Bank Loan Case Study

Final Project – 2

Charmy Raj

Project Description

The primary objective of this study is to discover apparent trends that serve as indicators of potential challenges faced by customers in meeting their instalment payment obligations. This data can be utilised to make determinations such as declining the loan, decreasing the loan amount, or extending a loan to high-risk applicants at a higher interest rate. The organisation seeks to get insight into the primary determinants contributing to loan default to enhance its loan approval decision-making process.

Tech-Stack

This analytical study is conducted by employing Microsoft Excel, leveraging its extensive range of statistical functions including COUNT, Pivot Charts, and other advanced tools. In addition to this, a data analysis toolkit was utilised to successfully complete this job. The utilisation of my expertise in deciphering this data will contribute to the recognition of practical patterns and trends, thereby providing guidance to the business in making better informed and efficient decisions regarding recruitment.

Excel file

PROJECT_TRAINITY
.xlsx

https://docs.google.com/spreadsheets/d/1eUibd1gTAAsmNo4P9yD35jFmhpP7U9ww/edit?usp=sharing&ouid=115236894826816815181&rtpof=true&sd=true

Table of Contents

A. Data Cleaning:

STATMENT:

Missing data was observed in the loan application dataset. Effectively managing missing data is crucial to maintain the correctness of the analysis.

APPROACH:

With the help of 2 functions one can find out total number of blank spaces in the given datasets. Both the functions are demonstrated in the excel file. The functions are:

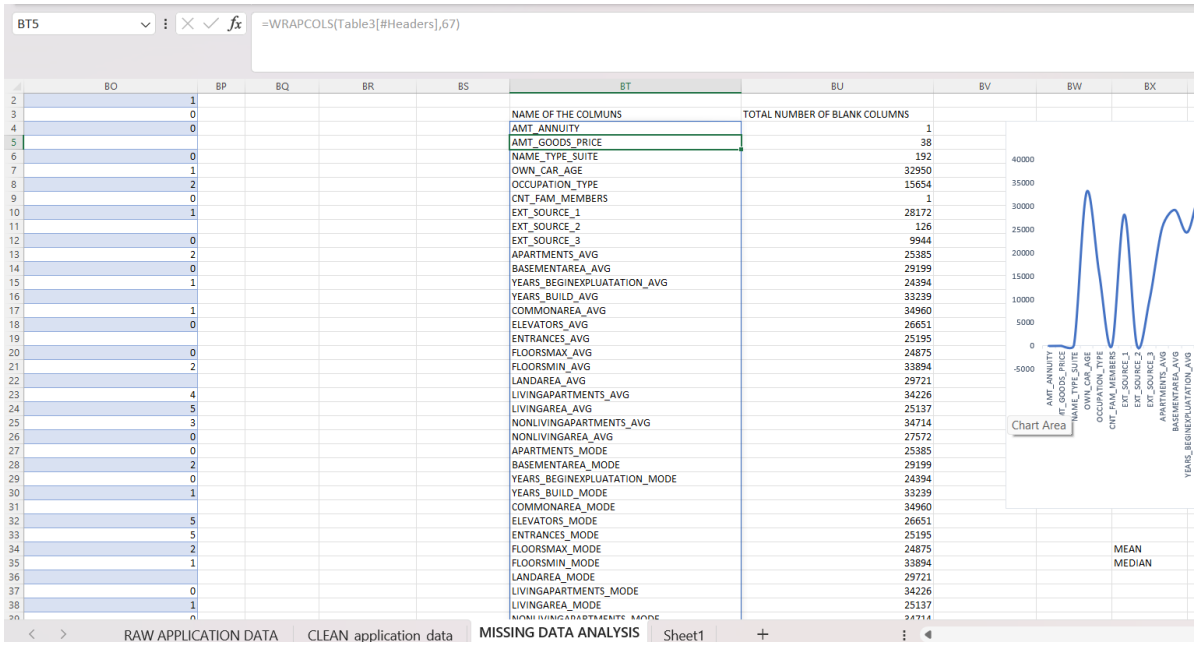


FIG 1 Total Number of Blank columns in the application dataset

- WRAPCOLS – To display all the columns name, this function is used.
- COUNTBLANK – This function is used to calculate the total number of blank cells in a particular column.

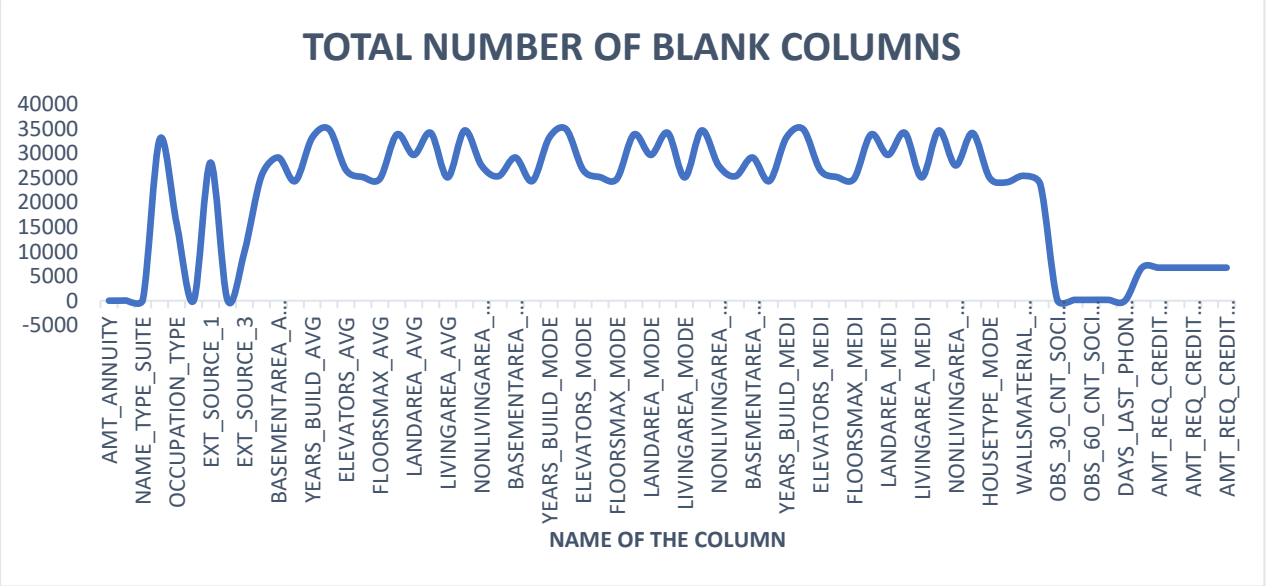
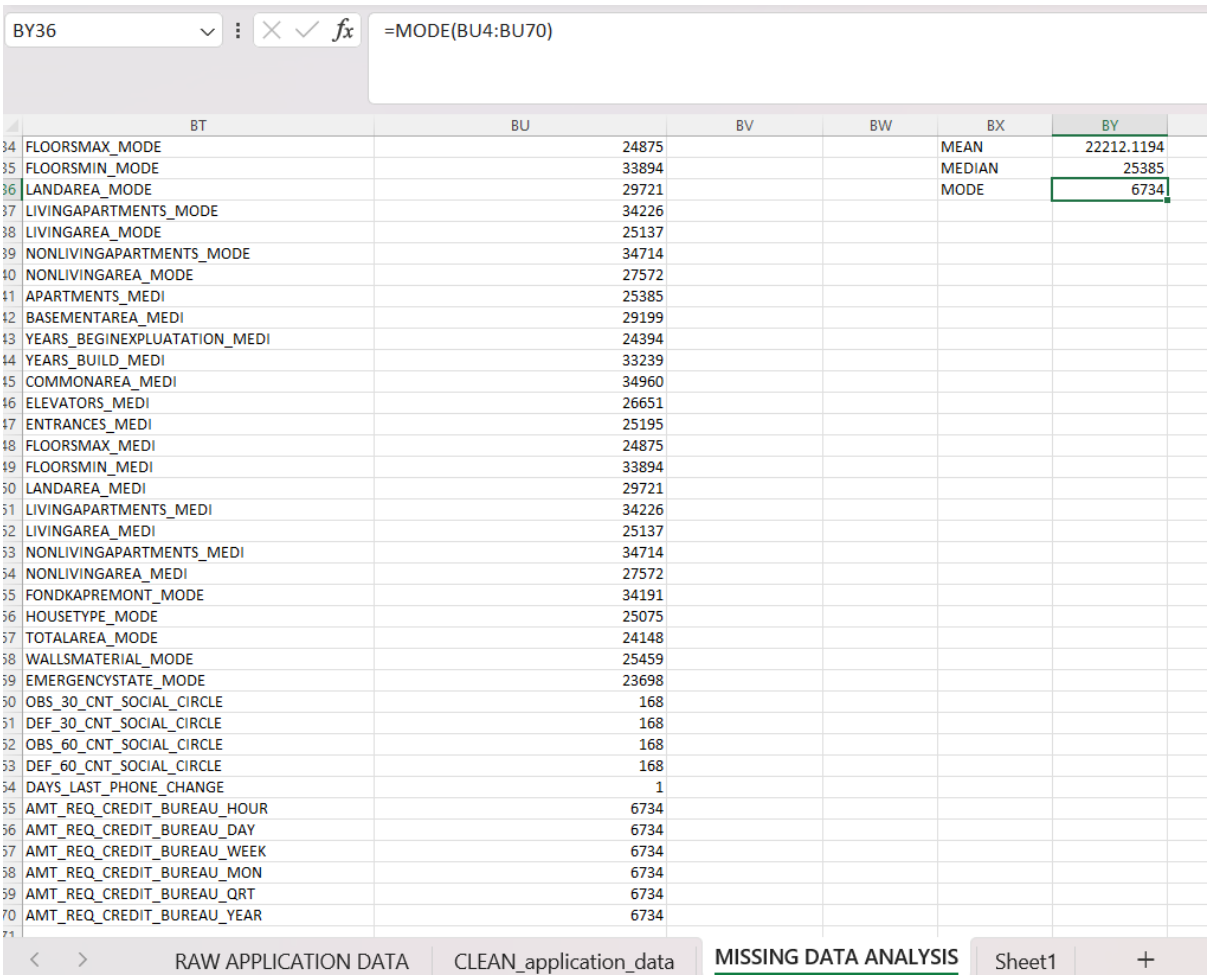


