Bank Loan Case Study

Project Description

The analysis aims to uncover patterns within a dataset containing information about loan applications, focusing on identifying factors influencing the likelihood of loan default. The dataset encompasses two distinct scenarios: when customers apply for loans and face either payment difficulties or all other cases.

In the former scenario, customers experienced late payments exceeding a certain threshold on initial instalments, while the latter encompasses cases where payments were made on time.

The project's primary objective is to employ Exploratory Data Analysis (EDA) techniques to discern how customer attributes and loan characteristics contribute to the probability of default. By understanding these patterns, the company can make informed decisions regarding loan approval, such as adjusting loan amounts, interest rates, or denying loans to high-risk applicants.

Key Business Objectives:

- Identify predictive factors indicating whether a customer is likely to encounter difficulties in repaying their loan instalments.
- Improve decision-making processes related to loan approval by leveraging insights gleaned from EDA.
- Mitigate financial risk by proactively identifying high-risk loan applicants and implementing appropriate measures, such as adjusting loan terms or denying loans altogether.

Approach

Upon reviewing the dataset description, the files were downloaded, including:

1. Application Data: The primary dataset containing comprehensive information about current loan applicants.
2. Previous Application Data: Historical data on past loan applicants.
3. Column Description: Detailed descriptions of each feature present in the dataset, aiding in understanding the data attributes.
4. Important Notes for Project 6: Providing insights and instructions pertinent to the project.

Following the download, each file was meticulously examined to grasp its contents and relevance to the project objectives. Here's a breakdown of the approach:

- Application Data: Focused on extracting insights related to current loan applicants.
- Previous Application Data: Analysis of past applicant behavior and trends.
- Column Description: Understanding the nuances of each feature to facilitate a deeper understanding of the dataset's attributes.
- Important Notes for Project 6: Utilized for crucial information and guidelines essential for executing the project effectively.

Additionally, supplementary research was conducted through relevant articles and resources on risk analysis, ensuring a comprehensive grasp of the project's objectives and methodologies. This preparatory phase equipped with the necessary insights to embark on the subsequent stages of data exploration and analysis effectively.

Tech Stack Used

Microsoft Excel 2021 version is used for this project due to its simplicity of use and extraordinary analysis and visualization capability.

A. **Identify Missing Data and Deal with it Appropriately:**

- In the application data file, several columns had null (NaN) values. Here are the new columns I introduced from the existing ones:

1. **Average EXT_SOURCE**: Created from EXT_SOURCE_1, EXT_SOURCE_2, and EXT_SOURCE_3. As these columns contained float scores, but with varying percentages of null data (56.35%, 0.25%, and 19.89% respectively), I computed their average to utilize all available data effectively.
2. **Total FLAG_DOCUMENT**: Consolidated from FLAG_DOCUMENT_2 to FLAG_DOCUMENT_21. While these columns had no null values, their unspecified document references prompted total aggregation, facilitating easy analysis of document submission patterns.
3. **Years Birth**: Derived from DAYS_BIRTH, converting days to years.
4. **Years Employed**: Derived from DAYS_EMPLOYED, converting days to years.
5. **Years Registration**: Derived from DAYS_REGISTRATION, converting days to years.
6. **Years ID_PUBLISH**: Derived from DAYS_ID_PUBLISH, converting days to years.

- Columns dropped due to insufficient data (with null values percentage):

OWN_CAR_AGE (65.90%), OCCUPATION_TYPE (31.31%), APARTMENTS_AVG (50.77%), BASEMENTAREA_AVG (58.40%), YEARS_BEGINEXPLUATATION_AVG (48.79%), YEARS_BUILD_AVG (66.48%), COMMONAREA_AVG (69.92%), ELEVATORS_AVG (53.30%), ENTRANCES_AVG (50.39%), FLOORSMAX_AVG (49.75%), FLOORSMIN_AVG (67.79%), LANDAREA_AVG (59.44%), LIVINGAPARTMENTS_AVG (68.45%), LIVINGAREA_AVG (50.28%), NONLIVINGAPARTMENTS_AVG (69.43%), NONLIVINGAREA_AVG (55.15%), APARTMENTS_MODE (50.77%), BASEMENTAREA_MODE (58.40%), YEARS_BEGINEXPLUATATION_MODE (48.79%), YEARS_BUILD_MODE (66.48%), COMMONAREA_MODE (69.92%),

ELEVATORS_MODE (53.30%), ENTRANCES_MODE (50.39%), FLOORSMAX_MODE (49.75%), FLOORSMIN_MODE (67.79%), LANDAREA_MODE (59.44%), LIVINGAPARTMENTS_MODE (68.45%), LIVINGAREA_MODE (50.28%), NONLIVINGAPARTMENTS_MODE (69.43%), NONLIVINGAREA_MODE (55.15%), APARTMENTS_MEDI (50.77%), BASEMENTAREA_MEDI (58.40%), YEARS_BEGINEXPLUATATION_MEDI (48.79%), YEARS_BUILD_MEDI (66.48%), COMMONAREA_MEDI (69.92%), ELEVATORS_MEDI (53.30%), ENTRANCES_MEDI (50.39%), FLOORSMAX_MEDI (49.75%), FLOORSMIN_MEDI (67.79%), LANDAREA_MEDI (59.44%), LIVINGAPARTMENTS_MEDI (68.45%), LIVINGAREA_MEDI (50.28%), NONLIVINGAPARTMENTS_MEDI (69.43%), NONLIVINGAREA_MEDI (55.15%), FONDKAPREMONT_MODE (68.38%), HOUSETYPE_MODE (50.15%), TOTALAREA_MODE (48.30%), WALLSMATERIAL_MODE (50.92%), EMERGENCYSTATE_MODE (47.40%)

- ## Columns dropped for least relevance:

  - WEEKDAY_APPR_PROCESS_START
  - HOUR_APPR_PROCESS_START

- ## Columns with null values imputed:

  - AMT_GOODS_PRICE: Imputed with 90% of AMT_CREDIT values.
  - NAME_TYPE_SUITE: Imputed with the most frequently occurring value using Index Match Max COUNTIF function.
  - OBS_30_CNT_SOCIAL_CIRCLE, DEF_30_CNT_SOCIAL_CIRCLE, -
    OBS_60_CNT_SOCIAL_CIRCLE, DEF_60_CNT_SOCIAL_CIRCLE: Imputed with rounded averages.

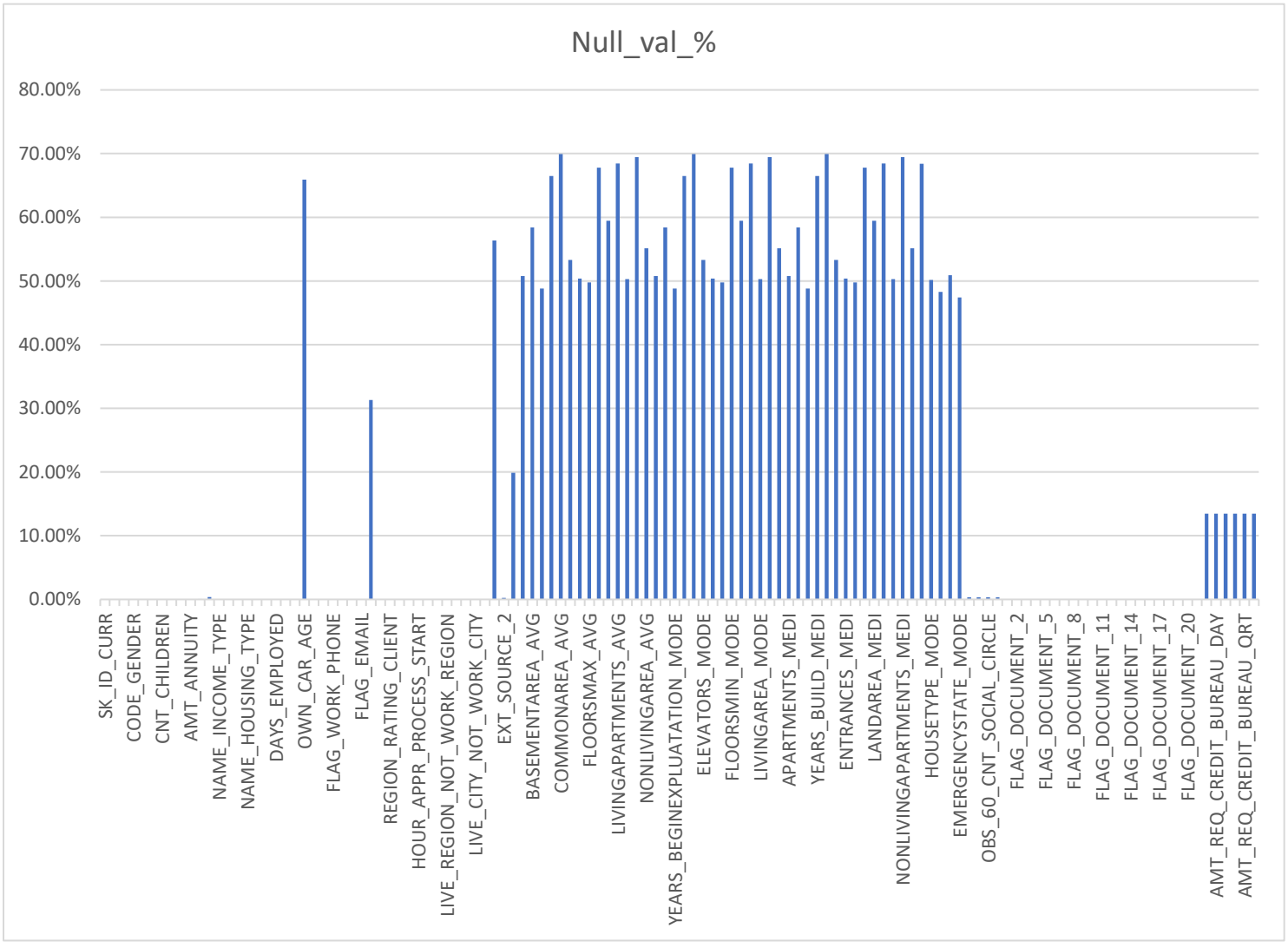
- ## Columns with null values removed:

  - All six columns related to credit bureau inquiries - namely AMT_REQ_CREDIT_BUREAU_HOUR, AMT_REQ_CREDIT_BUREAU_DAY, AMT_REQ_CREDIT_BUREAU_WEEK, AMT_REQ_CREDIT_BUREAU_MON, AMT_REQ_CREDIT_BUREAU_QRT, and AMT_REQ_CREDIT_BUREAU_YEAR - may not appear immediately pertinent to this dataset. However, in the context of creditor analysis and borrower evaluation, they hold significant importance.

- Understanding risk analysis underscores the critical role of an applicant's credit history, akin to the concept of a credit score. These columns capture the inquiries made by applicants to the credit bureau, which are typically regarded as requests for credit-related information.

- In risk assessment, a high frequency of such inquiries over a short period often indicates a "credit-hungry" behavior, suggesting potential financial instability or overreliance on credit. Therefore, despite their seemingly indirect relevance, these columns were deemed important. Consequently, null values were dropped to ensure a comprehensive analysis, aligning with the broader objective of identifying patterns indicative of loan repayment capability and default likelihood.

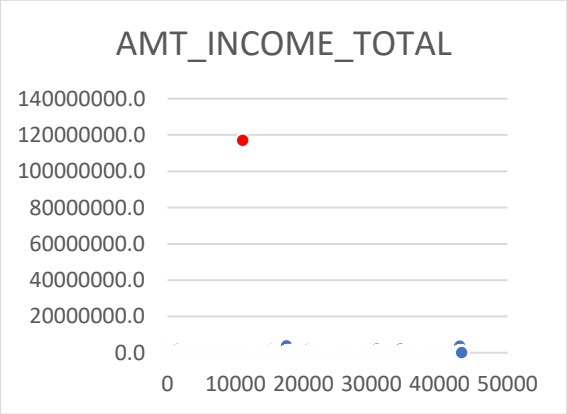
This graph indicates null values in various features:

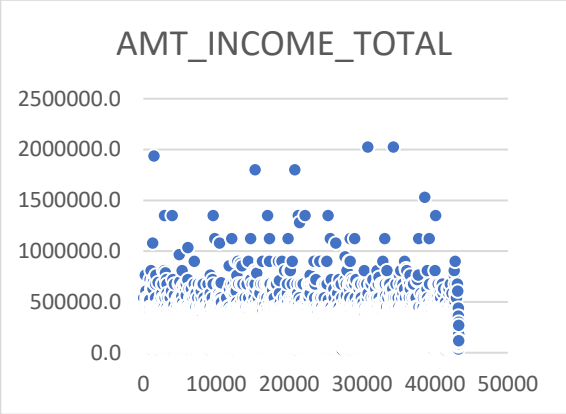


B. Identify Outliers in the Dataset:

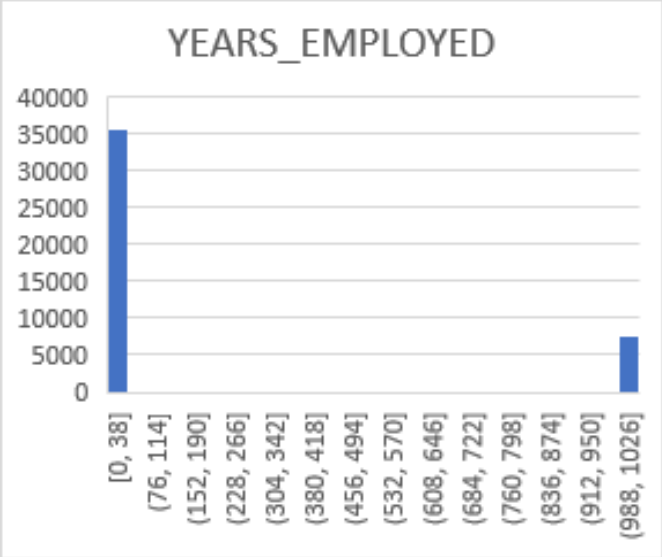
To identify outliers within the dataset, I employed the interquartile range (IQR) method, which involves computing the first quartile (Q1) and the third quartile (Q3) at the 25th and 75th percentiles, respectively. I then utilized a multiplier value (k) of 1.5 to establish thresholds for the lower and upper limits.

