Bank Loan Case Study

Project Description:

The **objective** of this Bank Loan Case Study is to use Exploratory Data Analysis (EDA) to identify key factors that influence loan default, enabling the finance company to make better decisions on loan approvals. Specifically, the goals are to:

1. **Reduce Financial Risk**: Identify customers who are likely to default on their loans so the company can either reject their applications, reduce loan amounts, or adjust the interest rates to mitigate risks.
2. **Maximize Business Opportunities**: Ensure that capable applicants are not wrongly rejected, allowing the company to approve loans for clients who are financially stable, thus maximizing profit.
3. **Understand Patterns**: Analyze customer and loan attributes to detect patterns related to loan default, such as income level, loan amount, credit history, and annuity payments.
4. **Improve Decision-Making**: Provide insights into which variables are the strongest predictors of loan default to enhance the company's loan approval process, reducing both approval risks and missed opportunities.

By achieving these objectives, the company aims to improve overall lending performance while maintaining financial health.

Approach:

To achieve the objective of identifying patterns that influence loan default, the approach will involve structured steps focusing on data exploration, analysis, and insights extraction using Exploratory Data Analysis (EDA).

Step 1: Data Understanding and Preparation

1. **Data Collection**:
   - Obtain the loan application dataset, which includes customer and loan attributes such as income, loan amount, loan status (TARGET), and payment history.
2. **Data Inspection**:
   - Examine the dataset's structure, types of variables (categorical and numerical), and content.

    - Ensure familiarity with key variables, particularly the TARGET variable that indicates loan default (1 for default, 0 for non-default).
3. **Data Cleaning**:
    - **Handle Missing Data**: Identify and deal with missing values using techniques like:
        - Removing rows or columns with excessive missing values.
        - Imputing missing data using averages, medians, or domain-specific values.
    - **Outlier Detection**: Identify outliers using statistical methods (e.g., interquartile range (IQR), z-scores) and assess whether they need to be removed or treated.

Tools: Excel functions like COUNTIF, IF, AVERAGE, MEDIAN, QUARTILE, and conditional formatting.

Step 2: Data Exploration (EDA)

1. **Univariate Analysis**:
    - **Goal**: Understand the distribution of individual variables (e.g., income, loan amount).
    - **Techniques**:
        - Use descriptive statistics (mean, median, mode, standard deviation).
        - Create histograms and bar charts to visualize the distribution of numerical and categorical variables.
2. **Segmented Univariate Analysis**:
    - **Goal**: Compare variable distributions for clients with payment difficulties (TARGET = 1) and without payment difficulties (TARGET = 0).
    - **Techniques**:
        - Use pivot tables and filtering to create comparisons between the two segments.
        - Visualize the differences using stacked or grouped bar charts.
3. **Bivariate Analysis**:
    - **Goal**: Explore relationships between variables and their impact on loan default.
    - **Techniques**:
        - Correlation analysis for numerical variables (e.g., correlation between income and loan amount).
        - Scatter plots or heatmaps to visualize relationships between variables.
        - Analyze categorical variables using cross-tabulations.

Step 3: Data Segmentation

1. **Segment Data by Loan Status**:
    - Create segments based on TARGET (loan default or non-default) and analyze each segment separately.
    - Calculate the proportion of each class (TARGET = 0 and TARGET = 1) to identify potential data imbalance.
2. **Identify Correlations within Segments**:
    - Perform correlation analysis within each segment to determine which variables (e.g., AMT_CREDIT, AMT_ANNUITY) have the strongest relationship with loan default.
    - Rank the variables by correlation strength to identify the top indicators of default for each segment.

Step 4: Analyze Data Imbalance

1. **Goal**: Assess if there is an imbalance between the number of default and non-default cases.
2. **Techniques**:
    - Calculate the ratio of default to non-default cases using COUNTIF.
    - Visualize the class distribution using pie charts or bar charts.

Step 5: Insights and Interpretation

1. **Top Correlations**:
    - Identify the variables that have the highest correlation with loan default in both segments (default and non-default).
    - Focus on strong positive and negative correlations to highlight key indicators of financial risk (e.g., loan amount, income level).
2. **Outlier Investigation**:
    - Investigate whether outliers (e.g., very high or low incomes) significantly affect default rates.

Step 6: Visualization

1. **Correlation Heatmaps**:

- Create heatmaps to visualize correlations across variables in both segments, highlighting strong correlations.

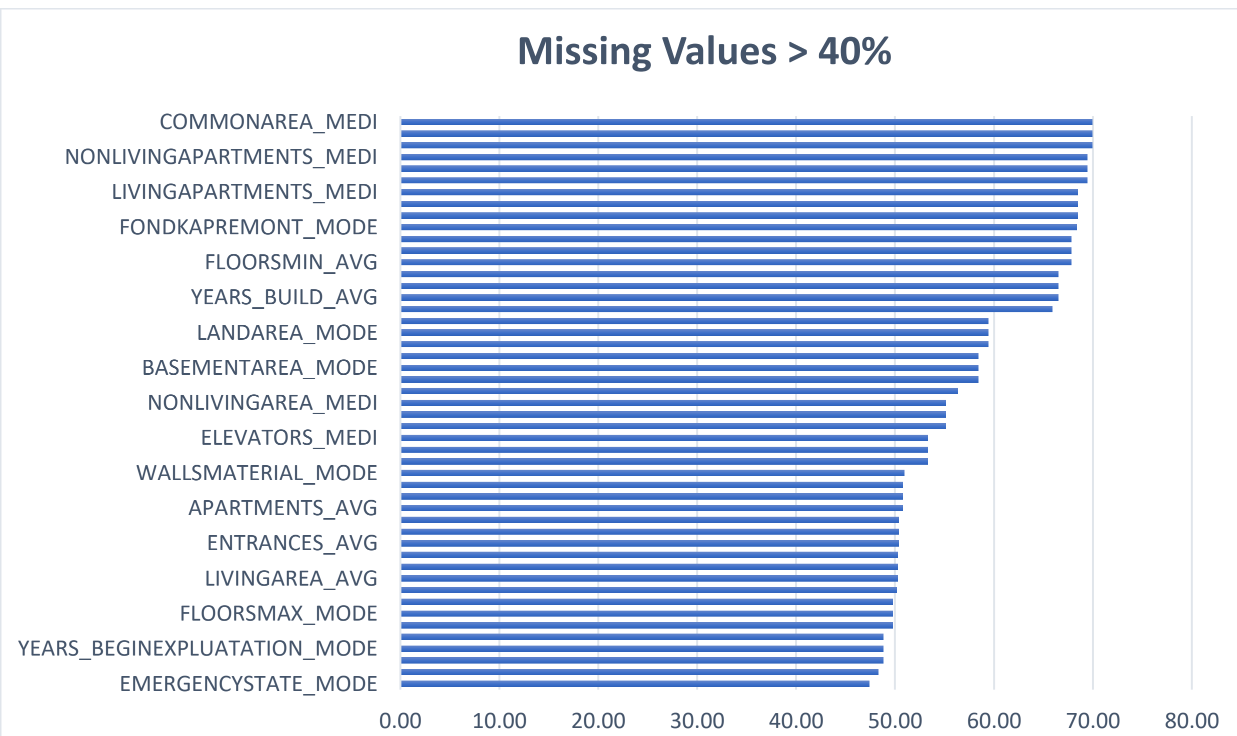
2. **Bar Charts and Box Plots**:
   - Visualize the distribution of key variables like loan amount, income, and loan status using bar charts or box plots to gain a better understanding of their influence on default.

Tech-Stack Used:

1. Microsoft excels

Insights:

A. Identify Missing Data and Deal with it Appropriately: Identified the missing data in the dataset and handle it with using excel function COUNT, ISBLANK, and IF. Also Perform imputation using excel function AVERAGE or MEDIAN.