FIG 2 Visual representation of Blank columns

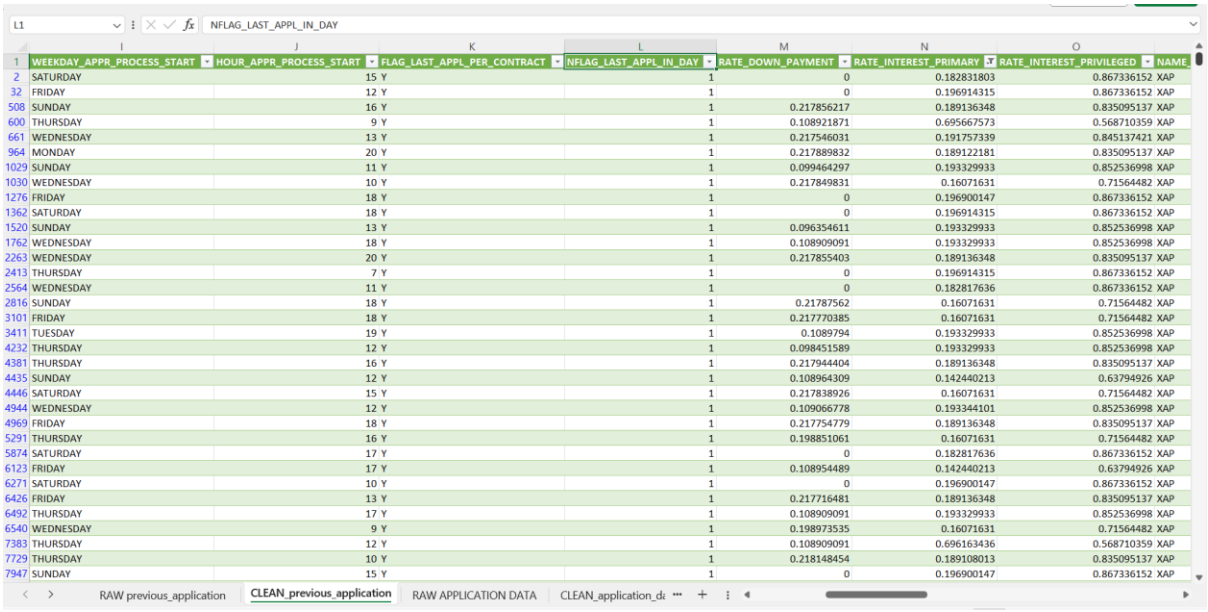
- Line chart was used for the representation of the data.
- From the graph it is readily apparent that, minimum number of blank columns were in independent variables like occupation_type, name_type_suite and many more.



	BT	BU	BV	BW	BX	BY
34	FLOORSMAX_MODE	24875			MEAN	22212.1194
35	FLOORSMIN_MODE	33894			MEDIAN	25385
36	LANDAREA_MODE	29721			MODE	6734
37	LIVINGAPARTMENTS_MODE	34226				
38	LIVINGAREA_MODE	25137				
39	NONLIVINGAPARTMENTS_MODE	34714				
40	NONLIVINGAREA_MODE	27572				
41	APARTMENTS_MEDI	25385				
42	BASEMENTAREA_MEDI	29199				
43	YEARS_BEGINEXPLUATATION_MEDI	24394				
44	YEARS_BUILD_MEDI	33239				
45	COMMONAREA_MEDI	34960				
46	ELEVATORS_MEDI	26651				
47	ENTRANCES_MEDI	25195				
48	FLOORSMAX_MEDI	24875				
49	FLOORSMIN_MEDI	33894				
50	LANDAREA_MEDI	29721				
51	LIVINGAPARTMENTS_MEDI	34226				
52	LIVINGAREA_MEDI	25137				
53	NONLIVINGAPARTMENTS_MEDI	34714				
54	NONLIVINGAREA_MEDI	27572				
55	FONDKAPREMONT_MODE	34191				
56	HOUSETYPE_MODE	25075				
57	TOTALAREA_MODE	24148				
58	WALLSMATERIAL_MODE	25459				
59	EMERGENCYSTATE_MODE	23698				
60	OBS_30_CNT_SOCIAL_CIRCLE	168				
61	DEF_30_CNT_SOCIAL_CIRCLE	168				
62	OBS_60_CNT_SOCIAL_CIRCLE	168				
63	DEF_60_CNT_SOCIAL_CIRCLE	168				
64	DAYS_LAST_PHONE_CHANGE	1				
65	AMT_REQ_CREDIT_BUREAU_HOUR	6734				
66	AMT_REQ_CREDIT_BUREAU_DAY	6734				
67	AMT_REQ_CREDIT_BUREAU_WEEK	6734				
68	AMT_REQ_CREDIT_BUREAU_MON	6734				
69	AMT_REQ_CREDIT_BUREAU_QRT	6734				
70	AMT_REQ_CREDIT_BUREAU_YEAR	6734				

FIG 3 Statistical Operations on the missing data

- Measures of central tendency were used on the dataset to find out the average and median value of blank data.
- Apart from this, the information of the credited amount has the same missing values, which I was able to find out with the help of MODE.



	WEEKDAY_APPR_PROCESS_START	HOUR_APPR_PROCESS_START	FLAG_LAST_APPL_PER_CONTRACT	NFLAG_LAST_APPL_IN_DAY	RATE_DOWN_PAYMENT	RATE_INTEREST_PRIMARY	RATE_INTEREST_PRIVILEGED	NAME
2	SATURDAY	15	Y	1	0	0.182831803	0.867336152	XAP
32	FRIDAY	12	Y	1	0	0.196914315	0.867336152	XAP
508	SUNDAY	16	Y	1	0.217856217	0.189136348	0.835095137	XAP
600	THURSDAY	9	Y	1	0.108921871	0.695667573	0.568710359	XAP
661	WEDNESDAY	13	Y	1	0.217546031	0.191757339	0.845137421	XAP
964	MONDAY	20	Y	1	0.217889832	0.189122181	0.835095137	XAP
1029	SUNDAY	11	Y	1	0.099464297	0.193329933	0.852536998	XAP
1030	WEDNESDAY	10	Y	1	0.217849831	0.16071631	0.71564482	XAP
1276	FRIDAY	18	Y	1	0	0.196900147	0.867336152	XAP
1362	SATURDAY	18	Y	1	0	0.196914315	0.867336152	XAP
1520	SUNDAY	13	Y	1	0.096354611	0.193329933	0.852536998	XAP
1762	WEDNESDAY	18	Y	1	0.108909091	0.193329933	0.852536998	XAP
2263	WEDNESDAY	20	Y	1	0.217855403	0.189136348	0.835095137	XAP
2413	THURSDAY	7	Y	1	0	0.196914315	0.867336152	XAP
2564	WEDNESDAY	11	Y	1	0	0.182817636	0.867336152	XAP
2816	SUNDAY	18	Y	1	0.21787562	0.16071631	0.71564482	XAP
3101	FRIDAY	18	Y	1	0.217770385	0.16071631	0.71564482	XAP
3411	TUESDAY	19	Y	1	0.1089794	0.193329933	0.852536998	XAP
4232	THURSDAY	12	Y	1	0.098451589	0.193329933	0.852536998	XAP
4381	THURSDAY	16	Y	1	0.217944404	0.189136348	0.835095137	XAP
4435	SUNDAY	12	Y	1	0.108964309	0.142440213	0.63794926	XAP
4446	SATURDAY	15	Y	1	0.217838926	0.16071631	0.71564482	XAP
4944	WEDNESDAY	12	Y	1	0.109066778	0.193344101	0.852536998	XAP
4969	FRIDAY	18	Y	1	0.217754779	0.189136348	0.835095137	XAP
5291	THURSDAY	16	Y	1	0.198851061	0.16071631	0.71564482	XAP
5874	SATURDAY	17	Y	1	0	0.182817636	0.867336152	XAP
6123	FRIDAY	17	Y	1	0.108954489	0.142440213	0.63794926	XAP
6271	SATURDAY	10	Y	1	0	0.196900147	0.867336152	XAP
6426	FRIDAY	13	Y	1	0.217716481	0.189136348	0.835095137	XAP
6492	THURSDAY	17	Y	1	0.108909091	0.193329933	0.852536998	XAP
6540	WEDNESDAY	9	Y	1	0.198973535	0.16071631	0.71564482	XAP
7383	THURSDAY	12	Y	1	0.108909091	0.696163436	0.568710359	XAP
7729	THURSDAY	10	Y	1	0.218148454	0.189108013	0.835095137	XAP
7947	SUNDAY	15	Y	1	0	0.196900147	0.867336152	XAP

FIG 4 Previous_application Dataset

- The procedure to clean this dataset was same as application dataset.
- As the information in this dataset was sensitive statistical operation were not carried down.
- For example, if the loan application was unused by the client, then that user won't have data related to the same. Hence, his column will be shown blank.