Upon examination of the column AMT_INCOME_TOTAL, an outlier was detected with a value of 117,000,000. This value stands out significantly from the rest of the dataset, as illustrated in the scatterplot graph that depicts the data both with and without the outlier [3825000, 3600000, 117000000]

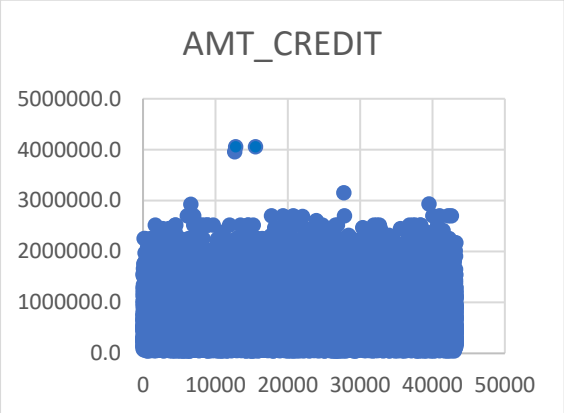


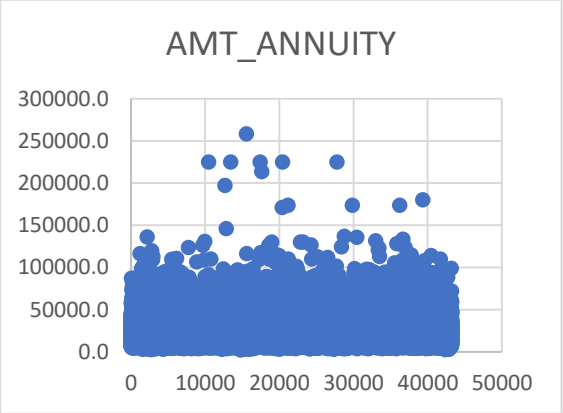


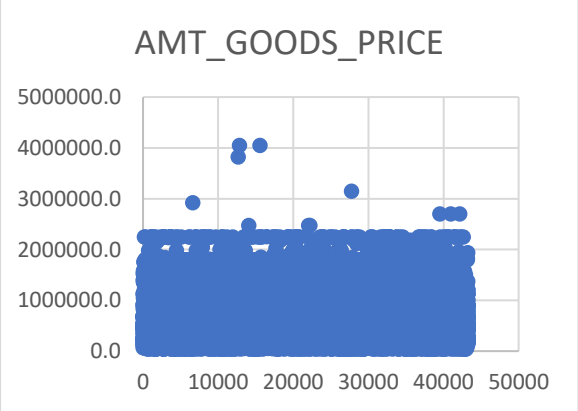
Similarly, in the column YEARS_EMPLOYED, outliers were observed, suggesting potential typographical errors. Notably, one entry indicates an applicant's employment duration of 1001 years, which is clearly implausible. This anomaly is evident in the accompanying graph.

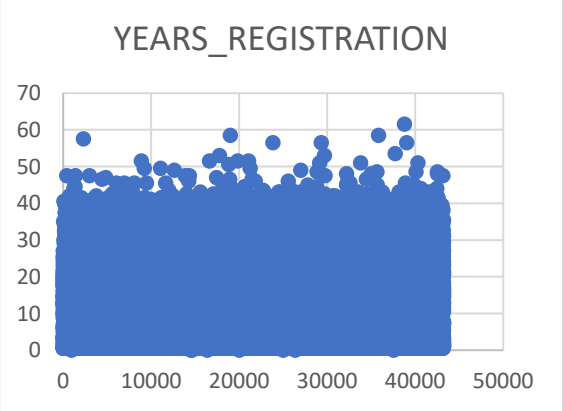


For further insight, scatterplot graphs were generated for columns such as AMT_CREDIT, AMT_ANNUITY, AMT_GOODS_PRICE, and YEARS_REGISTRATION. These visualizations aid in identifying any additional outliers and understanding their impact on the dataset's distribution and overall analysis.





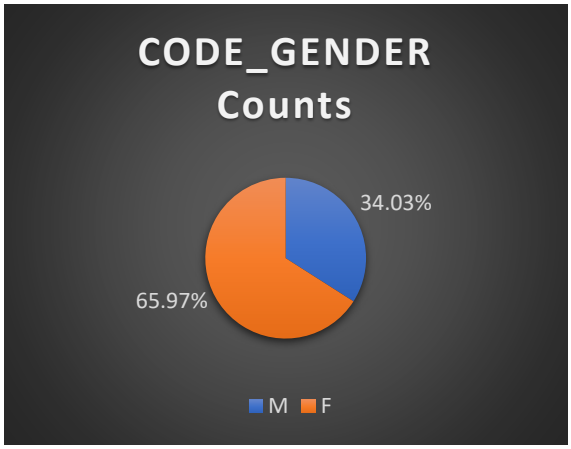


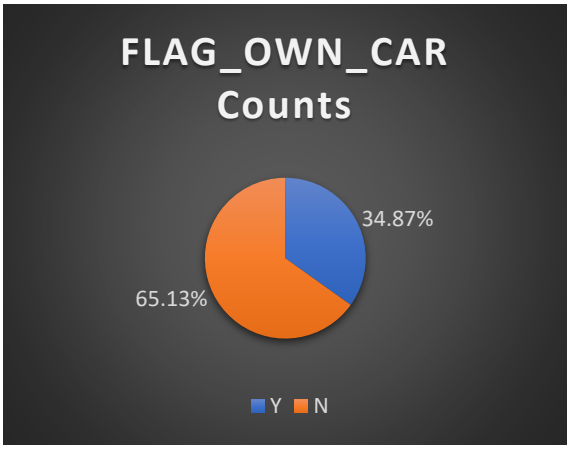


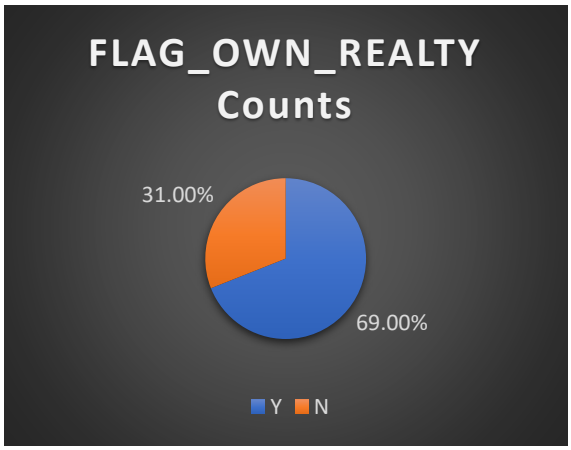
C. Analyze Data Imbalance:

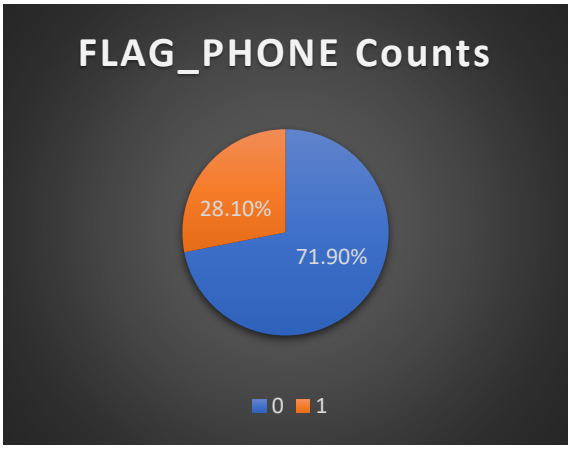
The dataset reveals a notable imbalance in certain columns, with some features displaying varying degrees of disproportion between different categories. Here's an examination of the data imbalance across three categories:

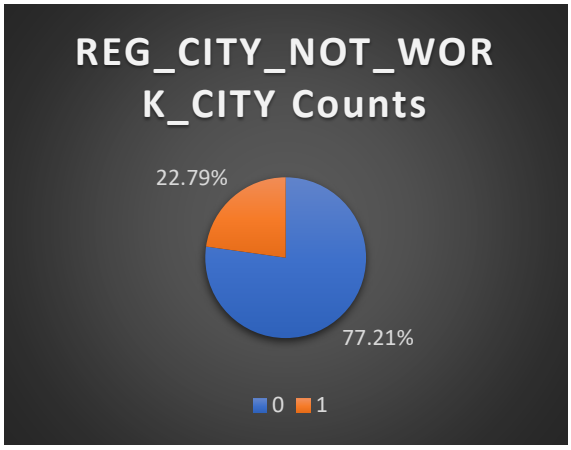
1. **Mild Imbalance [20-40% minority proportion]:**
    - CODE_GENDER: 34.03% Male against 65.97% Female
    - FLAG_OWN_CAR: 34.87% "N" Value against 65.13% "Y" Value
    - FLAG_OWN_REALTY: 31% "N" value against 69% "Y" value
    - FLAG_PHONE: 28.10% "1" value against 71.90% "0" value
    - REG_CITY_NOT_WORK_CITY: 22.79% "1" value against 77.21% "0" value



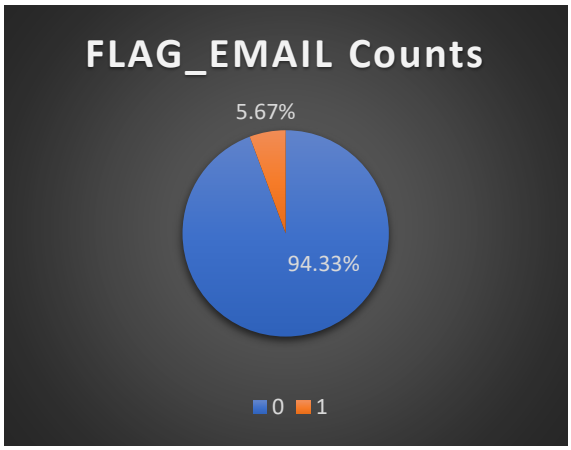


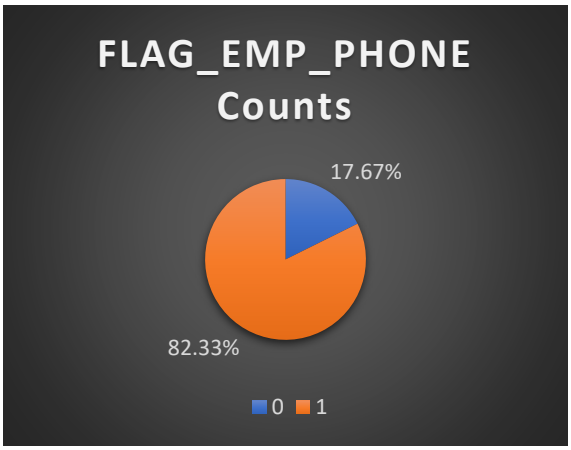


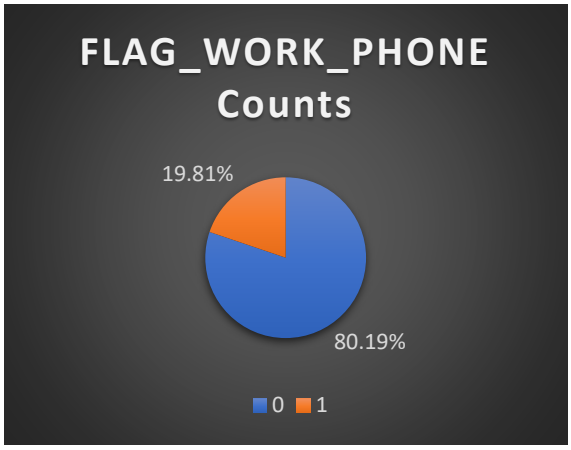


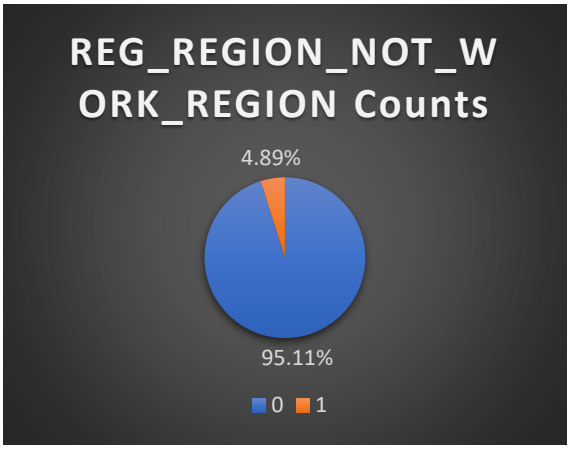


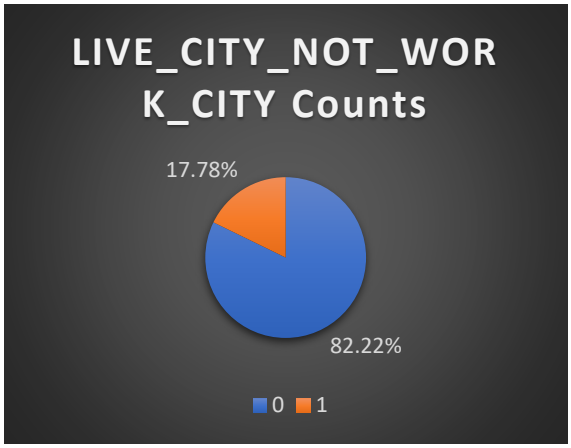
2. **Moderate Imbalance [1-20% minority proportion]:**
    - FLAG_EMAIL: 5.67% "1" value against 94.33% "0" value
    - FLAG_EMP_PHONE: 17.67% "1" value against 82.33% "0" value
    - FLAG_WORK_PHONE: 19.81% "1" value against 80.19% "0" value
    - REG_REGION_NOT_WORK_REGION: 4.89% "1" value against 95.11% "0" value
    - LIVE_CITY_NOT_WORK_CITY: 17.78% "1" value against 82.22% "0" value



