[TRAINITY PROJECT]

A. Identify Missing Data and Deal with it Appropriately.
B. Identify Outliers in the Dataset.
C. Analyze Data Imbalance.
D. Perform Univariate, Segmented Univariate, and Bivariate Analysis.
E. Identify Top Correlations for Different Scenarios.

BANK LOAN

CASE STUDY

By :- ABHAY SAXENA

PROJECT DESCRIPTION

The project is about to carry out an exploratory data analysis (EDA) for a bank. Here we are going to consider the different types of EDA (univariate, bivariate and multivariate analysis) in order to carry out the analysis and find the solution to problem statement.

When a customer applies for a loan, your company faces two risks:

If the applicant can repay the loan but is not approved, the company loses business.

If the applicant cannot repay the loan and is approved, the company faces a financial loss.

The dataset you'll be working with contains information about loan applications. It includes two types of scenarios:

Customers with payment difficulties: These are customers who had a late payment of more than X days on at least one of the first Y installments of the loan.

All other cases: These are cases where the payment was made on time.

When a customer applies for a loan, there are four possible outcomes:

Approved: The company has approved the loan application.

Cancelled: The customer cancelled the application during the approval process.

Refused: The company rejected the loan.

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

Your goal in this project is to use EDA to understand how customer attributes and loan attributes influence the likelihood of default.

APPROACH

My first approach was to analyse the dataset, clean the dataset finding the blanks and missing values, input the missing values with the appropriate method (mean, median, mode). Then I tried to find the outliers in the dataset, there are some anomalies such as negative values which need either to be deleted or standardized. After all these I used pivot tables and basic charts to visualise the data. Moreover, insights were drawn based on my understandings.

TECH-STACK USED

For this project I used Microsoft Excel to run my queries. Microsoft Excel is a spreadsheet developed by Microsoft for Windows, macOS, Android and iOS. It features calculation or computation capabilities, graphing tools, pivot tables, and a macro programming language called Visual Basic for Applications (VBA). Excel forms part of the Microsoft Office suite of software.

I used the Excel sheet provided and ran multiple functions to get the desired answers.

This project helped me in understanding the Excel Table at a much detailed manner and helped to improve my strength in extracting data from tables and visualize it in the forms of different graphs.

INSIGHTS

Following insights were drawn based on understanding and capabilities:

- Cash loans have higher percentage as compared to revolving loan.
- Most of the people have no problems regarding the loan payment.
- The defaulters of cash loans are higher than the revolving ones.
- Most of the labourers have problems in loan payment.

Excel files for data analysis of Bank Loan Case Study :

CURRENT APPLICATION

PREVIOUS APPLICATION

CURRENT APPLICATION

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

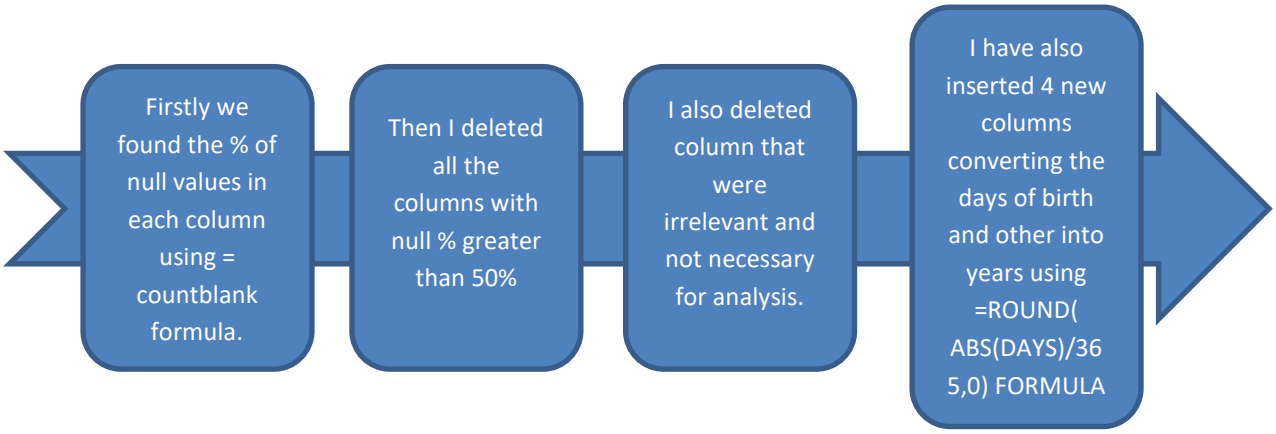
SOLUTION:

Initially we had:

COLUMNS - 122

ROWS - 50000

STEPS TO CLEAN DATA:



I then filled the null values:

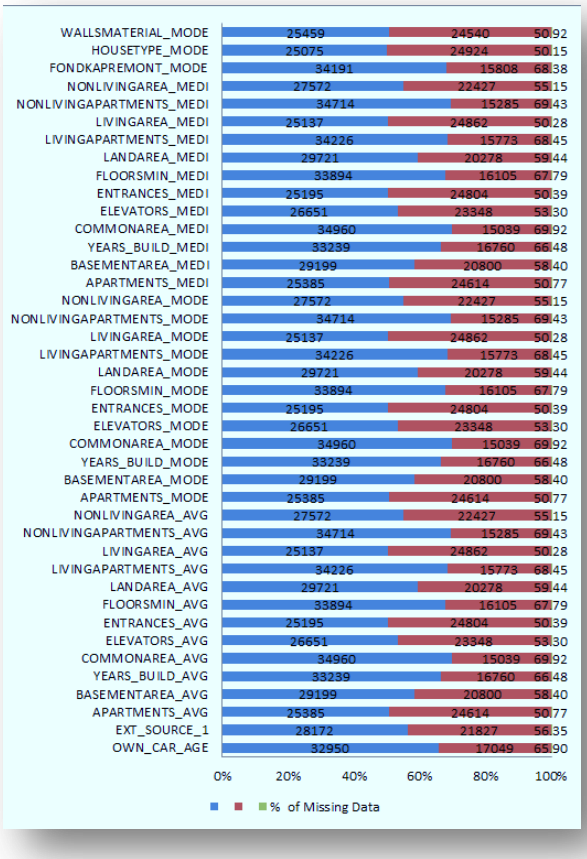
For numerical data I used median and filled the null values with the median because the data contain outliers so mean would be inappropriate.

For non-numerical data I deleted the entire rows instead of filling the cells haphazardly.

Formulas used:

=COUNTBLANK(C7:C50005)

Following is the graph of field wise missing data.



Columns to be deleted:

OWN_CAR_AGE	BASEMENTAREA_MODE	NONLIVINGAREA_MEDI
HOUR_APPR_PROCESS_START	YEARS_BEGINEXPLUATATION_M	FONDKAPREMONT_MODE
REG_REGION_NOT_LIVE_REGION	YEARS_BUILD_MODE	HOUSETYPE_MODE
REG_REGION_NOT_WORK_REGI	COMMONAREA_MODE	TOTALAREA_MODE
LIVE_REGION_NOT_WORK_REGI	ELEVATORS_MODE	WALLSMATERIAL_MODE
REG_CITY_NOT_LIVE_CITY	ENTRANCES_MODE	EMERGENCYSTATE_MODE
REG_CITY_NOT_WORK_CITY	FLOORSMAX_MODE	FLAG_DOCUMENT_2
LIVE_CITY_NOT_WORK_CITY	FLOORSMIN_MODE	FLAG_DOCUMENT_3
EXT_SOURCE_1	LANDAREA_MODE	FLAG_DOCUMENT_4
EXT_SOURCE_2	LIVINGAPARTMENTS_MODE	FLAG_DOCUMENT_5
EXT_SOURCE_3	LIVINGAREA_MODE	FLAG_DOCUMENT_6
APARTMENTS_AVG	NONLIVINGAPARTMENTS_MOD	FLAG_DOCUMENT_7
BASEMENTAREA_AVG	NONLIVINGAREA_MODE	FLAG_DOCUMENT_8
YEARS_BEGINEXPLUATATION_A	APARTMENTS_MEDI	FLAG_DOCUMENT_9
YEARS_BUILD_AVG	BASEMENTAREA_MEDI	FLAG_DOCUMENT_10
COMMONAREA_AVG	YEARS_BEGINEXPLUATATION_M	FLAG_DOCUMENT_11
ELEVATORS_AVG	YEARS_BUILD_MEDI	FLAG_DOCUMENT_12
ENTRANCES_AVG	COMMONAREA_MEDI	FLAG_DOCUMENT_13
FLOORSMAX_AVG	ELEVATORS_MEDI	FLAG_DOCUMENT_14
FLOORSMIN_AVG	ENTRANCES_MEDI	FLAG_DOCUMENT_15
LANDAREA_AVG	FLOORSMAX_MEDI	FLAG_DOCUMENT_16
LIVINGAPARTMENTS_AVG	FLOORSMIN_MEDI	FLAG_DOCUMENT_17
LIVINGAREA_AVG	LANDAREA_MEDI	FLAG_DOCUMENT_18
NONLIVINGAPARTMENTS_AVG	LIVINGAPARTMENTS_MEDI	FLAG_DOCUMENT_19
NONLIVINGAREA_AVG	LIVINGAREA_MEDI	FLAG_DOCUMENT_20
APARTMENTS_MODE	NONLIVINGAPARTMENTS_MED	FLAG_DOCUMENT_21

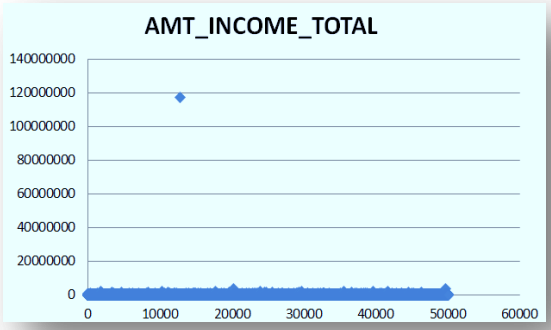
These are the 78 columns not needed for our analysis and hence we deleted them. Now we are left with cleaned data on which we can work upon.