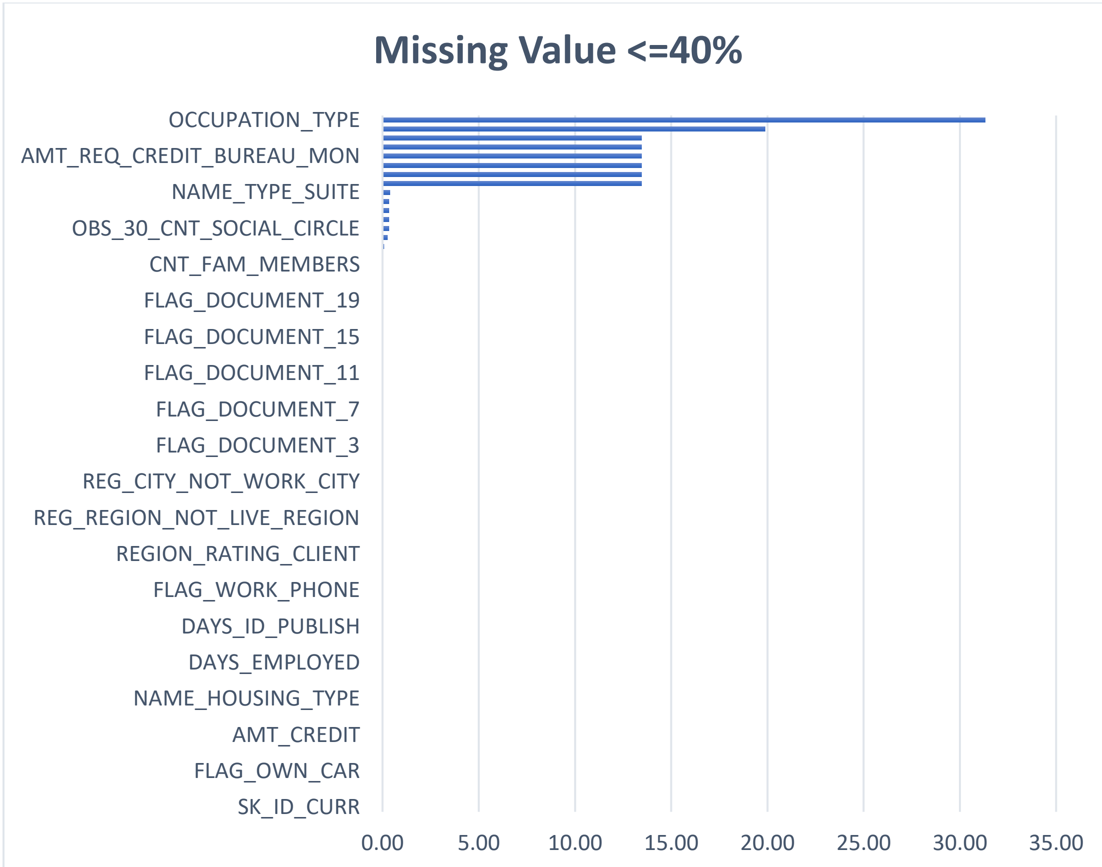


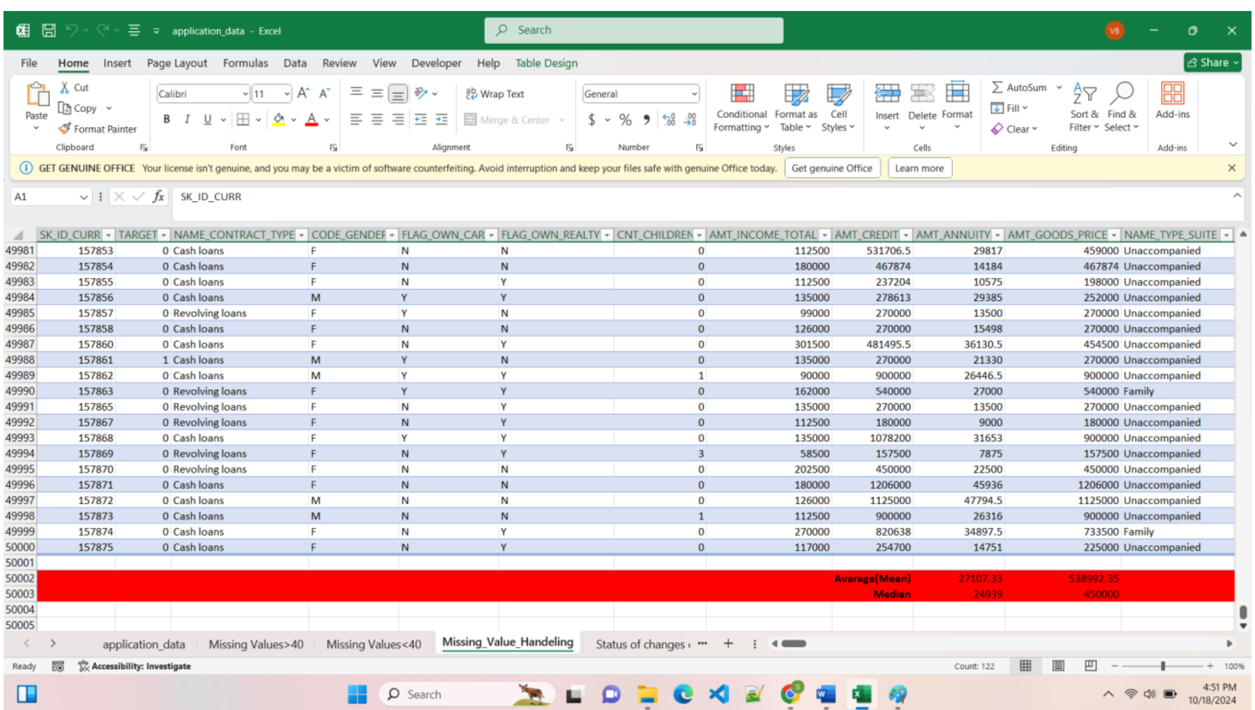
Column_Name	Blank_Percentage
EMERGENCYSTATE_MODE	47.40
TOTALAREA_MODE	48.30
YEARS_BEGINEXPLUATATION_AVG	48.79
YEARS_BEGINEXPLUATATION_MODE	48.79
YEARS_BEGINEXPLUATATION_MEDI	48.79
FLOORSMAX_AVG	49.75
FLOORSMAX_MODE	49.75
FLOORSMAX_MEDI	49.75
HOUSETYPE_MODE	50.15
LIVINGAREA_AVG	50.28
LIVINGAREA_MODE	50.28
LIVINGAREA_MEDI	50.28
ENTRANCES_AVG	50.39
ENTRANCES_MODE	50.39
ENTRANCES_MEDI	50.39
APARTMENTS_AVG	50.77
APARTMENTS_MODE	50.77
APARTMENTS_MEDI	50.77
WALLSMATERIAL_MODE	50.92
ELEVATORS_AVG	53.30
ELEVATORS_MODE	53.30
ELEVATORS_MEDI	53.30
NONLIVINGAREA_AVG	55.15
NONLIVINGAREA_MODE	55.15
NONLIVINGAREA_MEDI	55.15
EXT_SOURCE_1	56.35
BASEMENTAREA_AVG	58.40
BASEMENTAREA_MODE	58.40
BASEMENTAREA_MEDI	58.40
LANDAREA_AVG	59.44
LANDAREA_MODE	59.44
LANDAREA_MEDI	59.44
OWN_CAR_AGE	65.90
YEARS_BUILD_AVG	66.48
YEARS_BUILD_MODE	66.48
YEARS_BUILD_MEDI	66.48
FLOORSMIN_AVG	67.79
FLOORSMIN_MODE	67.79
FLOORSMIN_MEDI	67.79
FONDKAPREMONT_MODE	68.38
LIVINGAPARTMENTS_AVG	68.45
LIVINGAPARTMENTS_MODE	68.45
LIVINGAPARTMENTS_MEDI	68.45
NONLIVINGAPARTMENTS_AVG	69.43
NONLIVINGAPARTMENTS_MODE	69.43
NONLIVINGAPARTMENTS_MEDI	69.43
COMMONAREA_AVG	69.92
COMMONAREA_MODE	69.92
COMMONAREA_MEDI	69.92

Fig. Missing Value >40

Column_Name	Blank_Percentage
SK_ID_CURR	0.00
TARGET	0.00
NAME_CONTRACT_TYPE	0.00
CODE_GENDER	0.00
FLAG_OWN_CAR	0.00
FLAG_OWN_REALTY	0.00
CNT_CHILDREN	0.00
AMT_INCOME_TOTAL	0.00
AMT_CREDIT	0.00
NAME_INCOME_TYPE	0.00
NAME_EDUCATION_TYPE	0.00
NAME_FAMILY_STATUS	0.00
NAME_HOUSING_TYPE	0.00
REGION_POPULATION_RELATIVE	0.00
DAYS_BIRTH	0.00
CUSTOMER_AGE	0.00
DAYS_EMPLOYED	0.00
EMPLOYMENT	0.00
DAYS_REGISTRATION	0.00
REGISTRATION_DAYS	0.00
DAYS_ID_PUBLISH	0.00
ID_PUBLISED_DAYS	0.00
FLAG_MOBIL	0.00
FLAG_EMP_PHONE	0.00
FLAG_WORK_PHONE	0.00
FLAG_CONT_MOBILE	0.00
FLAG_PHONE	0.00
FLAG_EMAIL	0.00
REGION_RATING_CLIENT	0.00
REGION_RATING_CLIENT_W_CITY	0.00
WEEKDAY_APPR_PROCESS_START	0.00
HOUR_APPR_PROCESS_START	0.00
REG_REGION_NOT_LIVE_REGION	0.00
REG_REGION_NOT_WORK_REGION	0.00
LIVE_REGION_NOT_WORK_REGION	0.00
REG_CITY_NOT_LIVE_CITY	0.00
REG_CITY_NOT_WORK_CITY	0.00
LIVE_CITY_NOT_WORK_CITY	0.00
ORGANIZATION_TYPE	0.00
FLAG_DOCUMENT_2	0.00
FLAG_DOCUMENT_3	0.00
FLAG_DOCUMENT_4	0.00
FLAG_DOCUMENT_5	0.00
FLAG_DOCUMENT_6	0.00
FLAG_DOCUMENT_7	0.00
FLAG_DOCUMENT_8	0.00
FLAG_DOCUMENT_9	0.00
FLAG_DOCUMENT_10	0.00
FLAG_DOCUMENT_11	0.00
FLAG_DOCUMENT_12	0.00
FLAG_DOCUMENT_13	0.00
FLAG_DOCUMENT_14	0.00
FLAG_DOCUMENT_15	0.00
FLAG_DOCUMENT_16	0.00
FLAG_DOCUMENT_17	0.00
FLAG_DOCUMENT_18	0.00
FLAG_DOCUMENT_19	0.00
FLAG_DOCUMENT_20	0.00
FLAG_DOCUMENT_21	0.00
AMT_ANNUITY	0.00
CNT_FAM_MEMBERS	0.00
DAYS_LAST_PHONE_CHANGE	0.00
AMT_GOODS_PRICE	0.08
EXT_SOURCE_2	0.25
OBS_30_CNT_SOCIAL_CIRCLE	0.34
DEF_30_CNT_SOCIAL_CIRCLE	0.34
OBS_60_CNT_SOCIAL_CIRCLE	0.34
DEF_60_CNT_SOCIAL_CIRCLE	0.34
NAME_TYPE_SUITE	0.38
AMT_REQ_CREDIT_BUREAU_HOUR	13.47
AMT_REQ_CREDIT_BUREAU_DAY	13.47
AMT_REQ_CREDIT_BUREAU_WEEK	13.47
AMT_REQ_CREDIT_BUREAU_MON	13.47
AMT_REQ_CREDIT_BUREAU_QRT	13.46826937
AMT_REQ_CREDIT_BUREAU_YEAR	13.46826937
EXT_SOURCE_3	19.89
OCCUPATION_TYPE	31.31