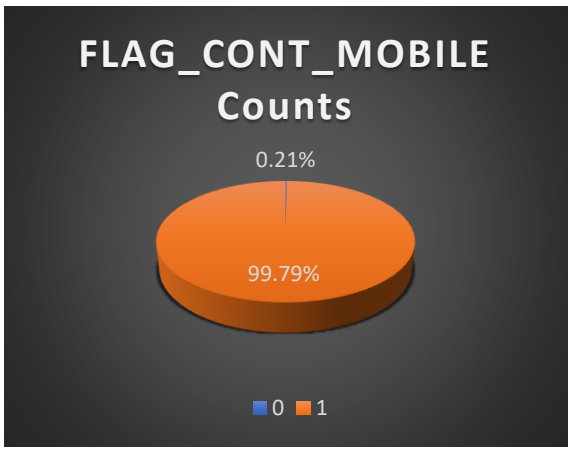








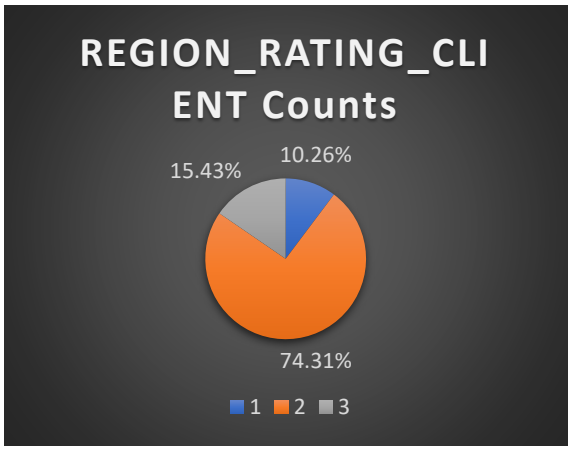
3. **Extreme Imbalance [less than 1% minority proportion]:**
    - FLAG_CONT_MOBILE: 0.21% "0" value against 99.79% "1" value

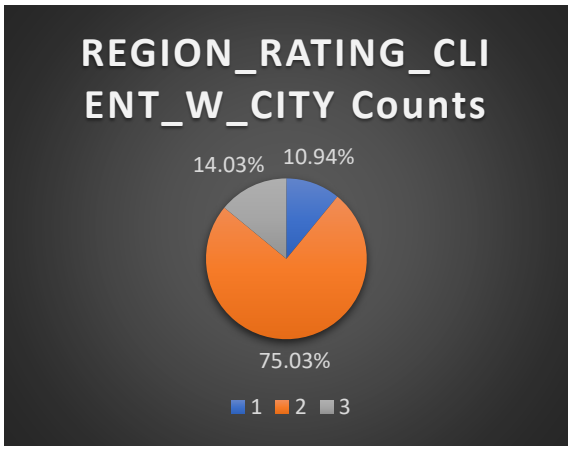


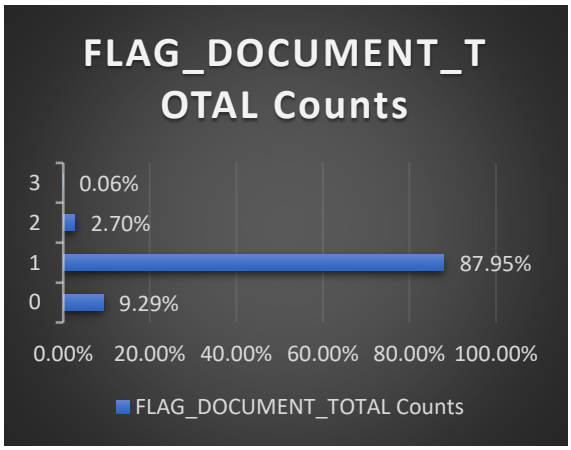
now here are those imbalances that has more than two variables

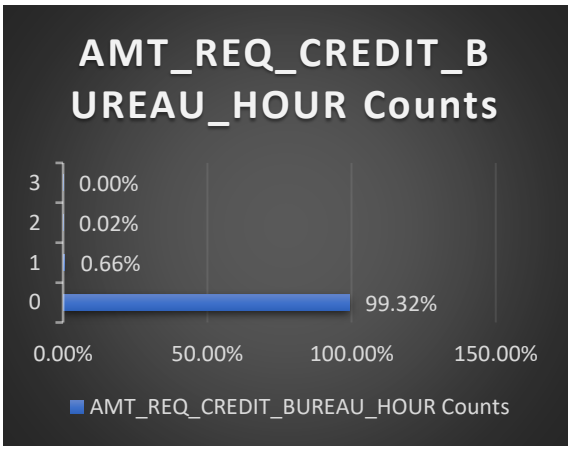
- REGION_RATING_CLIENT has 10.26% of "1" value, 74.31% of "2" value, and 15.43% of "3" value.
- REGION_RATING_CLIENT_W_CITY has 10.94% of "1" value, 75.03% of "2" value, and 14.03% of "3" value.

- FLAG_DOCUMENT_TOTAL has 9.29% of "0" value, 87.95% of "1" value, 2.70% of "2" value, and 0.06% of "3" value.
- AMT_REQ_CREDIT_BUREAU_HOUR has 99.32% of "0" value, 0.66% of "1" value, 0.02% of "2" value, and 0.00% of "3" value.







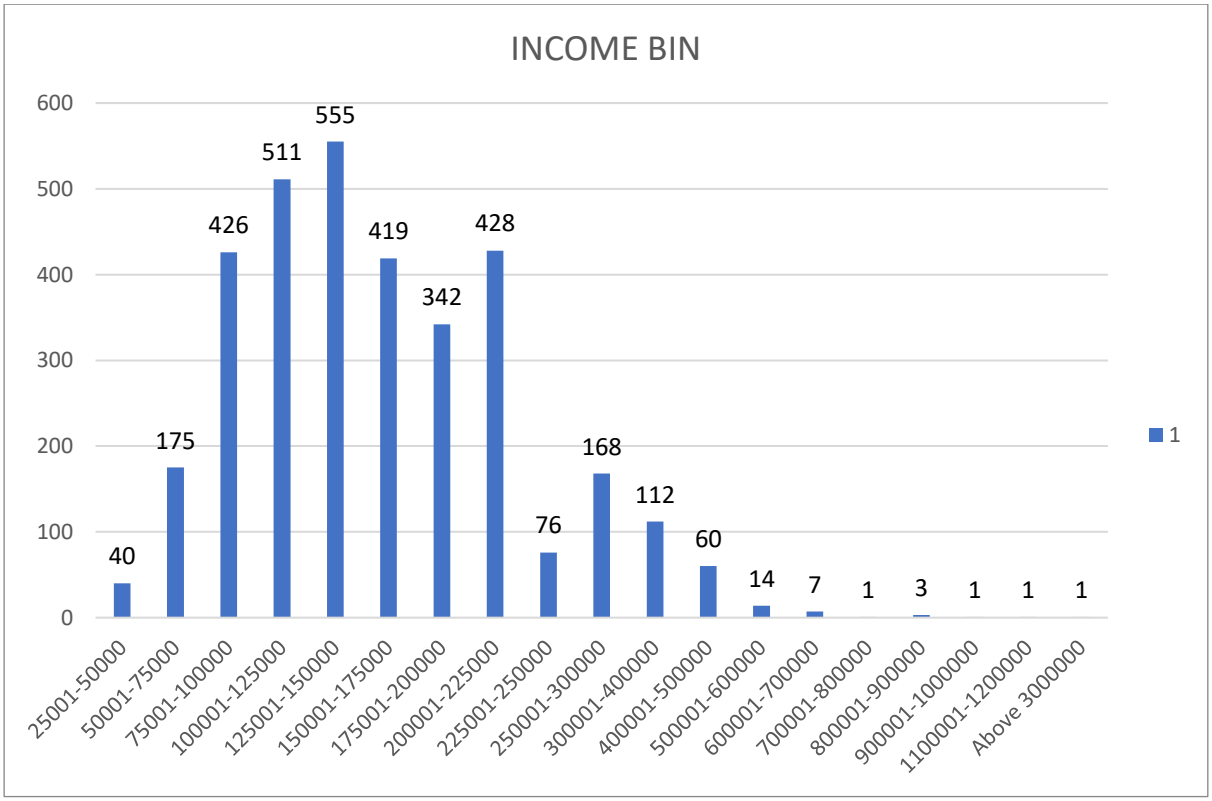


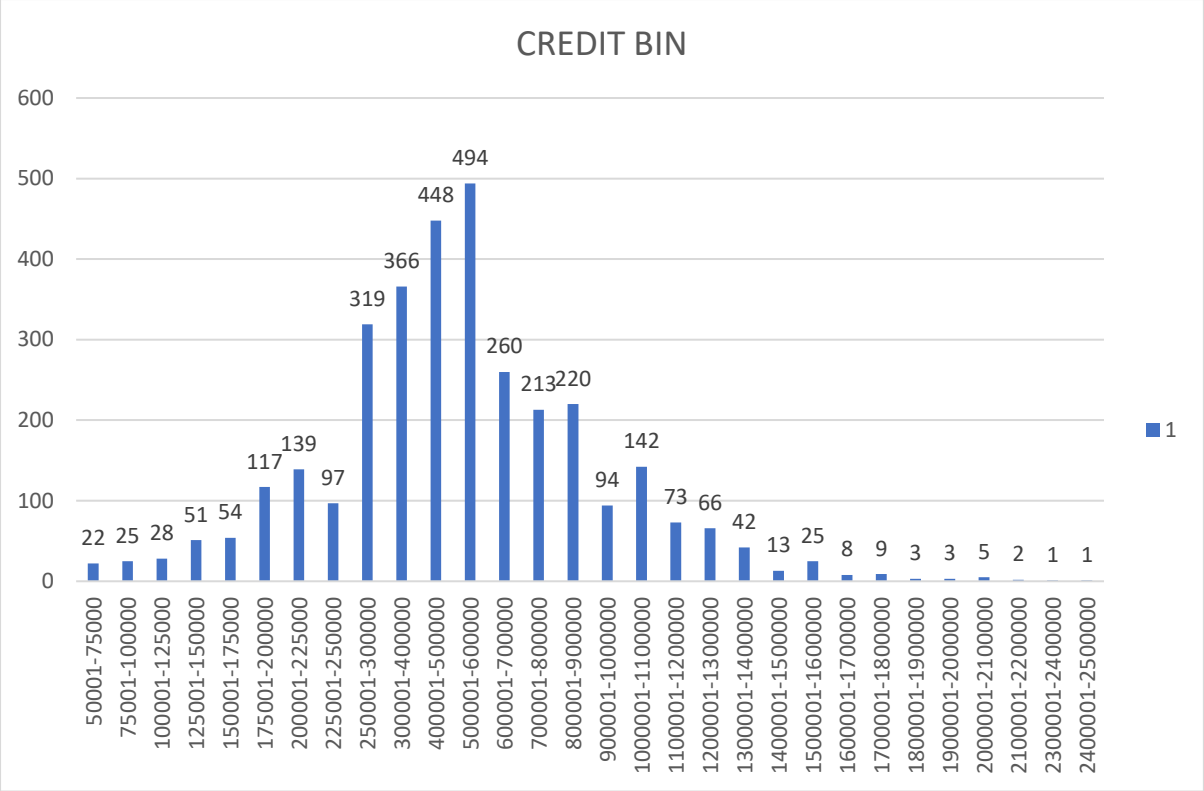
D. Perform Univariate, Segmented Univariate, and Bivariate Analysis

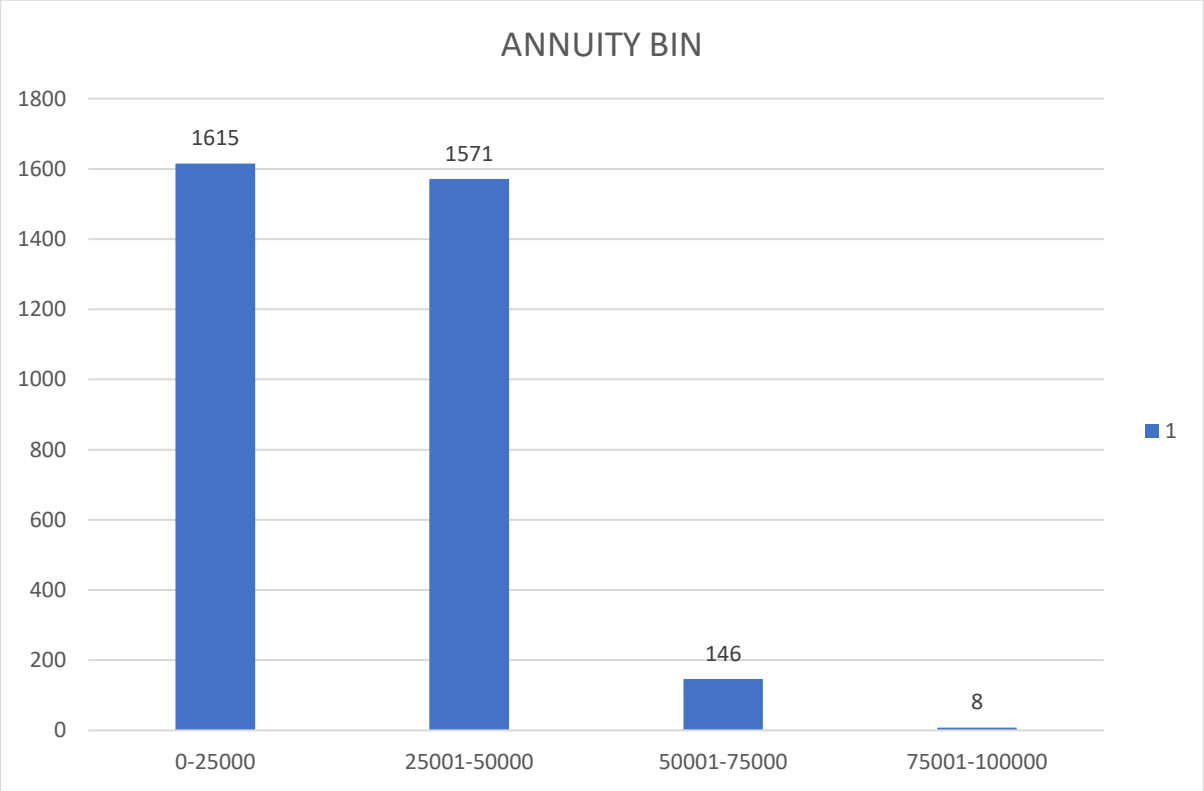
To conduct the analysis effectively, several key columns were selected, including AMT_INCOME_TOTAL, AMT_CREDIT, AMT_ANNUITY, AMT_GOODS_PRICE, Avg_EXT_SOURCE, and YEARS_EMPLOYED.