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

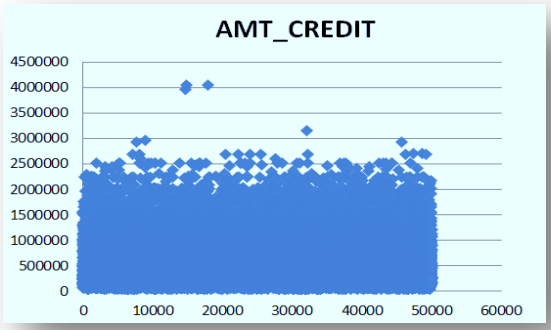
SOLUTION:

Outliers in a dataset are data points that significantly deviate from the rest of the data points. They are values that are either unusually high or unusually low compared to the majority of the data. Outliers can arise due to various reasons such as measurement errors, data entry mistakes, natural variability, or even as valid data points that represent rare occurrences or anomalies.

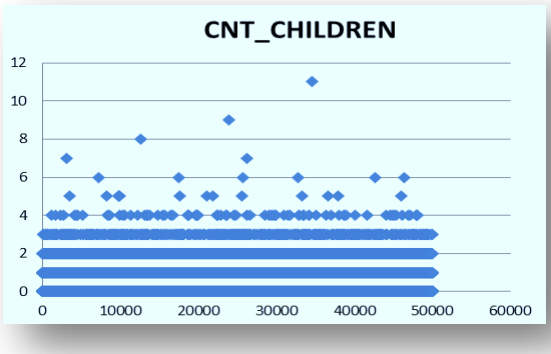
We used box and whisker and sometimes scatter plot chart to find the outliers.



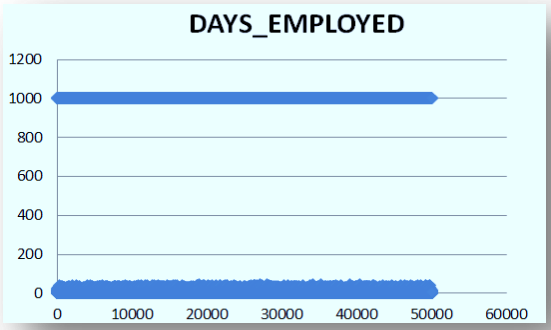
Analysis: we can clearly see that most of the values are clustered below 40,00,000. However only one value is 11,70,00,000. Which doesn't seem to be realistic. So this is a outliers as it will affect all the measures of central tendency.



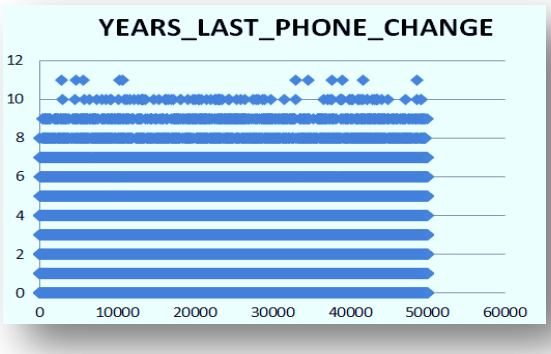
Though the median is around 5,00,000 yet we can see some figures above 25,00,000. Which is above normal credit and this would affect our analysis.



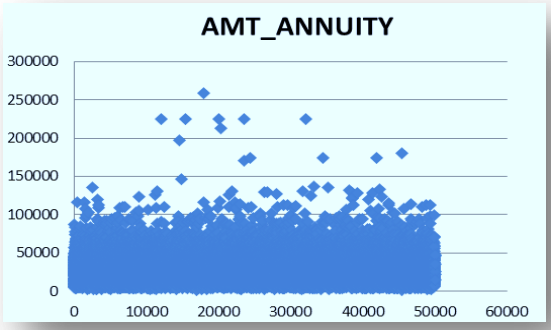
In few cases there can be seen to be errors in data entry as the data shows that some clients have more than 5 children which is again not too realistic in today's day n age. And some have count upto 11 children which is not proper.



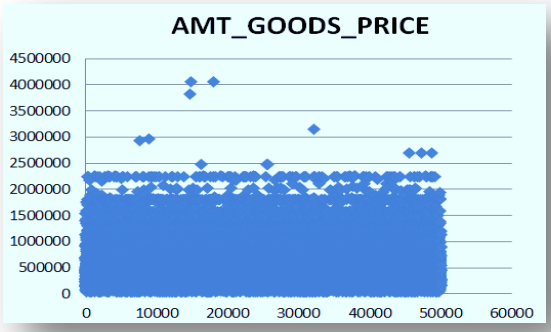
We found that some of the employees have been employed for more than 1000 years which is clearly not possible and thus an outlier. Rest all the values are realistic and proper.



The graph in this suggests that people have been using the same phone for almost 11 years which is a little too difficult with the advancement in the technology these days. So these are the outliers and will be affecting our analysis.



In this, there are few amounts which are exorbitantly high and above the median so they are outliers as there are very few chances of having the annuity more than 1,00,000.



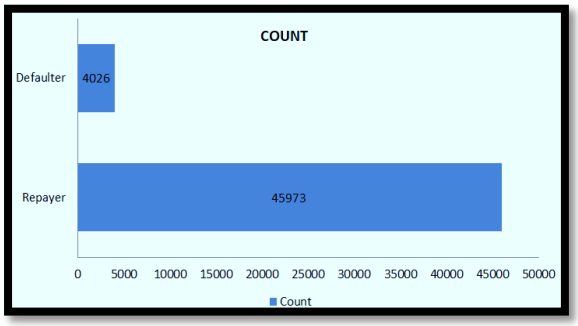
Now in this the prices of few items for which the client is applying for loan is extremely high and not reasonable and thus those are the outliers.

C.) Analyze Data Imbalance: Data imbalance can affect the accuracy of the analysis, especially for binary classification problems. Understanding the data distribution is crucial for building reliable models.

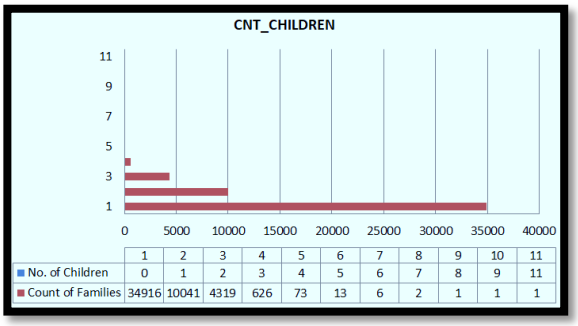
Task: Determine if there is data imbalance in the loan application dataset and calculate the ratio of data imbalance using Excel functions.

SOLUTION:

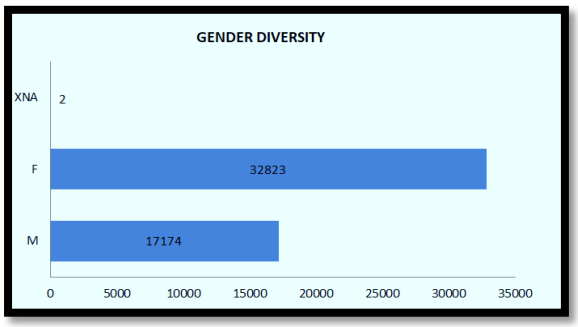
Data imbalance occurs when data is disseminated in an unequal manner. data imbalance refers to a situation where the distribution of classes in a classification problem is not equal or nearly equal. In other words, one class has significantly more or fewer instances compared to the other classes. This imbalance can pose challenges when developing predictive models or performing data analysis, as it can lead to biased or inaccurate results.



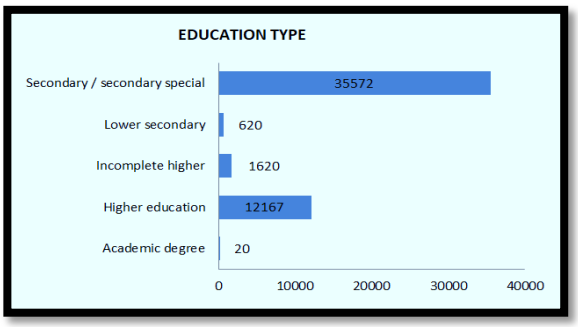
The dataset exhibits a significant data imbalance between repayees and defaulters. The majority class (repayers) vastly outweighs the minority class (defaulters). Addressing this imbalance is crucial to ensure accurate model performance and to avoid biases toward the majority class during analysis the ratio is 11.41 : 1.



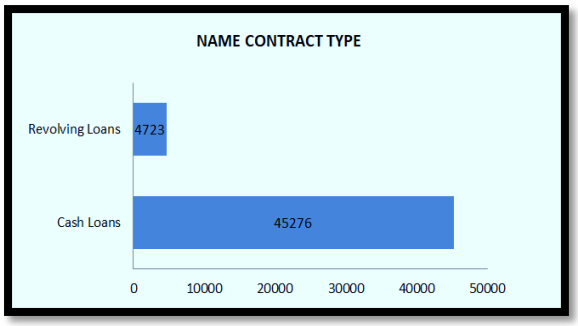
The dataset displays an imbalanced distribution in the number of children feature. Clients with no children have a notably higher frequency, while others are less represented. The ratio is 2.32 : 1 ,i.e., the number of clients with no children is 2.32 times the others.



