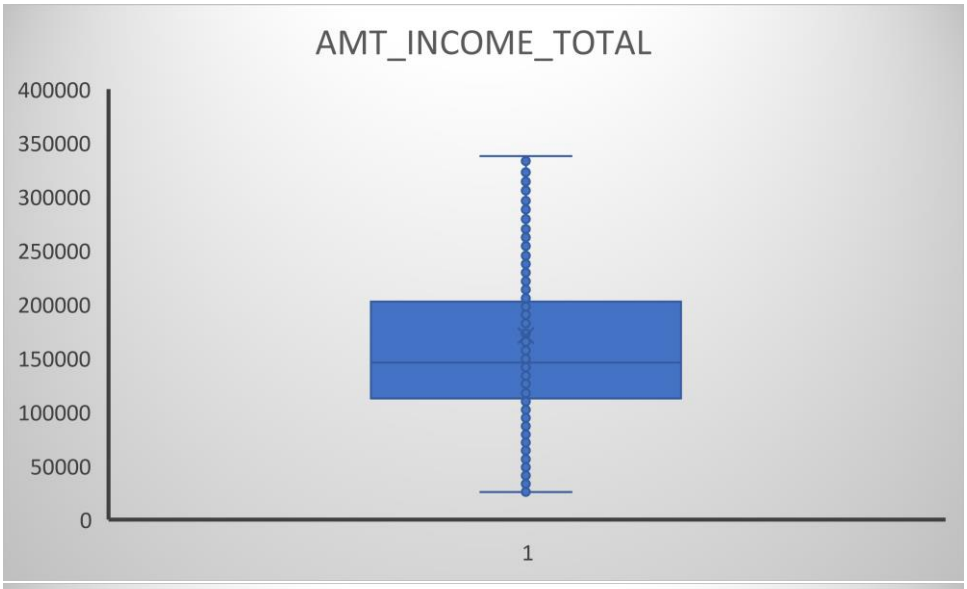
   https://docs.google.com/spreadsheets/d/1JYdPdrTVUm-1J8ab20tFfIjHCyqwYMkv/edit?usp=sharing&ouid=106942457558004201317&rtpof=true&sd=true

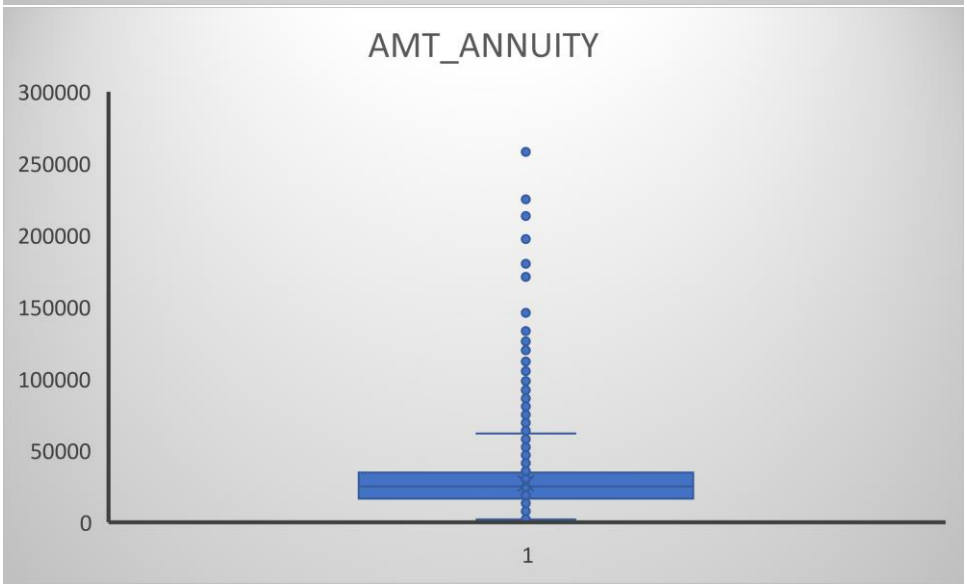
	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY
lower bound	112500	270000	16456.5
upper bound	202500	808650	34596

AMT_INCOME_TOTAL	AMT_CREDIT	A
202500	406597.5	
270000	1293502.5	
67500	135000	
135000	312682.5	
121500	513000	
99000	490495.5	
171000	1560726	
360000	1530000	
112500	1019610	
135000	405000	
112500	652500	
38419.155	148365	
67500	80865	
225000	918468	
189000	773680.5	
157500	299772	
108000	509602.5	
81000	270000	
112500	157500	
90000	544491	
135000	427500	
202500	1132573.5	
450000	497520	
83250	239850	
135000	247500	

Graph:







3. Analyze Data Imbalance:

   Function:-

   First, I found the age of people using the following function:
   =ROUND(R2:R50000/365*-1,0)

   Then I found occurrences of each unique elements in different column.
   =UNIQUE(C2:C50000)
   =COUNTIF(C$2:C$50000,BX2)

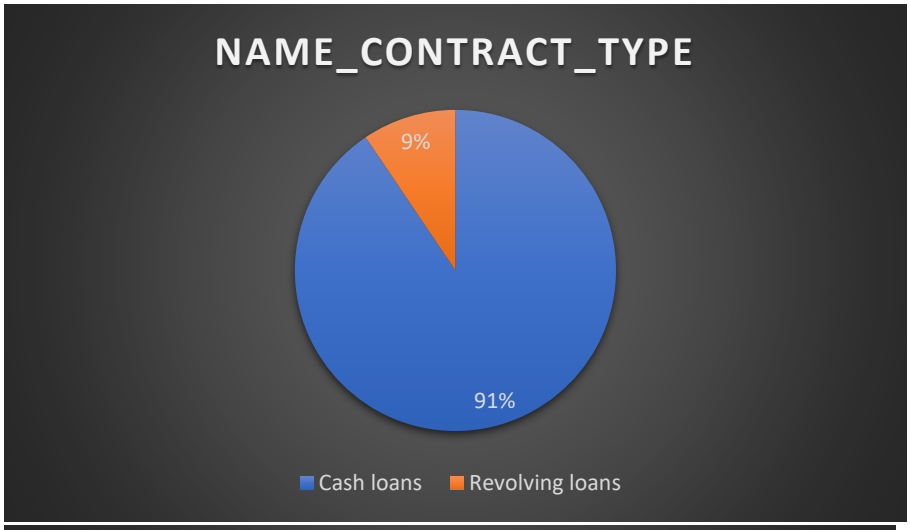
   After that I plotted the graphs.

