TRAINITY

PROJECT: BANK LOAN CASE STUDY

PRESENTED BY: MUDIT VYAS

Video Link: Click Here!

Aim: Use Excel Workbench to pre-process data (handle missing values, remove duplicates, data type conversion, etc.), explore the data to understand relation between different variables and provide valuable insights by using visualizations to help tell stories which assist to elaborate the finding in much better way.

Project Overview: Before lending money to someone we think numerous times and make sure the lend money will be returned back based on trust, surety and guarantee. Think about Banks, Online marketplaces and Insurance Companies who have to lend Crores of Money. For that they make sure Client is well capable of paying instalments at time which surely takes multiple factors into consideration like Client's Occupation, Assets, Income, Locality, credit score, and many such parameters.
In this Exploratory Data Analysis we look upon such parameters; how they affect the loan repayment.

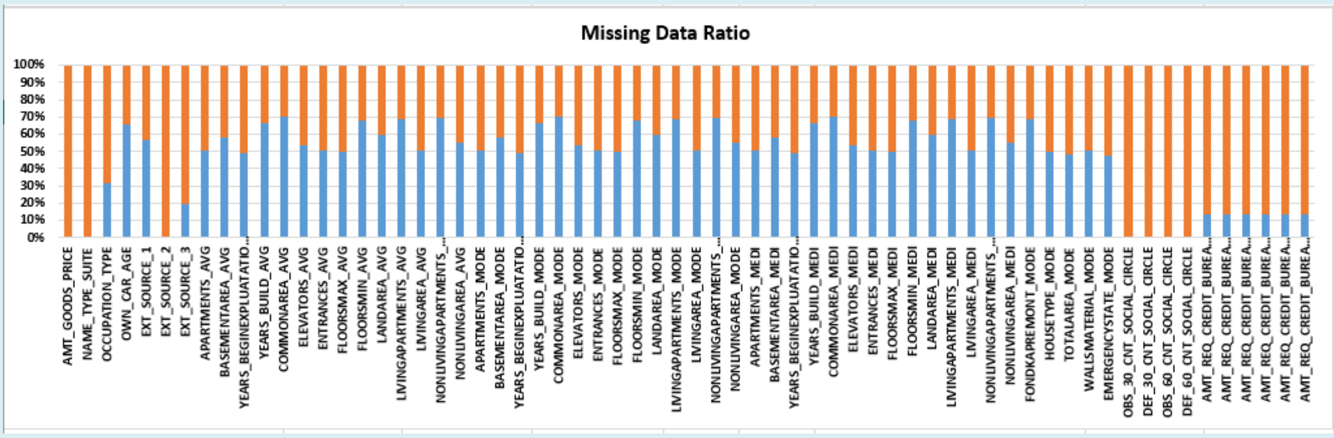
Tech-Stack Used: Microsoft Excel 2016 is used however it's recommended to use latest Version of Excel, but the 2016 version fulfilled the requirements as per the project.

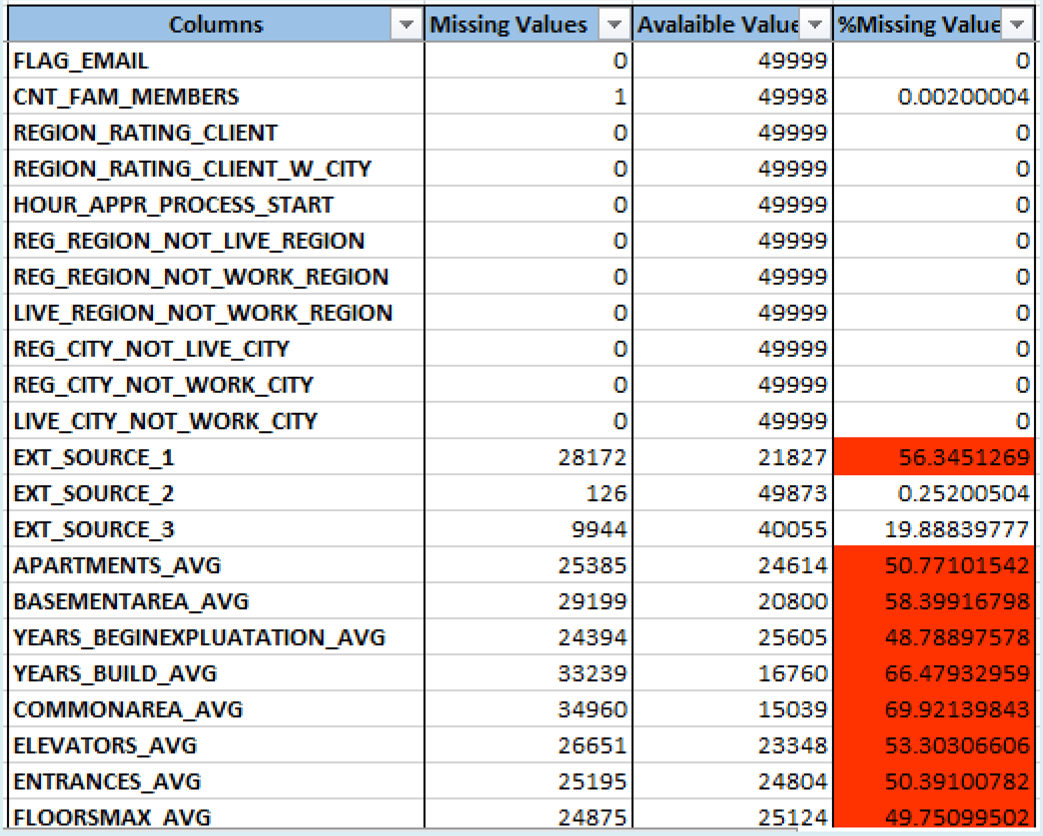
Data Pre-processing/Data Cleaning: Before Analysis we check the dataset provided. Certainly it needs to be cleaned and missing values to be handled.

- We observe each column contains unique data and missing values in each column is different in counts.
- A lot of columns are clean with no missing values, while some columns seems not to have much data
- So a proper dedicated cleaning, missing data handling and Outliers handling has to be conducted for précised results and Analysis.
- For easy understanding the data is consolidated in Categories such as General Information, Loan Amount Details, Documents Submitted, etc.
- The given dataset of 'application_data' have 50000 record count and 122 variable count.
- **We choose to deal with 'application_data' only as this is the latest data which needs to be examined and give important insights to use in future. Yes it is good to check previous data but there is not much relation with 'application data'.**