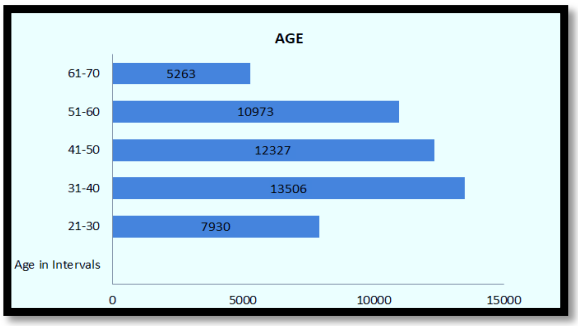
The dataset reveals a gender diversity imbalance, with a higher representation of females compared to males. This disparity might affect gender-related analyses and requires careful consideration to ensure fair and accurate insights from the data. The ratio is 1.91 : 1. Means 1 male for every 1.9 females application. For default males have higher percentage.



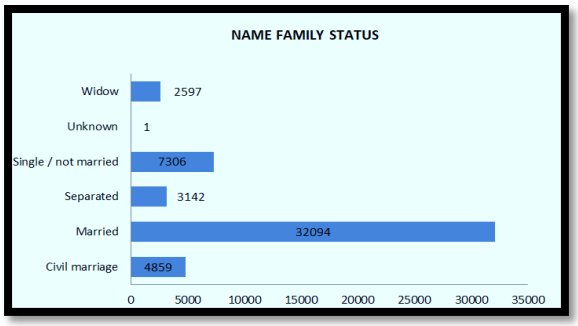
The majority of the clients have Secondary/secondary special education, followed by clients with Higher education. Only a very small number have an academic degree The Lower secondary category, although rare, have the largest rate of not returning the loan (11.8%). People with Academic degrees have less than a 0% defaulting rate.



Revolving loans are just a small fraction (10.4%) of the total number of loans at the same time, a larger amount of Revolving loans, compared with their frequency, are not repaid. The ratio is 9.58 : 1.



Highest number of people who have applied for loan are from 31-40 age group. However the default % is highest for the clients between 21-30 age group. People above 60 age group have lowest probability of default.



Most of the people who have taken loans are married, followed by Single/not married and civil marriage. In terms of the percentage of not repayment of loans, single/not married has the highest percentage of not repayment (10%), with Widows the lowest (exception being Unknown). The ratio is 1.8 : 1.

D.) Perform Univariate, Segmented Univariate, and Bivariate Analysis: To gain insights into the driving factors of loan default, it is important to conduct various analyses on consumer and loan attributes.

Task: Perform univariate analysis to understand the distribution of individual variables, segmented univariate analysis to compare variable distributions for different scenarios, and bivariate analysis to explore relationships between variables and the target variable using Excel functions and features.

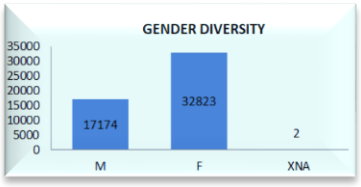
SOLUTION:

Univariate analysis:

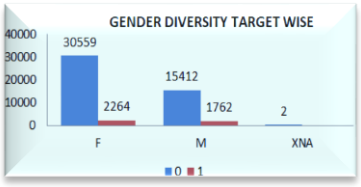
Univariate analysis is an analysis of a row or a column with itself. It is a statistical method used in data analysis to examine and describe a single variable at a time. It is often the first step in understanding the characteristics and patterns of individual variables within a dataset.

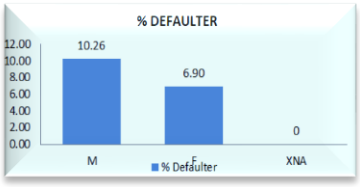
In segmented univariate analysis I have classified the data of various variables into 2 segments i.e., repayer and default where 0 stands for repayer and 1 stands for defaulter.

Univariate analysis



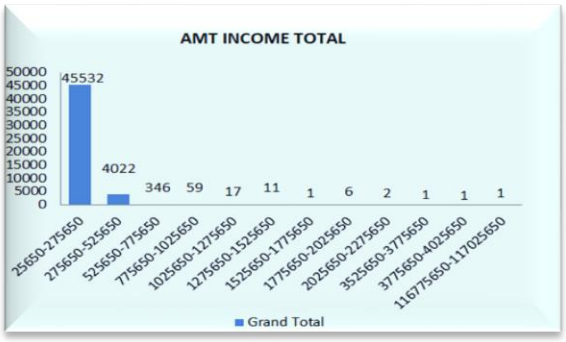
Segmented univariate analysis



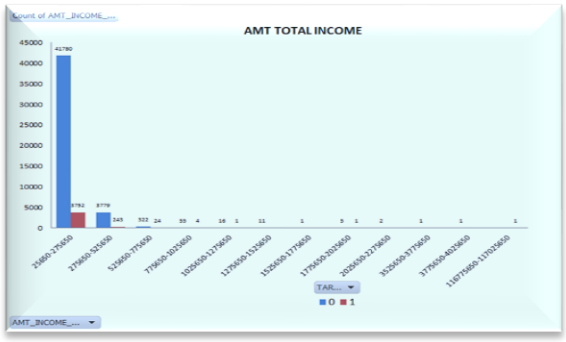


In the univariate analysis of gender, the dataset shows a higher frequency of females compared to males. This observation provides insight into the gender distribution within the studied population. Also the default rate is higher in males.

Univariate analysis

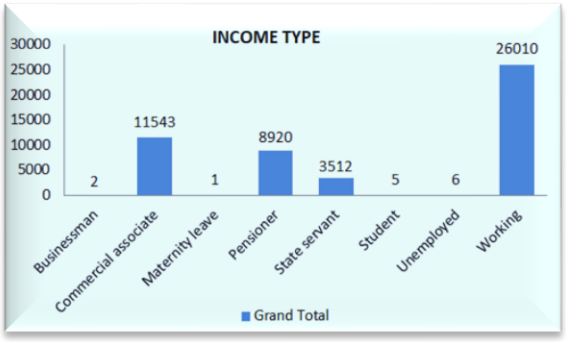


Segmented univariate analysis

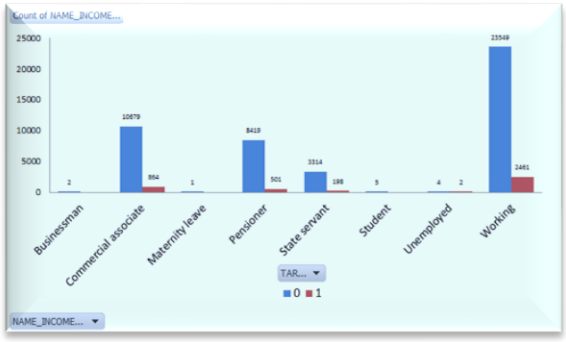


We can clearly see that most of the clients applying for loan are from income group 25000-275000. These may be small scale clients with limited finance requirement. The default is also higher in them.

Univariate analysis

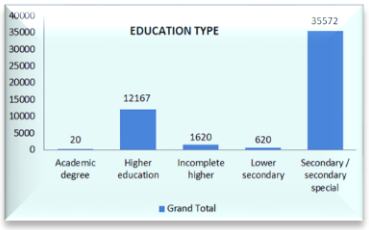


Segmented univariate analysis

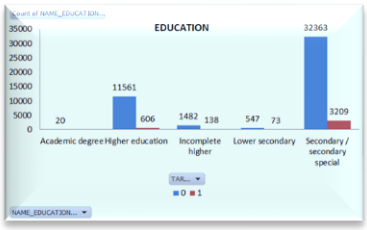


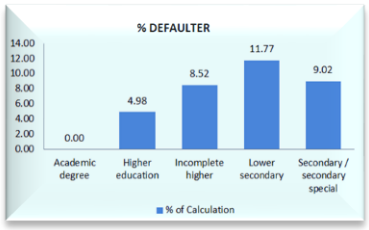
Most of the clients applying for loan are from working category as they are the one most in need of the loans followed by commercial associate. The default percentage is also higher in working.

Univariate analysis



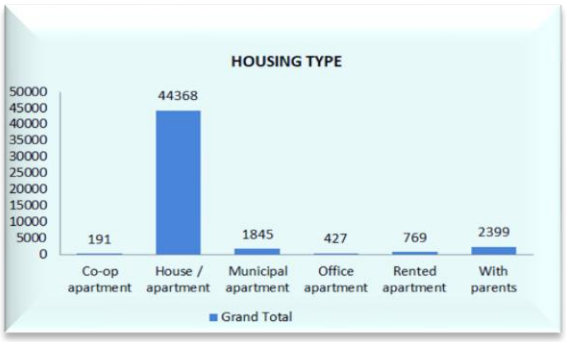
Segmented univariate analysis



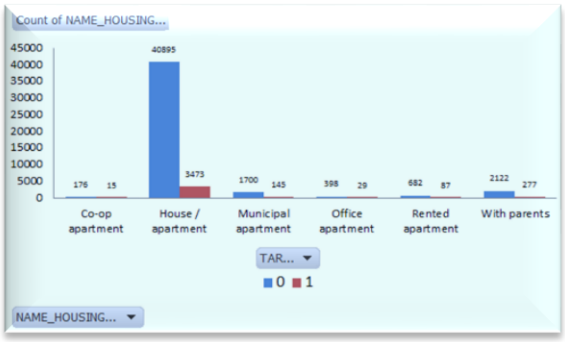


For education type the maximum clients were secondary/ secondary special pass out only followed by higher education . the possible reason for this could be that they are small business man. The default % is highest in lower secondary educated clients.