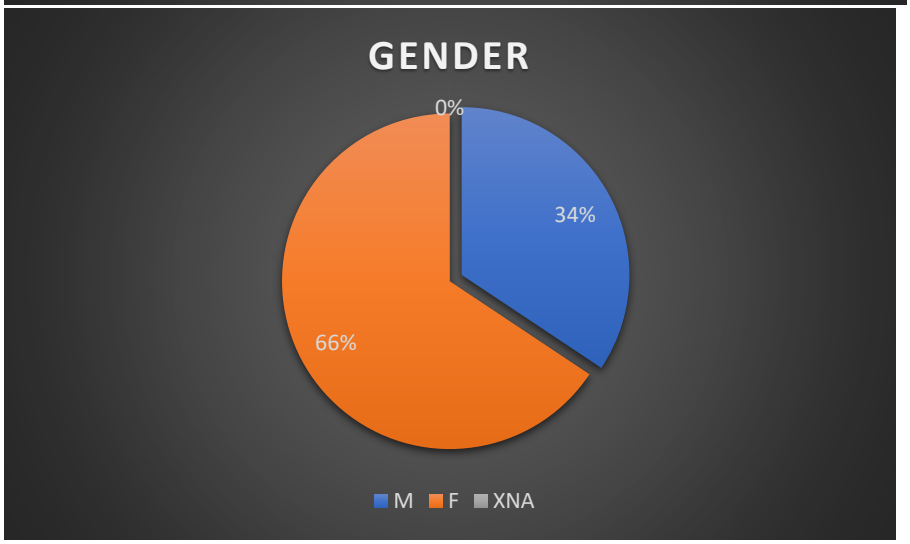
   Output:-

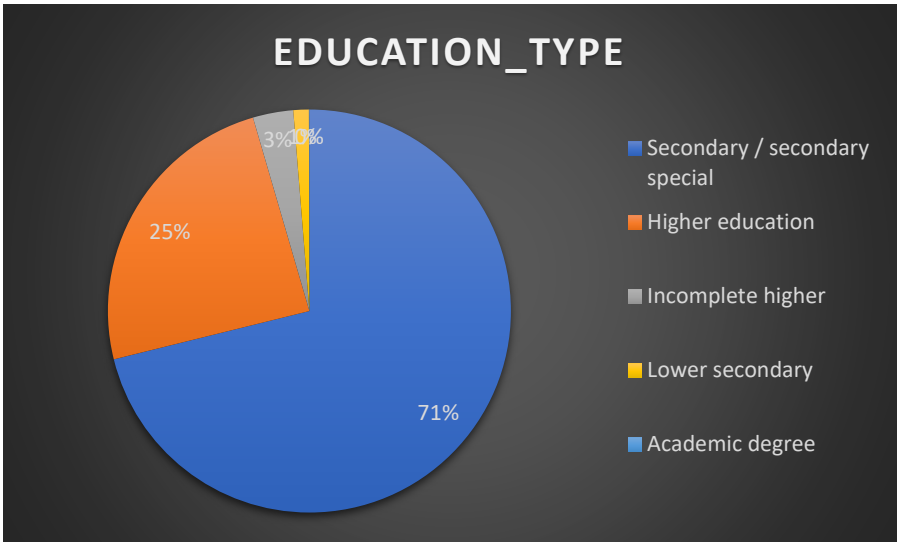
https://docs.google.com/spreadsheets/d/1qaTJqUwWq5NDogJFvwFH9y_-6ote5GFk/edit?usp=sharing&ouid=106942457558004201317&rtpof=true&sd=true

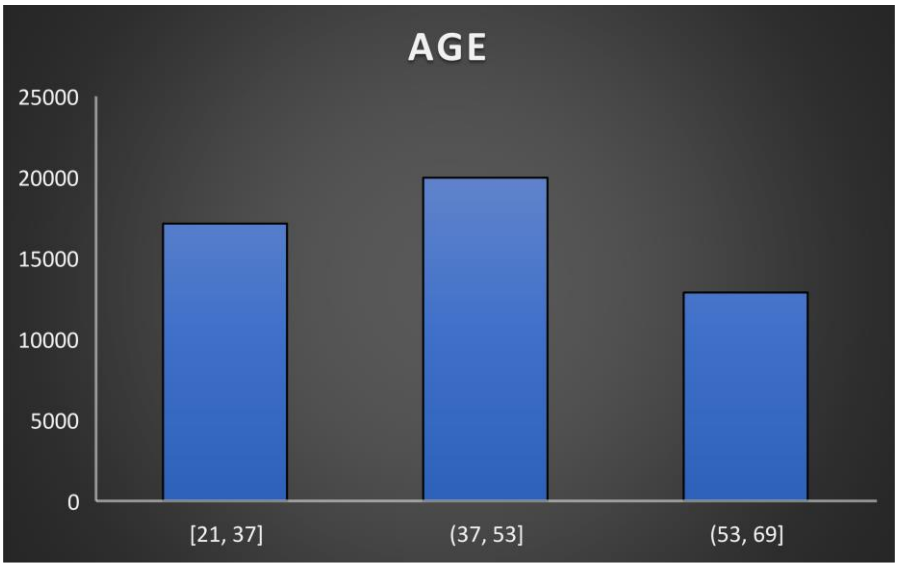
NAME_CONTRACT_TYPE	Occurrence	GENDER	Occurrence	EDUCATION_TYPE	Occurrence
Cash loans	45276	M	17174	Secondary / secondary special	35572
Revolving loans	4723	F	32823	Higher education	12167
		XNA	2	Incomplete higher	1620
				Lower secondary	620
				Academic degree	20

Graph:









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

   Function:-

   =AVERAGE(H2:H50000)
   =MEDIAN(H2:H50000)
   =MODE(H2:H50000)

   Then I found unique values for HOUSING_TYPE, FAMILY_STATUS, INCOME_TYPE and occurrences of each unique value.

=UNIQUE(P2:P50000)
=COUNTIF(P$2:P$50000,CD2)

Then I plotted graphs for different variables

Output:-

https://docs.google.com/spreadsheets/d/1zK1mjbyOD_t4vjcyNqw5a7TYdXmONp7k/edit?usp=sharing&ouid=106942457558004201317&rtpof=true&sd=true

	AMT_INCOME_TOTAL	AMT_CREDIT	AGE
AVERAGE	170767.5905	599700.5815	43.8975
MEDIAN	145800	514777.5	43
MODE	135000	450000	39

HOUSING_TYPE	OCCURRENCE	TARGET_0	TARGET_1
House / apartment	44368	40895	3473
Rented apartment	769	682	87
With parents	2399	2122	277
Municipal apartment	1845	1700	145
Office apartment	427	398	29
Co-op apartment	191	176	15

FAMILY_STATUS	OCCURRENCE	TARGET_0	TARGET_1
Single / not marrie	7306	6577	729
Married	32094	29699	2395
Civil marriage	4859	4377	482
Widow	2597	2449	148
Separated	3142	2870	272
Unknown	1	1	0

INCOME_TYPE	OCCURRENCE	TARGET_0	TARGET_1
Working	26010	23549	2461
State servant	3512	3314	198
Commercial asso	11543	10679	864
Pensioner	8920	8419	501
Unemployed	6	4	2
Student	5	5	0
Businessman	2	2	0
Maternity leave	1	1	0

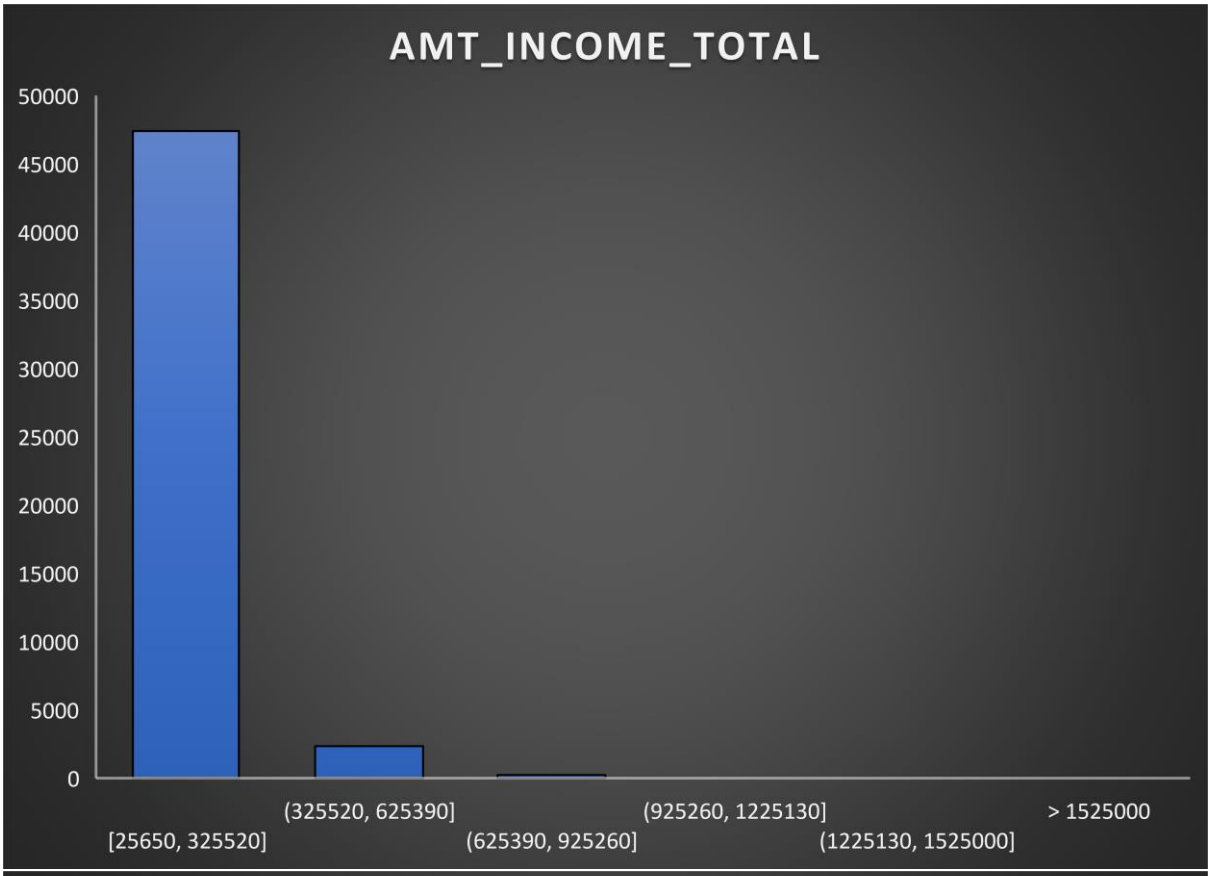
NAME_CLIENT_TYPE	Approvec	Refused	Canceled	Unused offer	
Repeater	19814	7761	7924	668	
New	8944	407	115	82	
Refreshed	3110	480	529	108	
XNA	17	12	27	1	