A) Missing Data Handling:-

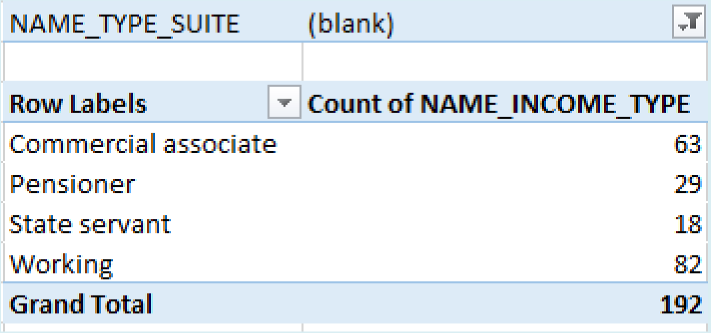
- Using COUNTBLANK () number of blank cells are counted in each column.
- A lot of columns have no missing data, so we ignored them for this task. For columns with Missing Data, We create a column chart to visualise its ratio as per total data, using conditional Formatting we identify those columns in dataset.
- Variable with more than 30% of missing data is decided not to be used for Analysis. Variable such as Average, Mode and Median of Area/Apartment client lives have missing data more than 45% which might not account into any right results.
- Although dimension of apartment/locality can be a factor to look upon, but it's significance is not as much as other factors. So we can decide to drop those columns. Similarly age of Car owned can also be dropped.
- Some parameters like 'External Source' 1 rating and 'Occupation type' are kept as there significance is relatively more. After removal of variables the total variable count drops to 65.



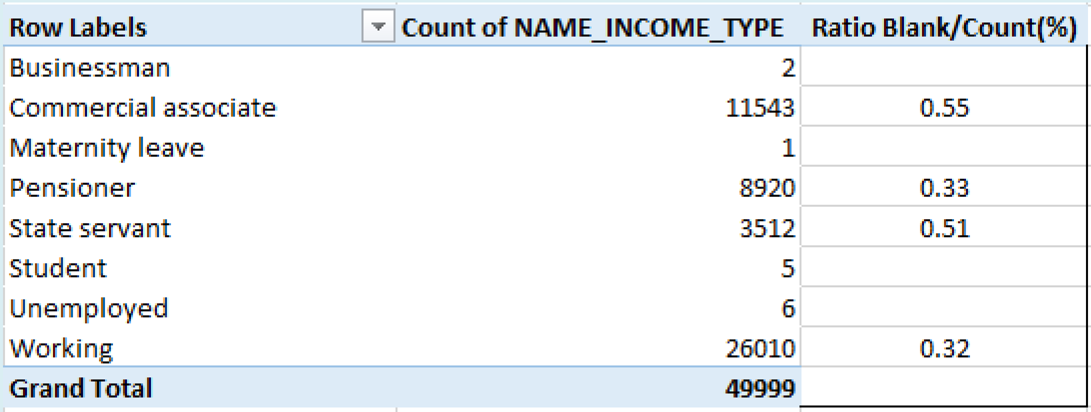


Columns	Missing Values	Avalaible Value	%Missing Value
FLAG_EMAIL	0	49999	0
CNT_FAM_MEMBERS	1	49998	0.00200004
REGION_RATING_CLIENT	0	49999	0
REGION_RATING_CLIENT_W_CITY	0	49999	0
HOUR_APPR_PROCESS_START	0	49999	0
REG_REGION_NOT_LIVE_REGION	0	49999	0
REG_REGION_NOT_WORK_REGION	0	49999	0
LIVE_REGION_NOT_WORK_REGION	0	49999	0
REG_CITY_NOT_LIVE_CITY	0	49999	0
REG_CITY_NOT_WORK_CITY	0	49999	0
LIVE_CITY_NOT_WORK_CITY	0	49999	0
EXT_SOURCE_1	28172	21827	56.3451269
EXT_SOURCE_2	126	49873	0.25200504
EXT_SOURCE_3	9944	40055	19.88839777
APARTMENTS_AVG	25385	24614	50.77101542
BASEMENTAREA_AVG	29199	20800	58.39916798
YEARS_BEGINEXPLUATATION_AVG	24394	25605	48.78897578
YEARS_BUILD_AVG	33239	16760	66.47932959
COMMONAREA_AVG	34960	15039	69.92139843
ELEVATORS_AVG	26651	23348	53.30306606
ENTRANCES_AVG	25195	24804	50.39100782
FLOORSMAX_AVG	24875	25124	49.75099502

- The annuity amount is depends on various banking factors and is decided by bank itself. So it’s hard to presume any value and use. For analysis being we take average of Ratio of Annuity and Total and use it to calculate missing Annuity data.
- The Data for Pensioners, Students and Unemployed is missing in Occupation type, which is well understood, so we fill those missing data with N/A. After this update the remaining data’s missing to available ratio is low enough to approach for further analysis.

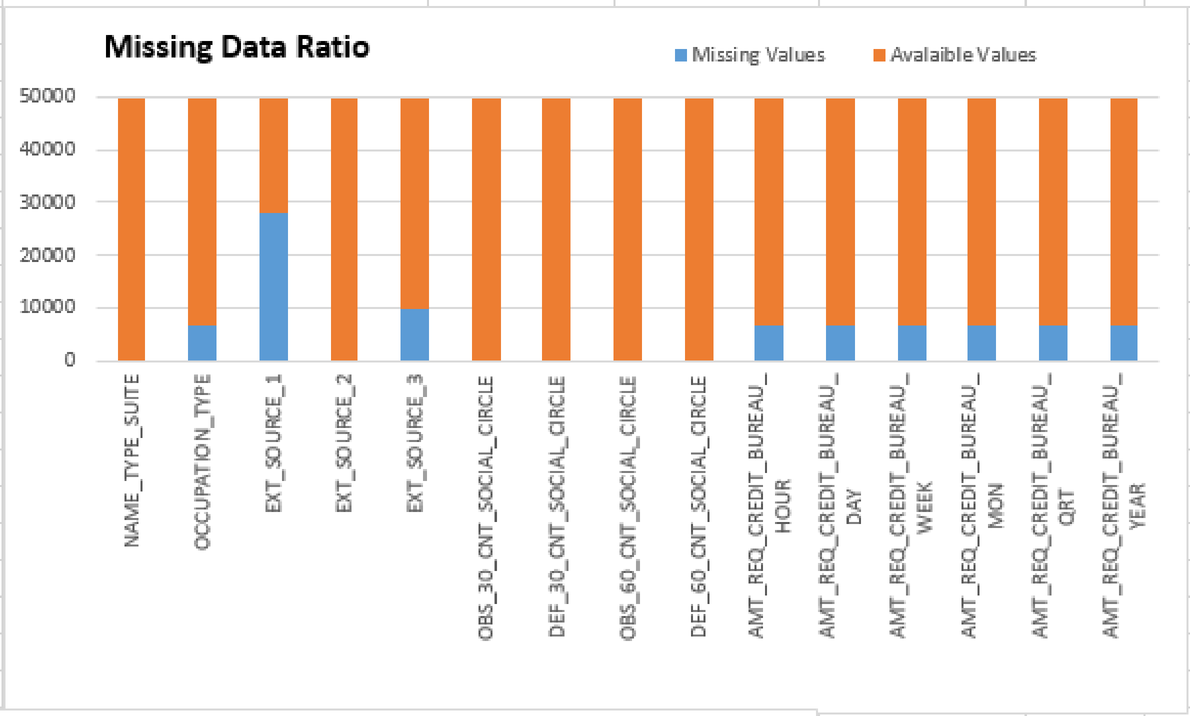


NAME_TYPE_SUITE	(blank)
Row Labels	**Count of NAME_INCOME_TYPE**
Commercial associate	63
Pensioner	29
State servant	18
Working	82
Grand Total	**192**