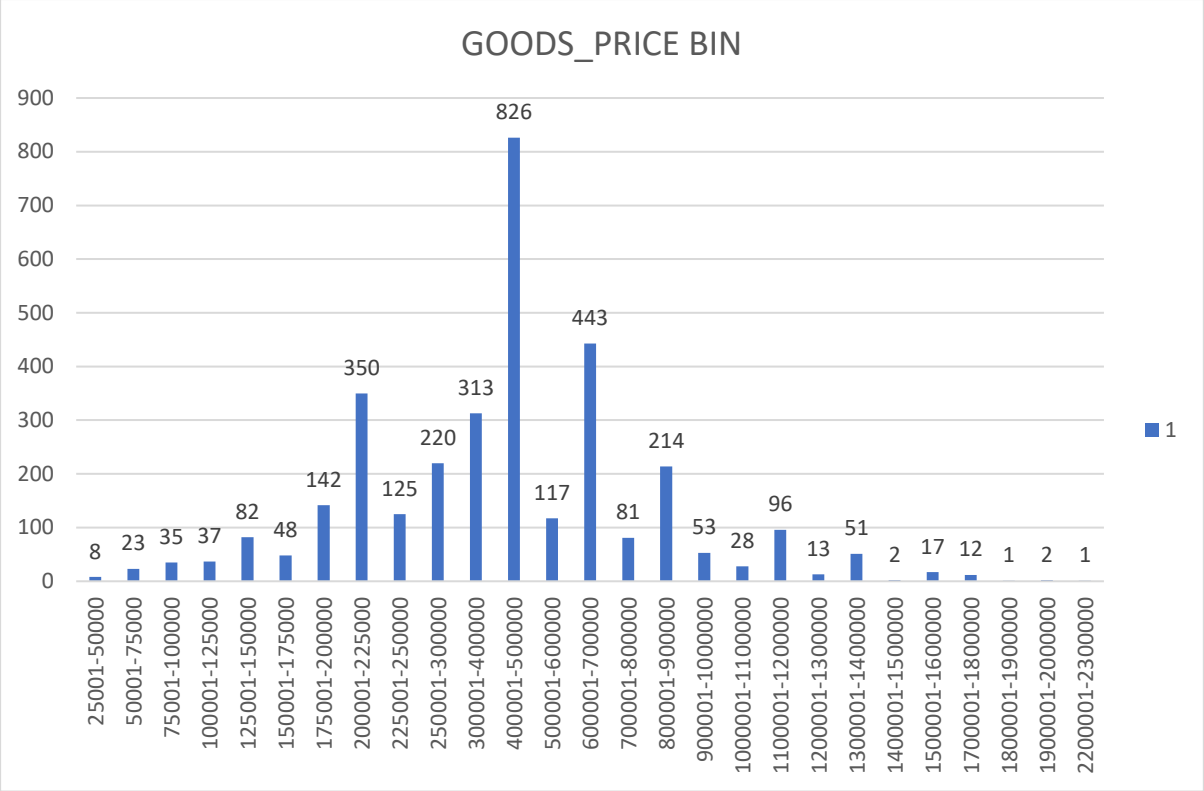
Next, the values in these columns were categorized into bins using nested IF and AND functions, resulting in the creation of six new columns representing the bin values.

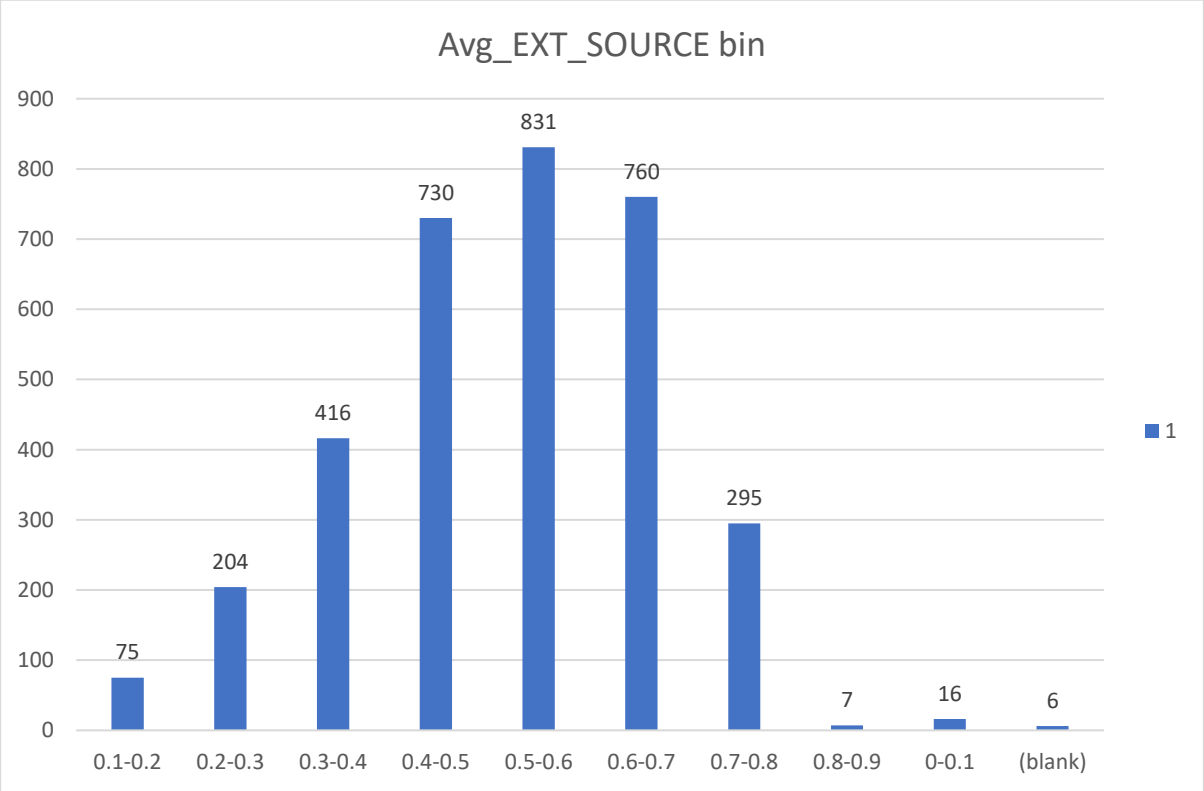
Subsequently, pivot tables were generated from these columns to perform **univariate analysis**, comparing each of the six columns with the Target value of 1. This allowed for an understanding of how each column correlates with loan default.

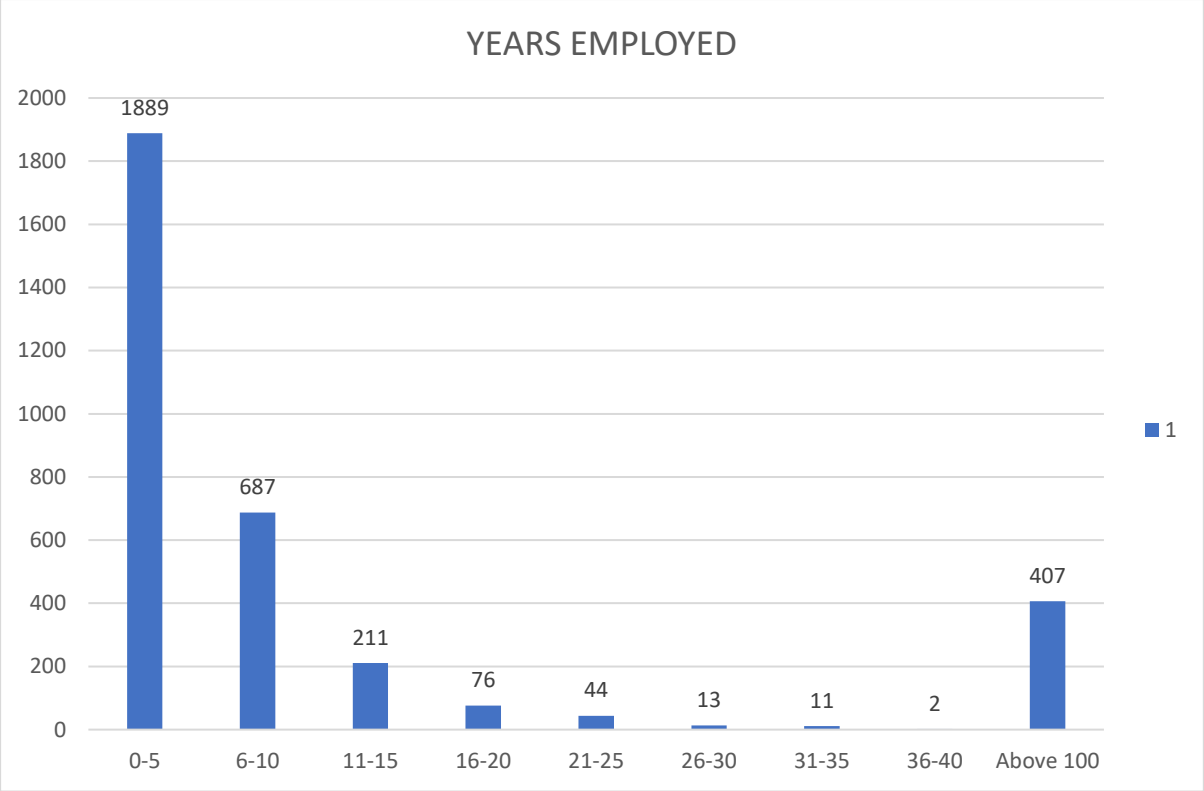



CREDIT BIN
600
500
400
300
200
100
0
22 25 28 51 54 117 139 97 319 366 448 494 260 213 220 94 142 73 66 42 13 25 8 9 3 3 5 2 1 1
50001-75000
75001-100000
100001-125000
125001-150000
150001-175000
175001-200000
200001-225000
225001-250000
250001-300000
300001-400000
400001-500000
500001-600000
600001-700000
700001-800000
800001-900000
900001-1000000
1000001-1100000
1100001-1200000
1200001-1300000
1300001-1400000
1400001-1500000
1500001-1600000
1600001-1700000
1700001-1800000
1800001-1900000
1900001-2000000
2000001-2100000
2100001-2200000
2300001-2400000
2400001-2500000
1