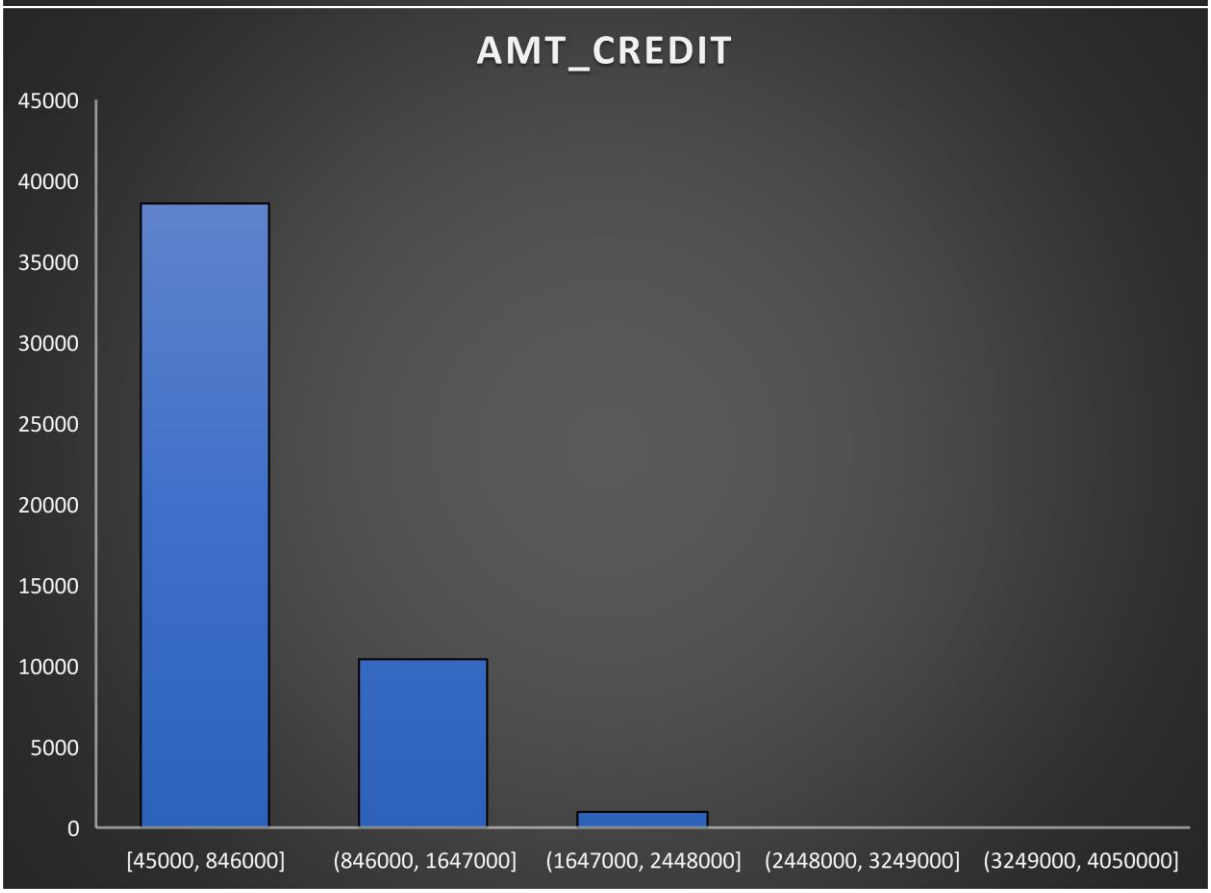
NAME_CONTRA	Approved	Refused	Canceled	Unused offer	
Consumer loar	6813	2468	51	842	
Cash loans	8899	4741	7199	17	
Revolving loar	2837	1451	1337	0	
XNA	0	0	8	0	

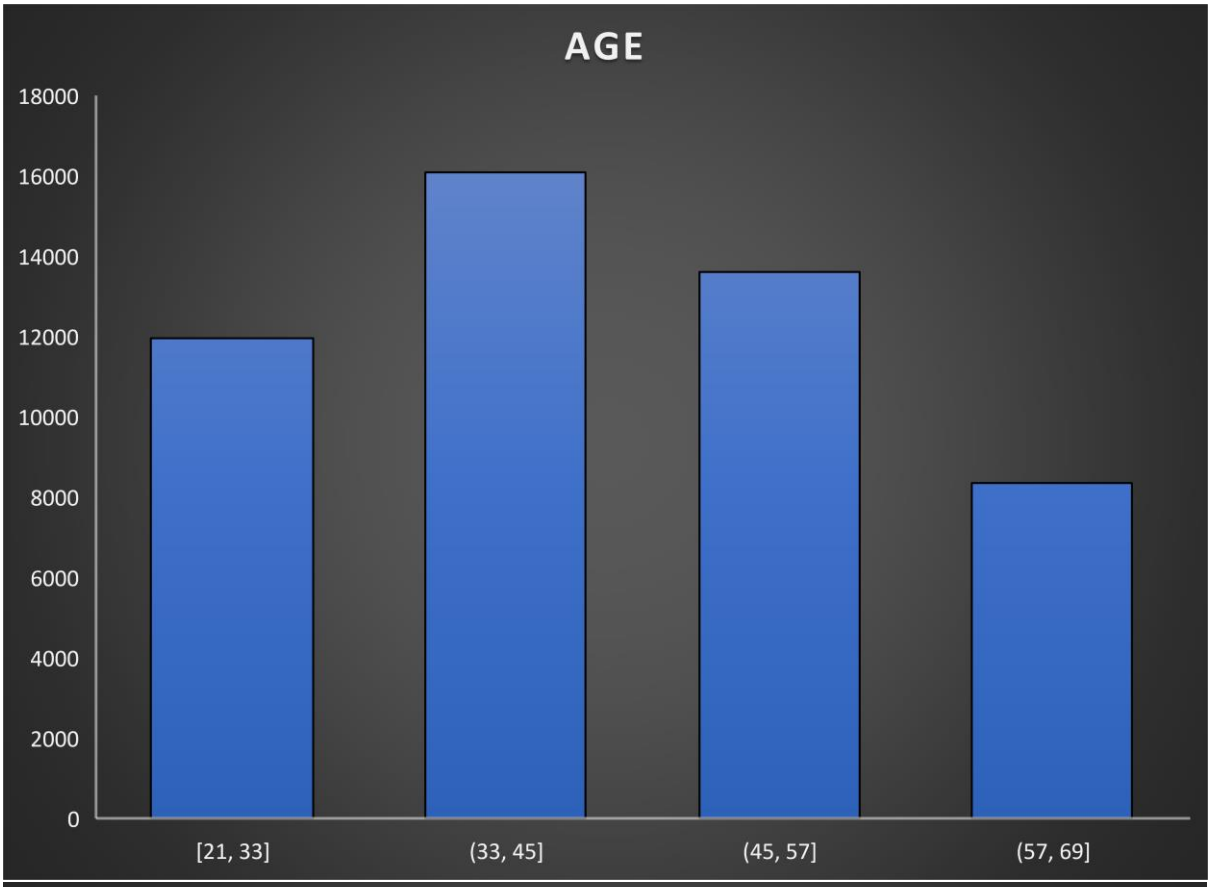
NAME_SELLER_INDUSTRY	Approved	Refused	Canceled	Unused offer
Connectivity	6813	1213	40	717
XNA	9793	5418	8543	17
Consumer electronics	11184	1635	9	114
Industry	544	48	1	2
Clothing	720	53	0	1
Furniture	1684	165	2	3
Construction	881	102	0	5
Jewelry	76	7	0	0
Auto technology	151	14	0	0
MLM partners	28	4	0	0
Tourism	11	1	0	0