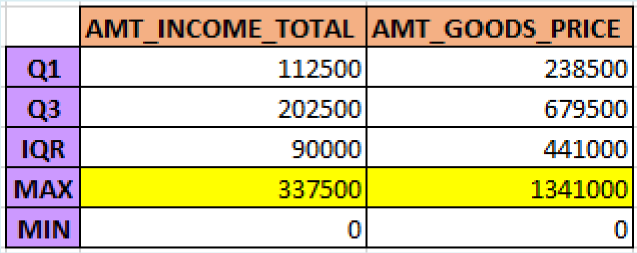
Row Labels	Count of NAME_INCOME_TYPE	Ratio Blank/Count(%)
Businessman	2	
Commercial associate	11543	0.55
Maternity leave	1	
Pensioner	8920	0.33
State servant	3512	0.51
Student	5	
Unemployed	6	
Working	26010	0.32
Grand Total	**49999**	

- For Remaining missing data, we keep the data as it is and deal with it during further analysis.



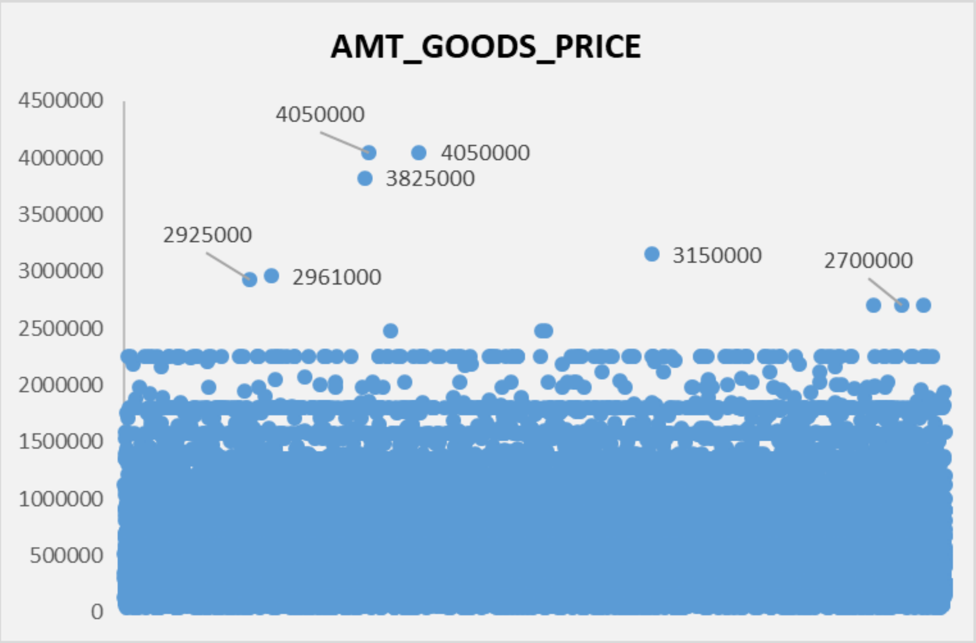
B) Outliers:-

- Wechoose‘CNT_CHILDREN’,’AMNT_INCOME_TOTAL’,’AMT_GOODS_PRICE’,’DAYS_EMPLOYED’, columns for Outliers check.
- Potential outliers are found first by data visualisation using scatter plot chart. Using QUARTILE, IQR, MAX and MIN functions the extreme values of columns are calculated.
- The limit values of Total income and goods price is calculated and plot is checked. Although there are multiple values which stands out of given limits, but on practical notes some clients can have loan demands which goes beyond the normally asked values. So we actually don’t consider them as outliers and keep them for analysis.