Univariate analysis

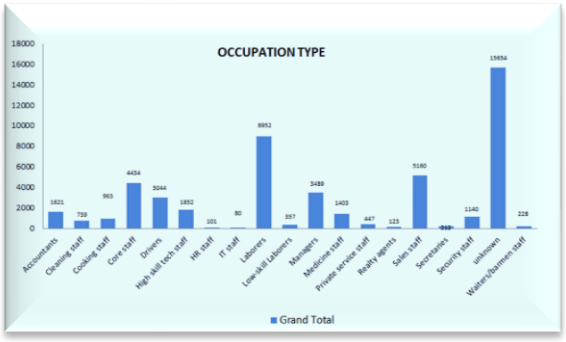


Segmented univariate analysis

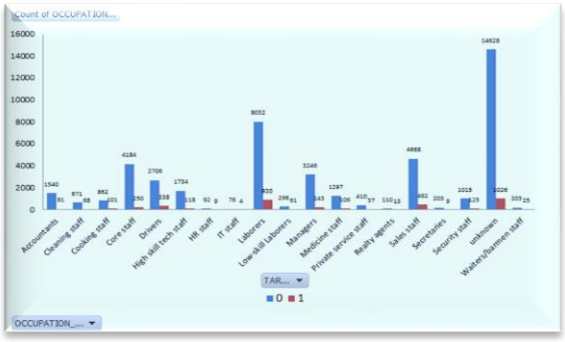


Most of the clients live in the house apartments and it is relatively much higher than other clients. The default % also would be higher as they comprise of almost 88%.

Univariate analysis



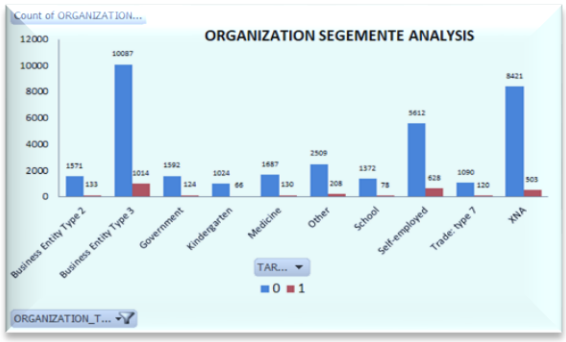
Segmented univariate analysis



For occupation of most of the clients are unknown. The possible reason could be data entry at the time of evaluating loan application. However, majority of the clients were laborers as they may be the one most in need of loan since they belong to lower income groups.

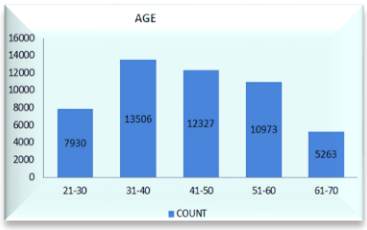
Univariate analysis

Segmented univariate analysis

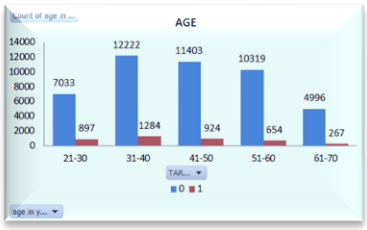


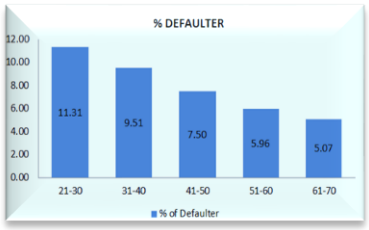
The maximum loans were applied by business entity type 3 followed by XNA. Business entity comprised of almost 22% of all the application under observation. The default rate is also higher. This could be due to failure of business.

Univariate analysis



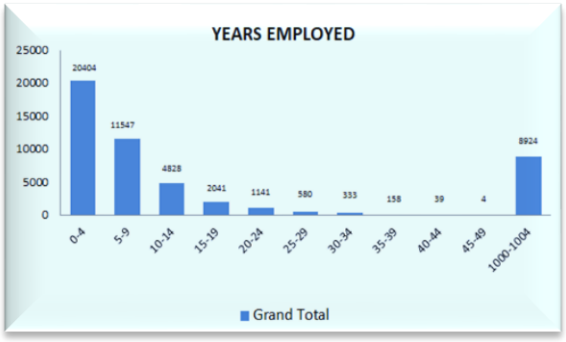
Segmented univariate analysis





Highest number of people who have applied for loan are from 31-40 age group. However the default % is highest for the clients between 21-30 age group. People above 60 age group have lowest probability of default.

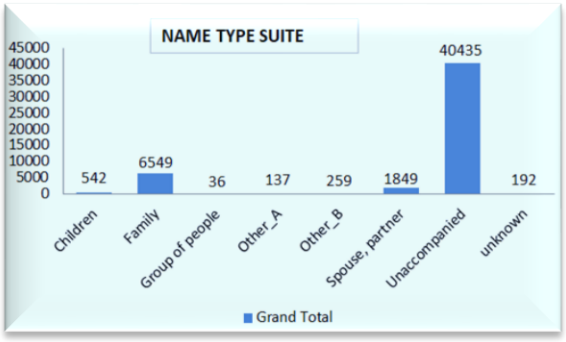
Univariate analysis



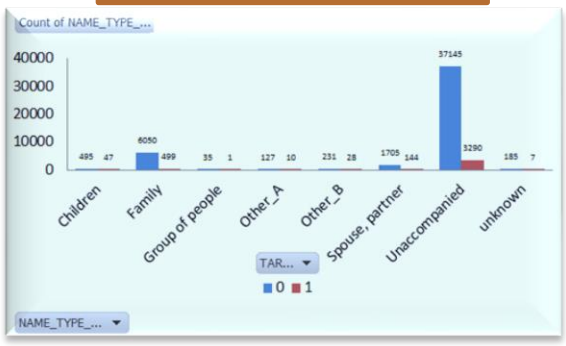
Segmented univariate analysis

There were 8924 rows with number of years employed to be 1004 and they were wrong data entry and also highlighted in the outliers points. However, most of the clients were employed for 0-4 years only and also the default rate is higher in them.

Univariate analysis

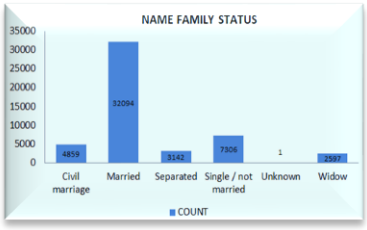


Segmented univariate analysis

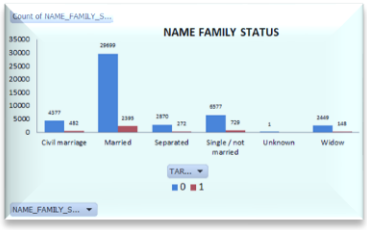


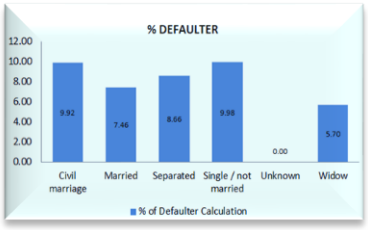
Though it does not matter with whom the client for applying for loan but still if we do analysis then we can say that most of the clients were unaccompanied which justifies our earlier statement. The default rate is also higher in them.

Univariate analysis



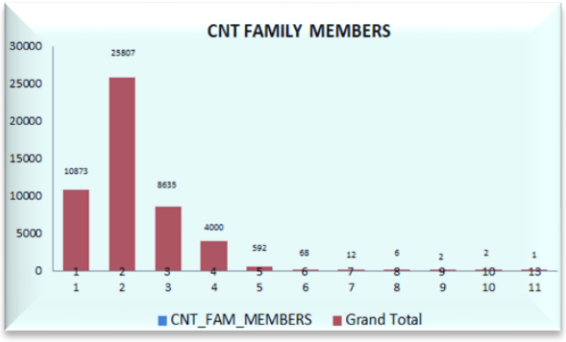
Segmented univariate analysis



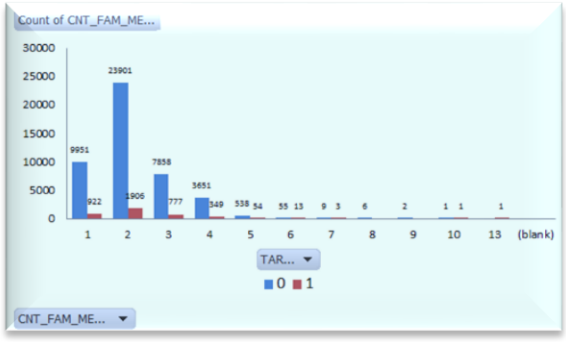


Most of the people who have taken loans are married, followed by Single/not married and civil marriage. In terms of the percentage of not repayment of loans, single/not married has the highest percentage of not repayment (10%), with Widows the lowest (exception being Unknown). The possible reason for higher married loans could be due to family responsibility.