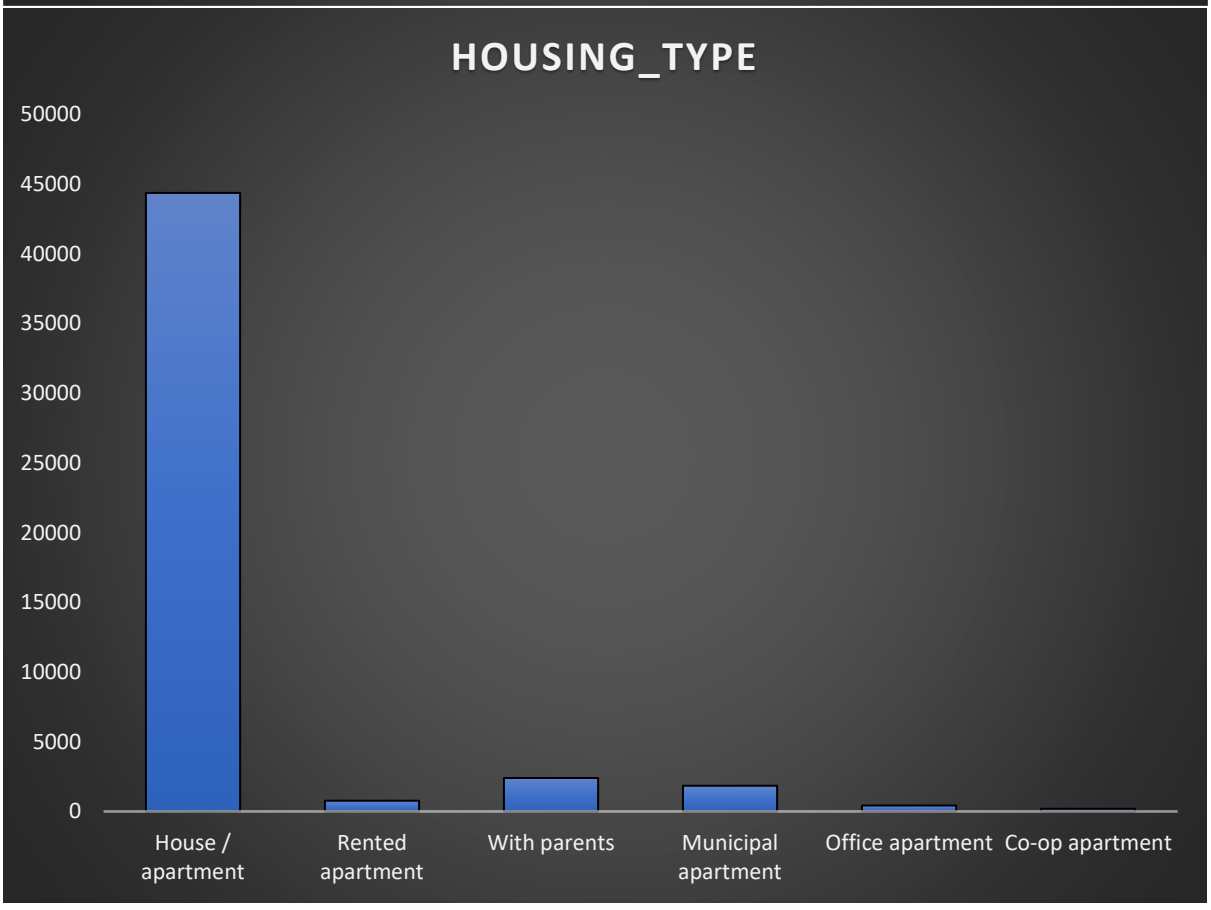
Graph:

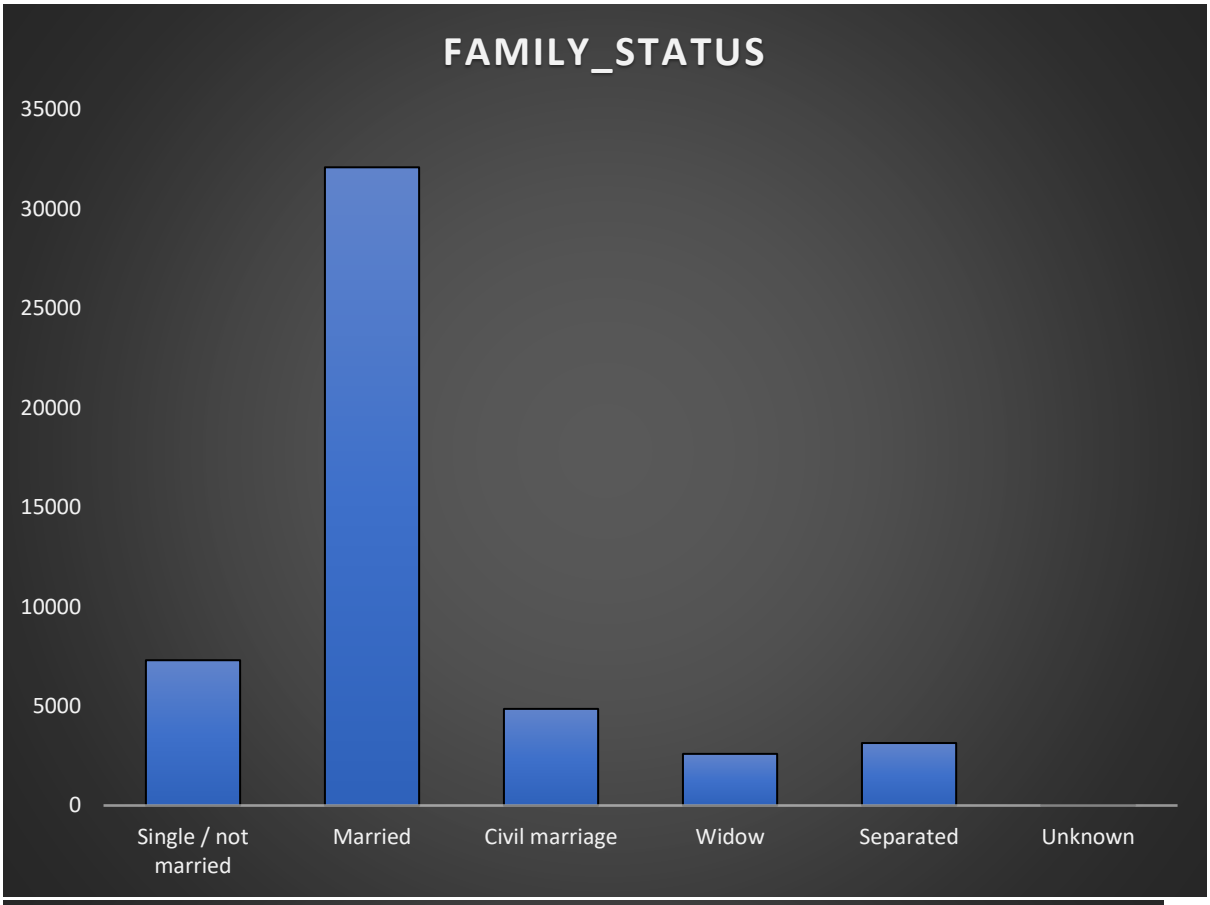
Univariate

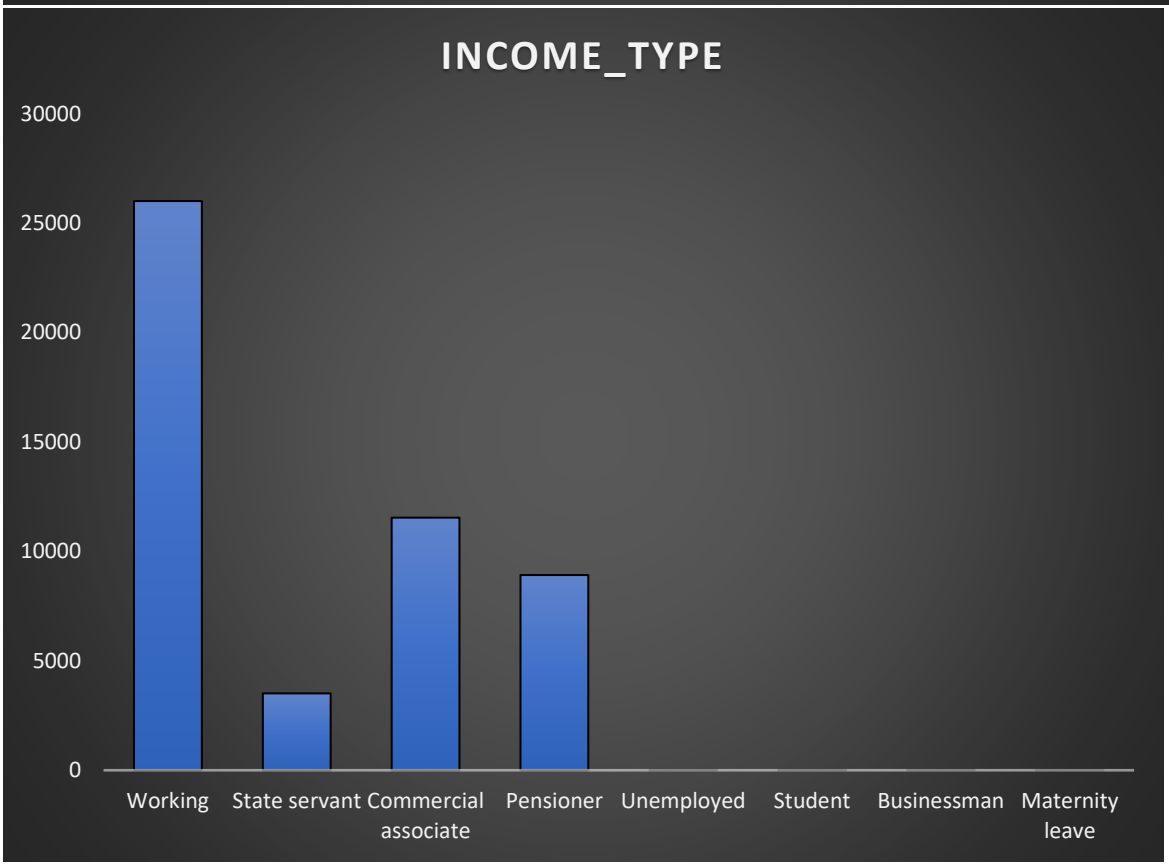

AMT_INCOME_TOTAL
50000
45000
40000
35000
30000
25000
20000
15000
10000
5000
0
[25650, 325520]
(325520, 625390]
(625390, 925260]
(925260, 1225130]
(1225130, 1525000]
> 1525000


AMT_CREDIT
45000
40000
35000
30000
25000
20000
15000
10000
5000
0
[45000, 846000]
(846000, 1647000]
(1647000, 2448000]
(2448000, 3249000]
(3249000, 4050000]

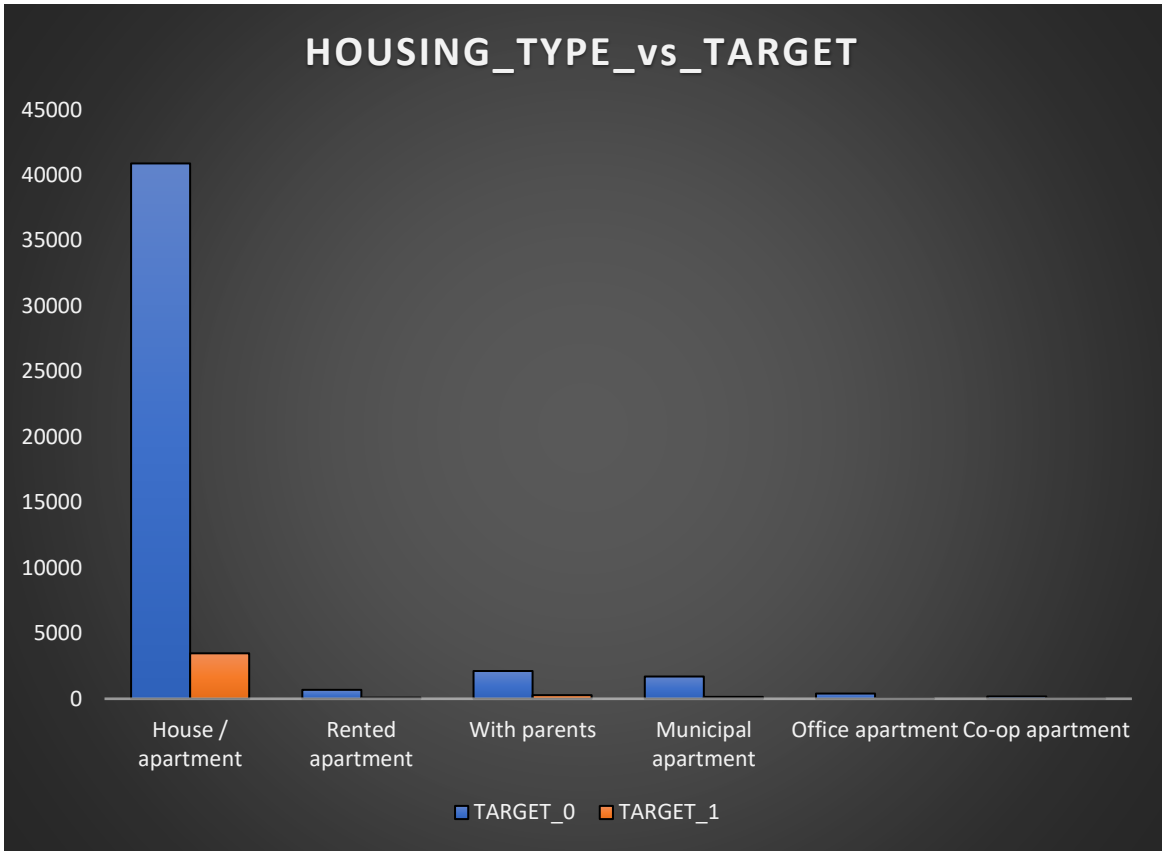

AGE
18000
16000
14000
12000
10000
8000
6000
4000
2000
0
[21, 33]
(33, 45]
(45, 57]
(57, 69]