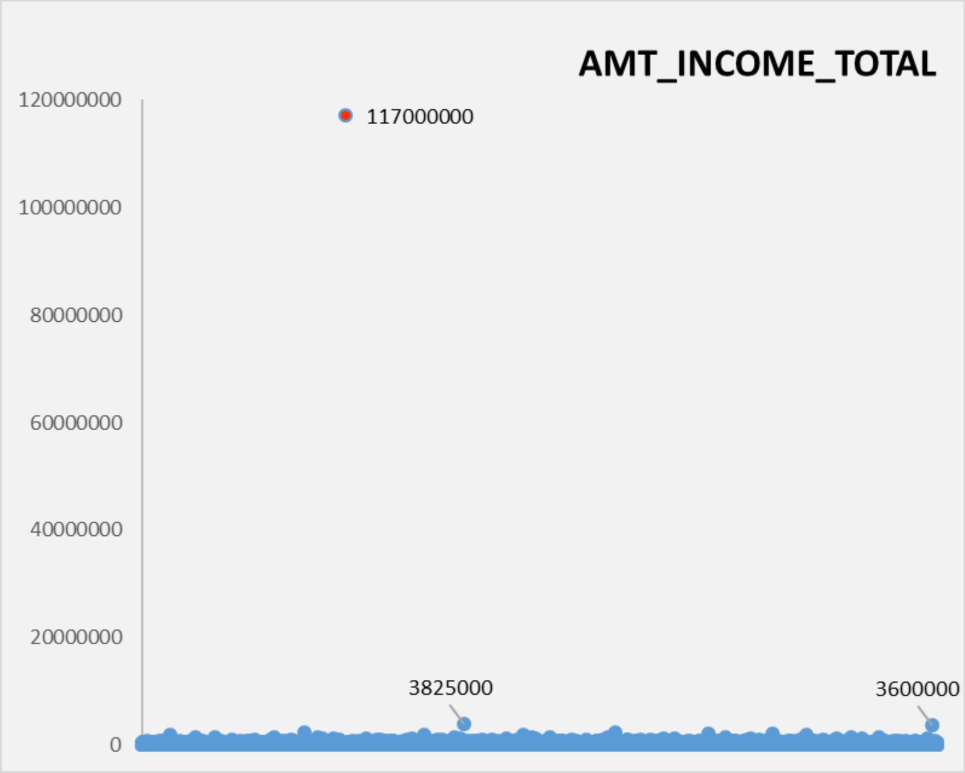


	AMT_INCOME_TOTAL	AMT_GOODS_PRICE
Q1	112500	238500
Q3	202500	679500
IQR	90000	441000
MAX	337500	1341000
MIN	0	0

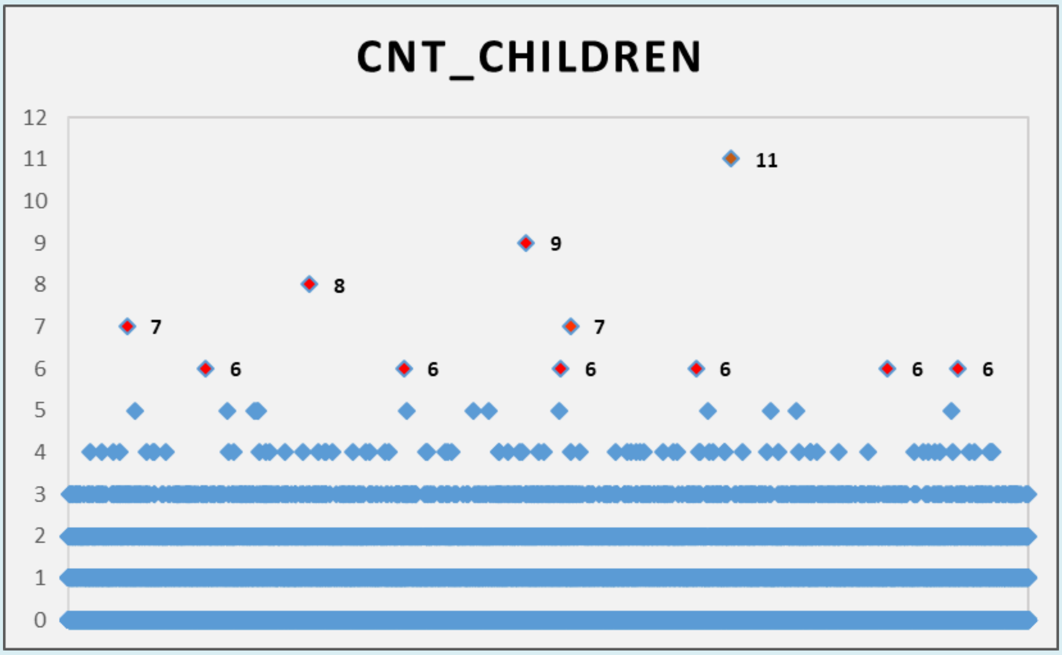
(Since Minimum Salary and Loan cannot be negative, we put the minimum value as 0)



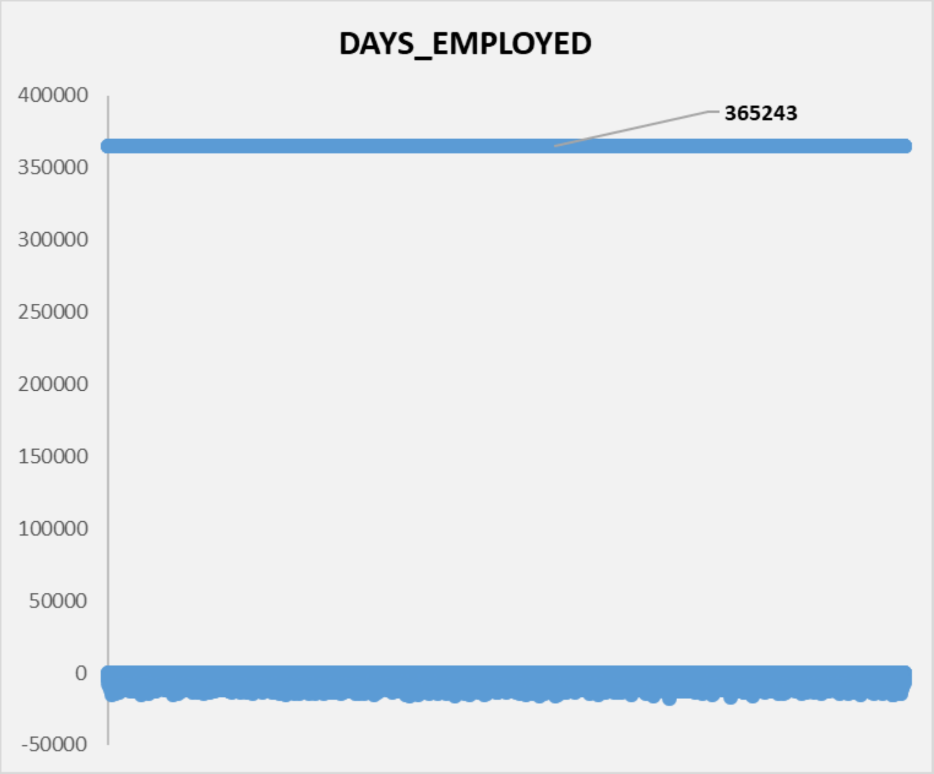
- There is one value Income value which stands out in the dataset, hence that outlier is removed. Although there are a few values which stands out of data limit, but on checking applied loan amount it’s pretty justified that client with such a high salary can also apply for Loan.



- The ‘CNT_CHILDREN’ column is plotted. Now in reality it’s unusual to have more than 5 children. Where having 4 children gives bit of a thinking, having 8/11 children is something which should be ignored or removed from analysis. We mark those data points with more than 5 children and check later during analysis.



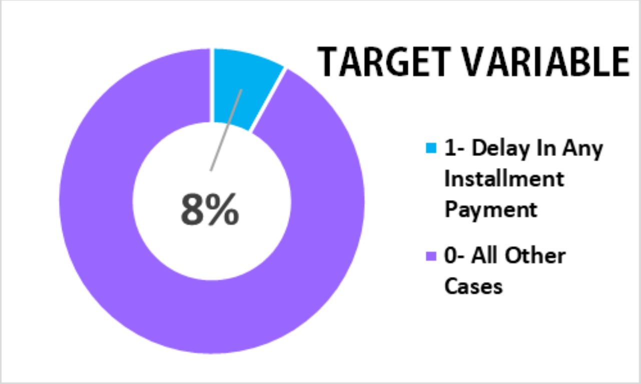
- When ‘DAYS_EMPLOYED’ column is plotted an unusual value of 36256 is observed for a lot of clients, while other values are negative. On a close inspection, these values are for Pensioners. It makes sense as pensioners are retired and 36256 days means more than 1000 years! In order not to have any ambiguity in column values we change it to 0.



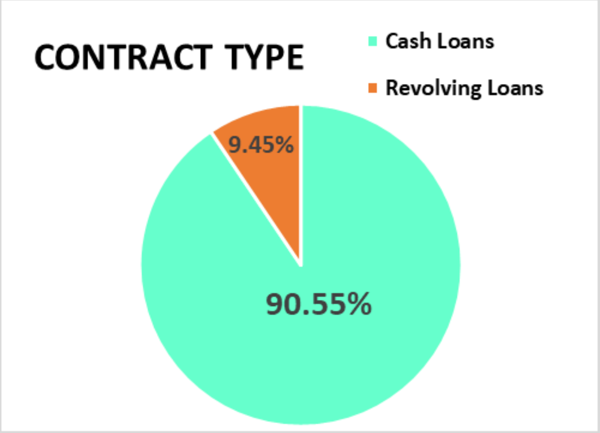
C) Data Imbalance:-

Our Primary focus is Target Variable i.e. is there delay in payment of any Instalment by client. So our first focus goes on factors which should not have any flag values or Data Imbalance.