Fig. Missing Value<40%





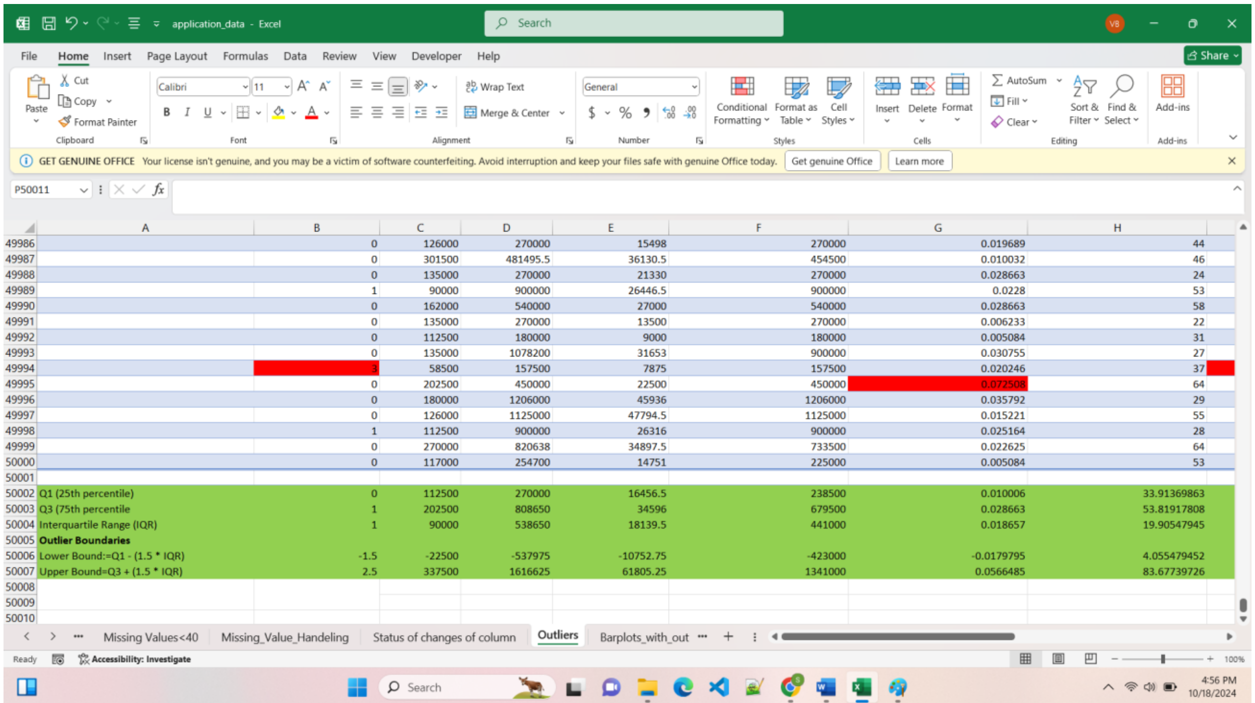
	SK_ID_CURR	TARGET	NAME_CONTRACT_TYPE	CODE_GENDEF	FLAG_OWN_CAR	FLAG_OWN_REALTY	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRICE	NAME_TYPE_SUITE
49981	157853	0	Cash loans	F	N	N	0	112500	531706.5	29817	459000	Unaccompanied
49982	157854	0	Cash loans	F	N	N	0	180000	467874	14184	467874	Unaccompanied
49983	157855	0	Cash loans	F	N	Y	0	112500	237204	10575	198000	Unaccompanied
49984	157856	0	Cash loans	M	Y	Y	0	135000	278613	29385	252000	Unaccompanied
49985	157857	0	Revolving loans	F	Y	N	0	99000	270000	13500	270000	Unaccompanied
49986	157858	0	Cash loans	F	N	N	0	126000	270000	15498	270000	Unaccompanied
49987	157860	0	Cash loans	F	N	Y	0	301500	481495.5	36130.5	454500	Unaccompanied
49988	157861	1	Cash loans	M	Y	N	0	135000	270000	21330	270000	Unaccompanied
49989	157862	0	Cash loans	M	Y	Y	1	90000	900000	26446.5	900000	Unaccompanied
49990	157863	0	Revolving loans	F	Y	Y	0	162000	540000	27000	540000	Family
49991	157865	0	Revolving loans	F	N	Y	0	135000	270000	13500	270000	Unaccompanied
49992	157867	0	Revolving loans	F	N	Y	0	112500	180000	9000	180000	Unaccompanied
49993	157868	0	Cash loans	F	Y	Y	0	135000	1078200	31653	900000	Unaccompanied
49994	157869	0	Revolving loans	F	N	Y	3	58500	157500	7875	157500	Unaccompanied
49995	157870	0	Revolving loans	F	N	N	0	202500	450000	22500	450000	Unaccompanied
49996	157871	0	Cash loans	F	N	N	0	180000	1206000	45936	1206000	Unaccompanied
49997	157872	0	Cash loans	M	N	N	0	126000	1125000	47794.5	1125000	Unaccompanied
49998	157873	0	Cash loans	M	N	N	1	112500	900000	26316	900000	Unaccompanied
49999	157874	0	Cash loans	F	N	Y	0	270000	820638	34897.5	733500	Family
50000	157875	0	Cash loans	F	N	Y	0	117000	254700	14751	225000	Unaccompanied
50001												
50002									Avarage(Mean)	27107.33	538992.35	
50003									Median	24939	450000	
50004												
50005												

Fig. Missing value Handling

Category	Count of missing value	Status Of Missing Data
AMT_ANNUITY	1	Replace with median value 24939
AMT_GOODS_PRICE	38	Calculate median and replace null values to that median
NAME_TYPE_SUITE	192	On the basis of majority "Unaccompanied", null value replaced with it
OWN_CAR_AGE	32950	Not found relevent data in previous loan sheet so it was deleted
OCCUPATION_TYPE	15654	replaced with "NA"
CNT_FAM_MEMBERS	1	Replaced with 0 (ZERO)
EXT_SOURCE_1	28172	Calculate median and replace null values to that median
EXT_SOURCE_2	126	Replace with median value
EXT_SOURCE_3	9944	Replace with median value
APARTMENTS_AVG	25385	Replace with median value
BASEMENTAREA_AVG	29199	Replace with median value
YEARS_BEGINEXPLUATATION_AVG	24394	Replace with median value
YEARS_BUILD_AVG	33239	Replace with median value
COMMONAREA_AVG	34960	Replace with median value
ELEVATORS_AVG	26651	Replace with median value
ENTRANCES_AVG	25195	Replace with median value
FLOORSMAX_AVG	24875	Replace with median value
FLOORSMIN_AVG	33894	Replace with median value
LANDAREA_AVG	29721	Replace with median value
LIVINGAPARTMENTS_AVG	34226	Replace with median value
LIVINGAREA_AVG	25137	Replace with median value
NONLIVINGAPARTMENTS_AVG	34714	Replace with median value
NONLIVINGAREA_AVG	27572	Replace with median value
APARTMENTS_MODE	25385	Replace with median value
BASEMENTAREA_MODE	29199	Replace with median value
YEARS_BEGINEXPLUATATION_MODE	24394	Replace with median value
YEARS_BUILD_MODE	33239	Replace with median value
COMMONAREA_MODE	34960	Replace with median value
ELEVATORS_MODE	26651	Replace with median value
ENTRANCES_MODE	25195	Replace with median value
FLOORSMAX_MODE	24875	Replace with median value
FLOORSMIN_MODE	33894	Replace with median value
LANDAREA_MODE	29721	Replace with median value
LIVINGAPARTMENTS_MODE	34226	Replace with median value
LIVINGAREA_MODE	25137	Replace with median value
NONLIVINGAPARTMENTS_MODE	34714	Replace with median value
NONLIVINGAREA_MODE	27572	Replace with median value
APARTMENTS_MEDI	25385	Replace with median value
BASEMENTAREA_MEDI	29199	Replace with median value
YEARS_BEGINEXPLUATATION_MEDI	24394	Replace with median value
YEARS_BUILD_MEDI	33239	Replace with median value
COMMONAREA_MEDI	34960	Replace with median value
ELEVATORS_MEDI	26651	Replace with median value
ENTRANCES_MEDI	25195	Replace with median value
FLOORSMAX_MEDI	24875	Replace with median value
FLOORSMIN_MEDI	33894	Replace with median value
LANDAREA_MEDI	29721	Replace with median value
LIVINGAPARTMENTS_MEDI	34226	Replace with median value
LIVINGAREA_MEDI	25137	Replace with median value
NONLIVINGAPARTMENTS_MEDI	34714	Replace with median value
NONLIVINGAREA_MEDI	27572	Replace with median value
FONDKAPREMONT_MODE	34191	Column deleted
HOUSETYPE_MODE	25075	Column deleted