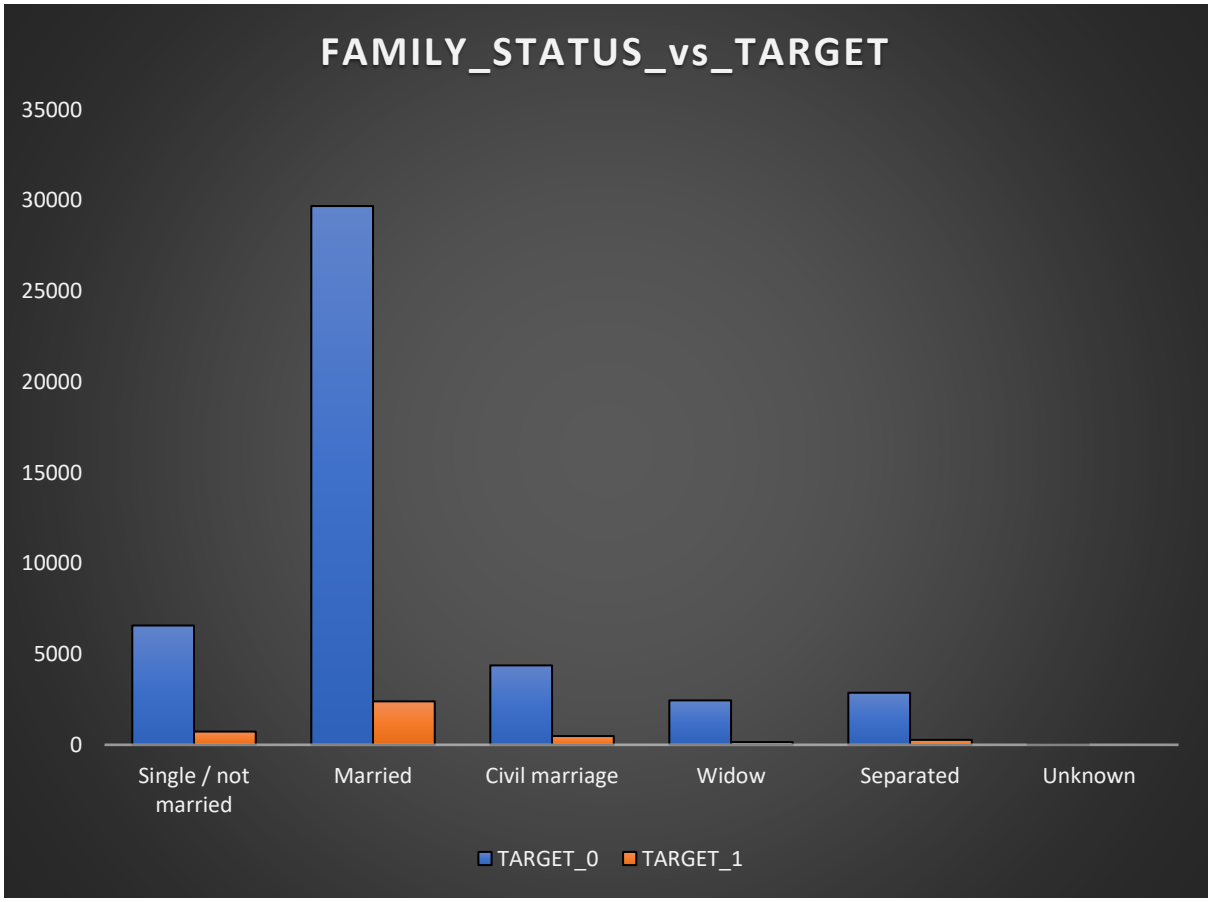

HOUSING_TYPE
50000
45000
40000
35000
30000
25000
20000
15000
10000
5000
0
House / apartment
Rented apartment
With parents
Municipal apartment
Office apartment
Co-op apartment

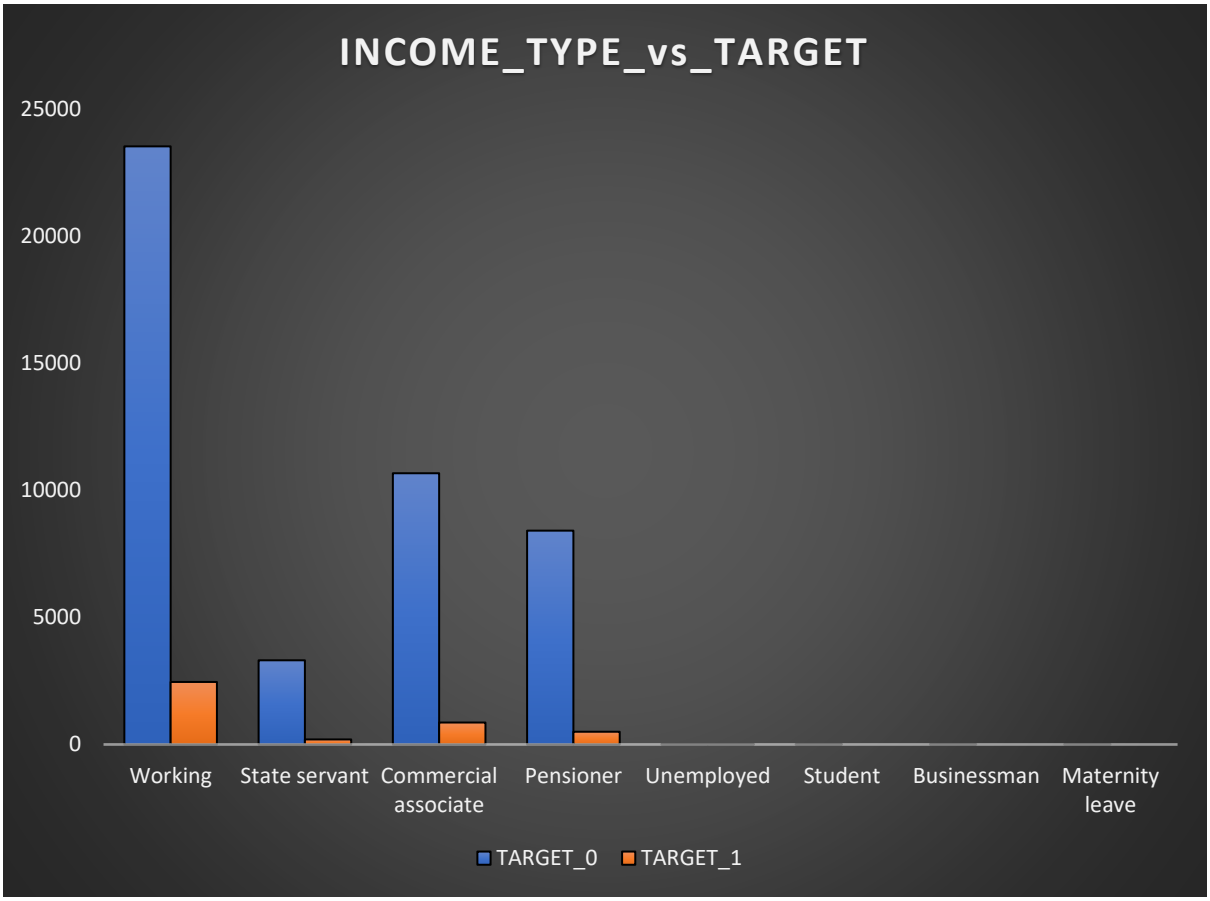

FAMILY_STATUS
35000
30000
25000
20000
15000
10000
5000
0
Single / not married
Married
Civil marriage
Widow
Separated
Unknown