INTERPRETATION:

- From the above images we can see that, it is important to clean the data for better credibility of the insights.
- Whereas, with the help of sorting, filtering, and other statistical concepts the problem was taken care.

B. Outliers:

STATMENT:

The presence of outliers might have a substantial influence on the study and introduce distortions to the obtained results. It is necessary to detect outliers within the loan application dataset.

The identification and detection of outliers within a dataset can be accomplished through the utilisation of statistical functions and features available in Excel. This analysis mostly concentrates on numerical variables.

APPROACH:

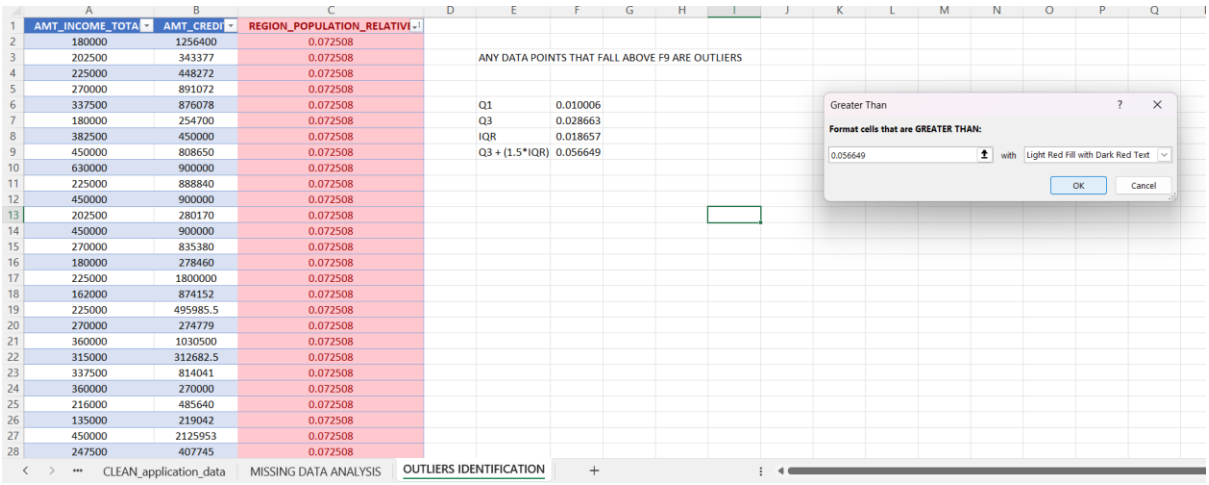


FIG 5 Data Sorting – Cell Formatting

- Regional population column, I used to analysis because based on the region the loan application can vary.
- For example, in urban area the number of loan application can be more compare than rural area. Therefore, the applicant approval rate is more in urban area.
- With the help of cell formatting, I coloured the data by using conditional formatting.
- I used median and IQR to find out outliers for the same.

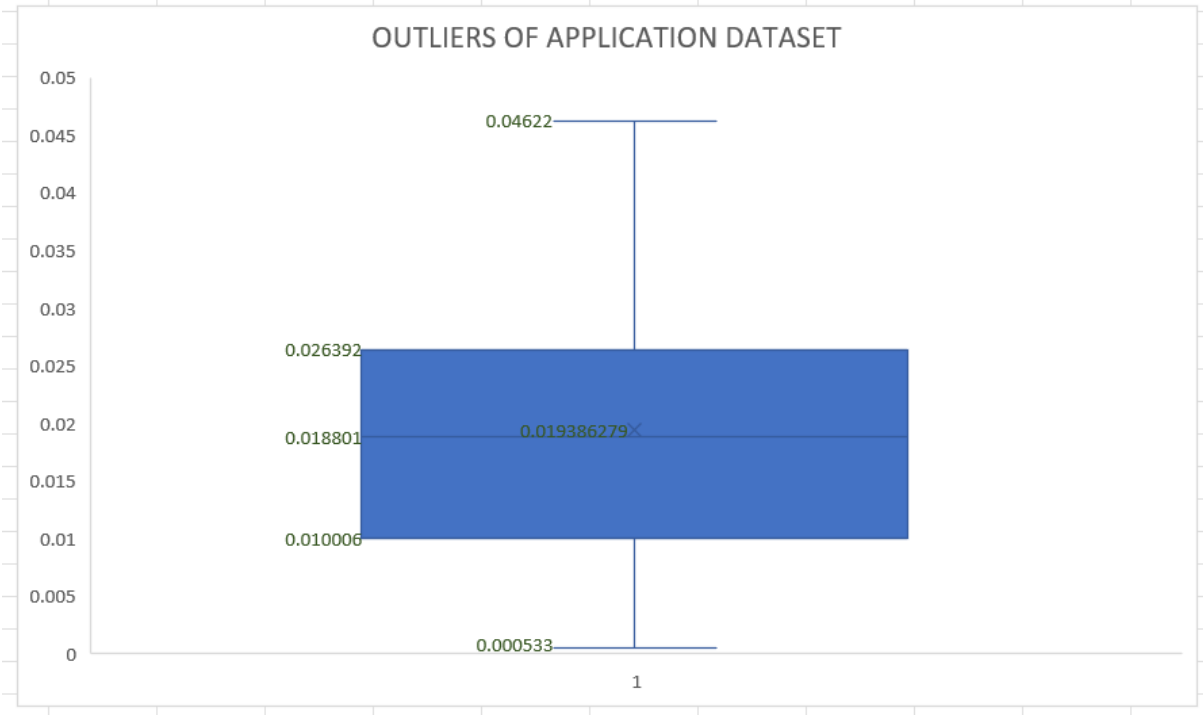


FIG 6 Box Plot Graph

- Box plot graph was used for visual representation of outliers of above-mentioned column.
- From the above image we can clearly see the mean, median, lower quartile, and upper quartile of the data.

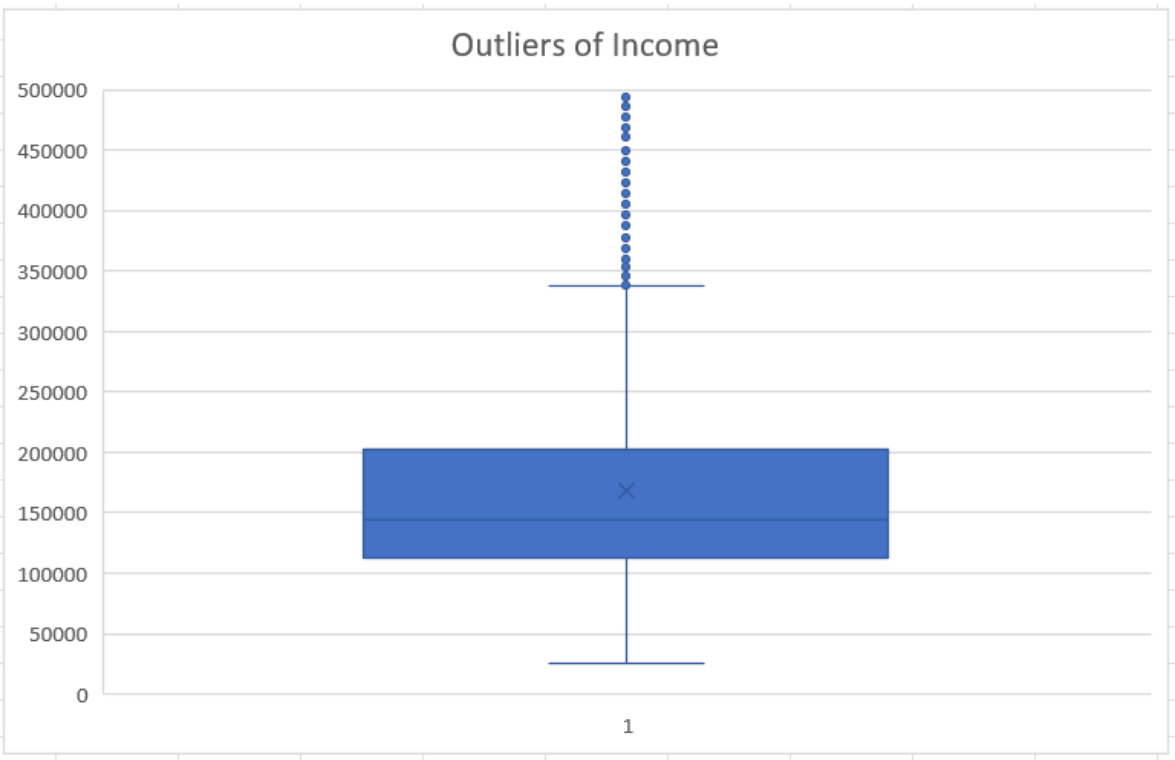


FIG 7 BOX- PLOT for income column representation

- I also consider analysing income column of the client based on the region.
- From that I was able to find out the applicants that falls under the outliers.
- In the FIG 7, all the datapoints that are above the maximum value are the outliers.

INTERPRETATION:

- Some extreme datapoints were observed, while most of the population falls between some specific class interval. In such situation there are high chances of misleading insights because of imbalance in the dataset. So further investigation is required for the same.

C. Data Imbalance Analysis:

STATMENT:

The presence of data imbalance has the potential to significantly impact the accuracy of the analysis, particularly in the context of binary classification problems. The objective of this analysis is to assess the presence of data imbalance within the loan application dataset. To accomplish this, I have utilised Excel functions and Pivot table to calculate the ratio of data imbalance.