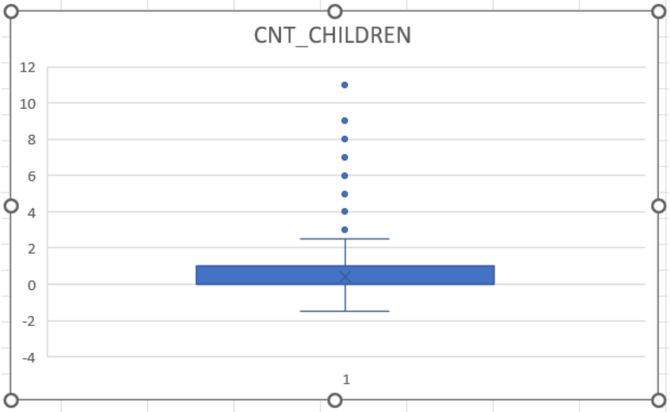
Fig. Missing value status

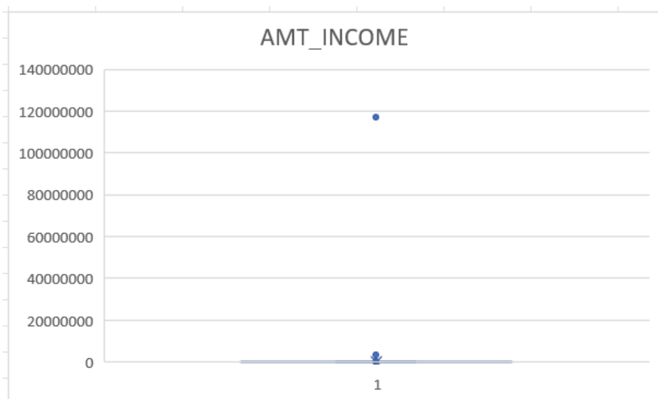
B. Identify Outliers in the Dataset: Identified the outliers using the IQR method, which is a common statistical technique. The IQR represents the middle 50% of the data, and outliers are typically values that fall below Q1 - 1.5 * IQR or above Q3 + 1.5 * IQR.

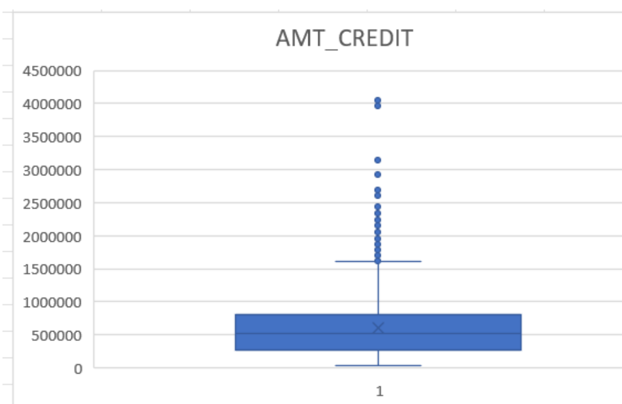


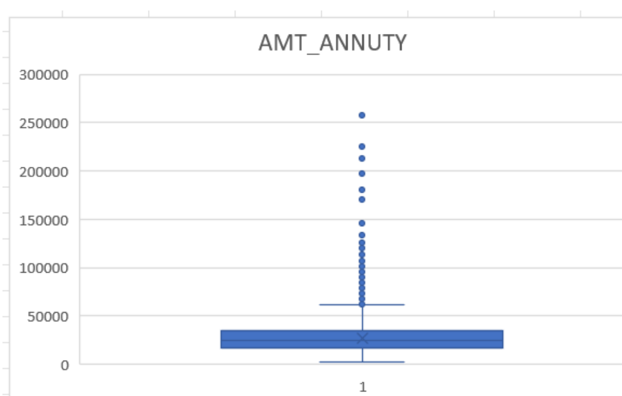
	A	B	C	D	E	F	G	H
49986		0	126000	270000	15498	270000	0.019689	44
49987		0	301500	481495.5	36130.5	454500	0.010032	46
49988		0	135000	270000	21330	270000	0.028663	24
49989		1	90000	900000	26446.5	900000	0.0228	53
49990		0	162000	540000	27000	540000	0.028663	58
49991		0	135000	270000	13500	270000	0.006233	22
49992		0	112500	180000	9000	180000	0.005084	31
49993		0	135000	1078200	31653	900000	0.030755	27
49994		3	58500	157500	7875	157500	0.020246	37
49995		0	202500	450000	22500	450000	0.072508	64
49996		0	180000	1206000	45936	1206000	0.035792	29
49997		0	126000	1125000	47794.5	1125000	0.015221	55
49998		1	112500	900000	26316	900000	0.025164	28
49999		0	270000	820638	34897.5	733500	0.022625	64
50000		0	117000	254700	14751	225000	0.005084	53
50001								
50002	Q1 (25th percentile)	0	112500	270000	16456.5	238500	0.010006	33.91369863
50003	Q3 (75th percentile	1	202500	808650	34596	679500	0.028663	53.81917808
50004	Interquartile Range (IQR)	1	90000	538650	18139.5	441000	0.018657	19.90547945
50005	**Outlier Boundaries**							
50006	Lower Bound:=Q1 - (1.5 * IQR)	-1.5	-22500	-537975	-10752.75	-423000	-0.0179795	4.055479452
50007	Upper Bound=Q3 + (1.5 * IQR)	2.5	337500	1616625	61805.25	1341000	0.0566485	83.67739726
50008								
50009								
50010								

Fig. Outlier Detection

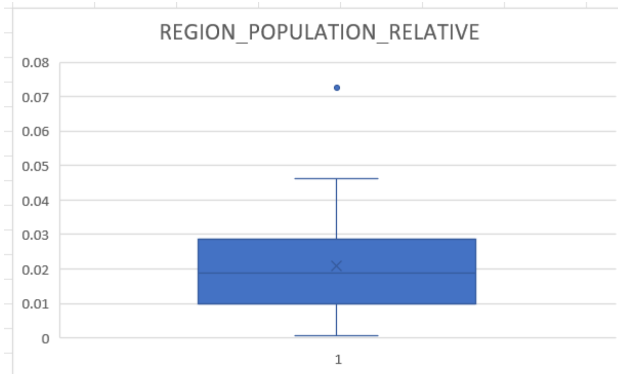

CNT_CHILDREN
12
10
8
6
4
2
0
-2
-4
1


AMT_INCOME
140000000
120000000
100000000
80000000
60000000
40000000
20000000
0
1

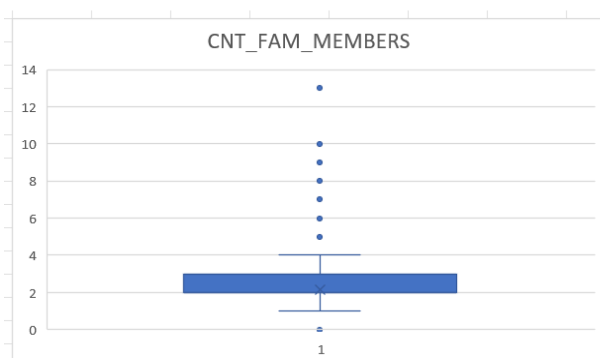

AMT_CREDIT
4500000
4000000
3500000
3000000
2500000
2000000
1500000
1000000
500000
0
1

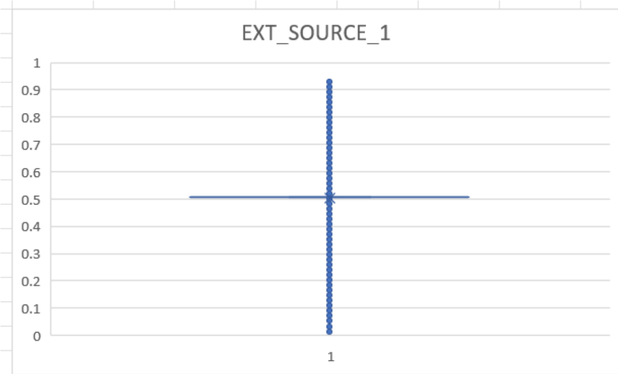

AMT_ANNUTY
300000
250000
200000
150000
100000
50000
0
1

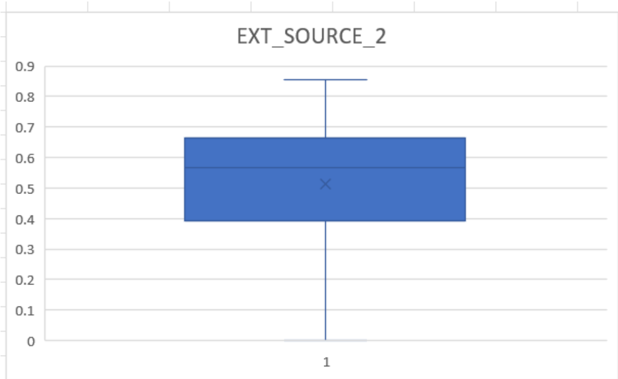

AMT_GOODS_PRICE
4500000
4000000
3500000
3000000
2500000
2000000
1500000
1000000
500000
0
1

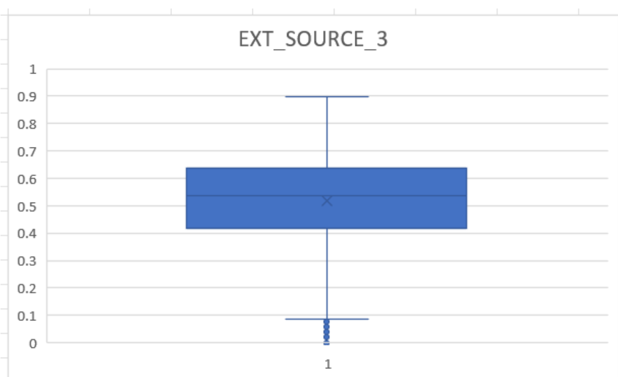

REGION_POPULATION_RELATIVE
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0
1