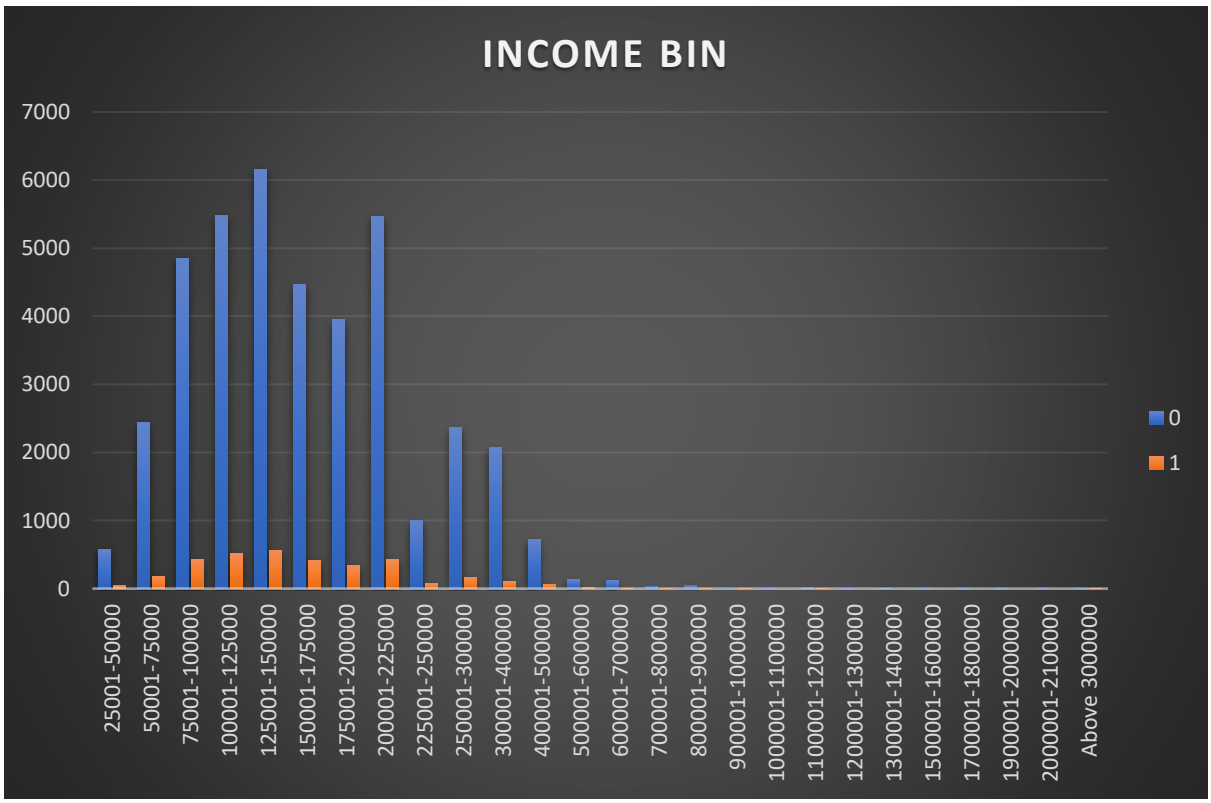

ANNUITY BIN
1800
1600
1400
1200
1000
800
600
400
200
0
1615
1571
146
8
0-25000
25001-50000
50001-75000
75001-100000
1

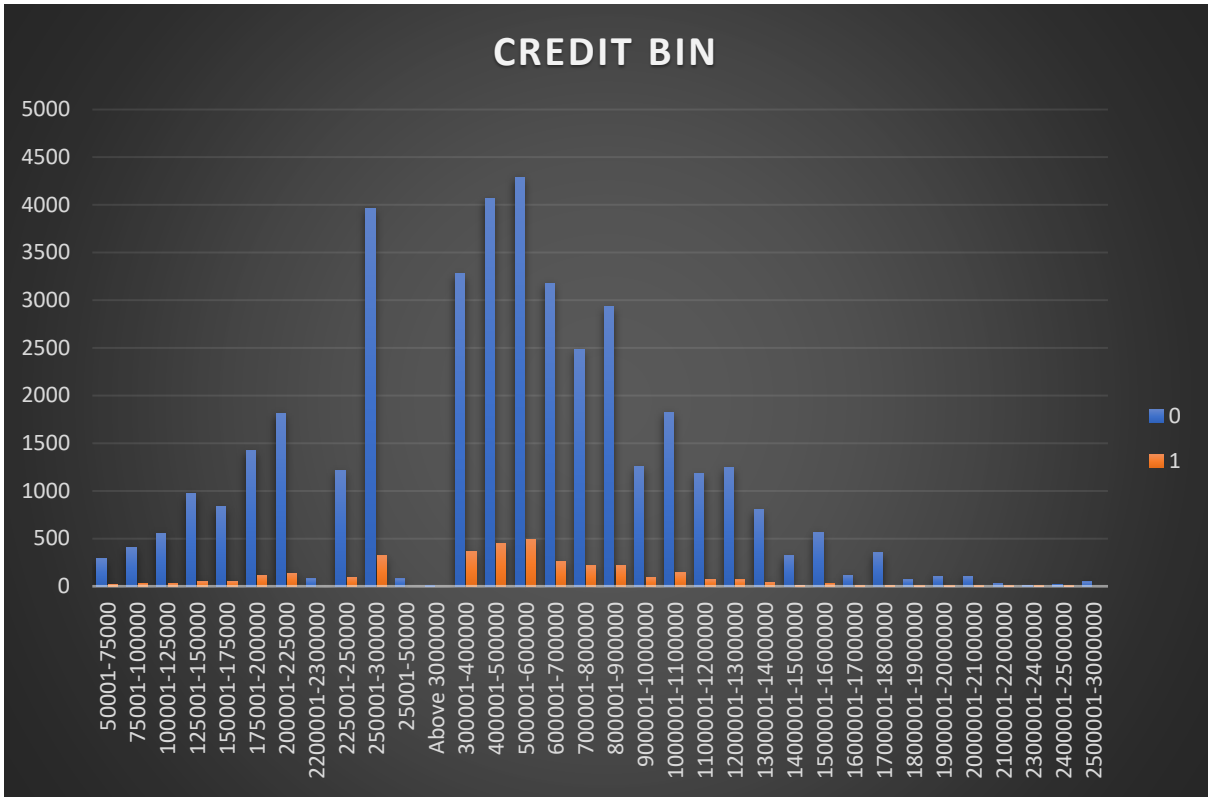

GOODS_PRICE BIN
900
800
700
600
500
400
300
200
100
0
8
23
35
37
82
48
142
350
125
220
313
826
117
443
81
214
53
28
96
13
51
2
17
12
1
2
1
25001-50000
50001-75000
75001-100000
100001-125000
125001-150000
150001-175000
175001-200000
200001-225000
225001-250000
250001-300000
300001-400000
400001-500000
500001-600000
600001-700000
700001-800000
800001-900000
900001-1000000
1000001-1100000
1100001-1200000
1200001-1300000
1300001-1400000
1400001-1500000
1500001-1600000
1700001-1800000
1800001-1900000
1900001-2000000
2200001-2300000
1

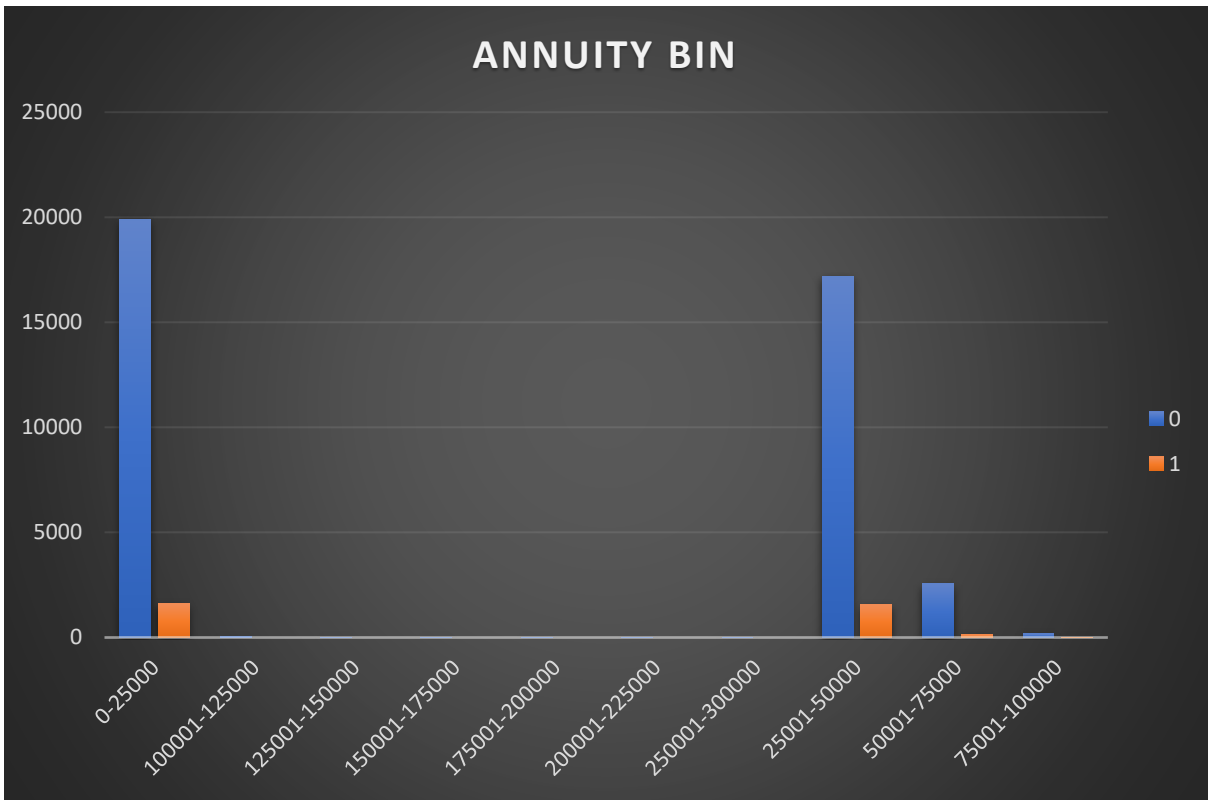

Avg_EXT_SOURCE bin
900
800
700
600
500
400
300
200
100
0
75
204
416
730
831
760
295
7
16
6
0.1-0.2
0.2-0.3
0.3-0.4
0.4-0.5
0.5-0.6
0.6-0.7
0.7-0.8
0.8-0.9
0-0.1
(blank)
1



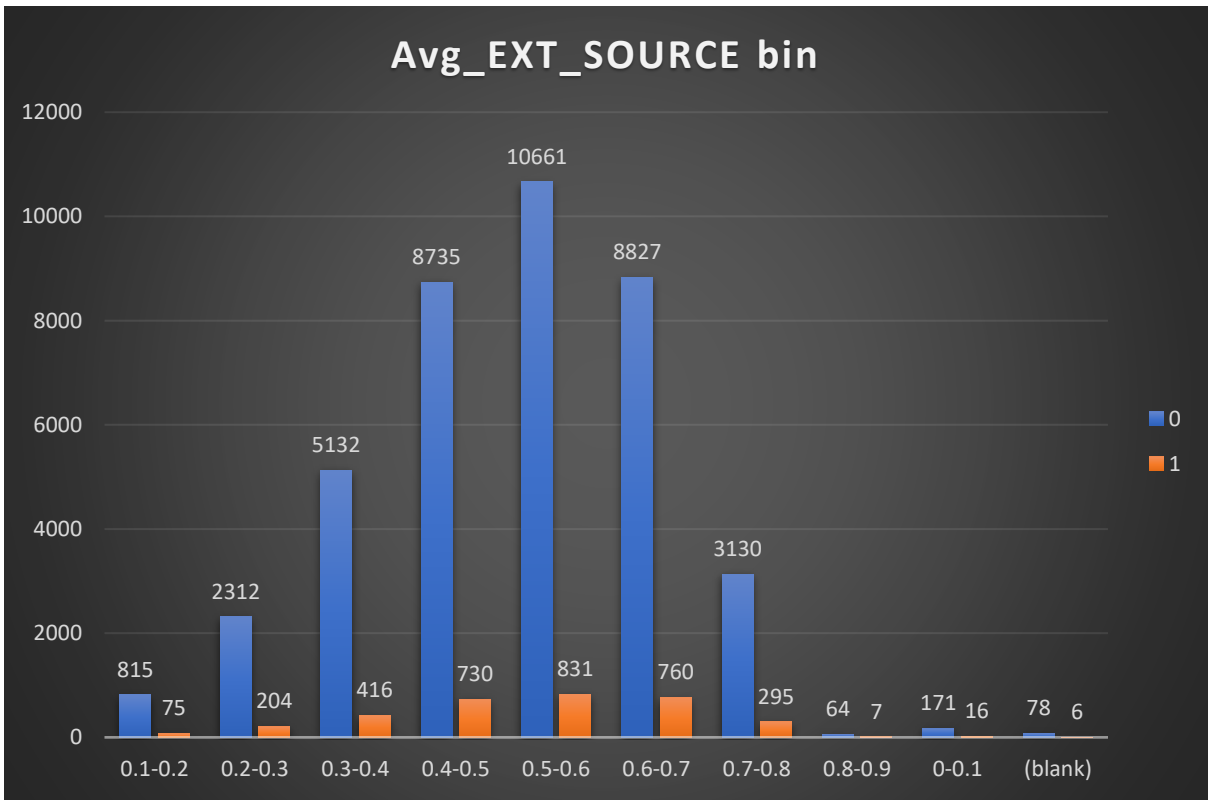
Following the univariate analysis, **segmented univariate analysis** was conducted, enabling both Target 0 and 1 against each of the six columns individually. This segmentation provided insights into how different values within each column impact loan default rates.

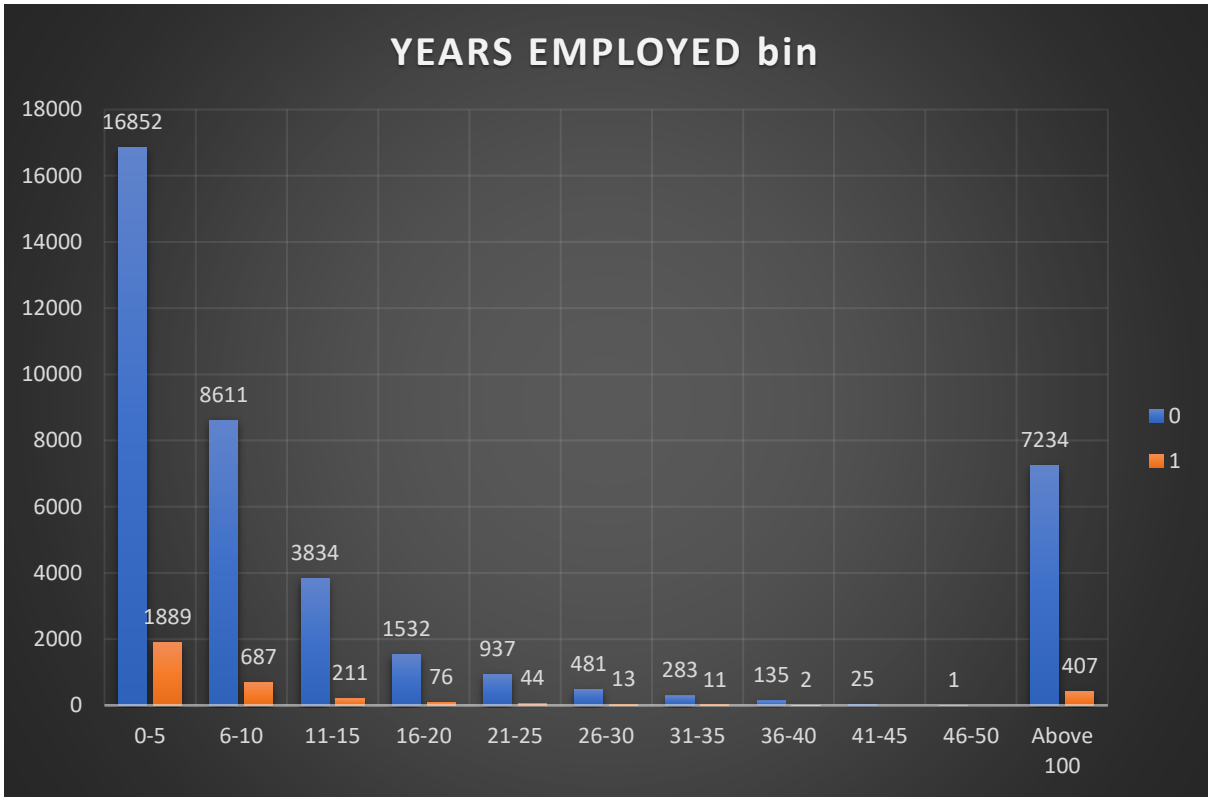



CREDIT BIN
5000
4500
4000
3500
3000
2500
2000
1500
1000
500
0
50001-75000
75001-100000
100001-125000
125001-150000
150001-175000
175001-200000
200001-225000
2200001-2300000
225001-250000
250001-300000
25001-50000
Above 3000000
300001-400000
400001-500000
500001-600000
600001-700000
700001-800000
800001-900000
900001-1000000
1000001-1100000
1100001-1200000
1200001-1300000
1300001-1400000
1400001-1500000
1500001-1600000
1600001-1700000
1700001-1800000
1800001-1900000
1900001-2000000
2000001-2100000
2100001-2200000
2300001-2400000
2400001-2500000
2500001-3000000
0
1