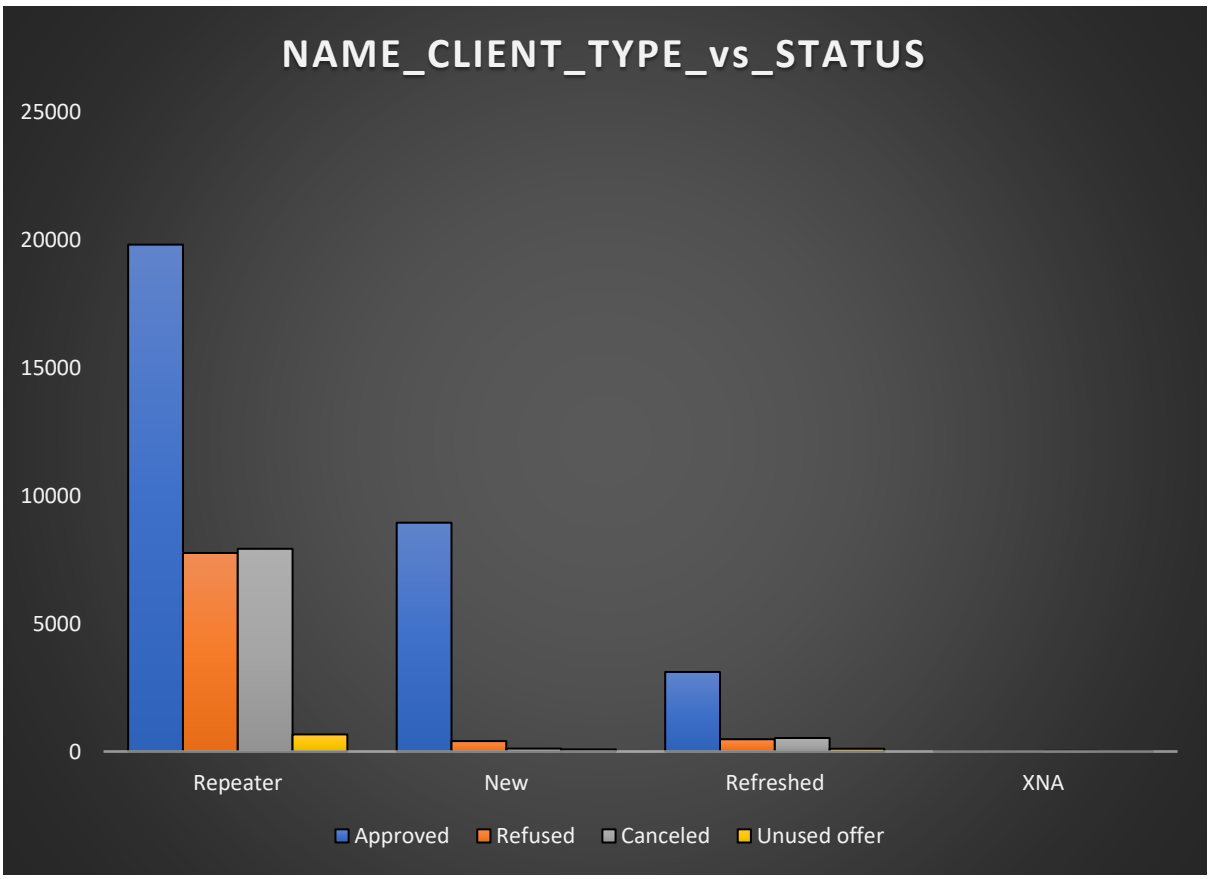

INCOME_TYPE
30000
25000
20000
15000
10000
5000
0
Working
State servant
Commercial associate
Pensioner
Unemployed
Student
Businessman
Maternity leave