Univariate analysis

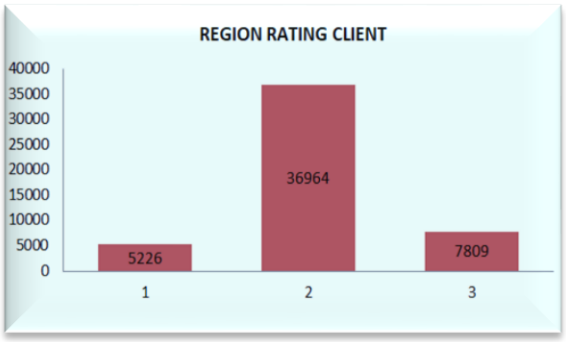


Segmented univariate analysis

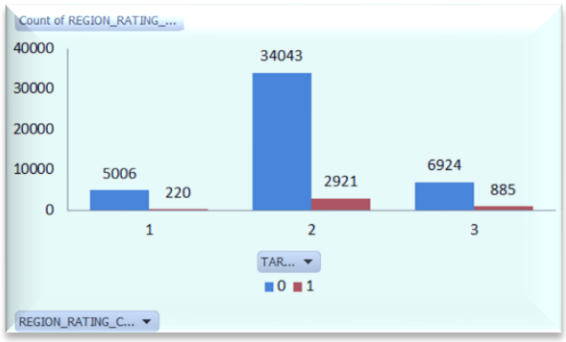


Again the count of family members does not affect much of the successful sanction of loan but we can see that those clients with only 2 family members are higherst. And also as a result their default percentage is also higher.

Univariate analysis



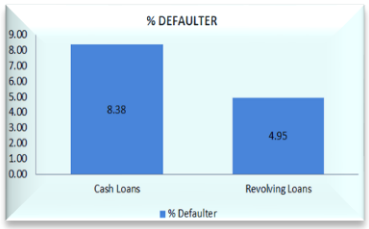
Segmented univariate analysis



Most of the clients were from region whose rating in bank records is 2. Also, the defaulters were highest from this region only i.e., 2921

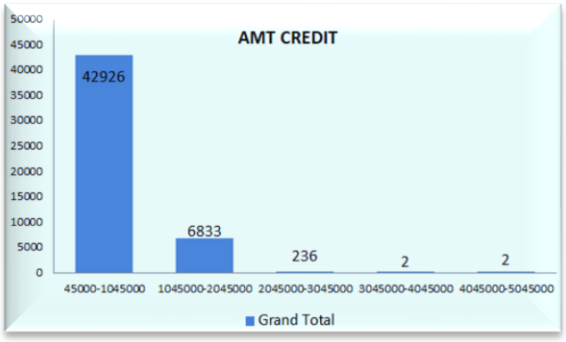
Univariate analysis

Segmented univariate analysis

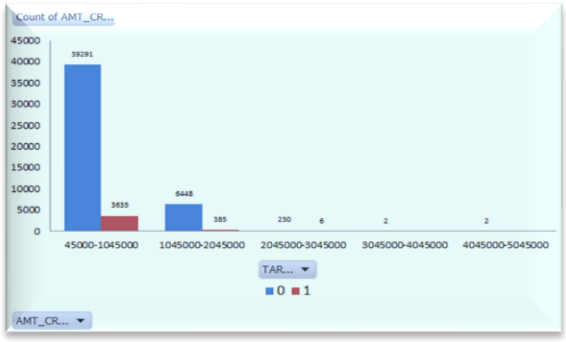


Revolving loans are just a small fraction (10.4%) of the total number of loans at the same time, a larger amount ofcash loans, compared with their frequency, are not repaid.

Univariate analysis



Segmented univariate analysis



We can see that most the clients have the requirement of loan amount upto 10,00,000 only as there could be much possibility that these are low income group people and also they have high chances of getting loan. The default percentage is also higher in them.

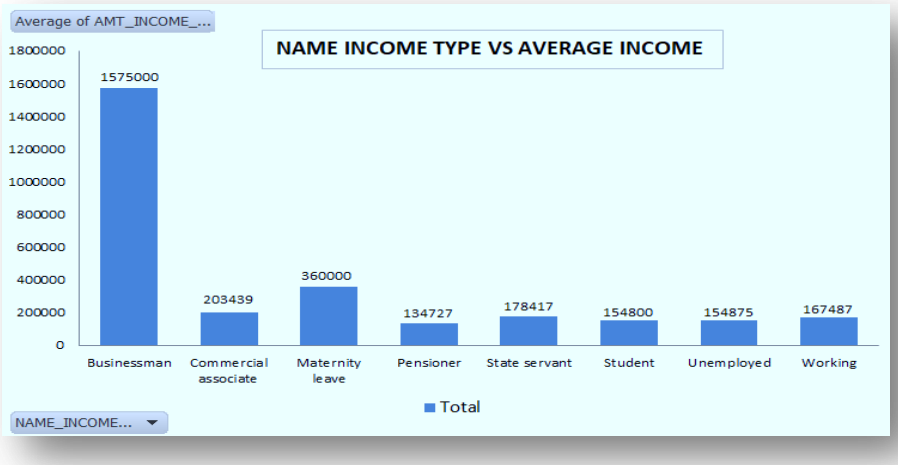
Other insights:

• Weekday app process start are irrelevant for analysis as bank don't have any specific days or don't prefer to given loan on specific days and not on other.

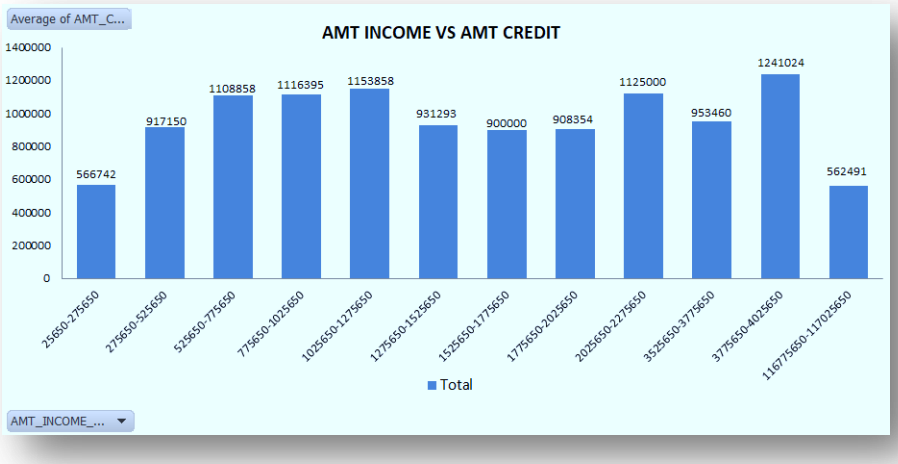
• Also count of children is irrelevant as whether a client is able or not to repay the loans doesn't much depend on the number of children the client has.

Bivariate analysis:

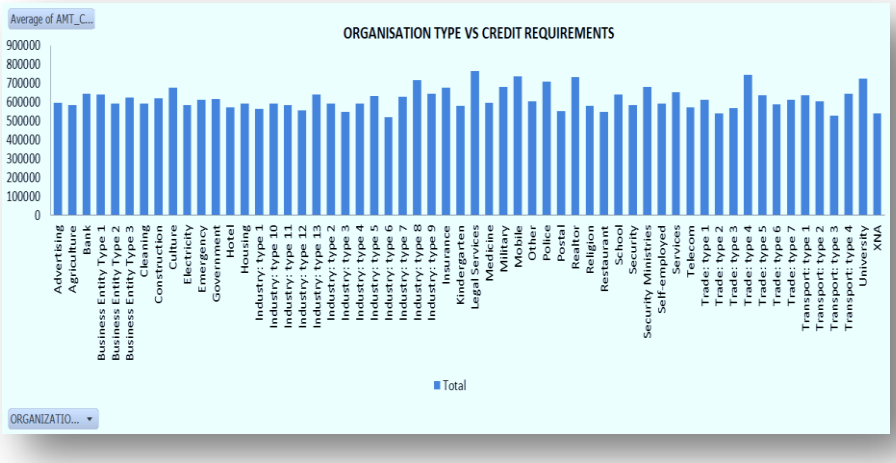
Bivariate analysis is a statistical method used in data analytics to analyze the relationship between two variables. Unlike univariate analysis, which focuses on a single variable, bivariate analysis examines how two variables vary together and whether there is a correlation or association between them. It helps to understand how changes in one variable may influence changes in another.



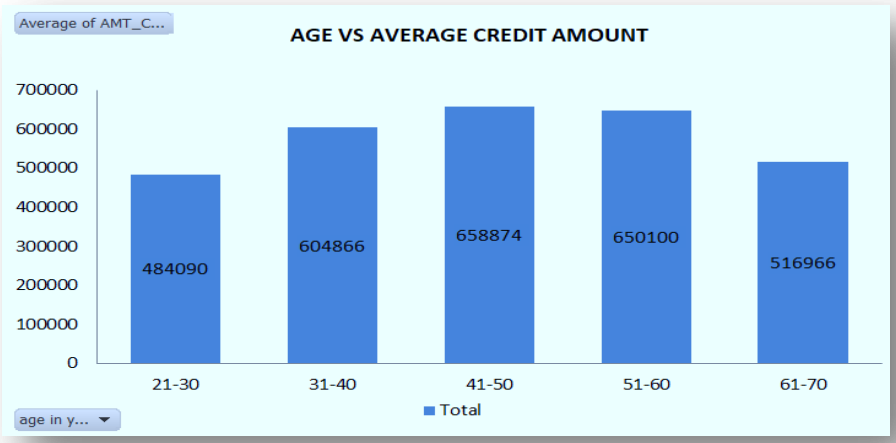
The average income of businessman is highest and the outlier though in working group of amount of 117000000 is not making it highest.