CUSTOMER_AGE
80
70
60
50
40
30
20
10
0
1

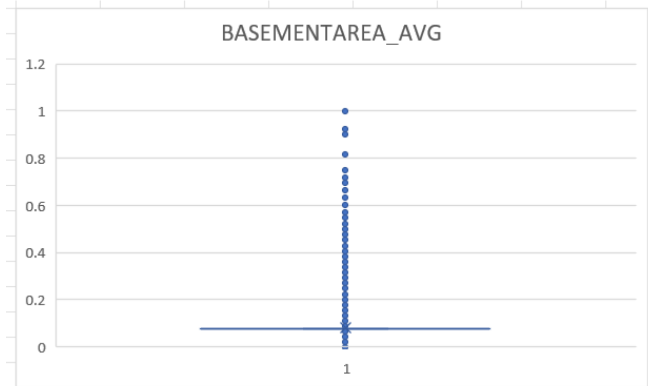

CNT_FAM_MEMBERS
14
12
10
8
6
4
2
0
1

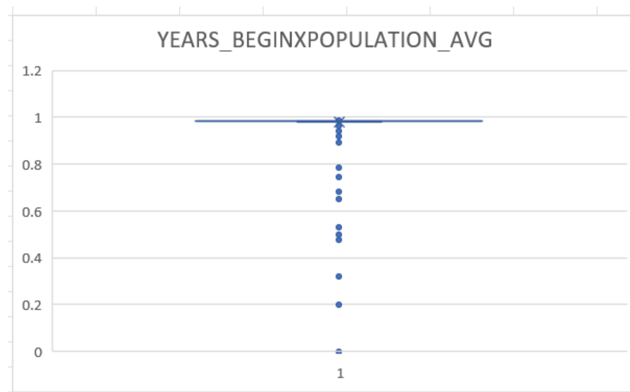

EXT_SOURCE_1
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
1

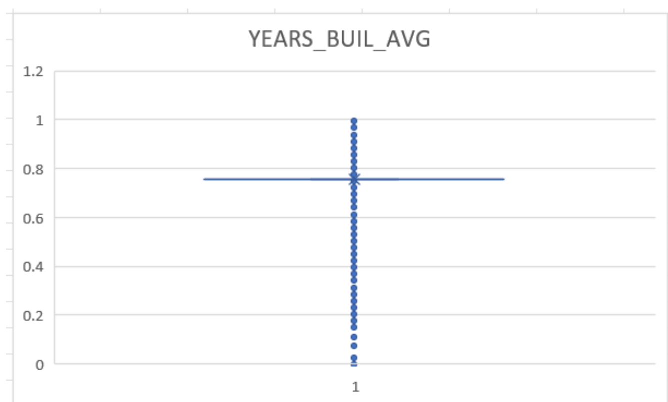

EXT_SOURCE_2
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
1

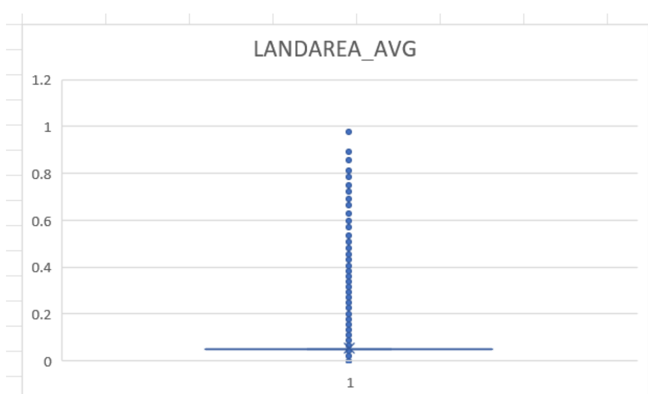

EXT_SOURCE_3
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
1


APPARTMENT_AVG
1.2
1
0.8
0.6
0.4
0.2
0
1


BASEMENTAREA_AVG
1.2
1
0.8
0.6
0.4
0.2
0
1


YEARS_BEGINXPOPULATION_AVG
1.2
1
0.8
0.6
0.4
0.2
0
1


YEARS_BUIL_AVG
1.2
1
0.8
0.6
0.4
0.2
0
1


LANDAREA_AVG
1.2
1
0.8
0.6
0.4
0.2
0
1

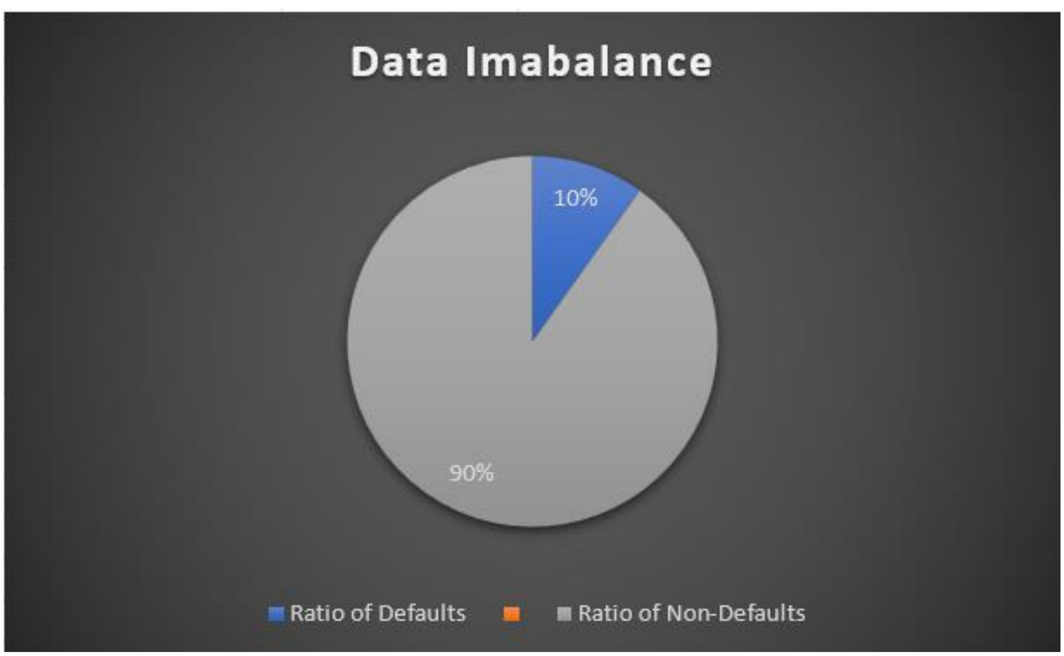
C. Analyze Data Imbalance: Determined data imbalance in the loan application dataset and calculated the ratio of data imbalance using Excel function COUNTIF and SUM to calculate the proportions of each class.

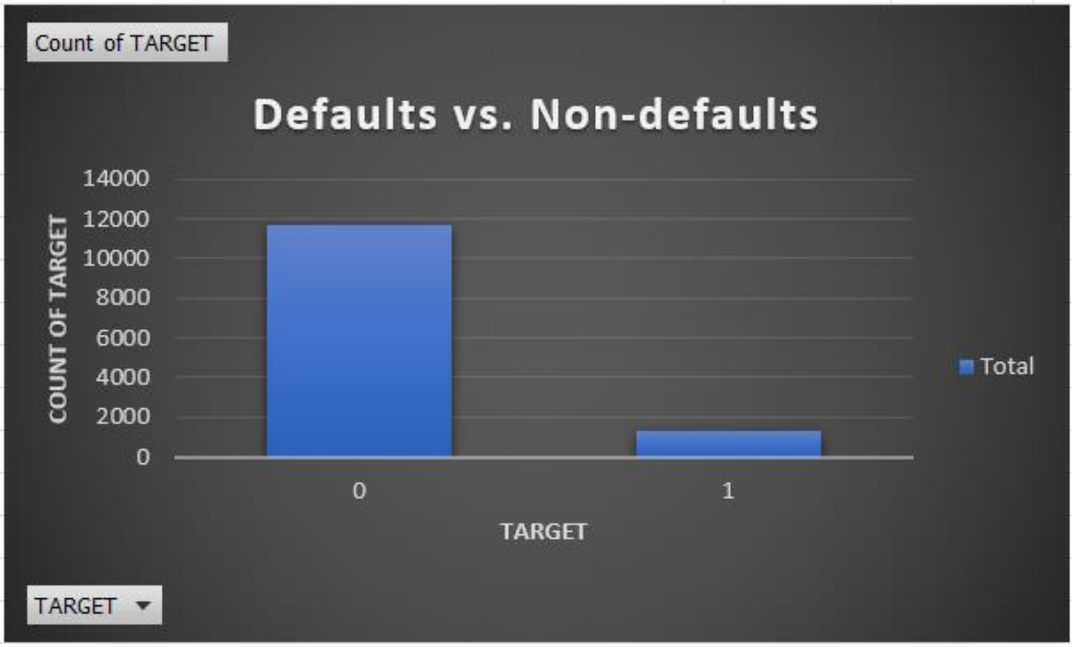
Class Proportions

Row Labels	Count of TARGET
0	11698
1	1277
Grand Total	**12975**

Note: 1 represents customers with repayment difficulties (defaulted)
0 represents customers without repayment difficulties (non-defaulted)