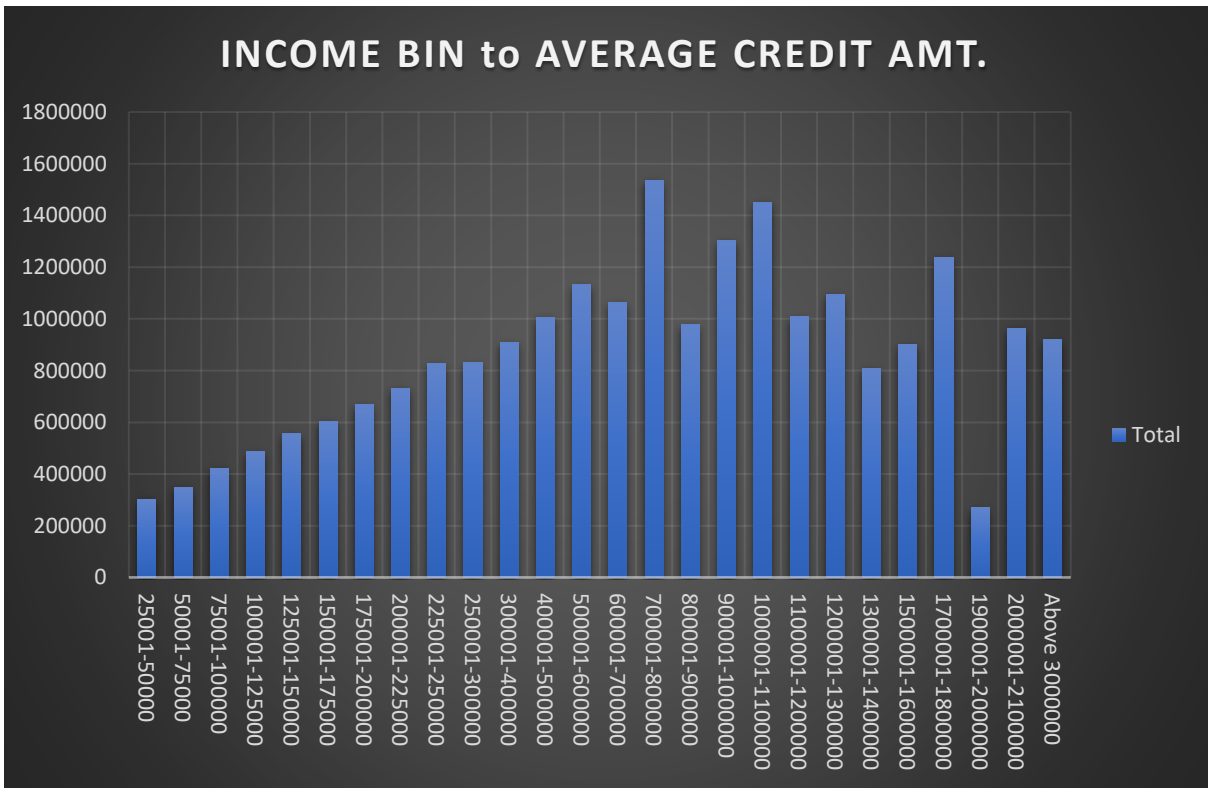

ANNUITY BIN
25000
20000
15000
10000
5000
0
0-25000
100001-125000
125001-150000
150001-175000
175001-200000
200001-225000
250001-300000
25001-50000
50001-75000
75001-100000
0
1

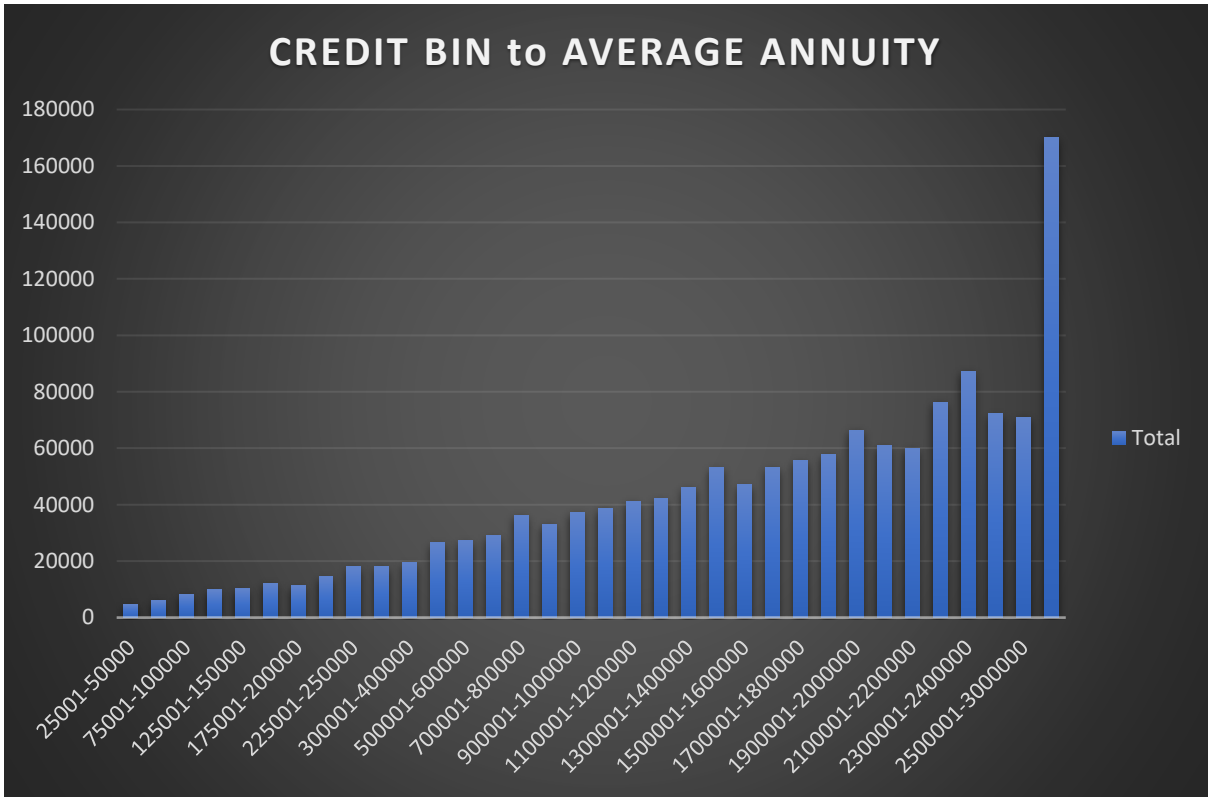

GOODS_PRICE BIN
8000
7000
6000
5000
4000
3000
2000
1000
0
25001-50000
50001-75000
75001-100000
100001-125000
125001-150000
150001-175000
160001-1700000
175001-200000
2000001-2100000
200001-225000
2100001-2200000
225001-250000
2400001-2500000
2500001-3000000
250001-300000
300001-400000
400001-500000
500001-600000
600001-700000
700001-800000
800001-900000
900001-1000000
1000001-1100000
1100001-1200000
1200001-1300000
1300001-1400000
1400001-1500000
1500001-1600000
Above 3000000
1700001-1800000
1800001-1900000
1900001-2000000
2200001-2300000
0
1

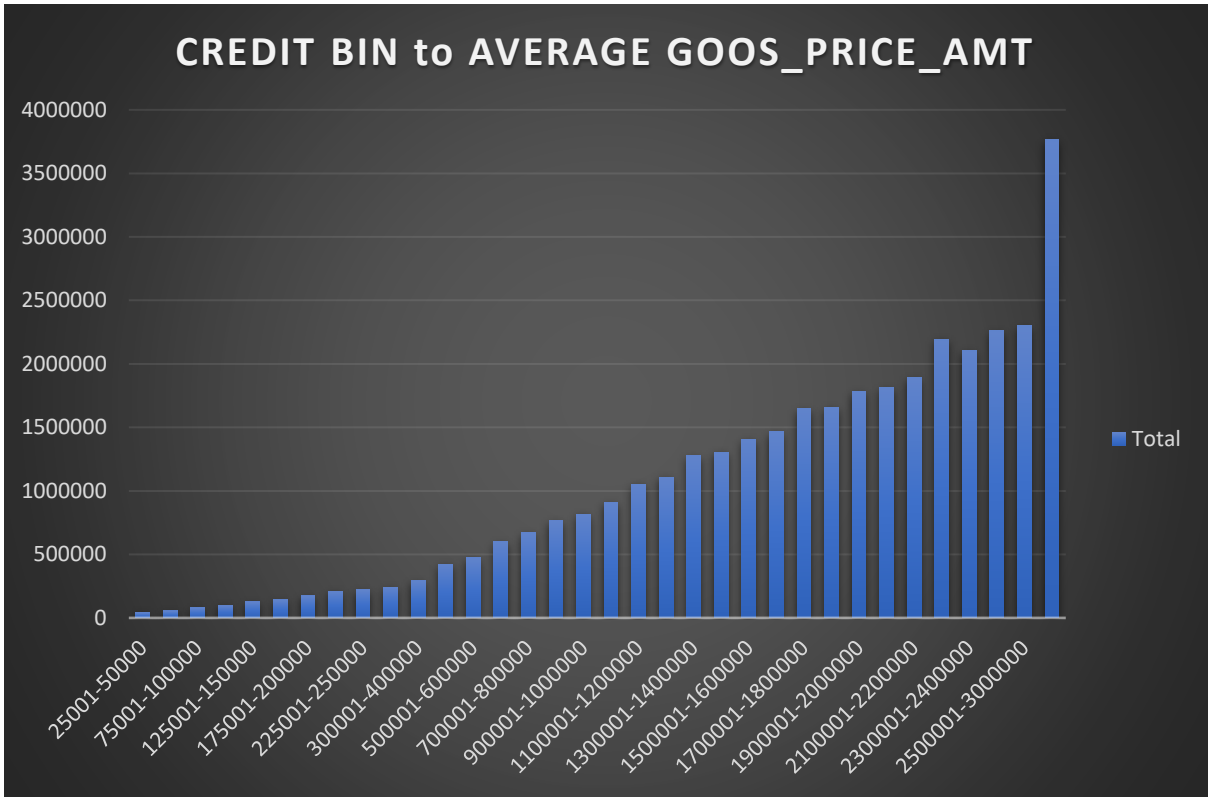

Avg_EXT_SOURCE bin
12000
10000
8000
6000
4000
2000
0
815
75
2312
204
5132
416
8735
730
10661
831
8827
760
3130
295
64
7
171
16
78
6
0.1-0.2
0.2-0.3
0.3-0.4
0.4-0.5
0.5-0.6
0.6-0.7
0.7-0.8
0.8-0.9
0-0.1
(blank)
0
1



Finally, **bivariate analysis** was conducted by examining relationships between certain column pairs. For instance, the relationship between income bin and average credit amount, credit bin and average annuity, and credit bin and average goods price amount were analyzed.







This comprehensive approach helped gain deeper insights into the factors influencing loan default and enabled informed decisions regarding loan approval processes.

E. Identify Top Correlations for Different Scenarios:

Top correlations were identified through three iterations of analysis:

1. **All Applicants:**
    - Avg_EXT_SOURCE: -0.226
    - YEARS_BIRTH: -0.080
    - REGION_RATING_CLIENT_W_CITY: 0.065
    - REGION_RATING_CLIENT: 0.064
    - DAYS_LAST_PHONE_CHANGE: 0.054

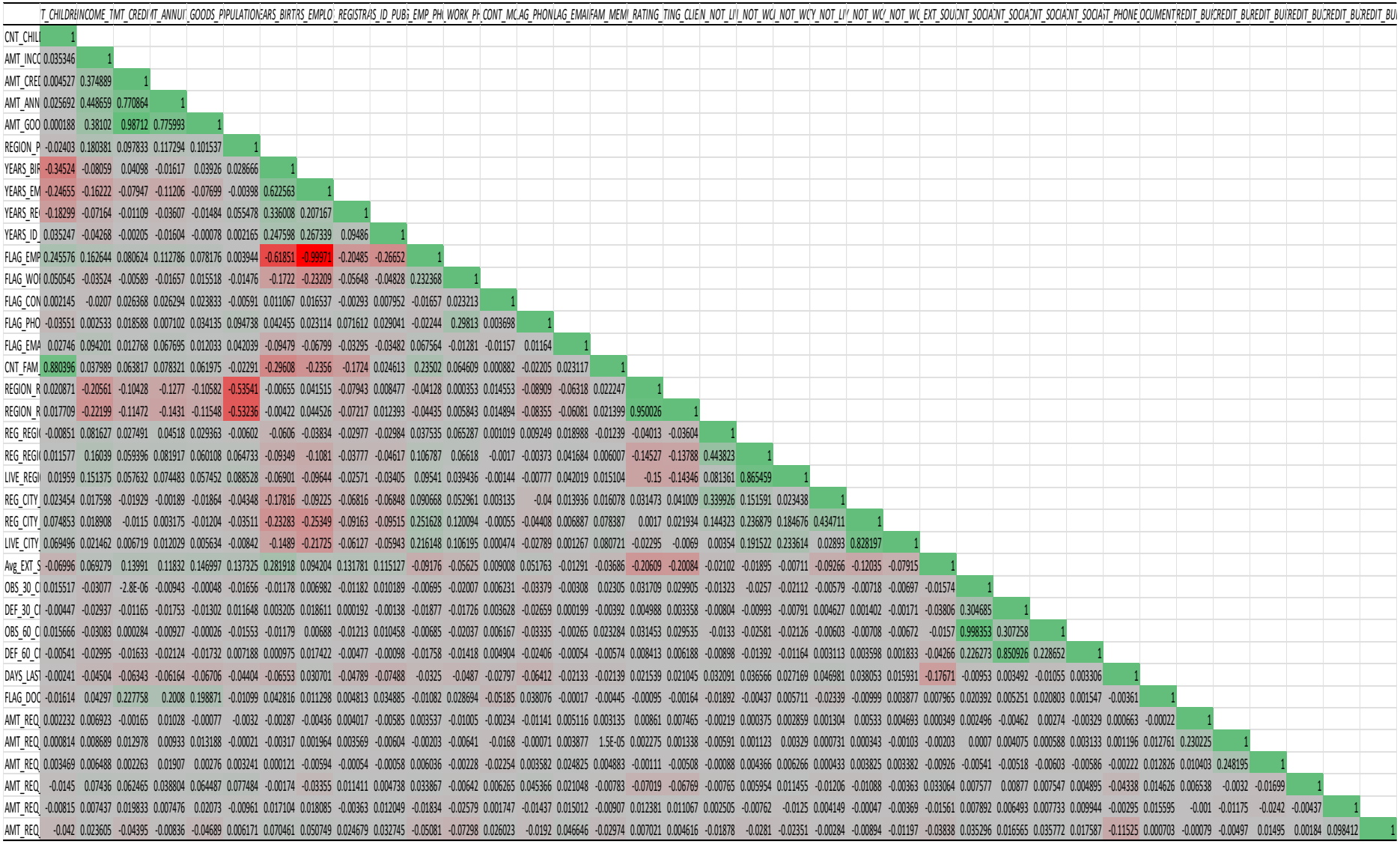
2. **Applicants with Payment Difficulties:**
    - FLAG_EMP_PHONE vs. YEARS_EMPLOYED: -0.999
    - OBS_60_CNT_SOCIAL_CIRCLE vs. OBS_30_CNT_SOCIAL_CIRCLE: 0.998
    - AMT_GOODS_PRICE vs. AMT_CREDIT: 0.982
    - REGION_RATING_CLIENT_W_CITY vs. REGION_RATING_CLIENT: 0.949
    - FLAG_EMP_PHONE vs. YEARS_BIRTH: -0.586
3. **All Other Cases:**
    - FLAG_EMP_PHONE vs. YEARS_EMPLOYED: -0.999
    - OBS_60_CNT_SOCIAL_CIRCLE vs. OBS_30_CNT_SOCIAL_CIRCLE: 0.998
    - AMT_GOODS_PRICE vs. AMT_CREDIT: 0.982
    - REGION_RATING_CLIENT_W_CITY vs. REGION_RATING_CLIENT: 0.949
    - FLAG_EMP_PHONE vs. YEARS_BIRTH: -0.586

These correlations provide valuable insights into the relationships between different variables under different scenarios, aiding in better decision-making processes.