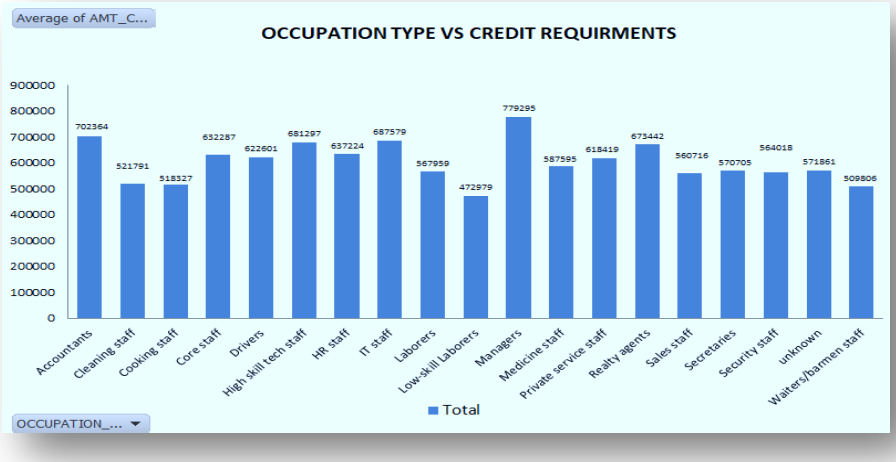
The average credit requirement of clients in the income group 3375000-4025000 is highest and those above 1167750000 is lowest which leads us to outlier.



The maximum average credit requirements is of legal services. This finding underscores the significance of legal needs within the studied population. Adequate resource allocation and tailored financial solutions are crucial to meet the specific credit demands associated with legal services.



The clients in the age group of 41-50 are need of more funds as compared to other age groups. Though the average credit requirements is more or less same for all the age groups.



We can clearly see that the average credit requirements of managers is highest. Followed by accountants. This suggests that individuals in managerial roles often seek larger credit amounts, potentially for personal or professional needs. Tailoring financial products to accommodate these requirements could cater effectively to this specific group's credit needs.

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

SOLUTION:

The top 10 correlation between different variable and the clients who are repayer is:

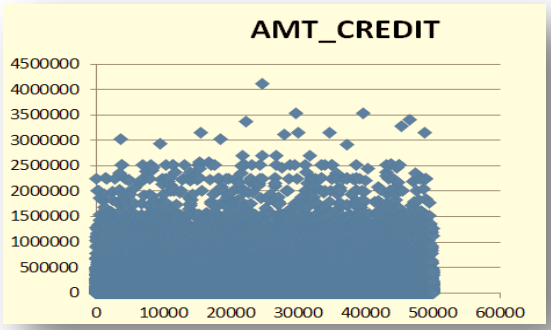
variable 1	variable 2	coreelation
OBS_60_CNT_SOCIAL_CIRCLE	OBS_30_CNT_SOCIAL_CIRCLE	0.998355381
AMT_GOODS_PRICE	AMT_INCOME_TOTAL	0.987244066
CNT_FAM_MEMBERS	CNT_CHILDREN	0.87923936
DEF_60_CNT_SOCIAL_CIRCLE	DEF_30_CNT_SOCIAL_CIRCLE	0.850953626
AMT_GOODS_PRICE	AMT_ANNUITY	0.776141898
AMT_ANNUITY	AMT_INCOME_TOTAL	0.770772965
YEARS_EMPLOYED2	age in years	0.62324914
AMT_ANNUITY	AMT_INCOME_TOTAL	0.451135696
AMT_GOODS_PRICE	AMT_INCOME_TOTAL	0.384675092
AMT_CREDIT	AMT_INCOME_TOTAL	0.377965752

The top 10 correlation between different variable and the clients who are defaulter is:

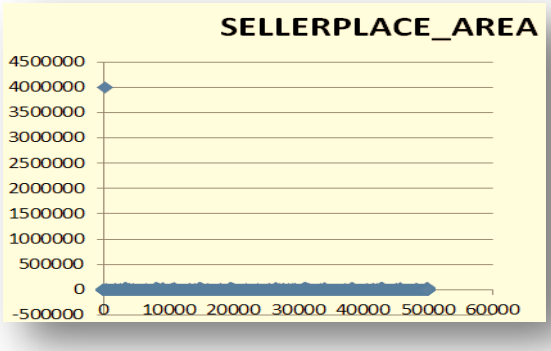
variable 1	variable 2	correlation
OBS_60_CNT_SOCIAL_CIRCLE	OBS_30_CNT_SOCIAL_CIRCLE	0.998064837
AMT_GOODS_PRICE	AMT_CREDIT	0.982432318
CNT_FAM_MEMBERS	CNT_CHILDREN	0.892521875
DEF_60_CNT_SOCIAL_CIRCLE	DEF_30_CNT_SOCIAL_CIRCLE	0.890496348
AMT_ANNUITY	AMT_CREDIT	0.749665201
AMT_GOODS_PRICE	AMT_ANNUITY	0.749705184
YEARS_EMPLOYED2	age in years	0.587858433
OBS_60_CNT_SOCIAL_CIRCLE	DEF_30_CNT_SOCIAL_CIRCLE	0.367887288
DEF_30_CNT_SOCIAL_CIRCLE	OBS_30_CNT_SOCIAL_CIRCLE	0.364900071
DEF_60_CNT_SOCIAL_CIRCLE	OBS_60_CNT_SOCIAL_CIRCLE	0.301256981

PREVIOUS APPLICATIONS

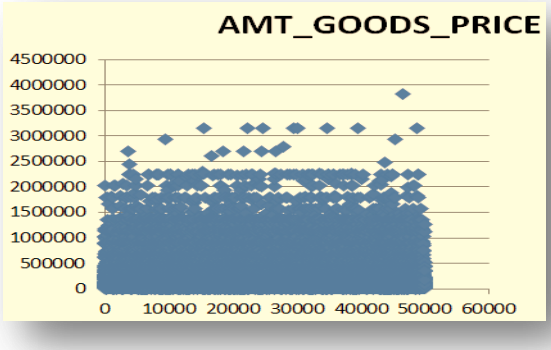
1.) Identify Outliers in the Dataset: Outliers can significantly impact the analysis and distort the results. You need to identify outliers in the loan application dataset.



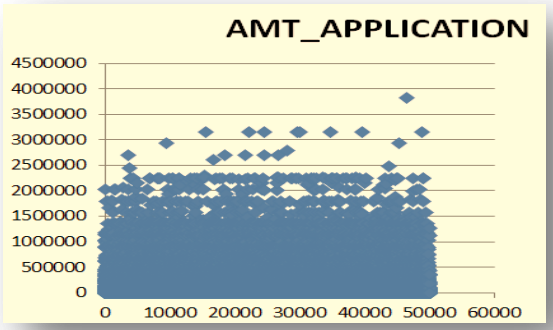
We can see significant outliers in amt credit. This will affect the measures of central tendency. Though the median lies around 78000 there are few values which are above 4000000.



Though the median is 10 yet there is few values that touched 4000000. This will affect our analysis.

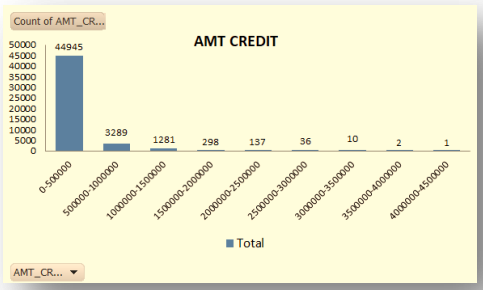


In amt goods price there are few values exorbitantly higher than the median. And this touched around 3700000.

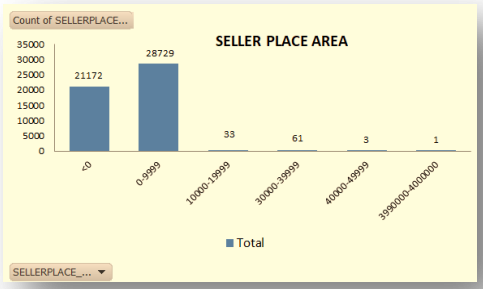


The presence of outliers in the loan amount applications indicates instances where requested loan amounts deviate significantly from the norm. Addressing these outliers is vital to ensure accurate risk assessment and fair lending practices.

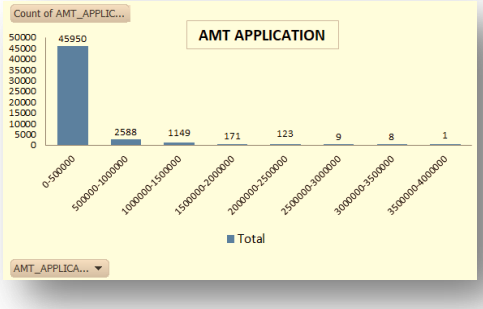
2.) Analyze Data Imbalance: Data imbalance can affect the accuracy of the analysis, especially for binary classification problems. Understanding the data distribution is crucial for building reliable models.



The majority of clients belong to credit group of 0- 500000. This constitutes of almost 90% of loan applications.

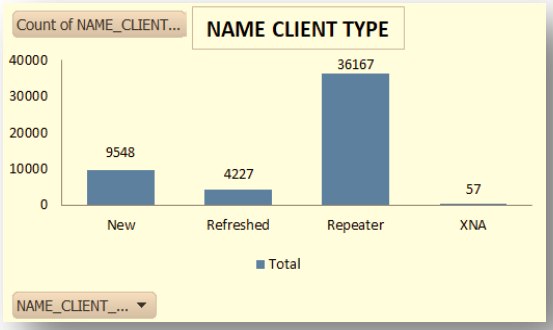


The seller place area contains outliers and also is highly imbalance. Area between 0-9999 constitutes maximum.

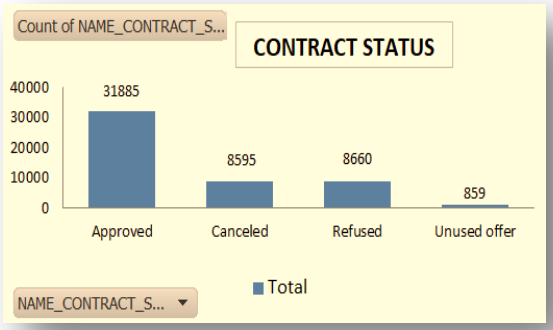


The amt application data is highly imbalanced. Amt application between 0-500000 constitues again upto almost 90% of data.