Target	Ratio	Interpret Data Imbalance
Ratio of Defaults	10%	the ratio is significantly skewed (e.g., 90% non-defaults and 10% defaults), the dataset is imbalanced.
Ratio of Non-Defaults	90%	





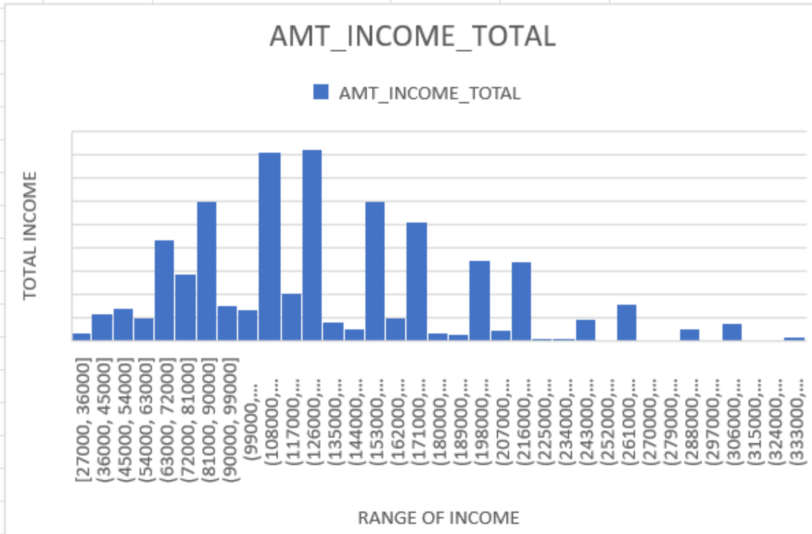
D. Perform Univariate, Segmented Univariate, and Bivariate Analysis:

1. **Univariate Analysis:** Univariate analysis involves examining the distribution of a single variable at a time to understand its characteristics, such as central tendency, variability, and shape.

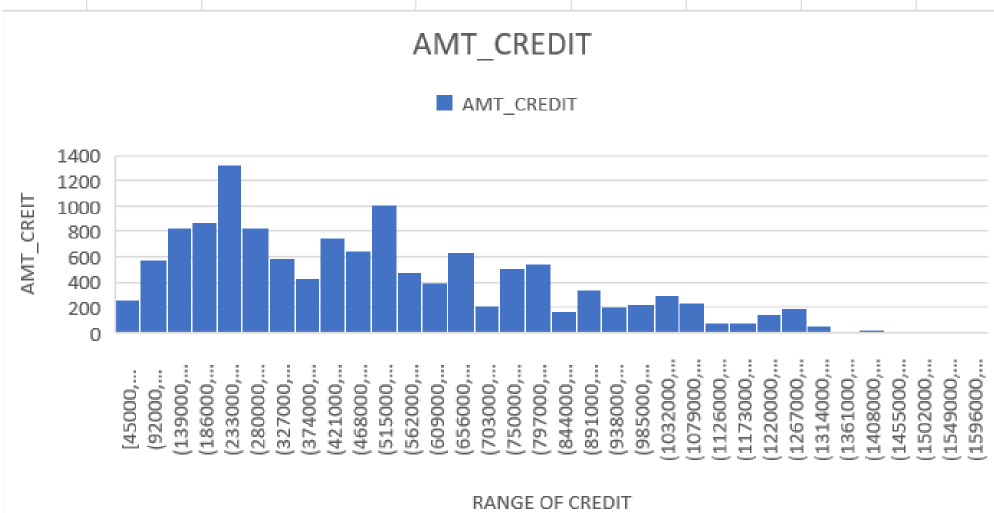
AMT_INCOME_TOTAL	AMT_CREDIT	COAD_GENDER	NAME_CONTRACT_TYPE
67500	135000	M	Revolving loans
135000	312682.5	F	Cash loans
121500	513000	M	Cash loans
99000	490495.5	M	Cash loans
135000	405000	M	Revolving loans
81000	270000	F	Revolving loans
90000	544491	F	Cash loans
112500	327024	M	Cash loans
202500	604152	F	Cash loans
202500	661702.5	M	Cash loans
90000	180000	F	Revolving loans
202500	305221.5	F	Cash loans
99000	260640	F	Cash loans
67500	298728	F	Cash loans
157500	755190	M	Cash loans
135000	675000	F	Cash loans
202500	1288350	F	Cash loans
112500	135000	F	Revolving loans
81000	252000	F	Cash loans
157500	760225.5	M	Cash loans
225000	270000	M	Revolving loans
72000	450000	F	Cash loans
126000	263686.5	F	Cash loans
135000	391194	M	Cash loans

Fig. Categorial and numerical data for Univariate analysis

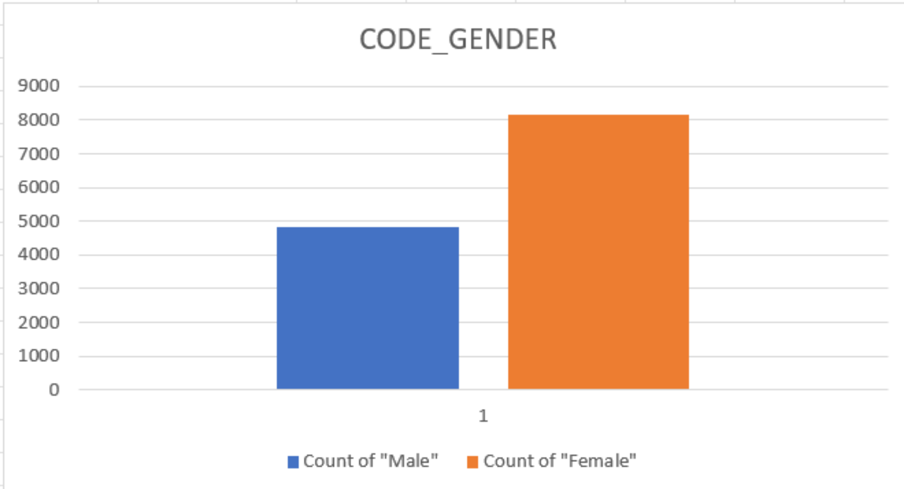
AMT_INCOME_TOTAL	
Count of entries	12975
Mean	140177.8
Median	135000
Standard Deviation	60543.95



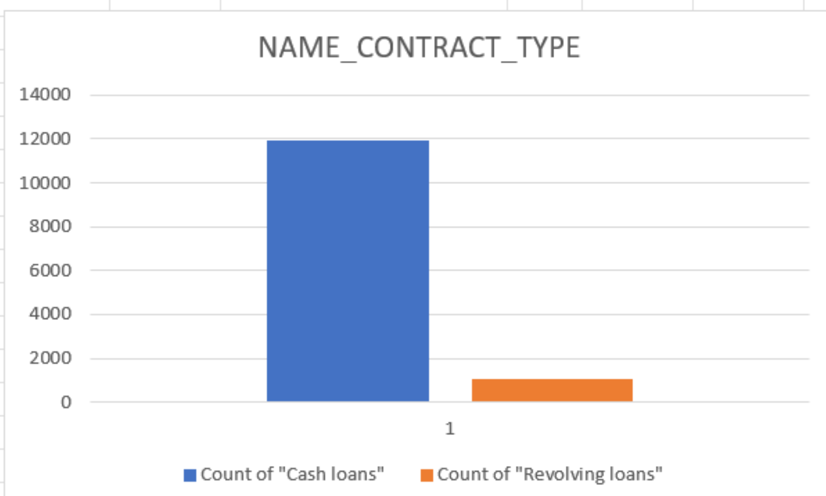
AMT_CREDIT	
Count of entries	12975
Mean	521347.6741
Median	472500
Standard Deviation	314813.3162



COAD_GENDER	
Count of "Male"	4833
Count of "Female"	8141



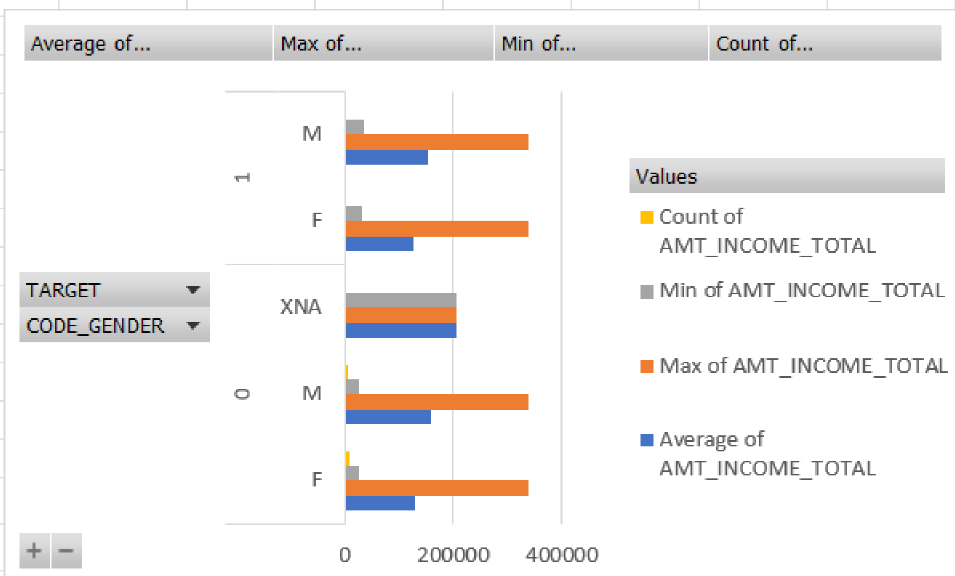
NAME_CONTRACT_TYPE	
Count of "Cash loans"	11936
Count of "Revolving loans"	1038



2. **Segmented Univariate Analysis:** Segmented univariate analysis compares the distribution of a variable across different groups or scenarios, such as comparing AMT_INCOME_TOTAL for loan defaults vs. non-defaults.