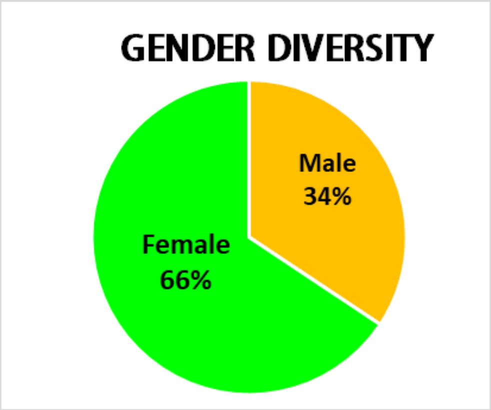
- About 8% of client's delayed in timely instalment payment. Now this figure has to be closely analysed to know what could be probable reason of delay in Loan Payment. Therefore we shall check behaviour of loan payment with other factors.



- About 9.5 % of loans applied are Revolving Loans. It's Important to know as Revolving Loans Holds more risk in loan payment than Cash Loans.



- 2/3 of Applicants are Female. This is not actually an Imbalance but gender of Applicant can effect on approval of loan. In Indian a lot of benefits such as priority lending, Subsidized Interest rates, schemes, etc. However there are possibilities this could lead to delay in instalment payments.



- If correct or reachable contact details are not provided by client, there are high probability of fraudster. The applicant might simply vanish after getting the loan.

Mobile Phone Reachable?	Count	Ratio(%)
Yes	49898	99.8
No	101	0.2

Email Provided By Client?	Count	Ratio(%)
Yes	47216	94.43
No	2783	5.57

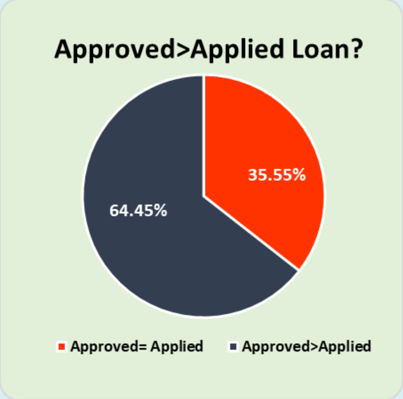
Mobile is unreachable for 0.2% clients. In % this looks less but in numbers- 101 clients Contact is not reachable which is concerning. We also see about 2800 clients have not Provided email address.

D) Factors, Variations and Effect - Univariate, Segmented Univariate and Bivariate Analysis: -

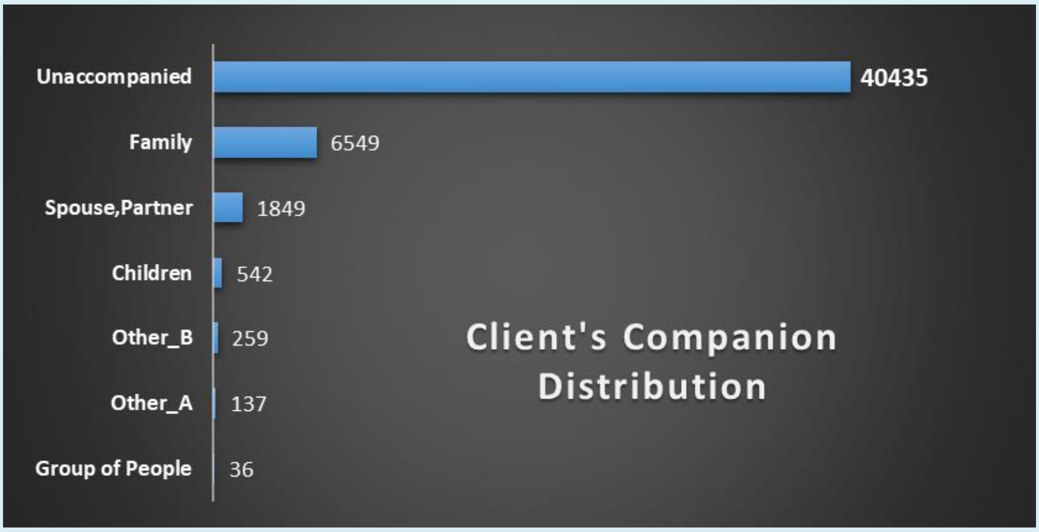
Our target variable is loan instalment payment. Let's explore how target variable behaves with other variables. Let's also look upon distribution of each variable.

	Average	Median	Mode	Standard Deviation	Max	Min
Client Income	168430.9124	145800	135000	99165.41372	3825000	25650
Loan Applied	599700.5815	514777.5	450000	402411.4096	4050000	45000
Loan Approved	538907.2974	450000	450000	369829.4108	4050000	45000

- The average Income is around 1.7 Lakh while Applied Loan average is around 3.5 times of average Income.
- Even if 65% of loan approved values are more than their corresponding applied values, the average of Approved values is less than Average of Applied values.

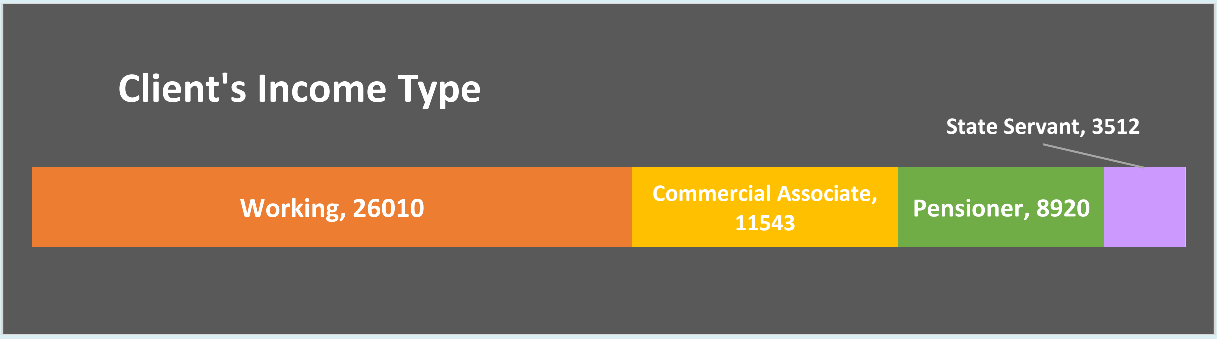


- During Loan Application client's Companion also makes difference. Generally Nominees or Witness for surety are preferred, hence having accompany shall be beneficial for both bank as well bank as a matter of trust.