APPROACH:

Target Variable	Count of TARGET
0	45973
1	4026
Grand Total	**49999**

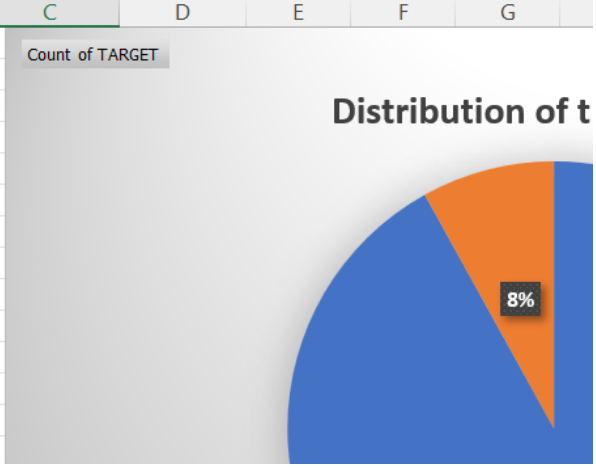


FIG 8 Pivot table for Target Variable

- The target variable refers to the variable that is being predicted or estimated in a statistical or machine learning model.
- A client is classified as 1, [having payment troubles] if they have a late payment of more than X days on at least one of the first Y instalments of the loan.
- All other cases are denoted as 0.
- To analyse the same by using pivot table.
- I took target column and utilized count value to get total number of 0 and 1.

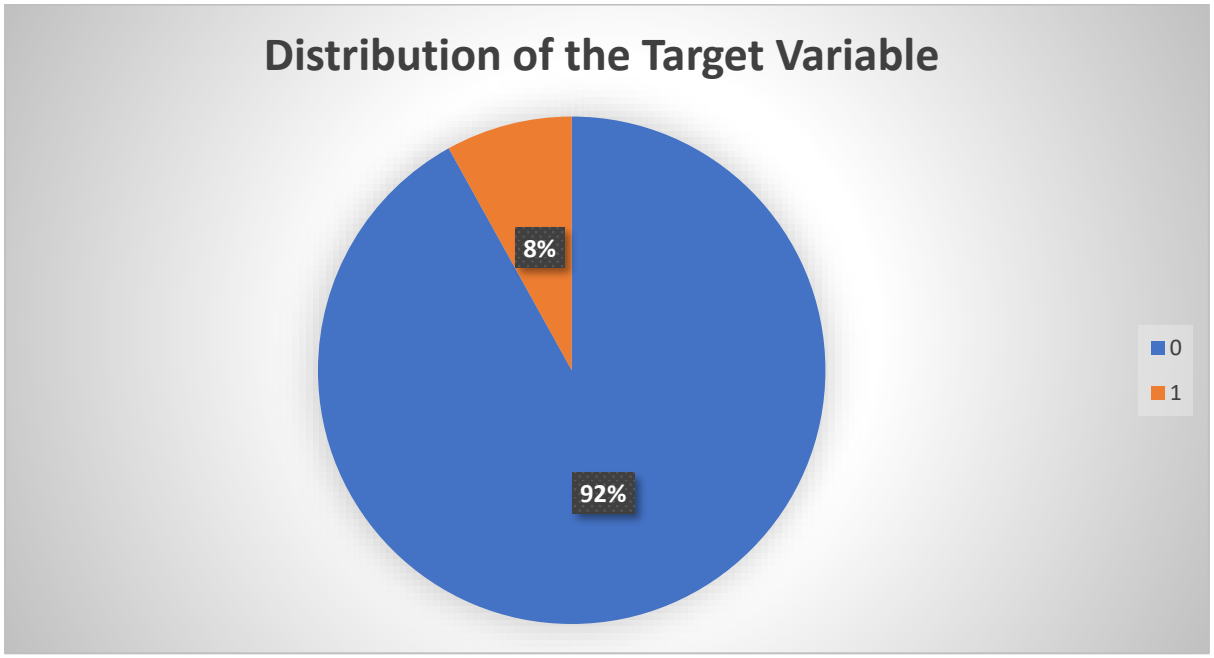


FIG 9 Visual representation of Distribution of the Target Variable

- As we can see, there only 8% clients that were regular in paying the instalments without creating any problem.
- Whereas majority faced one or the other problem in repayment of the loan amount, reasons can be due to some family emergency, shut down of their business and likewise.

Comparison of 2 variables	Target Variable		
Contract Type	0	1	Grand Total
Cash loans	41484	3792	45276
Revolving loans	4489	234	4723
Grand Total	**45973**	**4026**	**49999**

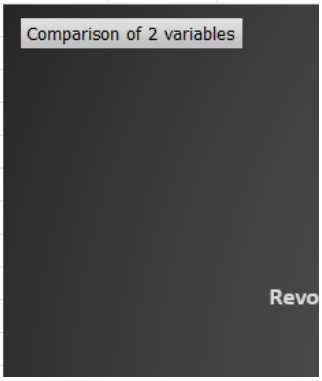


FIG 10 Pivot table for Analysis of Data Imbalance

- To analyse the ratio of data imbalance, I again took help of pivot table.
- I took target column and contract type column, to do comparison between these two variables that gave answer to my question:

  How many people faced or not faced problem at the time of paying instalments of cash loans and revolving loans?

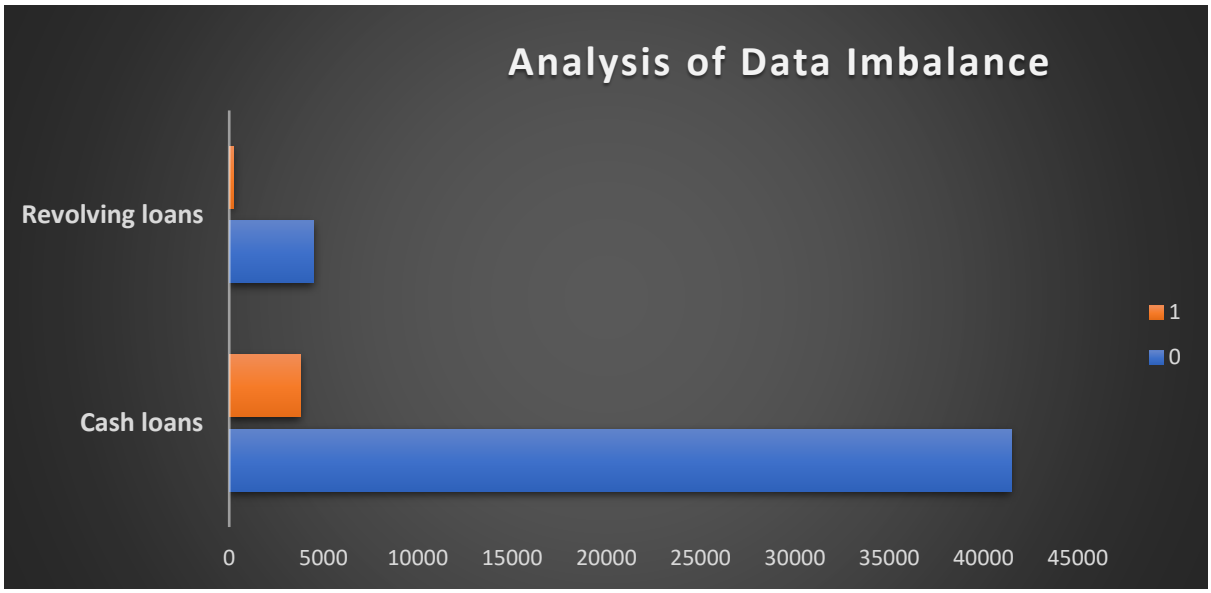


FIG 11 Visual Representation for Data Imbalance

- Bar chart was used for visual representation of Data imbalance between two types of loans:
- Revolving loans and cash loans.
- It is readily apparent that, the proportion of allotted cash loans is more than revolving loans. Hence, we can see a huge imbalance because of this allotment system.

INTERPRETATION:

- The success ratio in revolving loans is very less. As clients are getting more credit in this type of loan, there are high chances that may take bank for granted and not pay the dues on time.

- Talking about cash loans, the proportion of cash loans is pretty good, though there are only 2% people who doesn't face any problem in paying.
- Bank should decrease the rate of interest or the EMI amount so clients can easily pay their instalments without facing any problem and this imbalance of data can also be improved if both the types of loans are granted in equal proportion.

D. Various Analysis:

STATMENT:

In order to obtain a deeper understanding of the underlying causes of loan default, it is important to undertake a comprehensive analysis of both consumer characteristics and loan qualities. Conducting univariate, segmented univariate, and bivariate analysis using Excel can yield significant insights into the determinants of loan defaults. Utilise the included charts and tables as visual aids to facilitate the effective visualisation and interpretation of the data.

APPROACH:

Univariate Analysis:

Conducting a univariate analysis is essential in order to get insights into the distribution of individual variables. The data is structured and arranged within an Excel spreadsheet, considering variables that are of particular importance. There are customer attributes such as "Income," loan attributes such as "Loan Amount," and a binary target variable denoted as "Target" (with a value of 0 indicating default and 1 indicating no default).