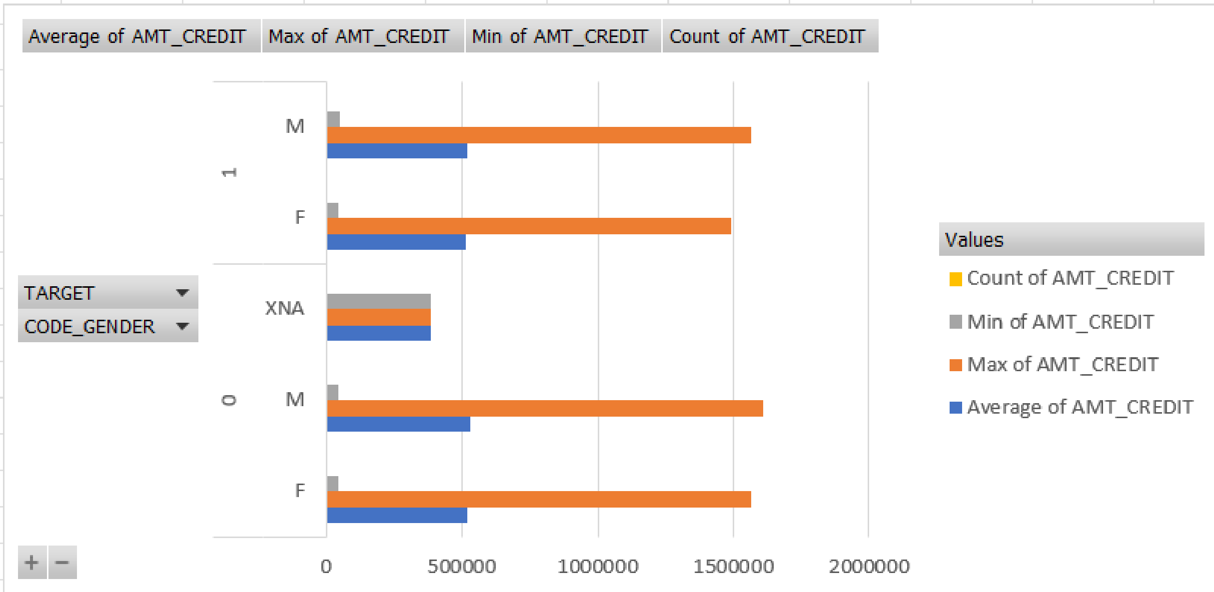
Comparison of Total Income with Gender and Default and Non Default Loan Payer

Row Labels	Average of AMT_INCOME_TOTAL	Max of AMT_INCOME_TOTAL	Min of AMT_INCOME_TOTAL	Count of AMT_INCOME_TOTAL
0				
F	130259.7776	337500	27000	7480
M	157826.719	337500	27000	4217
XNA	207000	207000	207000	1
1				
F	127000.8722	337500	31500	661
M	153822.2143	337500	36000	616
Grand Total	**140177.8398**	**337500**	**27000**	**12975**



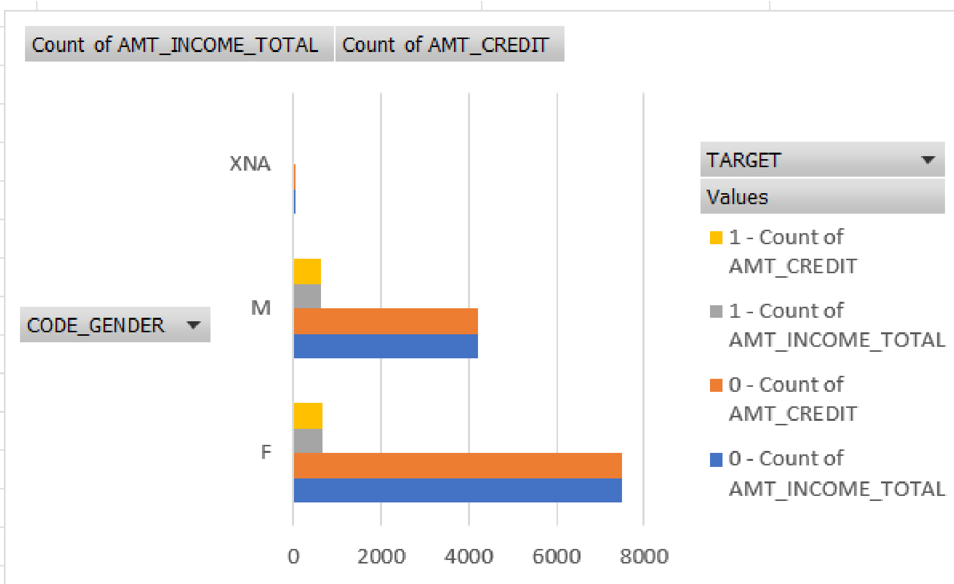
Comparison of Loan Amount with Gender and Default and Non Default Loan Payer

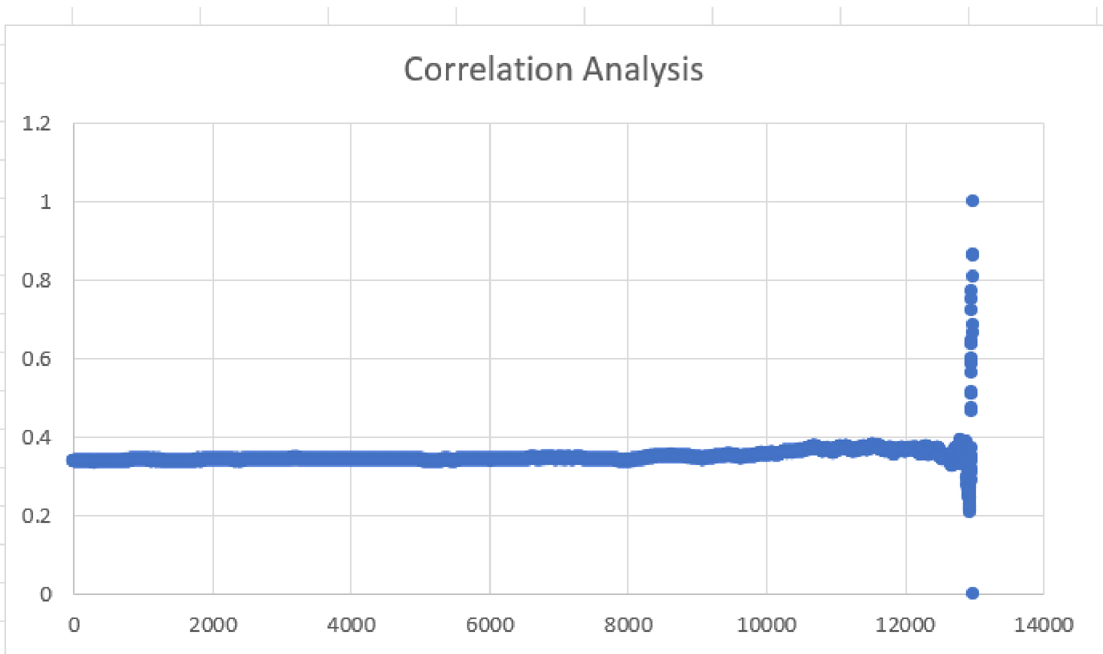
Row Labels	Average of AMT_CREDIT	Max of AMT_CREDIT	Min of AMT_CREDIT	Count of AMT_CREDIT
0	**522012.3406**	**1609272**	**45000**	**11698**
F	517249.8475	1566909	45000	7480
M	530493.0046	1609272	45000	4217
XNA	382500	382500	382500	1
1	**515258.9753**	**1563291**	**45000**	**1277**
F	512187.1611	1494486	45000	661
M	518555.1916	1563291	50940	616
Grand Total	**521347.6741**	**1609272**	**45000**	**12975**



3. **Bivariate Analysis:** Bivariate analysis examines the relationship between two variables. For this, you'll explore how various customer and loan attributes relate to the target variable (loan default).