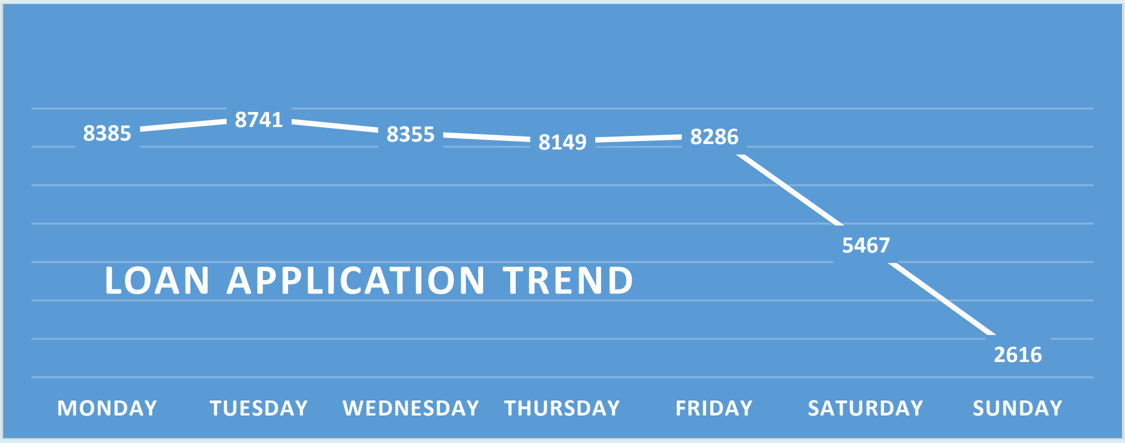


Here Data shows more than 90% clients are unaccompanied. Other parameters have to be considered majorly.

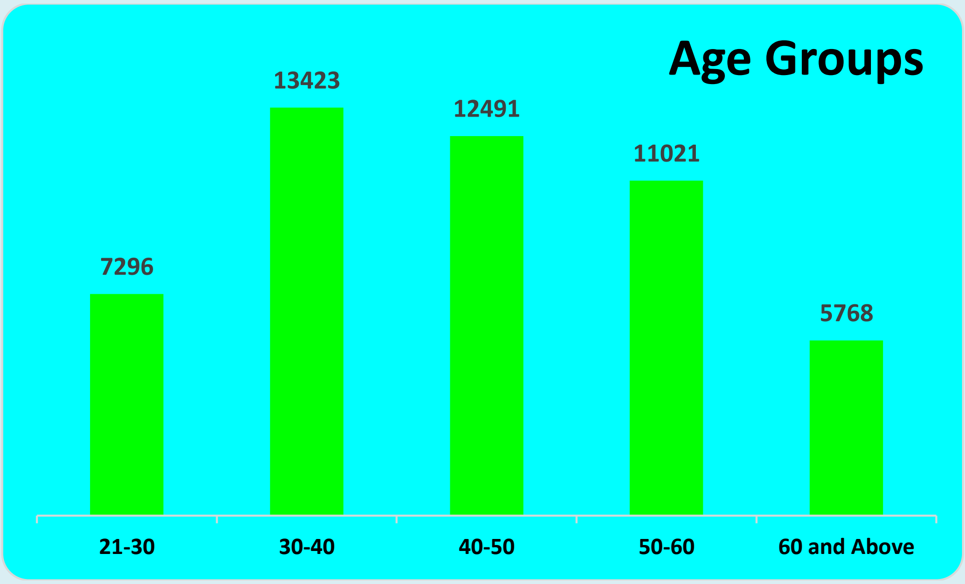
- Income type of a client also matters a lot. A client with a steady and constant source of Income is assumed to pay Instalments more likely than other cases. State Servant and Pensioners are out of focus yet it is good to know their total Income. A good number of students were expected for study loans but there number is in single digits. A large number of applicants have Working and Commercial Associate Income Type. The only risk here would be sudden job loss or loss in market.



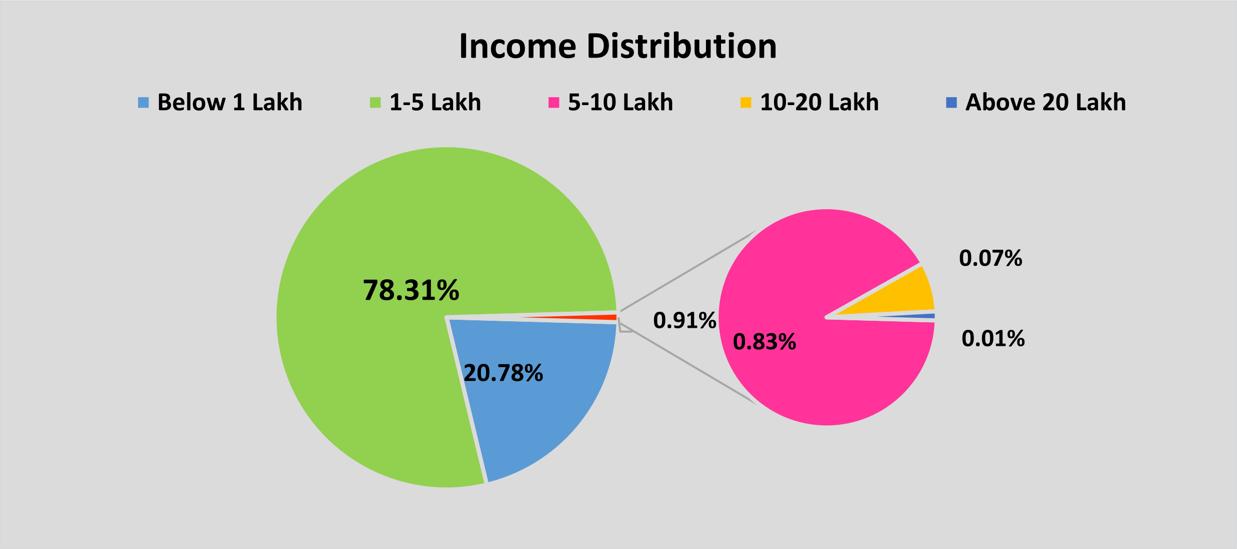
- On which days loan applications was highest? It was assumed that on weekends the applications would be higher than other days, but the count is lowest those days. People probably believe banks might be efficient on weekdays hence higher chances of getting loan approved.