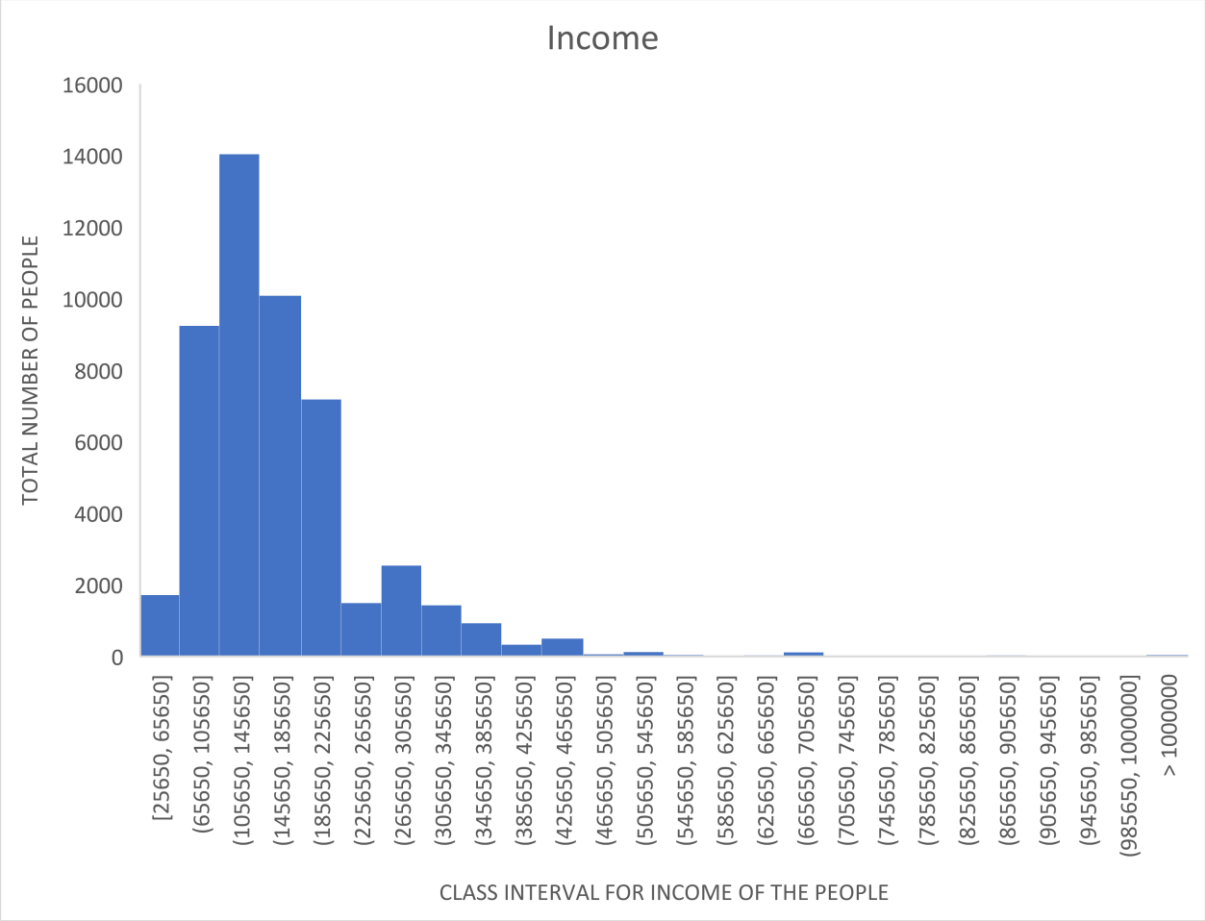


FIG 12 Visual Representation of Income Distribution

- As in income column outliers were observed so, I have set a condition on the upper limit of the graph for better creditability.
- From this we can observe that, maximum number of people falls under the class interval of 1,05,650 – 1,45,650.

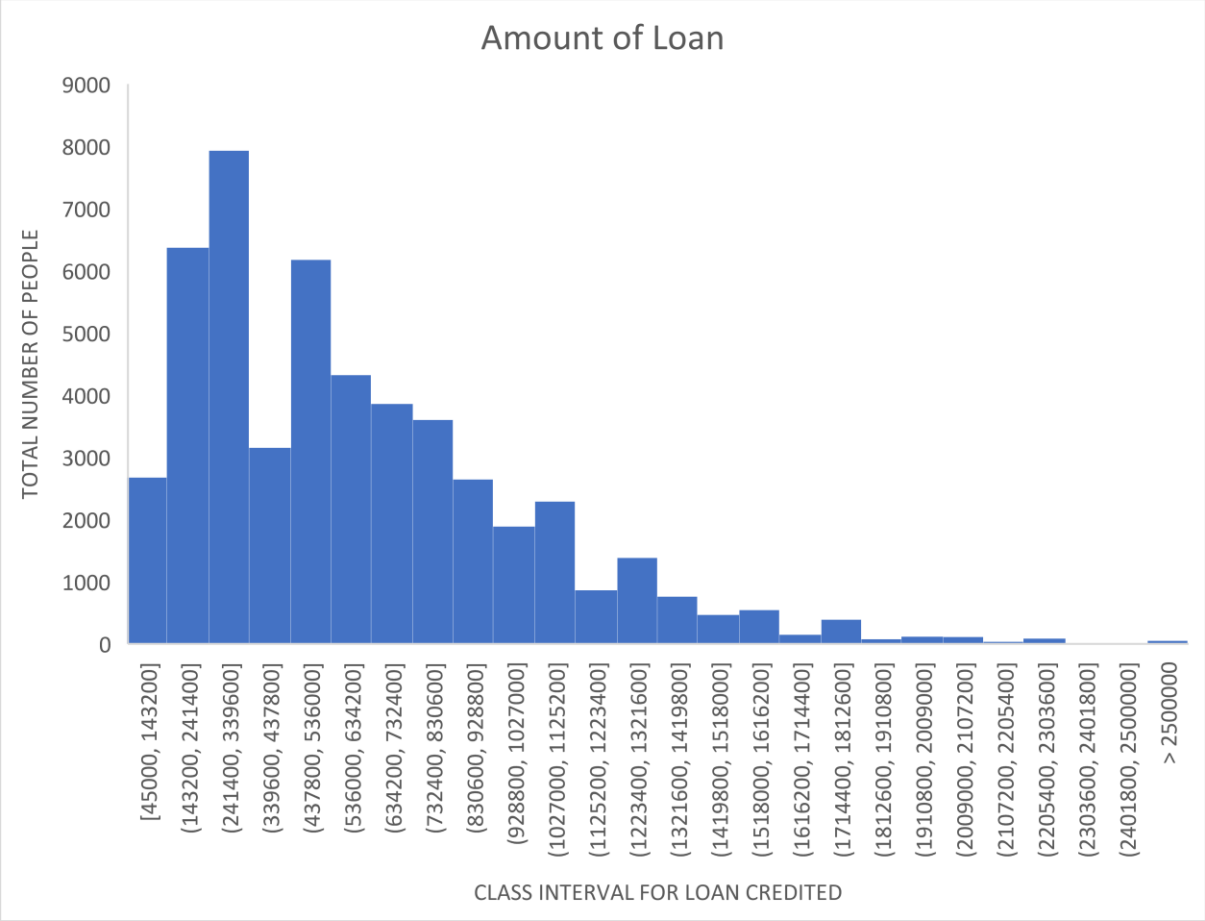


FIG 13 Visual Representation of Distribution Loan Amount

- From loan amount column outliers were observed so, I set a condition on the upper limit of the graph for better creditability.
- From this we can observe that, maximum number of people got loan between the class interval of 2,41,400– 3,39,600.
- In order to get insight into the distribution patterns of individual variables, histograms are constructed for continuous variables such as "Income" and "Loan Amount."
- People are getting the amount of loan more than the income they earn.
- Hence, to avoid this situation bank should provide the loan amount while considering the income of an individual.

	A	B	C	D	E	F	G	H
1	TARGET	INCOME	LOAN AMOUNT	Cash loans				
2	1	31500	45000	Revolving loans				
3	0	32400	45000	Cash loans	MAX VALUE OF INCOME	117000000		
4	0	36000	45000	Cash loans	MIN VALUE OF INCOME	25650		
5	0	45000	45000	Cash loans				
6	0	45000	45000	Cash loans				
7	0	49500	45000	Cash loans	MAX VALUE OF LOAN AMOUNT	4050000		
8	0	54000	45000	Cash loans	MIN VALUE OF LOAN AMOUNT	45000		
9	0	54000	45000	Revolving loans				
10	0	67500	45000	Cash loans				
11	0	67500	45000	Cash loans				
12	0	67500	45000	Cash loans		*DESCRIPTIVE STATSTICS*		
13	0	72000	45000	Cash loans				
14	0	81000	45000	Cash loans		Mean	599700.5815	
15	0	81000	45000	Cash loans		Standard Error	1799.674528	
16	0	81000	45000	Cash loans		Median	514777.5	
17	0	90000	45000	Revolving loans		Mode	450000	
18	0	90000	45000	Revolving loans		Standard Deviation	402415.4339	
19	0	90000	45000	Cash loans		Sample Variance	1.61938E+11	
20	0	90000	45000	Revolving loans		Kurtosis	1.917459058	
21	0	90000	45000	Cash loans		Skewness	1.223668739	
22	0	90000	45000	Cash loans		Range	4005000	
23	0	99000	45000	Cash loans		Minimum	45000	
24	0	99000	45000	Cash loans		Maximum	4050000	
25	0	103500	45000	Cash loans		Sum	29984429376	
26	0	103500	45000	Cash loans		Count	49999	
27	0	108000	45000	Cash loans				
28	0	112500	45000	Revolving loans				
29	0	112500	45000	Cash loans				
30	0	112500	45000	Cash loans				

FIG 14 Descriptive Statistics

- To support the above information, even descriptive statistics played a major role in displaying the above interpretation with the help of facts and figure.

Bivariate Analysis:

Bivariate analysis is employed to investigate interactions between variables and the goal variable.