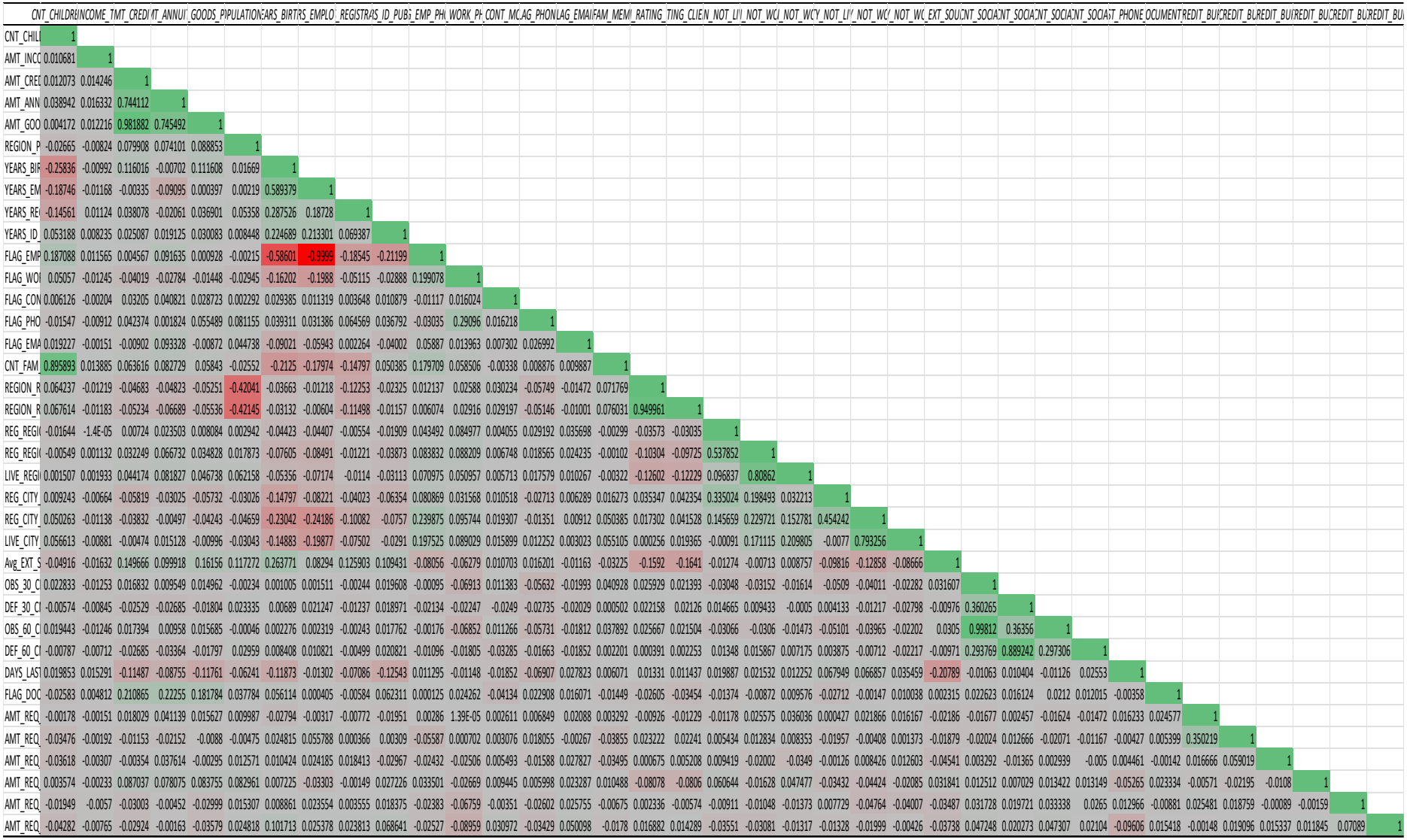
All variable with complete dataset

	TARGET	T_CHILDRE	INCOME_TM	T_CREDIT	T_ANNUIT	GOODS_PR	PULATION_	ARS_BIRTH	RS_EMPLO	_REGISTRA	S_ID_PUBL	_EMP_PH	WORK_PH	CONT_MO	AG_PHONE	LAG_EMAI	FAM_MEM	_RATING_	TING_CLIE	N_NOT_LI	_NOT_WO	NOT_WO	Y_NOT_LI	_NOT_WO	_NOT_WO	EXT_SOUR	NT_SOCIA	NT_SOCIA	NT_SOCIA	NT_SOCIA	T_PHONE	OCUMENT	REDIT_BU	REDIT_BU	REDIT_BU	REDIT_BU	REDIT_BU	REDIT_BU
TARGET	1	0.02883	0.013293	-0.03282	-0.01295	-0.04212	-0.03858	-0.07982	-0.04265	-0.03914	-0.04318	0.04156	0.017453	0.007608	-0.03515	-0.00083	0.016732	0.064078	0.065049	0.009161	-0.00098	-0.00615	0.0353	0.044985	0.031545	-0.22635	0.01749	0.043394	0.017325	0.04564	0.053953	0.021495	0.003258	0.011957	0.005731	-0.01136	-0.00081	0.02365
CNT_CHIL	0.02883	1	0.009269	0.004098	0.026168	-0.00075	-0.02524	-0.33888	-0.24249	-0.1807	0.035493	0.241545	0.051013	0.002567	-0.03484	0.026711	0.881809	0.026272	0.023709	-0.00901	0.010091	0.01792	0.023021	0.073763	0.069149	-0.07276	0.016654	-0.00335	0.016485	-0.00435	0.000897	-0.01609	0.001963	-0.00213	0.000181	-0.01353	-0.00917	-0.04136
AMT_INCO	0.013293	0.009269	1	0.064063	0.07689	0.06447	0.026987	-0.01667	-0.02976	-0.00928	-0.00517	0.029787	-0.00899	-0.00368	-0.00238	0.01479	0.010436	-0.03559	-0.03812	0.01315	0.0262	0.024958	0.001132	0.000243	0.001244	0.003087	-0.00826	-0.00696	-0.00824	-0.00664	-0.00275	0.008366	0.000718	0.000982	0.000264	0.011521	-0.00044	0.001894
AMT_CRE	-0.03282	0.004098	0.064063	1	0.769379	0.986809	0.097991	0.04803	-0.07376	-0.00686	0.001141	0.074958	-0.00874	0.026308	0.021134	0.011384	0.06304	-0.10241	-0.11251	0.02562	0.057603	0.056951	-0.02326	-0.0148	0.004863	0.143936	0.000538	-0.01407	0.000845	-0.01853	-0.06808	0.226006	-0.00042	0.010845	0.001691	0.064058	0.016477	-0.04361
AMT_ANN	-0.01295	0.026168	0.07689	0.769379	1	0.774177	0.115326	-0.01451	-0.11023	-0.0346	-0.01315	0.110971	-0.01752	0.026769	0.007234	0.069267	0.078187	-0.12315	-0.13868	0.043334	0.080867	0.074963	-0.00447	0.002011	0.011798	0.116622	-0.00838	-0.01869	-0.00823	-0.02264	-0.06375	0.201541	0.012316	0.007004	0.020182	0.040963	0.006668	-0.00819
AMT_GOO	-0.04212	-0.00075	0.06447	0.986809	0.774177	1	0.102311	0.046781	-0.07087	-0.01014	0.003004	0.072119	0.012768	0.023667	0.036795	0.010744	0.060777	-0.10481	-0.11403	0.027334	0.05844	0.056979	-0.02285	-0.01595	0.003219	0.153125	-0.00021	-0.01511	3.76E-05	-0.01907	-0.07205	0.196892	0.000182	0.011139	0.002146	0.065871	0.017378	-0.04698
REGION_P	-0.03858	-0.02524	0.026987	0.097991	0.115326	0.102311	1	0.030865	-0.00198	0.056775	0.004244	0.00199	-0.01641	-0.00584	0.095109	0.042182	-0.02367	-0.52864	-0.52581	-0.00569	0.061607	0.08702	-0.04358	-0.03752	-0.01118	0.140705	-0.01623	0.010843	-0.01514	0.007295	-0.04712	-0.00928	-0.0024	-0.001	0.00364	0.078138	-0.00784	0.006538
YEARS_BIF	-0.07982	-0.33888	-0.01667	0.04803	-0.01451	0.046781	0.030865	1	0.62105	0.334463	0.248281	-0.61699	-0.17218	0.011267	0.044881	-0.09408	-0.28953	-0.01381	-0.01136	-0.05971	-0.09183	-0.06722	-0.17735	-0.2351	-0.15077	0.290203	-0.01216	5.89E-05	-0.01206	-0.00195	-0.07313	0.041673	-0.00508	-0.00188	0.000439	-0.0003	0.016474	0.07074
YEARS_EM	-0.04265	-0.24249	-0.02976	-0.07376	-0.11023	-0.07087	-0.00198	0.62105	1	0.207224	0.264799	-0.99972	-0.22997	0.015943	0.025083	-0.06728	-0.23154	0.035154	0.038299	-0.03901	-0.10635	-0.0945	-0.09245	-0.25354	-0.21652	0.100349	0.005845	0.016818	0.005814	0.014705	0.025667	0.009798	-0.0044	0.005308	-0.00413	-0.033	0.018456	0.047823
YEARS_RE	-0.03914	-0.1807	-0.00928	-0.00686	-0.0346	-0.01014	0.056775	0.334463	0.207224	1	0.094513	-0.20491	-0.05669	-0.00291	0.072419	-0.03038	-0.17082	-0.08473	-0.07751	-0.02811	-0.03588	-0.02448	-0.06693	-0.09385	-0.06343	0.136549	-0.0118	-0.00264	-0.01208	-0.00654	-0.05141	0.003363	0.002994	0.002846	0.000622	0.011099	-0.00305	0.023646
YEARS_ID	-0.04318	0.035493	-0.00517	0.001141	-0.01315	0.003004	0.004244	0.248281	0.264799	0.094513	1	-0.26391	-0.0474	0.007723	0.031069	-0.03515	0.026045	0.003261	0.007708	-0.02924	-0.04551	-0.03353	-0.06937	-0.09524	-0.05813	0.121314	0.010182	-0.00124	0.010286	-0.0007	-0.08062	0.035528	-0.00712	-0.00578	-0.00315	0.00662	0.012588	0.034643
FLAG_EMP	0.04156	0.241545	0.029787	0.074958	0.110971	0.072119	0.00199	-0.61699	-0.99972	-0.20491	-0.26391	1	0.23023	-0.01597	-0.02435	0.066847	0.230999	-0.03501	-0.03821	0.038215	0.105065	0.093505	0.090866	0.251638	0.215376	-0.09774	-0.0058	-0.01701	-0.00577	-0.01491	-0.02752	-0.00934	0.003616	-0.00538	0.004214	0.03334	-0.01871	-0.04791
FLAG_WO	0.017453	0.051013	-0.00899	-0.00874	-0.01752	0.012768	-0.01641	-0.17218	-0.22997	-0.05669	-0.0474	0.23023	1	0.022819	0.296595	-0.01067	0.064344	0.003479	0.008806	0.067196	0.067907	0.040186	0.051481	0.118603	0.105173	-0.05929	-0.02383	-0.01698	-0.02405	-0.01373	-0.0449	0.028699	-0.00914	-0.00558	-0.00405	-0.00791	-0.02931	-0.07395
FLAG_CON	0.007608	0.002567	-0.00368	0.026308	0.026769	0.023667	-0.00584	0.011267	0.015943	-0.00291	0.007723	-0.01597	0.022819	1	0.004005	-0.01057	0.000769	0.015722	0.015999	0.001251	-0.00128	-0.00114	0.003804	0.000886	0.001553	0.007076	0.006594	0.00208	0.006527	0.002638	-0.027	-0.05122	-0.00204	-0.01549	-0.02093	0.006301	0.001453	0.026287
FLAG_PHO	-0.03515	-0.03484	-0.00238	0.021134	0.007234	0.036795	0.095109	0.044881	0.025083	0.072419	0.031069	-0.02435	0.296595	0.004005	1	0.012758	-0.02019	-0.08887	-0.08333	0.010515	-0.0021	-0.00582	-0.04001	-0.04316	-0.02581	0.055597	-0.03604	-0.02802	-0.03569	-0.02476	-0.06621	0.036398	-0.01012	0.000327	0.001952	0.043439	-0.0152	-0.02114
FLAG_EMA	-0.00083	0.026711	0.01479	0.011384	0.069267	0.010744	0.042182	-0.09408	-0.06728	-0.03038	-0.03515	0.066847	-0.01067	-0.01057	0.012758	1	0.021989	-0.05947	-0.05691	0.020406	0.040355	0.039709	0.013192	0.007027	0.001386	-0.01228	-0.00443	-0.00182	-0.00389	-0.0024	-0.01788	0.000796	0.006393	0.003322	0.025054	0.021161	0.015871	0.046885
CNT_FAM	0.016732	0.881809	0.010436	0.06304	0.078187	0.060777	-0.02367	-0.28953	-0.23154	-0.17082	0.026045	0.230999	0.064344	0.000769	-0.02019	0.021989	1	0.027315	0.026921	-0.01135	0.005403	0.013531	0.016671	0.076503	0.078873	-0.03928	0.02488	-0.00271	0.024823	-0.00407	-0.01832	-0.00473	0.003203	-0.00328	0.001576	-0.00676	-0.00888	-0.02827
REGION_R	0.064078	0.026272	-0.03559	-0.10241	-0.12315	-0.10481	-0.52864	-0.01381	0.035154	-0.08473	0.003261	-0.03501	0.003479	0.015722	-0.08887	-0.05947	0.027315	1	0.950228	-0.03905	-0.14187	-0.14836	0.033944	0.005854	-0.01897	-0.211	0.032306	0.009372	0.032041	0.01047	0.024348	-0.00106	0.007373	0.004743	-0.0006	-0.07131	0.011512	0.009303
REGION_R	0.065049	0.023709	-0.03812	-0.11251	-0.13868	-0.11403	-0.52581	-0.01136	0.038299	-0.07751	0.007708	-0.03821	0.008806	0.015999	-0.08333	-0.05691	0.026921	0.950228	1	-0.03486	-0.1346	-0.14203	0.04325	0.026392	-0.00267	-0.20692	0.030306	0.007871	0.029966	0.008713	0.02381	-0.0022	0.006076	0.003837	-0.00391	-0.06897	0.009663	0.006915
REG_REGI	0.009161	-0.00901	0.01315	0.02562	0.043334	0.027334	-0.00569	-0.05971	-0.03901	-0.02811	-0.02924	0.038215	0.067196	0.001251	0.010515	0.020406	-0.01135	-0.03905	-0.03486	1	0.45157	0.082467	0.339489	0.144701	0.003409	-0.02179	-0.0146	-0.00515	-0.01468	-0.00597	0.031512	-0.01828	-0.00304	-0.00474	8.57E-05	-0.00299	0.001448	-0.02008
REG_REGI	-0.00098	0.010091	0.0262	0.057603	0.080867	0.05844	0.061607	-0.09183	-0.10635	-0.03588	-0.04551	0.105065	0.067907	-0.00128	-0.0021	0.040355	0.005403	-0.14187	-0.1346	0.45157	1	0.861246	0.155467	0.235913	0.189637	-0.01728	-0.02618	-0.00806	-0.0262	-0.01087	0.035396	-0.00464	0.002417	0.002081	0.002455	0.004593	-0.00785	-0.02833
LIVE_REGI	-0.00615	0.01792	0.024958	0.056951	0.074963	0.056979	0.08702	-0.06722	-0.0945	-0.02448	-0.03353	0.093505	0.040186	-0.00114	-0.00582	0.039709	0.013531	-0.14836	-0.14203	0.082467	0.861246	1	0.023909	0.181528	0.231234	-0.0043	-0.02085	-0.00745	-0.02086	-0.01002	0.025796	0.005794	0.005407	0.003614	0.003159	0.01362	-0.01258	-0.02284
REG_CITY	0.0353	0.023021	0.001132	-0.02326	-0.00447	-0.02285	-0.04358	-0.17735	-0.09245	-0.06693	-0.06937	0.090866	0.051481	0.003804	-0.04001	0.013192	0.016671	0.033944	0.04325	0.339489	0.155467	0.023909	1	0.437368	0.026416	-0.09869	-0.00937	0.006094	-0.00959	0.004807	0.050397	-0.02273	0.001334	-0.00083	0.000478	-0.01399	0.004452	-0.003