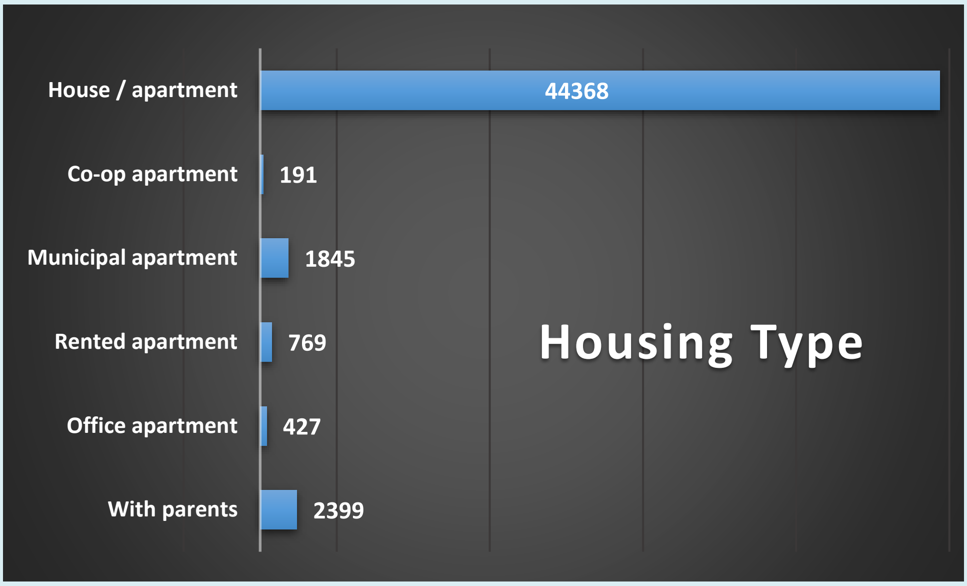
- Using TRUNC() function Age in Years of each client were calculated and categorized. Count of applicants in each category was obtained to understand the trend of loan application among people. Most client's age falls between 30-40 years followed by 40-50 years. This is the time when people come up with multiple responsibilities or plan to buy assets like property, vehicles, establish business or any other relevant reason.



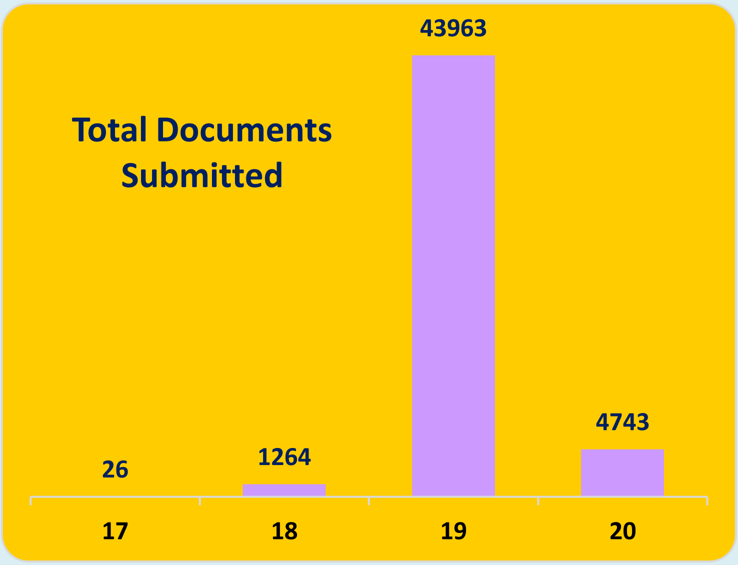
- Salary of client plays important role in deciding annuity, interest rates, loan to be credited and surety of repayment. More than 98% of applicants have Income less than 5LPA. So while approving Loan amount, it has to be stamped if applicant can really repay the amount in given duration.



- Housing type can be relate to loan repayment. If client lives in rented house a portion of their income shall go into rent payment. Those living with parents are either not able to afford housing or they are settled enough to live with them. Such confirmations can be made by knowing their salary caps and occupation type.

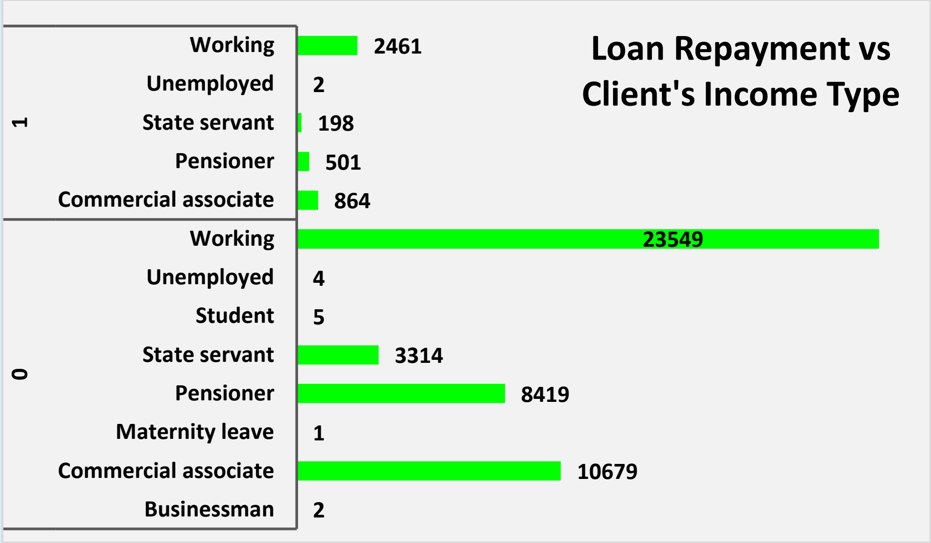


- Documents are necessary to monitor different steps of loan process and to have real evidence/requirement for security and safety purpose. The bank requires 20 additional documents other than the main form. A lot of applicants had submitted 19 forms, we can assume the 20th form a non-essential document such as passport.

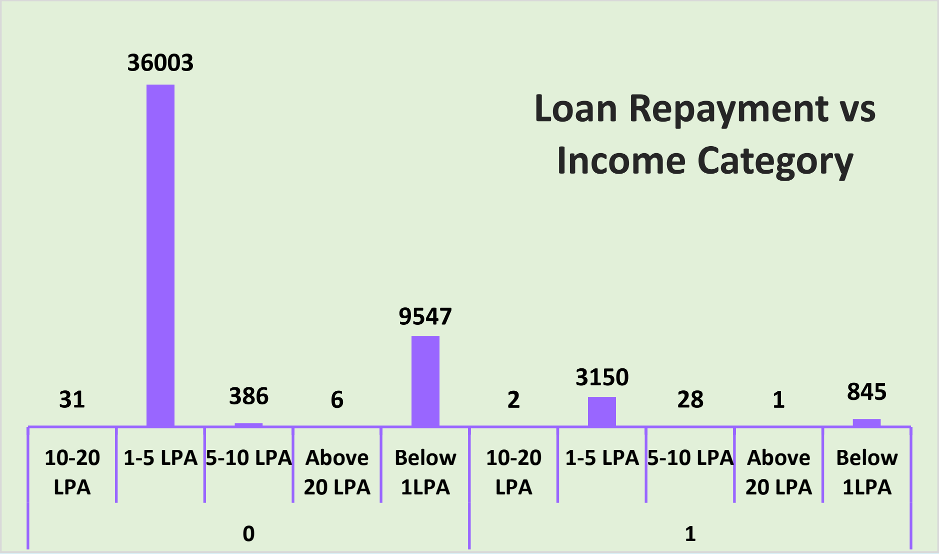


Likewise other parameters can also be explored and useful insights can be figured out from them. But our major focus still lies on Target Variable. So let’s have a look on effect of other categories on loan repayment.