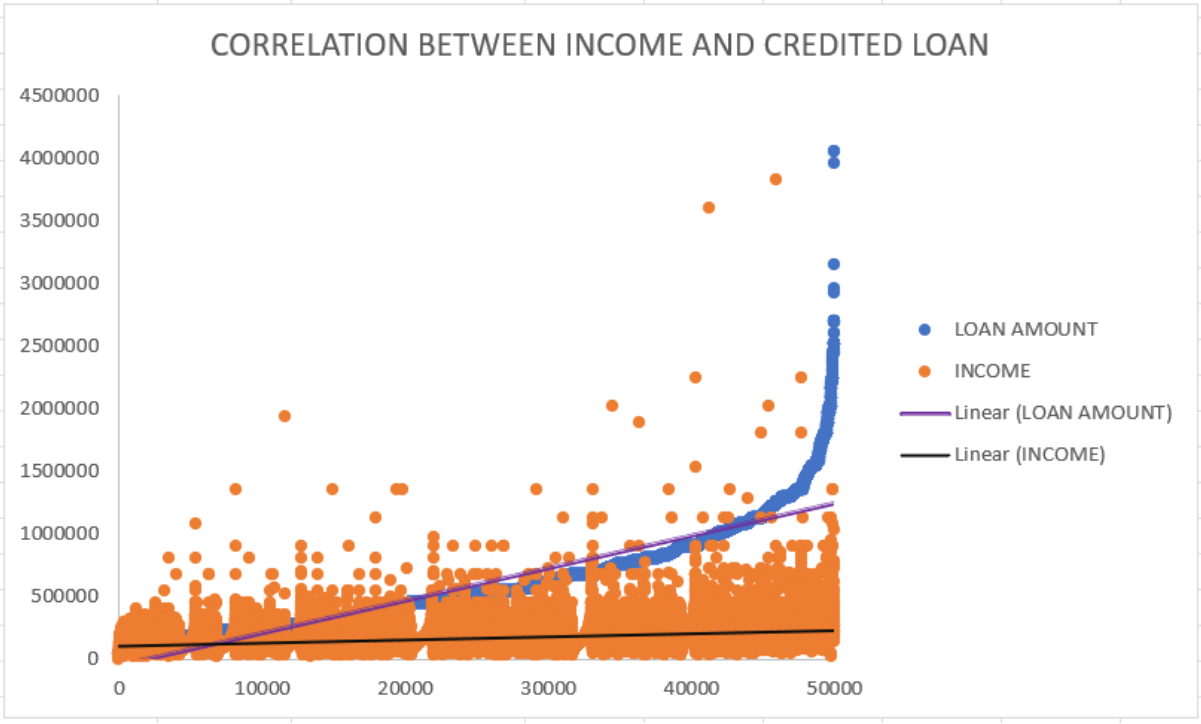


FIG 15 Correlation between two variables

- To explore the relationship between two continuous variables, such as "Income" and "Loan Amount," scatter plot is created.
- The outliers are ignored for better insights.
- From the FIG 15 we can say that, both the variables are negatively correlated with each other.

INCOME	Count of INCOME	Sum of TARGET
25650	2	1
27000	9	1
28350	1	0
28575	1	0
28800	1	0
29250	8	0
30600	1	0
31500	31	3
31531.5	1	0
31815	1	0
31860	1	0
32139	1	0
32400	4	1
32850	1	0
33300	3	0
33750	5	0
34650	1	0
35100	1	0
35550	1	0
36000	64	9
36450	3	0
36900	2	0
37350	3	0
37800	3	0
37854	1	0
38250	6	0
38419.155	1	0
38457	1	0
787500	1	0
810000	20	2
828000	1	0
855000	3	0
877500	1	0
882000	1	0
900000	29	1
945000	1	1
967500	1	0
1035000	1	0
1080000	3	0
1125000	13	1
1282500	1	0
1350000	10	0
1530000	1	0
1800000	2	0
1890000	1	1
1935000	1	0
2025000	2	0
2250000	2	0
3600000	1	0
3825000	1	0
117000000	1	1
Grand Total	**49999**	**4026**

FIG 16 Pivot Table

- A pivot table was generated in order to compute the default rate across several circumstances, such as loan kinds and income ranges.
- The variable "Loan Default" was placed in the "Values" section, whereas the variable "Income Range" was assigned to either the "Rows" or "Columns" section.

Segmented Univariate Analysis:

Employing segmented univariate analysis allows for the comparison of variable distributions across different scenarios.

FIG 17 Visual Representation for Data Imbalance

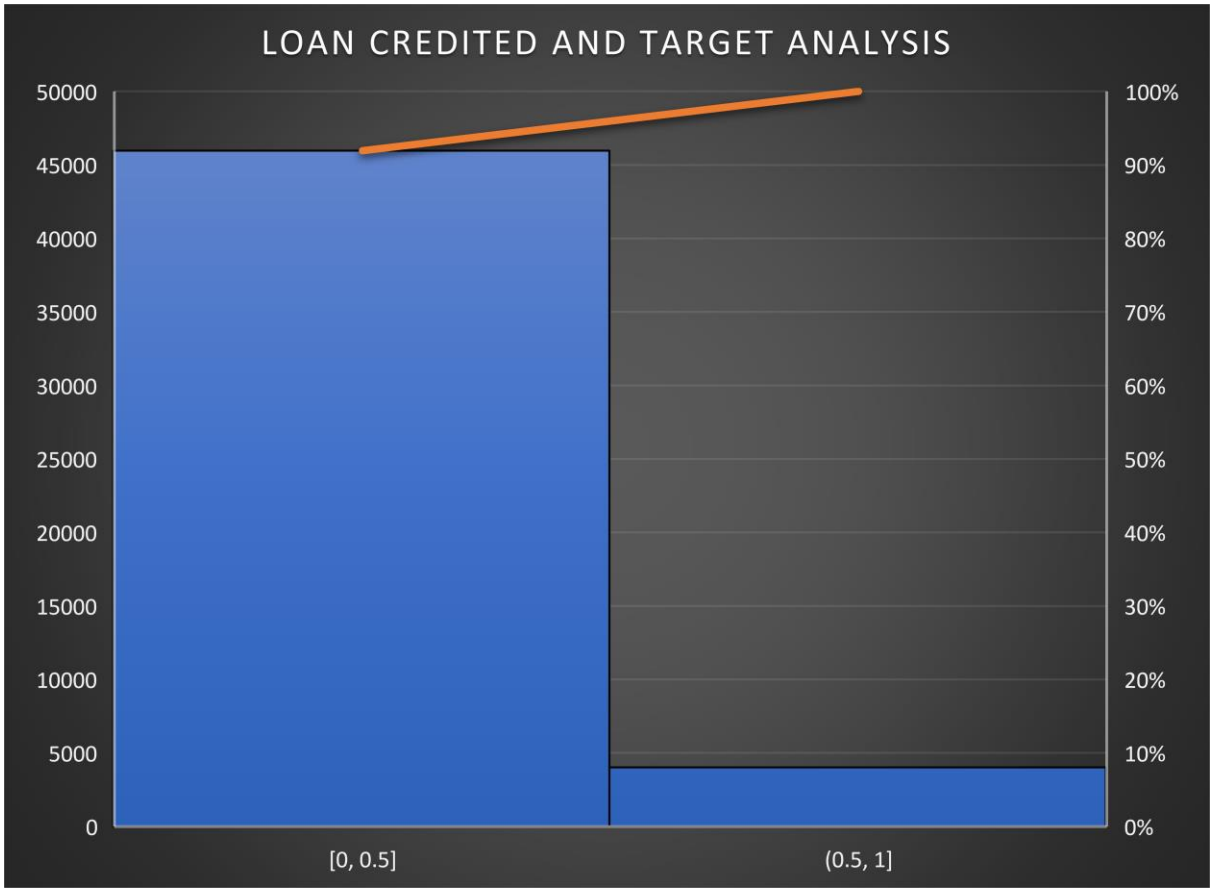


FIG 18 Loan Credited and Target Analysis

To compare variable distributions of distribution of loan, create a grouped column chart.

E. Analysis of Top Correlations for Different Scenarios:

STATMENT:

Gaining an understanding of the relationship between variables and the objective variable might yield valuable insights regarding accurate indicators of loan default.

APPROACH:

1st Case Scenario: APPROVED

The company has approved the loan application.

	AMT_ANNUITY	AMT_APPLICATION	AMT_CREDIT	AMT_DOWN_PAYMENT	AMT_GOODS_PRICE	HOUR_APPR_PROCESS_START	NFLAG_LAST_APPL_IN_DAY	RATE_DOWN_PAYMENT
AMT_ANNUITY	1							
AMT_APPLICATION	0.819442549	1						
AMT_CREDIT	0.828924357	0.964228219	1					
AMT_DOWN_PAYMENT	0.247912133	0.420398246	0.244155749	1				
AMT_GOODS_PRICE	0.834687121	1	0.993660062	0.420398246	1			
HOUR_APPR_PROCESS_START	-0.038599441	-0.047132193	-0.05838139	0.018588653	-0.0507164	1		
NFLAG_LAST_APPL_IN_DAY	-0.009404359	-0.00680498	-0.01286177	-0.003252832	-0.011343703	0.000654348	1	
RATE_DOWN_PAYMENT	-0.104948105	-0.096632115	-0.22027269	0.555691071	-0.096632115	0.024060191	9.3349E-05	1
RATE_INTEREST_PRIMARY	0.083675978	0.057901649	0.07940949	-0.089358982	0.057901649	-0.125514913		-0.180027065
RATE_INTEREST_PRIVILEGED	-0.210112242	-0.225897138	-0.20063767	-0.217518588	-0.225897138	-0.006410129		-0.210633923
DAYS_DECISION	0.249231893	0.257067641	0.262220431	-0.037932089	0.252054097	0.007255202	-0.016921953	-0.216304932
SELLERPLACE_AREA	-0.009457068	-0.006406942	-0.00750197	-0.001324741	-0.006838135	0.0168728	0.000578644	-0.007845926
CNT_PAYMENT	0.335467343	0.63672432	0.621113338	0.006754844	0.62781871	-0.057732159	-0.004826182	-0.292691787
DAYS_FIRST_DRAWING	0.043106564	0.063194406	-0.04374657	-0.01025256	-0.028766268	0.019580801	0.017012903	-0.02735958
DAYS_FIRST_DUE	-0.064115705	-0.049767098	-0.00631812	-0.019384196	-0.026878243	-0.005861845	0.005864614	-0.047690914
DAYS_LAST_DUE_1ST_VERSION	-0.06809935	-0.075519574	0.042861919	0.007675589	0.014372082	-0.022527466	-0.010835276	0.006199851
DAYS_LAST_DUE	0.085598361	0.186577617	0.23104536	-0.045663196	0.219164393	-0.017207514	-0.004142614	-0.153467937
DAYS_TERMINATION	0.07439619	0.165094272	0.223076927	-0.043758059	0.217062378	-0.019710262	-0.005783993	-0.150304397
TARGET	0.283828146	0.262838448	0.270447512	-0.04183805	0.249716587	-0.11949571	-0.001479351	0.000341506