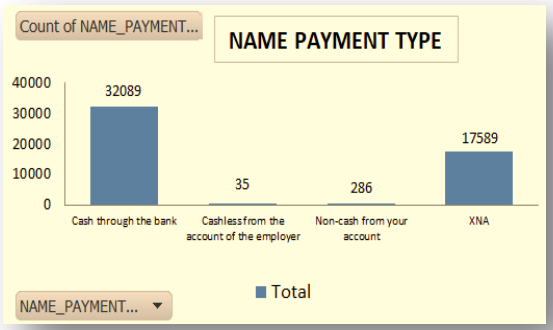
3.) Perform Univariate, Segmented Univariate, and Bivariate Analysis: To gain insights into the driving factors of loan default, it is important to conduct various analyses on consumer and loan attributes.



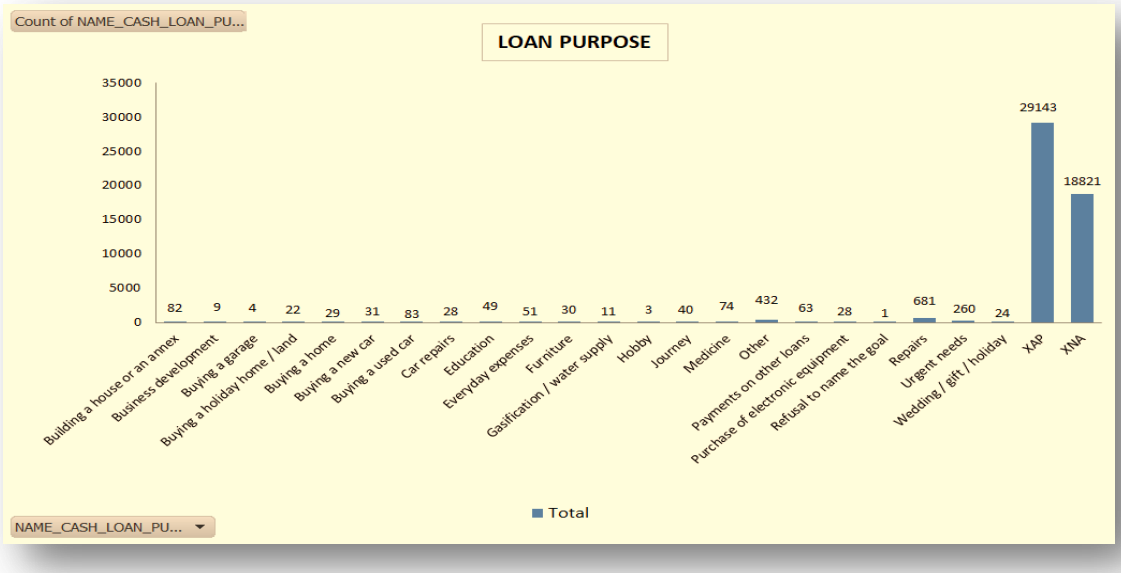
Most of the clients are repeaters and as a result will be easy for them to get the loan as they are aware of the process.



The maximum loan application were approved. This shows that most of the applicants are bonafide.

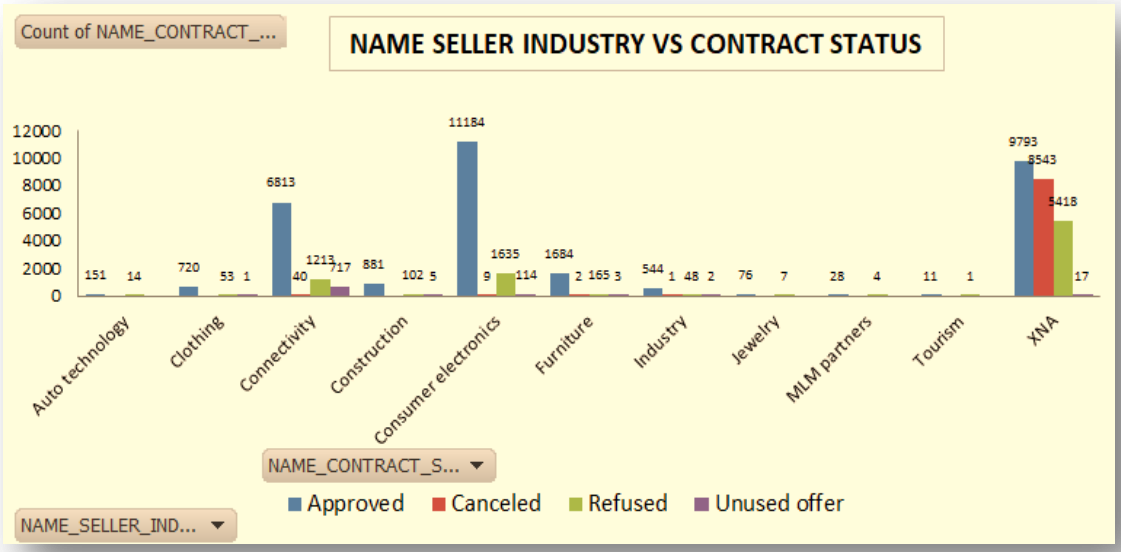


We can clearly see that most of the payment has been made by cash through the bank . this shows that people are promoting banking lines of payment giving boost to digital india.

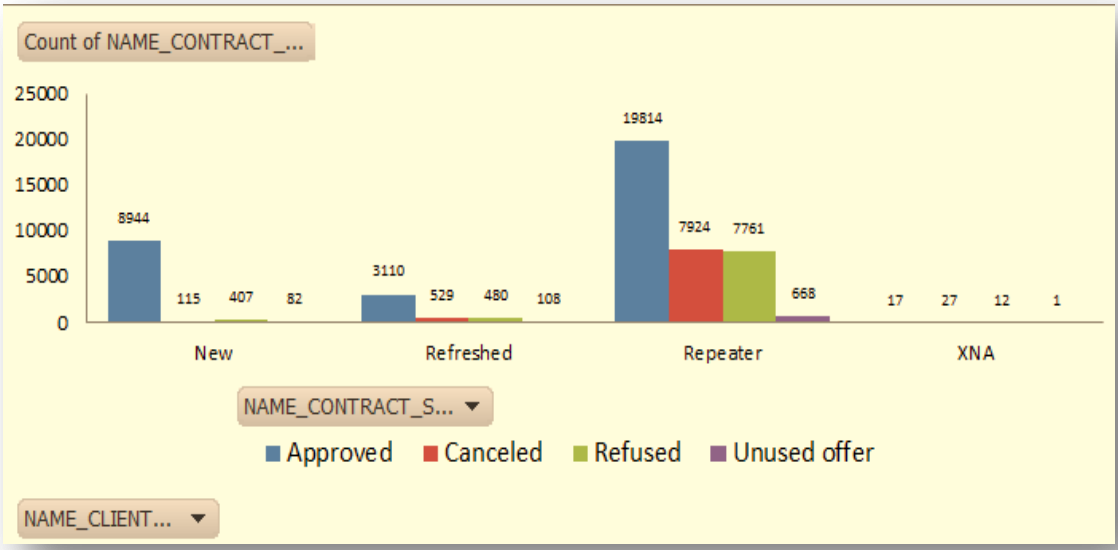


The loan purpose of most of the clients is XAP followed by XNA. They both constitutes almost 95% of loan application purpose.

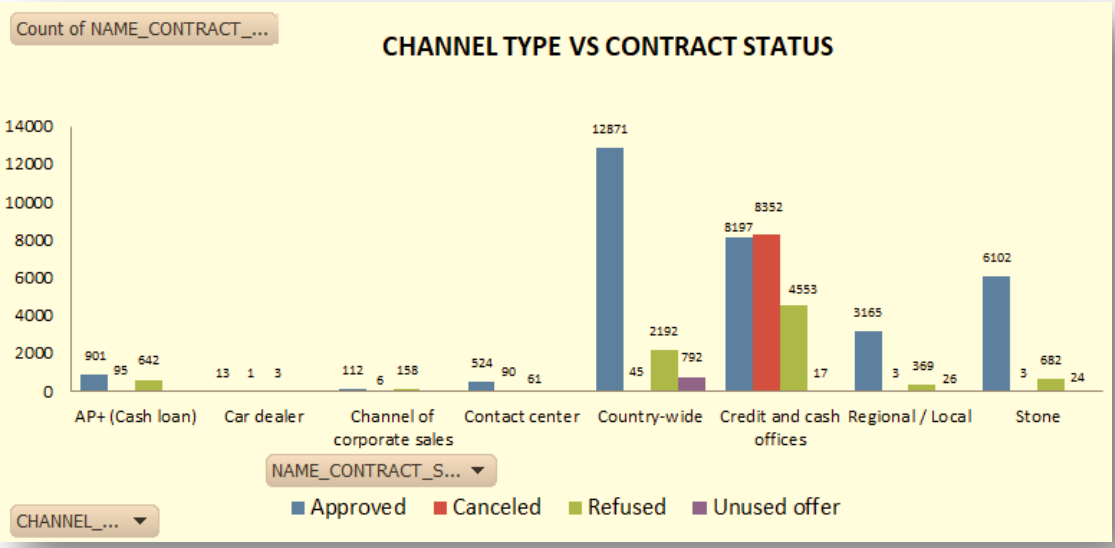
Bivariate analysis:



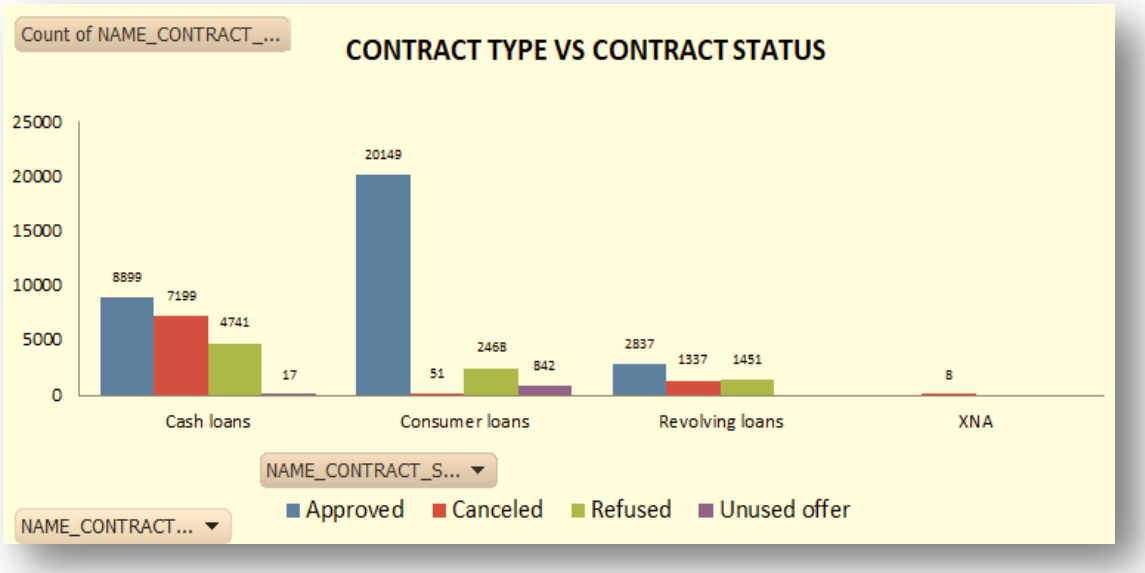
Most of the consumer electronics industry gets the loan approved followed by XNA and then connectivity. There are very few other data visible in the graph.



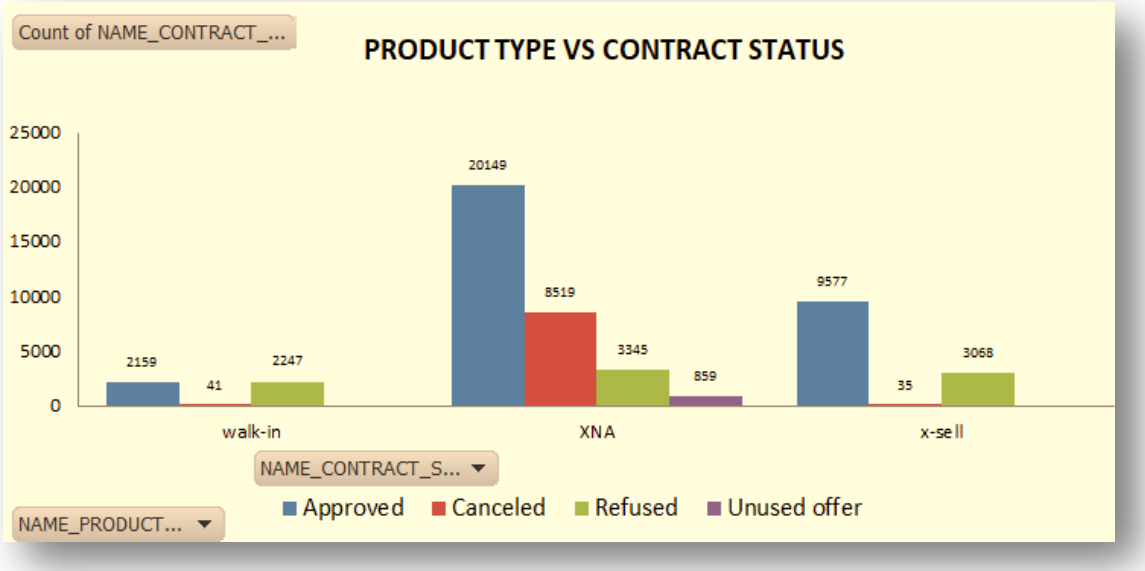
Most of the clients are repeater and hence have the edge over loan sanction as they as well versed with the process. So the rate of approval is also higher and this is indicated by the graph.



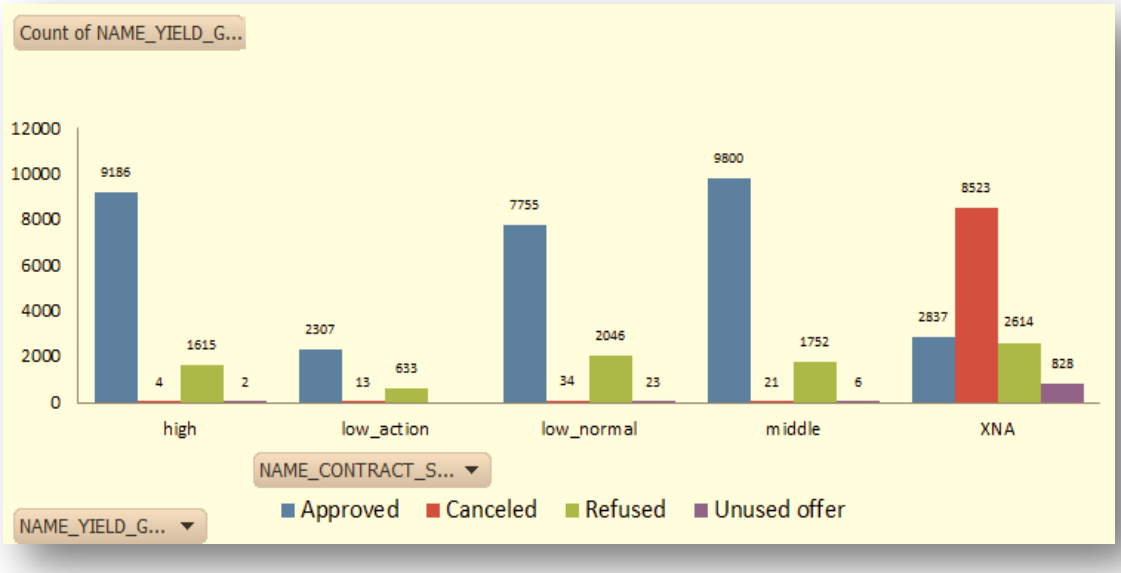
Majority of the loans approved were of country wide channel followed by credit and cash offices and the stone. Cash and credit offices also have maximum rejections.



Most of the consumer loans were approved and none were canceled. The cash loans are the one which were canceled and refused for maximum.



XNA were highly approved and the x-sell were second. The maximum cancelled were also XNA.



The group interest rate into small medium and high shows that the middle rate is highest approved followed by high and others are more or less same.

4.) Identify Top Correlations for Different Scenarios:Understanding the correlation between variables and the target variable can provide insights into strong indicators of loan default.

	AMT_APPLICATION	AMT_CREDIT	AMT_GOODS_PRICE	HOUR_APPR_PROCESS_START	DAYS_DECISION	SELLERPLACE_AREA	CNT_PAYMENT	NFLAG_INSURED_ON_APPROVAL
AMT_APPLICATION	1							
AMT_CREDIT	0.975771049	1						
AMT_GOODS_PRICE	0.999902539	0.993495986	1					
HOUR_APPR_PROCESS_START	-0.02175004	-0.03055564	-0.054489969	1				
DAYS_DECISION	0.13399106	0.137431211	0.278987358	-0.03546658	1			
SELLERPLACE_AREA	-0.003965725	-0.004949463	-0.007737876	0.015409673	-0.008624778	1		
CNT_PAYMENT	0.671340368	0.666621175	0.663684765	-0.065270021	0.228256641	-0.001015924	1	
NFLAG_INSURED_ON_APPROVAL	0.262838448	0.270447512	0.249716587	-0.11949571	-0.042417365	-0.010168303	0.311455155	1

Top 10 :

Variable 1	Variable 2	Correlation
AMT_GOODS_PRICE	AMT_APPLICATION	0.999902539
AMT_GOODS_PRICE	AMT_CREDIT	0.993495986
AMT_CREDIT	AMT_APPLICATION	0.975771049
CNT_PAYMENT	AMT_APPLICATION	0.671340368
CNT_PAYMENT	AMT_CREDIT	0.666621175
CNT_PAYMENT	AMT_GOODS_PRICE	0.663684765
NFLAG_INSURED_ON_APPROVAL	CNT_PAYMENT	0.311455155
DAYS_DECISION	AMT_GOODS_PRICE	0.278987358
NFLAG_INSURED_ON_APPROVAL	AMT_CREDIT	0.270447512
NFLAG_INSURED_ON_APPROVAL	AMT_APPLICATION	0.262838448

CONCLUSION:

After performing the analysis ,we can rectify whether a client will repay the loan or not. Also, the people who are likely to face problem in loan repayment are labourers .Also people with Secondary /secondary special education might face problem in loan repayment. Moreover, those who are living in house/apartment are facing difficulty in loan repayment (may be because of extra home loan, EMIs and so on).people opting for cash loan faces difficulty in doing the same.

THANK YOU

Excel files for data analysis : CURRENT APPLICATION

PREVIOUS APPLICATION

NAME : ABHAY SAXENA

EMAIL ID : abhaysaxena700@gmail.com