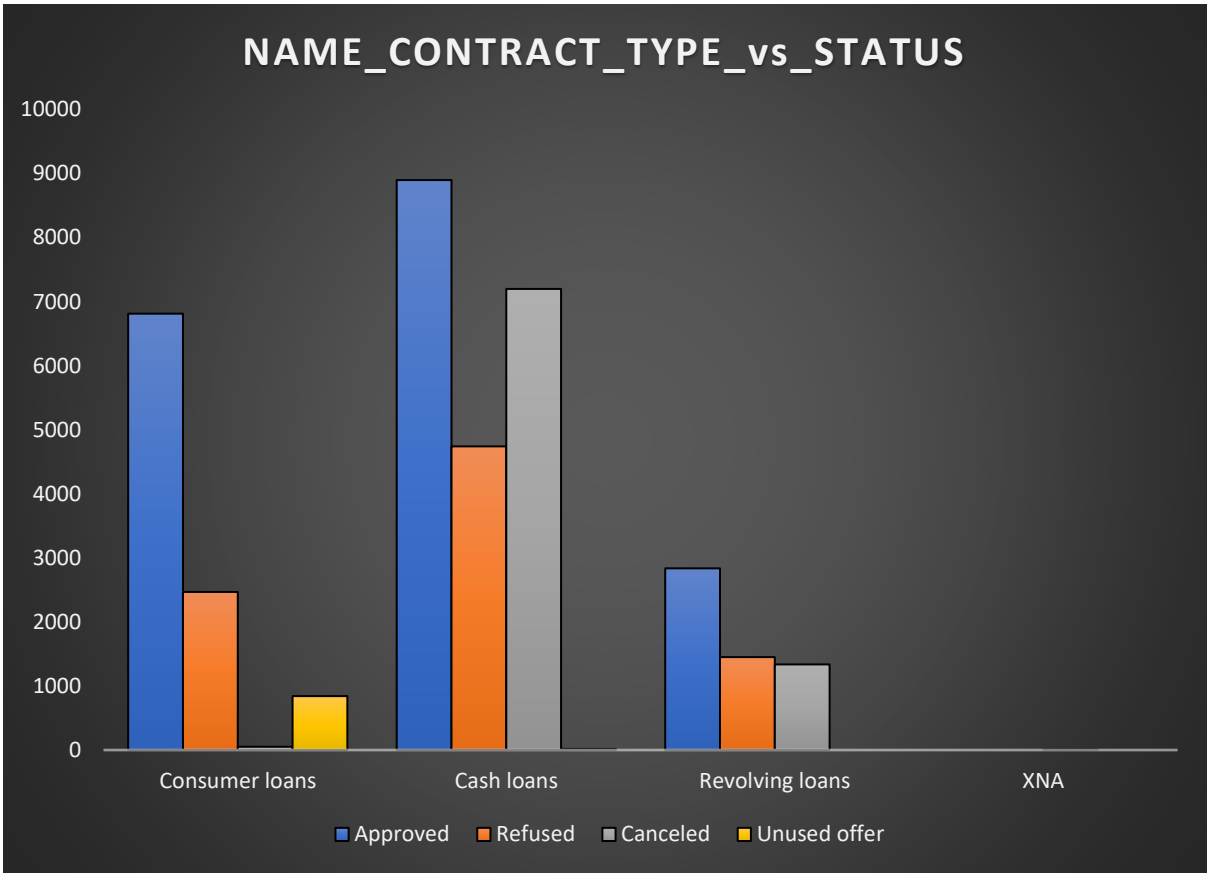
Bivariate

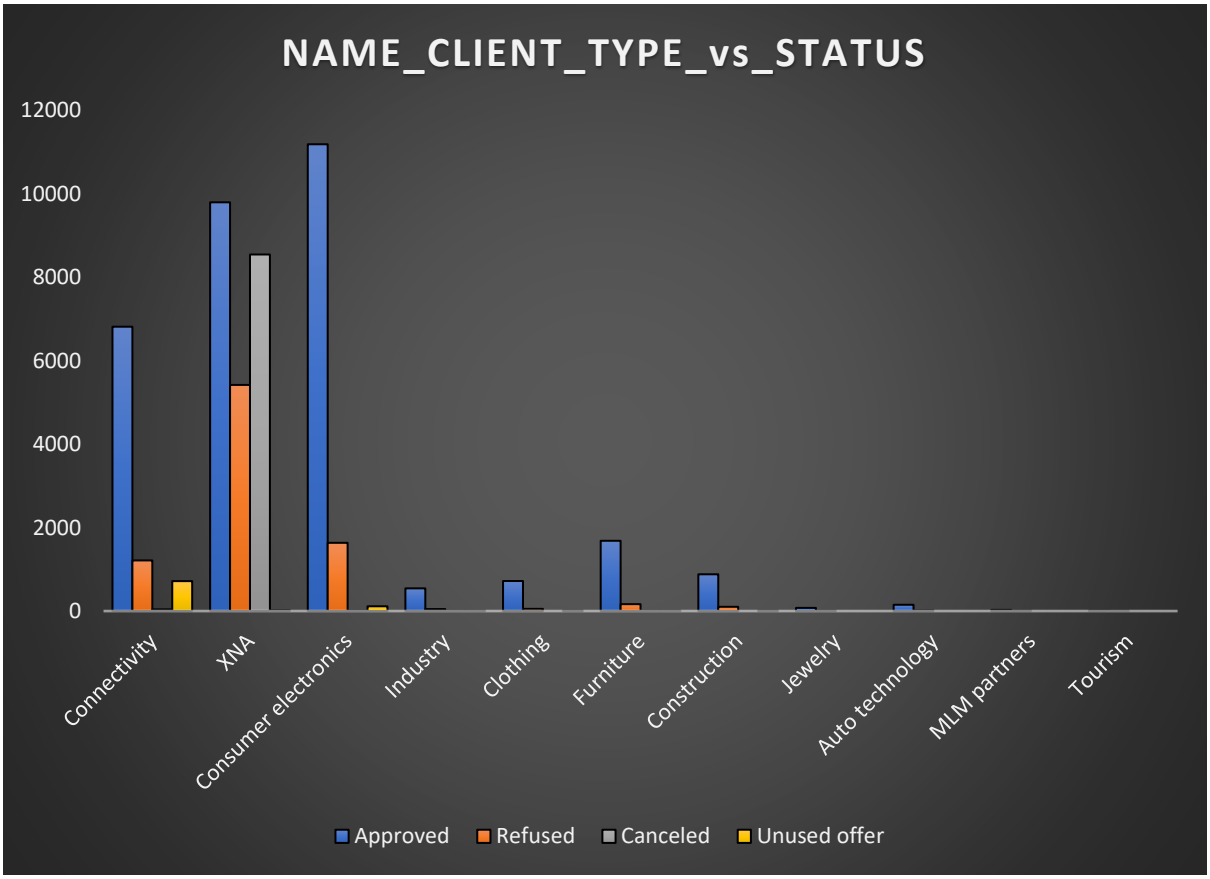

HOUSING_TYPE_vs_TARGET
45000
40000
35000
30000
25000
20000
15000
10000
5000
0
House / apartment
Rented apartment
With parents
Municipal apartment
Office apartment
Co-op apartment
TARGET_0
TARGET_1


FAMILY_STATUS_vs_TARGET
35000
30000
25000
20000
15000
10000
5000
0
Single / not married
Married
Civil marriage
Widow
Separated
Unknown
TARGET_0
TARGET_1


INCOME_TYPE_vs_TARGET
25000
20000
15000
10000
5000
0
Working
State servant
Commercial associate
Pensioner
Unemployed
Student
Businessman
Maternity leave
TARGET_0
TARGET_1


NAME_CLIENT_TYPE_vs_STATUS
25000
20000
15000
10000
5000
0
Repeater
New
Refreshed
XNA
Approved
Refused
Canceled
Unused offer


NAME_CONTRACT_TYPE_vs_STATUS
10000
9000
8000
7000
6000
5000
4000
3000
2000
1000
0
Consumer loans
Cash loans
Revolving loans
XNA
Approved
Refused
Canceled
Unused offer


NAME_CLIENT_TYPE_vs_STATUS
12000
10000
8000
6000
4000
2000
0
Connectivity
XNA
Consumer electronics
Industry
Clothing
Furniture
Construction
Jewelry
Auto technology
MLM partners
Tourism
Approved
Refused
Canceled
Unused offer

5. Identify Top Correlations for Different Scenarios:

   Function:-

   First, I found correlation between target and various columns by using following function:
   =CORREL(G2:G50000,B2:B50000)

   Then I found top five correlation among them:
   =INDEX(SORTBY(BX2:BY15,BY2:BY15,-1),SEQUENCE(5),{1,2})

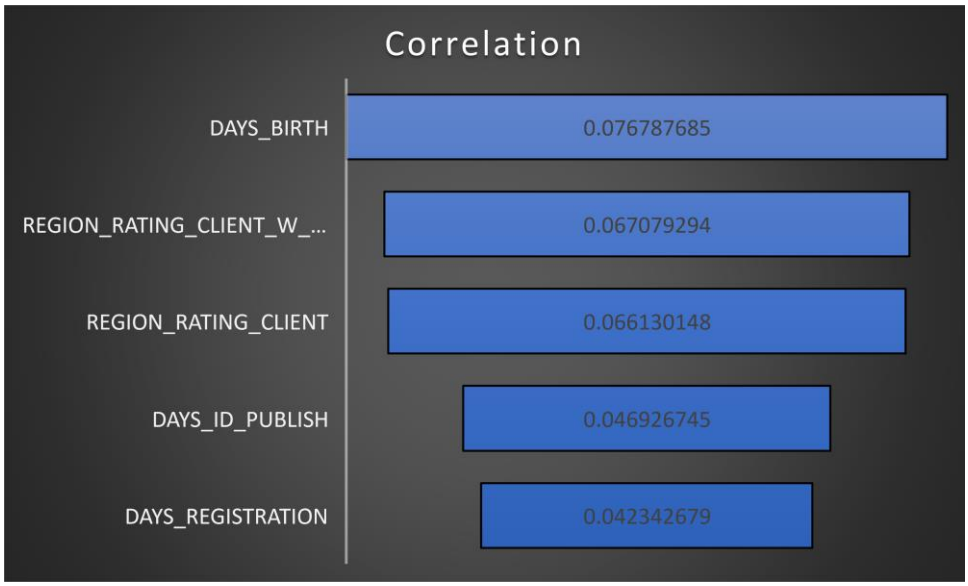
   Output:-

   https://docs.google.com/spreadsheets/d/1048NbwR9YLTf3777YbH3AvcKg0dcThUF/edit?usp=sharing&ouid=106942457558004201317&rtpof=true&sd=true

COLUMN_NAME	CORRELATION_WITH_TARGET
CNT_CHILDREN	0.026363931
AMT_INCOME_TOTAL	0.010893745
AMT_CREDIT	-0.032428347
AMT_ANNUITY	-0.012399094
AMT_GOODS_PRICE	-0.041306523
REGION_POPULATION_RELATIVE	-0.040799172
DAYS_BIRTH	0.076787685
DAYS_EMPLOYED	-0.040294905
DAYS_REGISTRATION	0.042342679
DAYS_ID_PUBLISH	0.046926745
CNT_FAM_MEMBERS	0.012992443
REGION_RATING_CLIENT	0.066130148
REGION_RATING_CLIENT_W_CITY	0.067079294

RANK	COLUMN_NAME	CORRELATION_WITH_TARGET
1	DAYS_BIRTH	0.076787685
2	REGION_RATING_CLIENT_	0.067079294
3	REGION_RATING_CLIENT	0.066130148
4	DAYS_ID_PUBLISH	0.046926745
5	DAYS_REGISTRATION	0.042342679

Graph:



Results:

2. Identify Outliers in the Dataset:
   There are many outliers in the data.

3. Analyze Data Imbalance:
   Data is imbalance in most of the columns.

4. Perform Univariate, Segmented Univariate, and Bivariate Analysis:
   People who are married, have low salary and live in house/apartment are most likely to take loan.

5. Identify Top Correlations for Different Scenarios:
   Highest correlation with TARGET is of DAYS_BIRTH .