AYMENT	RATE_INTEREST_PRIMARY	RATE_INTEREST_PRIVILEGED	DAYS_DECISION	SELLERPLACE_AREA	CNT_PAYMENT	DAYS_FIRST_DRAWING	DAYS_FIRST_DUE	DAYS_LAST_DUE_1ST_VERSION	DAYS_LAST_DUE	DAYS_TERMINATION	TARGET
1											
0027065	1										
0633923	-0.052793879	1									
6304932	-0.008231384	0.639582745	1								
7845926	0.034034531	-0.087714157	-0.002572071	1							
2691787	0.003756851	0.058947796	0.179444667	0.002293032	1						
2735958			-0.027323299	0.003621366	0.298224522	1					
7690914	0.07592197	0.161639739	0.185799722	-0.00085641	-0.205690509	-0.003072131	1				
6199851	-0.036573587	0.646164934	0.108025332	-0.003273509	-0.373308872	-0.788910637	0.532100342	1			
3467937	-0.019866847	0.257993025	0.457531327	-0.005247026	0.107395219	-0.246246726	0.40678918	0.417848007	1		
0304397	-0.0218543	0.268494698	0.417963136	-0.005383975	0.077542245	-0.37998028	0.331040041	0.479952423	0.931369398	1	
0341506			-0.042417365	-0.010168303	0.311455155	0.164544485	-0.115318634	-0.210025344	0.022265027	0.009675766	1

FIG 19 Correlation matrices for loan approved Analysis

The numerical values within the matrix denote the correlation coefficients that quantify the relationship between pairs of variables. The correlation coefficient is bounded between -1 and 1, with a value of 1 indicating a perfect positive correlation, where both variables increase linearly.

A value of -1 denotes a complete negative correlation, when an increase in one variable is accompanied by a linear drop in the other variable.

A value of 0 denotes the absence of a linear correlation, indicating that the variables under consideration are not linearly related.

INTERPRETATION

Presented below is an analysis of the correlation among several variables:

There exists a positive link between the variables AMT_ANNUITY and AMT_APPLICATION, AMT_CREDIT, as well as AMT_GOODS_PRICE. This implies that there is a positive relationship between the loan annuity amount and the application, credit, and products price.

There exists a positive association between the variable RATE_DOWN_PAYMENT and AMT_DOWN_PAYMENT, suggesting that an increase in the down payment rate is associated with a corresponding increase in the actual down payment amount.

There exists a positive connection between the variable DAYS_DECISION and the variables AMT_ANNUITY, AMT_APPLICATION, AMT_CREDIT, and AMT_GOODS_PRICE. This indicates that as the duration since the choice increases, there is a tendency for these financial variables to exhibit an increase.

The variable TARGET exhibits correlations, encompassing both positive and negative associations, with a range of other variables. It is noteworthy to acknowledge that the correlation coefficient of the variable TARGET with itself is 1, as anticipated due to the inherent correlation between a variable and itself.

Certain variables have low or near-zero correlations with one another, suggesting a minimal or non-existent linear association.

2ND Case Scenario: CANCELLED

The customer cancelled the application during the approval process.

	AMT_APPLICATION	*AMT_CREDIT*	*HOUR_APPR_PROCESS_START*	*DAYS_DECISION*	*SELLERPLACE_AREA*
AMT_APPLICATION	1				
AMT_CREDIT	0.996877133	1			
HOUR_APPR_PROCESS_START	0.015041571	0.015372588	1		
DAYS_DECISION	-0.104067559	-0.097225945	-0.009009041	1	
SELLERPLACE_AREA	0.106046578	0.09703496	0.024836775	-0.127631444	1

FIG 20 Correlation matrices for loan Cancelled Analysis

INTERPRTATION

The AMT_APPLICATION and AMT_CREDIT are two components that will be discussed in this analysis. The two variables exhibit a strong positive link, with a correlation coefficient of roughly 0.997. This observation implies a strong linear correlation between the loan amount requested (AMT_APPLICATION) and the loan amount approved (AMT_CREDIT) for loans that were ultimately cancelled. There is a positive correlation between the magnitude of the applied loan amount and the magnitude of the granted loan amount, indicating that higher loan applications are more likely to result in higher credit approvals, and conversely, lower loan applications are more likely to result in lower loan approvals.

The variables HOUR_APPR_PROCESS_START and DAYS_DECISION are being considered. The observed variables exhibit a negligible positive connection, with a coefficient of around 0.015. The data indicates a marginal positive linear correlation between the time of day when the loan application was handled and the duration it took to finalise the loan decision for loans that were subsequently cancelled. Nevertheless, the observed association has such a minimal magnitude that its practical significance may be deemed negligible.

The variables "SELLERPLACE_AREA" and "DAYS_DECISION" are being considered. The variables have a modest negative correlation, with a coefficient of roughly -0.128. These findings indicate the presence of a somewhat negative linear correlation between the size of the seller's location and the duration of time required to reach a loan decision for cancelled loans. To put it another, there may be a correlation between larger seller establishments and marginally reduced decision-making durations.

The observed correlations pertain exclusively to the category of cancelled loans and may potentially indicate distinct patterns or linkages within this particular sample of data. The strength of these linear associations is typically quite modest, and it's crucial to remember that correlation does not imply causation. There may be additional factors that have not been taken into account in this study, which could potentially impact the relationships between the variables under consideration. Therefore, it is advisable to do further investigation in order to gain a comprehensive understanding of the underlying factors that contribute to these observed correlations.

3rd Case Scenario: REFUSED

The company rejected the loan.

	AMT_ANNUITY	AMT_APPLICATION	AMT_CREDIT	AMT_DOWN_PAYMENT	AMT_GOODS_PRICE	HOUR_APPR_PROCESS_START	NFLAG_LAST_APPL_IN_DAY	RATE_DOWN_PAYMENT	DAYS_DECISION	SELLERPLACE_AREA	CNT_PAYMENT
AMT_ANNUITY	1										
AMT_APPLICATION	0.8280171	1									
AMT_CREDIT	0.83302337	0.982489914	1								
AMT_DOWN_PAYMENT	0.37810672	0.611281088	0.49802235	1							
AMT_GOODS_PRICE	0.83690891	0.999755516	0.99232654	0.611281088	1						
HOUR_APPR_PROCESS_START	-0.0178809	-0.025914324	-0.0361897	0.044378796	-0.045487261	1					
NFLAG_LAST_APPL_IN_DAY	0.08192932	0.058193076	0.00939201		0.003248274	0.027120167	1				
RATE_DOWN_PAYMENT	-0.08609928	-0.027220676	-0.1108877	0.468084352	-0.027220676	0.008376474		1			
DAYS_DECISION	0.28945799	0.25458892	0.26481792	-0.037918576	0.323691858	-0.062338318	0.051090021	-0.289041511	1		
SELLERPLACE_AREA	-0.09721993	-0.063031888	-0.077071	0.212038744	-0.084222	0.085369341	0.005175774	0.053921762	-0.135143034	1	
CNT_PAYMENT	0.43052249	0.665877219	0.67302326	0.113944835	0.654448489	-0.073130436	0.162719932	-0.23739201	0.296491261	-0.076638461	1

FIG 21 Correlation matrices for loan refused Analysis.

INTERPRETATION

The presented correlation matrix displays the correlation coefficients among multiple numerical factors within the context of loan refusal. For the analysis of denied loans, each variable is systematically compared to every other variable. The resulting correlation coefficients provide a quantitative measure of the magnitude and direction of the linear associations between the variables. The following analysis presents an interpretation of the observed correlations:

The variables AMT_ANNUITY, AMT_APPLICATION, and AMT_CREDIT is being considered. The aforementioned variables have robust positive connections among themselves. The variables AMT_APPLICATION and AMT_CREDIT exhibit a strong positive correlation of around 0.982, suggesting a close relationship between the loan application amount and the issued credit amount. The variable AMT_ANNUITY exhibits a significant positive correlation with both AMT_APPLICATION and AMT_CREDIT, suggesting a close association between the annuity amount and the loan application and credit amounts for rejected loans.

The variables AMT_DOWN_PAYMENT and AMT_GOODS_PRICE are being considered. The variables under consideration exhibit a correlation coefficient of around 0.611, indicating a reasonably strong positive relationship. This observation implies the presence of a substantial linear correlation

between the down payment amount and the price of products for loans that have been declined. There is a positive correlation between the price of items and the corresponding down payment amount.