REG_CITY	0.044985	0.073763	0.000243	-0.0148	0.002011	-0.01595	-0.03752	-0.2351	-0.25354	-0.09385	-0.09524	0.251638	0.118603	0.000886	-0.04316	0.007027	0.076503	0.005854	0.026392	0.144701	0.235913	0.181528	0.437368	1	0.825196	-0.12802	-0.00923	0.001914	-0.0091	0.004449	0.042551	-0.00842	0.006937	0.00048	0.004471	-0.01358	-0.00462	-0.00884
LIVE_CITY	0.031545	0.069149	0.001244	0.004863	0.011798	0.003219	-0.01118	-0.15077	-0.21652	-0.06343	-0.05813	0.215376	0.105173	0.001553	-0.02581	0.001386	0.078873	-0.01897	-0.00267	0.003409	0.189637	0.231234	0.026416	0.825196	1	-0.08486	-0.00778	-0.00311	-0.00749	0.000615	0.019078	0.004942	0.005808	-0.00043	0.004346	-0.00513	-0.00689	-0.01054
Avg_EXT_S	-0.22635	-0.07276	0.003087	0.143936	0.116622	0.153125	0.140705	0.290203	0.100349	0.136549	0.121314	-0.09774	-0.05929	0.007076	0.055597	-0.01228	-0.03928	-0.211	-0.20692	-0.02179	-0.01728	-0.0043	-0.09869	-0.12802	-0.08486	1	-0.0153	-0.0439	-0.01534	-0.04819	-0.18622	0.002501	-0.00231	-0.00616	-0.01332	0.034586	-0.01666	-0.04264
OBS_30_C	0.01749	0.016654	-0.00826	0.000538	-0.00838	-0.00021	-0.01623	-0.01216	0.005845	-0.0118	0.010182	-0.0058	-0.02383	0.006594	-0.03604	-0.00443	0.02488	0.032306	0.030306	-0.0146	-0.02618	-0.02085	-0.00937	-0.00923	-0.00778	-0.0153	1	0.31	0.998334	0.233083	-0.00864	0.020849	0.000927	-0.0009	-0.0046	0.007674	0.009857	0.036699
DEF_30_C	0.043394	-0.00335	-0.00696	-0.01407	-0.01869	-0.01511	0.010843	5.89E-05	0.016818	-0.00264	-0.00124	-0.01701	-0.01698	0.00208	-0.02802	-0.00182	-0.00271	0.009372	0.007871	-0.00515	-0.00806	-0.00745	0.006094	0.001914	-0.00311	-0.0439	0.31	1	0.312594	0.856026	0.006424	0.00696	-0.00373	0.005489	-0.00576	0.008053	0.007783	0.017922
OBS_60_C	0.017325	0.016485	-0.00824	0.000845	-0.00823	3.76E-05	-0.01514	-0.01206	0.005814	-0.01208	0.010286	-0.00577	-0.02405	0.006527	-0.03569	-0.00389	0.024823	0.032041	0.029966	-0.01468	-0.0262	-0.02086	-0.00959	-0.0091	-0.00749	-0.01534	0.998334	0.312594	1	0.235525	-0.00964	0.021141	0.0012	-0.00103	-0.00521	0.007705	0.009839	0.037132
DEF_60_C	0.04564	-0.00435	-0.00664	-0.01853	-0.02264	-0.01907	0.007295	-0.00195	0.014705	-0.00654	-0.0007	-0.01491	-0.01373	0.002638	-0.02476	-0.0024	-0.00407	0.01047	0.008713	-0.00597	-0.01087	-0.01002	0.004807	0.004449	0.000615	-0.04819	0.233083	0.856026	0.235525	1	0.007812	0.003346	-0.00436	0.002036	-0.00548	0.004987	0.011624	0.018961
DAYS_LAS	0.053953	0.000897	-0.00275	-0.06808	-0.06375	-0.07205	-0.04712	-0.07313	0.025667	-0.05141	-0.08062	-0.02752	-0.0449	-0.027	-0.06621	-0.01788	-0.01832	0.024348	0.02381	0.031512	0.035396	0.025796	0.050397	0.042551	0.019078	-0.18622	-0.00864	0.006424	-0.00964	0.007812	1	-0.00244	0.002002	0.001421	-0.00143	-0.04442	-0.00182	-0.11225
FLAG_DOC	0.021495	-0.01609	0.008366	0.226006	0.201541	0.196892	-0.00928	0.041673	0.009798	0.003363	0.035528	-0.00934	0.028699	-0.05122	0.036398	0.000796	-0.00473	-0.00106	-0.0022	-0.01828	-0.00464	0.005794	-0.02273	-0.00842	0.004942	0.002501	0.020849	0.00696	0.021141	0.003346	-0.00244	1	0.001443	0.012485	0.012036	0.0148	0.013987	0.002151
AMT_REQ	0.003258	0.001963	0.000718	-0.00042	0.012316	0.000182	-0.0024	-0.00508	-0.0044	0.002994	-0.00712	0.003616	-0.00914	-0.00204	-0.01012	0.006393	0.003203	0.007373	0.006076	-0.00304	0.002417	0.005407	0.001334	0.006937	0.005808	-0.00231	0.000927	-0.00373	0.0012	-0.00436	0.002002	0.001443	1	0.24084	0.010943	0.005682	0.001251	-0.00077
AMT_REQ	0.011957	-0.00213	0.000982	0.010845	0.007004	0.011139	-0.001	-0.00188	0.005308	0.002846	-0.00578	-0.00538	-0.00558	-0.01549	0.000327	0.003322	-0.00328	0.004743	0.003837	-0.00474	0.002081	0.003614	-0.00083	0.00048	-0.00043	-0.00616	-0.0009	0.005489	-0.00103	0.002036	0.001421	0.012485	0.24084	1	0.232114	-0.00457	-0.0091	-0.00256
AMT_REQ	0.005731	0.000181	0.000264	0.001691	0.020182	0.002146	0.00364	0.000439	-0.00413	0.000622	-0.00315	0.004214	-0.00405	-0.02093	0.001952	0.025054	0.001576	-0.0006	-0.00391	8.57E-05	0.002455	0.003159	0.000478	0.004471	0.004346	-0.01332	-0.0046	-0.00576	-0.00521	-0.00548	-0.00143	0.012036	0.010943	0.232114	1	-0.01663	-0.02229	0.015113
AMT_REQ	-0.01136	-0.01353	0.011521	0.064058	0.040963	0.065871	0.078138	-0.0003	-0.033	0.011099	0.00662	0.03334	-0.00791	0.006301	0.043439	0.021161	-0.00676	-0.07131	-0.06897	-0.00299	0.004593	0.01362	-0.01399	-0.01358	-0.00513	0.034586	0.007674	0.008053	0.007705	0.004987	-0.04442	0.0148	0.005682	-0.00457	-0.01663	1	-0.00418	0.002218
AMT_REQ	-0.00081	-0.00917	-0.00044	0.016477	0.006668	0.017378	-0.00784	0.016474	0.018456	-0.00305	0.012588	-0.01871	-0.02931	0.001453	-0.0152	0.015871	-0.00888	0.011512	0.009663	0.001448	-0.00785	-0.01258	0.004452	-0.00462	-0.00689	-0.01666	0.009857	0.007783	0.009839	0.011624	-0.00182	0.013987	0.001251	-0.0091	-0.02229	-0.00418	1	0.096027
AMT_REQ	0.02365	-0.04136	0.001894	-0.04361	-0.00819	-0.04698	0.006538	0.07074	0.047823	0.023646	0.034643	-0.04791	-0.07395	0.026287	-0.02114	0.046885	-0.02827	0.009303	0.006915	-0.02008	-0.02833	-0.02284	-0.003	-0.00884	-0.01054	-0.04264	0.036699	0.017922	0.037132	0.018961	-0.11225	0.002151	-0.00077	-0.00256	0.015113	0.002218	0.096027	1

All variable with datatset for Target 0



All variable with datatset for Target 1



Insights

- Within the application data file, a significant majority of applications, accounting for 90.68%, were registered for cash loans, whereas 9.32% of applications were earmarked for revolving loans. This distribution underscores the prevalent preference for cash loans among applicants.
- The Average External Sources (Avg EXT Sources) emerged as a pivotal feature in assessing loan defaults. Introducing additional features akin to Avg EXT Sources could enhance the model's robustness, enabling a more comprehensive analysis of loan repayment probabilities.
- An intriguing insight from the dataset's 'Name Income Type' column indicates that 17.66% of applicants identify as pensioners. This demographic segment represents a substantial portion of the applicant pool, warranting targeted strategies tailored to their unique financial circumstances.
- In loan borrower analysis, the region rating assumes paramount importance. Regions with lower ratings often face heightened scrutiny during the loan approval process, potentially leading to application rejections. Juxtaposing region rating with metrics like OBS_30_CNT_SOCIAL_CIRCLE and OBS_60_CNT_SOCIAL_CIRCLE could yield valuable insights. Such an approach could help identify applicants associated with groups of defaulters, thereby safeguarding against potential exploitation of financial institution resources.
- An intriguing observation stems from the 'DAYS_LAST_PHONE_CHANGE' column in the dataset. Among defaulted applications, a noteworthy 14.79% of individuals altered their phone numbers on the same day as their application, as indicated by the unique value '0' in this column. This finding suggests a potential correlation between immediate phone number changes and loan default, warranting further investigation into the underlying causes.

Results

A. The original dataset comprised 122 features. However, a significant portion of these features contained less than 50% of the data. Following a thorough evaluation, 49 columns were retained for in-depth analysis.

B. Outliers were identified primarily in the "Income Amount" and "Years Employed" features. Notably, an anomaly was detected in the "Years Employed" feature, with instances where the reported employment duration exceeded plausible limits, such as 1001 years, suggesting potential data entry errors.

C. The feature "FLAG_CONT_MOBILE" exhibited extreme data imbalance, with a minority proportion of only 0.21%.

D. Univariate, segmented univariate, and bivariate analyses yielded invaluable insights. These analyses involved the creation of new columns containing bin values derived from principal features. This approach facilitated a comprehensive examination of various data attributes and their relationships.

E. Correlation analysis with the target variable proved instrumental in gauging feature significance. Notably, the "Avg Ext Source" feature demonstrated noteworthy correlation performance, indicating its potential significance in predicting loan default probabilities.

For Excel File Please click here

For Loom Video Please click here