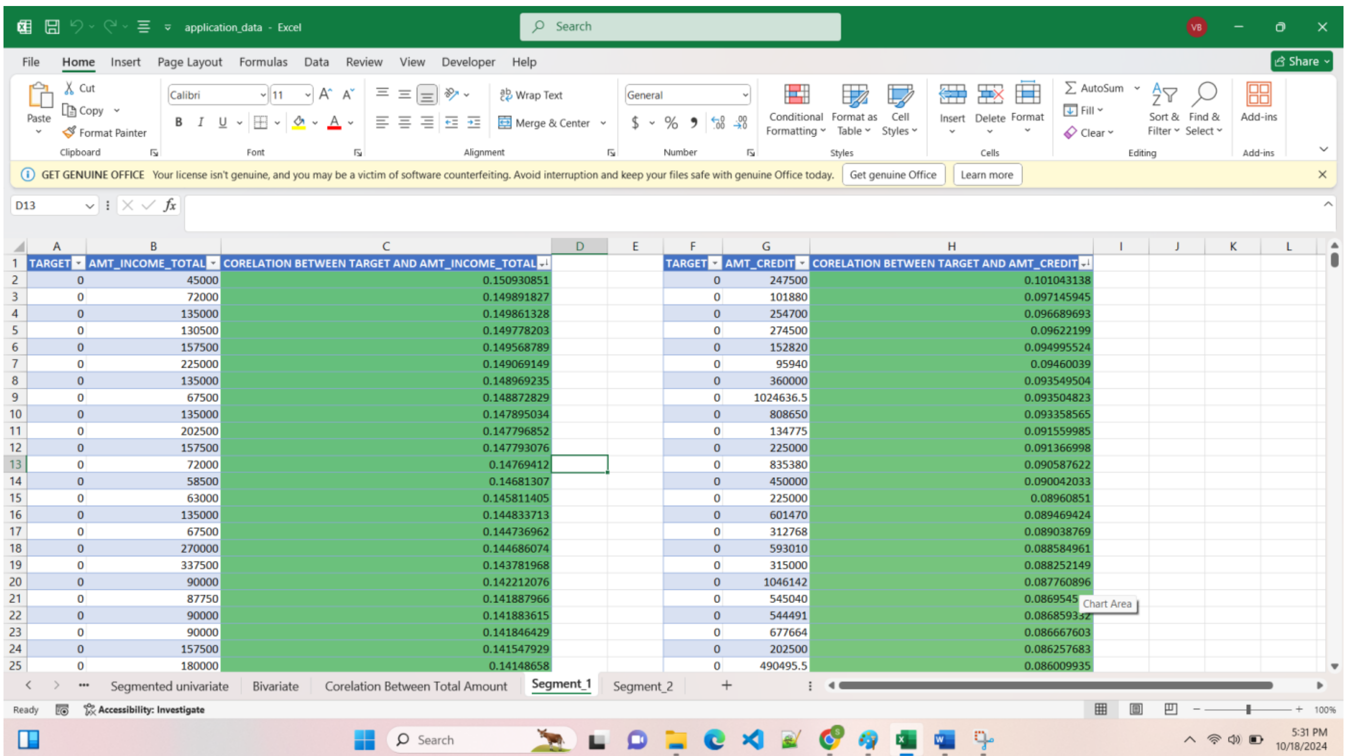
	Target(Default and Non-Default)			
	0		1	
CODE_GENDER	Count of AMT_INCOME_TOTAL	Count of AMT_CREDIT	Count of AMT_INCOME_TOTAL	Count of AMT_CREDIT
F	7480	7480	661	661
M	4217	4217	616	616
XNA	1	1		


Count of AMT_INCOME_TOTAL
Count of AMT_CREDIT
XNA
M
F
CODE_GENDER
0
2000
4000
6000
8000
TARGET
Values
1 - Count of AMT_CREDIT
1 - Count of AMT_INCOME_TOTAL
0 - Count of AMT_CREDIT
0 - Count of AMT_INCOME_TOTAL


Correlation Analysis
1.2
1
0.8
0.6
0.4
0.2
0
0
2000
4000
6000
8000
10000
12000
14000

E. Identify Top Correlations for Different Scenarios: Identified top correlations for different scenarios in the Bank Loan Case Study, segment the dataset based on different groups (e.g., clients with payment difficulties and clients without payment difficulties) and analyze which variables have the strongest correlations with the target variable (TARGET, where 1 = default, 0 = no default)

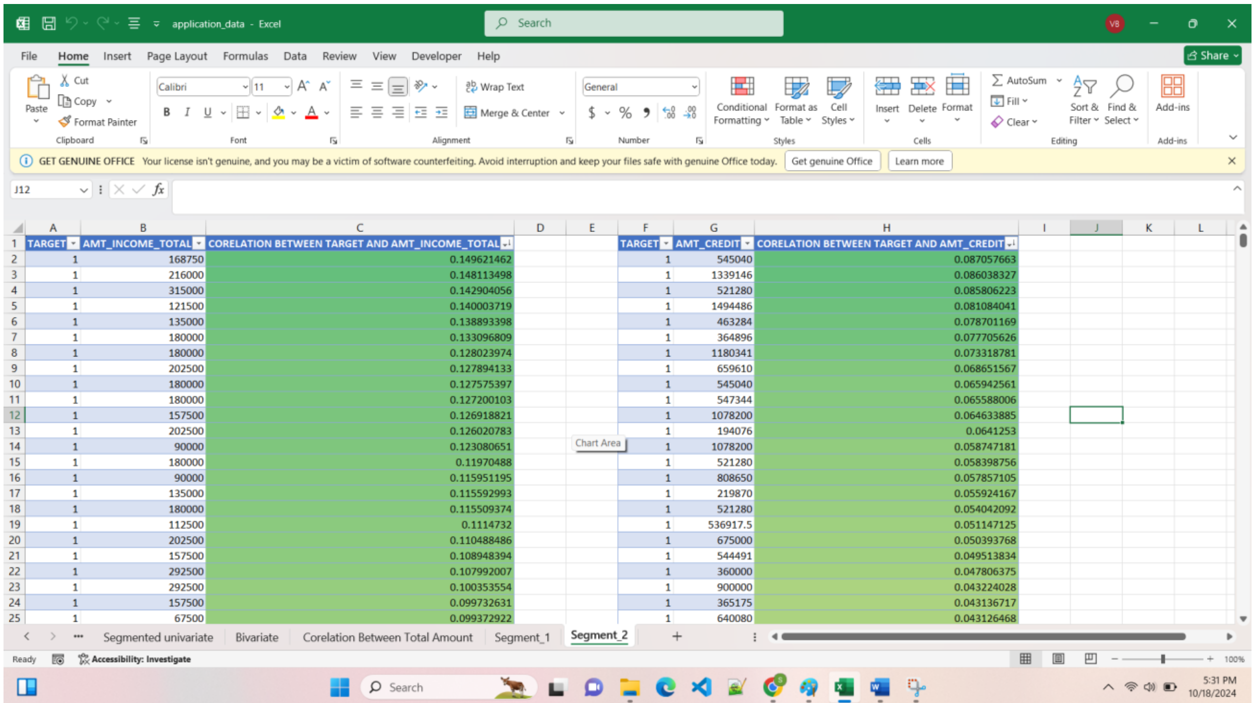
1. **Segment_1:**



TARGET	AMT_INCOME_TOTAL	CORELATION BETWEEN TARGET AND AMT_INCOME_TOTAL
0	45000	0.150930851
0	72000	0.149891827
0	135000	0.149861328
0	130500	0.149778203
0	157500	0.149568789
0	225000	0.149069149
0	135000	0.148969235
0	67500	0.148872829
0	135000	0.147895034
0	202500	0.147796852
0	157500	0.147793076
0	72000	0.14769412
0	58500	0.14681307
0	63000	0.145811405
0	135000	0.144833713
0	67500	0.144736962
0	270000	0.144686074
0	337500	0.143781968
0	90000	0.142212076
0	87750	0.141887966
0	90000	0.141883615
0	90000	0.141846429
0	157500	0.141547929
0	180000	0.14148658

TARGET	AMT_CREDIT	CORELATION BETWEEN TARGET AND AMT_CREDIT
0	247500	0.101043138
0	101880	0.097145945
0	254700	0.096689693
0	274500	0.09622199
0	152820	0.094995524
0	95940	0.09460039
0	360000	0.093549504
0	1024636.5	0.093504823
0	808650	0.093358565
0	134775	0.091559985
0	225000	0.091366998
0	835380	0.090587622
0	450000	0.090042033
0	225000	0.08960851
0	601470	0.089469424
0	312768	0.089038769
0	593010	0.088584961
0	315000	0.088252149
0	1046142	0.087760896
0	545040	0.0869545
0	544491	0.086859332
0	677664	0.086667603
0	202500	0.086257683
0	490495.5	0.086009935

2. **Segment_2:**



TARGET	AMT_INCOME_TOTAL	CORELATION BETWEEN TARGET AND AMT_INCOME_TOTAL	TARGET	AMT_CREDIT	CORELATION BETWEEN TARGET AND AMT_CREDIT
1	168750	0.149621462	1	545040	0.087057663
1	216000	0.148113498	1	1339146	0.086038327
1	315000	0.142904056	1	521280	0.085806223
1	121500	0.140003719	1	1494486	0.081084041
1	135000	0.138893398	1	463284	0.078701169
1	180000	0.133096809	1	364896	0.077705626
1	180000	0.128023974	1	1180341	0.073318781
1	202500	0.127894133	1	659610	0.068651567
1	180000	0.127575397	1	545040	0.065942561
1	180000	0.127200103	1	547344	0.065588006
1	157500	0.126918821	1	1078200	0.064633885
1	202500	0.126020783	1	194076	0.0641253
1	90000	0.123080651	1	1078200	0.058747181
1	180000	0.11970488	1	521280	0.058398756
1	90000	0.115951195	1	808650	0.057857105
1	135000	0.115592993	1	219870	0.055924167
1	180000	0.115509374	1	521280	0.054042092
1	112500	0.1114732	1	536917.5	0.051147125
1	202500	0.110488486	1	675000	0.050393768
1	157500	0.108948394	1	544491	0.049513834
1	292500	0.107992007	1	360000	0.047806375
1	292500	0.100353554	1	900000	0.043224028
1	157500	0.099732631	1	365175	0.043136717
1	67500	0.099372922	1	640080	0.043126468

Result:

The findings from this analysis can greatly improve the company's loan approval process by helping identify risky applicants before approval. By focusing on key variables such as income, annuity, and loan amounts, the company can better balance financial risk while capturing new business opportunities. Going forward, advanced modeling techniques can be implemented to create more precise predictions, further enhancing the company's ability to manage its loan portfolio effectively.