The variables HOUR_APPR_PROCESS_START and DAYS_DECISION are being considered. The variables have a weak negative connection, with a coefficient of around -0.063. The data suggests the presence of a weak negative linear correlation between the time of day when loan applications were processed and the duration it took to reach a decision for rejected loan applications. Nevertheless, the observed association has a rather low magnitude, suggesting that it may lack practical significance.

The variables NFLAG_LAST_APPL_IN_DAY, RATE_DOWN_PAYMENT, SELLERPLACE_AREA, and CNT_PAYMENT are of interest in this study. The variables in question exhibit weak correlations both among themselves and with the other variables present in the matrix. The observed correlations exhibit a general tendency towards weakness or proximity to zero.

4TH Case Scenario: UNUSED

The loan was approved but the customer did not use it.

	HOUR_APPR_PROCESS_START	AMT_ANNUITY	AMT_APPLICATION	AMT_CREDIT	AMT_GOODS_PRICE	DAYS_DECISION	SELLERPLACE_AREA	CNT_PAYMENT
HOUR_APPR_PROCESS_START	1							
AMT_ANNUITY	0.248766546	1						
AMT_APPLICATION	0.017704396	0.975212928	1					
AMT_CREDIT	0.017704396	0.975212928	1	1				
AMT_GOODS_PRICE	0.017704396	0.975212928	1	1	1			
DAYS_DECISION	0.03203004	0.267832306	0.17034769	0.17034769	0.17034769	1		
SELLERPLACE_AREA	0.032102424	0.175541484	0.260447429	0.260447429	0.260447429	-0.290186959	1	
CNT_PAYMENT	-0.314548582	0.006162416	0.212765131	0.212765131	0.212765131	0.026100983	0.188570867	1

FIG 22 Correlation matrices for loan unused Analysis

INTERPRETATION

The supplied correlation matrix displays the correlation coefficients across multiple numeric variables pertaining to loans that were approved but remained unused by the client. In this particular sample of loans, each variable is systematically compared to every other variable, and the correlation coefficients are utilised to measure the magnitude and direction of the linear associations between them. The following analysis presents an interpretation of the observed correlations:

The variable HOUR_APPR_PROCESS_START represents the hour at which a certain process begins. The observed correlation coefficient of 1 between a variable and itself is to be anticipated, as every variable will exhibit perfect correlation with itself.

The variables AMT_ANNUITY, AMT_APPLICATION, AMT_CREDIT, and AMT_GOODS_PRICE are being considered. The variables exhibit a high degree of positive correlation, approaching a value of 1. This suggests a strong correlation between the loan application amount, the authorised credit amount, the goods price, and the annuity amount when a loan is approved but remains unused by the client.

When one of these quantities has a high value, it is observed that the other quantities also tend to exhibit high values.

The variable "DAYS_DECISION" exhibits positive correlations with the financial variables "AMT_ANNUITY," "AMT_APPLICATION," "AMT_CREDIT," and "AMT_GOODS_PRICE." However, these relationships are quite weak in nature. The findings indicate a potential positive correlation between the duration of the loan decision-making process and the loan and financial factors pertaining to approved but unused loans.

The variable "SELLERPLACE_AREA" represents the area of the seller's place. The aforementioned variable exhibits a little positive association with certain financial variables (namely, AMT_ANNUITY, AMT_APPLICATION, AMT_CREDIT, and AMT_GOODS_PRICE), while displaying a slight negative correlation with CNT_PAYMENT. This implies that there may exist a modest correlation between the geographical location of the seller's establishment and the financial factors associated with sanctioned yet unutilised loans.

The concept of CNT_PAYMENT refers to the payment method used in a certain context. The aforementioned variable has a negative connection of low magnitude with HOUR_APPR_PROCESS_START, while displaying a positive correlation of negligible magnitude with other financial variables. The observed connection between the number of payments and other factors within this subset of loans is often found to be modest, suggesting a limited degree of association between them.

CONTRACT_STATUS	Count of STATUS	RANK
Approved	31885	4
Canceled	8595	2
Refused	8660	3
Unused offer	859	1
(blank)		
Grand Total	**49999**	

FIG 23 Pivot table for CONTRACT_STATUS

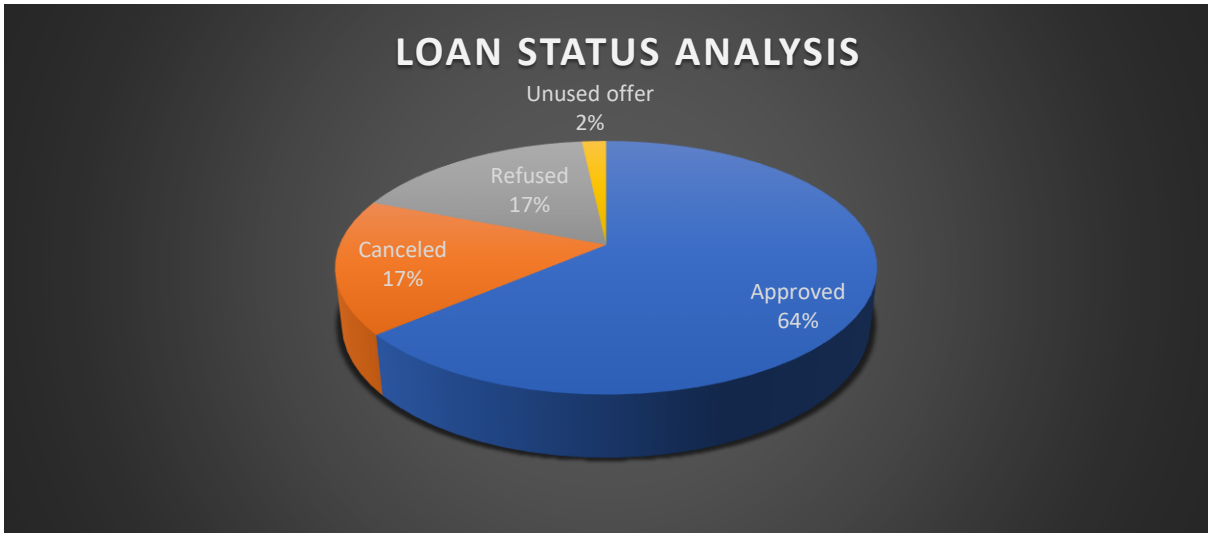


FIG 24 Visual Representation of Loan Status Analysis

INSIGHTS:

Distinct observations and potential measures that the organisation can do, as inferred from the correlation matrices that I have presented for various situations including loan approval, denial, unused loans, and cancellations.

Loan Approval:

- In the context of loan approval, the correlation matrix shows a strong link between variables like "AMT_APPLICATION," "AMT_CREDIT," "AMT_GOODS_PRICE," and "AMT_ANNUITY," which suggests a strong link between the requested loan amount, the amount of credit that was approved, the price of the goods, and the amount of the annuity.
- The loan amount requested by the applicant (called AMT_APPLICATION) should be carefully checked and evaluated based on their trustworthiness and the value of the things they want to buy.
- It's important to make sure that the approved credit amount (AMT_CREDIT) is in line with the applicant's income and the price of the things (AMT_GOODS_PRICE).

Loan Denial:

- In the context of loan denial, the correlation matrix indicates a strong correlation among the variables "AMT_APPLICATION," "AMT_CREDIT," "AMT_GOODS_PRICE," and "AMT_ANNUITY."
- Despite the denial of the loan, it is advisable to regularly monitor these financial factors since they continue to have a strong correlation. There exists potential for acquiring valuable insights pertaining to the reasons for the denial of specific loans, notwithstanding the presence of strong correlations among these variables.

Unused loans:

- They are the loans that have been approved but have not been utilised by the borrower. These loans represent a significant financial resource that has not been tapped into.
- The key variables that are associated with unused loans include the Within the framework of authorised yet unutilised loans, the correlation matrix indicates robust positive associations among the variables "AMT_APPLICATION," "AMT_CREDIT," "AMT_GOODS_PRICE," and "AMT_ANNUITY."
- The objective is to investigate the underlying reasons behind the non-utilization of loans by clients who have been granted approval. Examine potential factors contributing to this behaviour, such as alterations in financial conditions or unfulfilled expectations pertaining to the pricing of the goods.

Loan cancellations:

- These are a significant topic of interest in the academic realm. In order to comprehensively analyse this subject, it is crucial to identify and examine the key variables involved.
- In the context of loan cancellations, the correlation matrix reveals significant positive correlations among the variables "AMT_APPLICATION," "AMT_CREDIT," "AMT_GOODS_PRICE," and "AMT_ANNUITY." Additionally, there appears to be a weak negative association observed between the variables "SELLERPLACE_AREA" and "CNT_PAYMENT."

CONCLUSION

It is advisable to undertake an analysis of the underlying factors contributing to the cancellation of loans. Are borrowers terminating loan agreements due to differences between the approved loan amounts and their individual requirements or anticipated outcomes?

This analysis aims to assess the potential impact of revisions to loan terms or improved communication with clients on reducing the cancellation rate.

The objective is to analyse the potential relationship between the variables "SELLERPLACE_AREA" and "CNT_PAYMENT" and their influence on loan cancellations, with the intention of identifying any discernible patterns or trends that may emerge over a certain period.

In every instance, it is imperative for the organisation to employ predictive modelling and risk assessment methodologies to more accurately measure the influence of these pivotal variables on loan default, loan rejection, unused loans, and loan cancellations. Through this approach, the organisation may enhance the precision of its loan approval models and use data-driven strategies to effectively manage the potential hazards associated with loan defaults, all while upholding sound lending practises. Furthermore, the implementation of continuous monitoring and data enrichment techniques can contribute to the ongoing improvement and refinement of these insights.