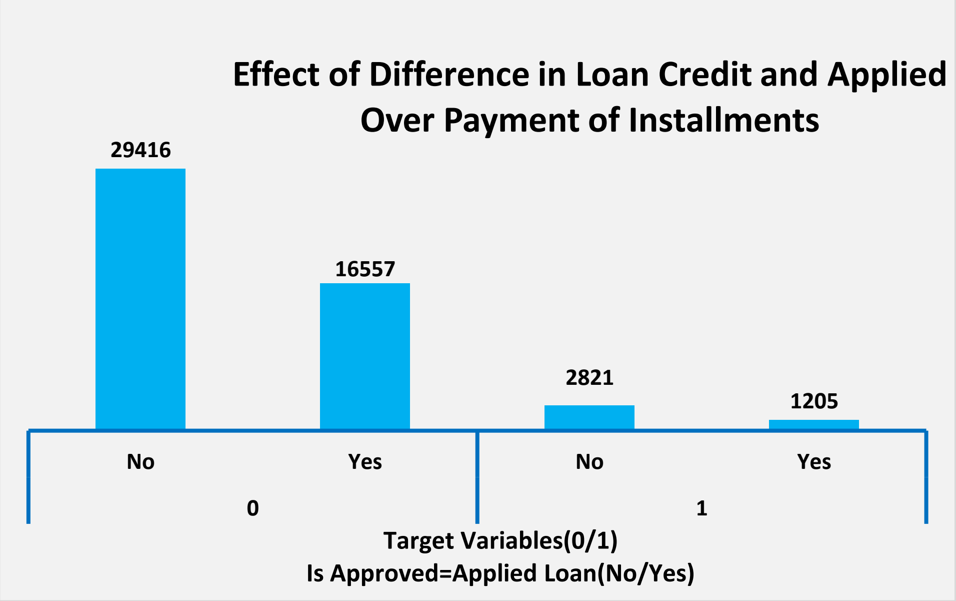
- We saw different Income types of clients. The proportion of Income type which delayed in instalment payment is good to know. Here number matters more than ratio as a single delay can cause significant loss to bank. Surprising to see State Servants and Pensioners in delay list. Around 9.5% of working clients failed to repay Instalments on time.



- We saw more than 98% of clients have salary less than 5LPA. Out of them 92% clients were able to pay instalments on time. There are 3 clients with more than 10 LPA salary and failed to pay at least 1 instalment on time. Generally with such higher salary delays are not expected.



- 65% of clients are provided loan amount more than the applied for. This could be a risk factor for Bank as client might not be able to repay the increased amount. About 2800 clients were not able to repay the instalment, which is a matter of concern.

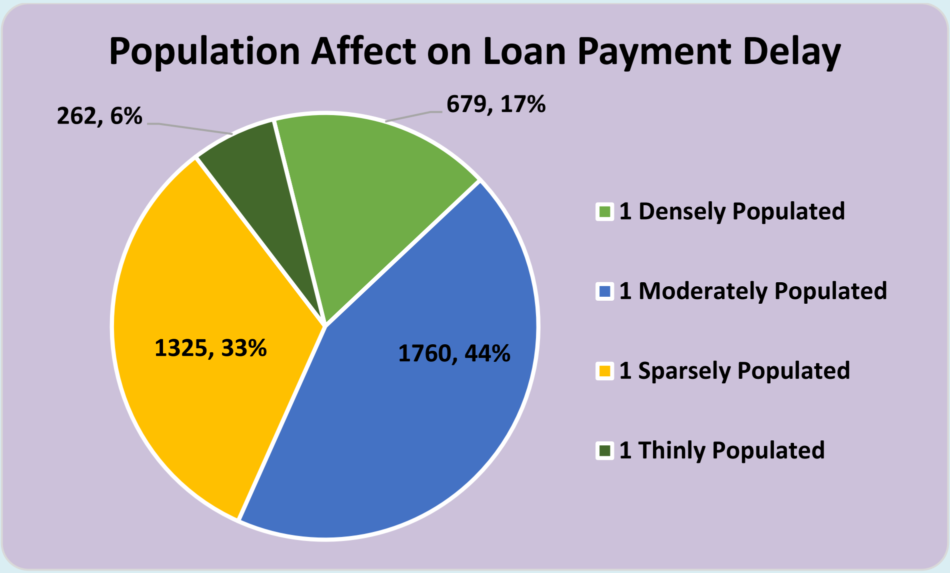


- Population density can somewhere tell us the status of person. If area is densely populated the cost of living could be very high- savings could be less and hence high chances of delay in repayment. Thinly populated could mean client lives in remote area or a village. Generally in village people do not have much income. Therefore along with population density one's income also should be accounted.

  Using IF() function we categorise Population based on per million scale. The categories are as follows:

  Thinly Populated: <5000 **Sparsely Populated: 5000 to 150000**
  Moderately Populated: 15000 to 30000 **Densely Populated: >30000**

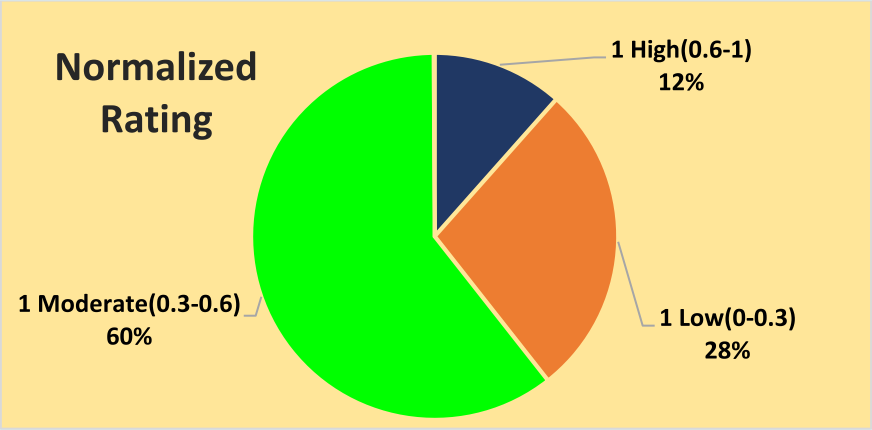
  The following chart shows proportion of people unable to pay instalment on time according to population density. Only 6% of total lies from thinly populated area.



- One of the prominent factor of Loan acceptance is Credit score or Normalized Rating. Banks do consider these ratings which depicts about financial balance of person. In the dataset given we have 3 columns of dataset given by external ratings. But External rating 1 have more than 50% of data missing.

  We decide to take average of all rating and take the final call. For that at least 2 of ratings should be present. If we drop 1st column then there will be a lot of ratings with only one or none numbers for average calculation. So, decision is made not to drop the first column. Ideally the column should be dropped but here let's give benefit of doubt to applicant.

  We categorize based on average rating. Now there are chances that client with high rating failed to repay instalment on time. Around 12% (465 in numbers) clients out of those who delayed in loan repayment have high rating.

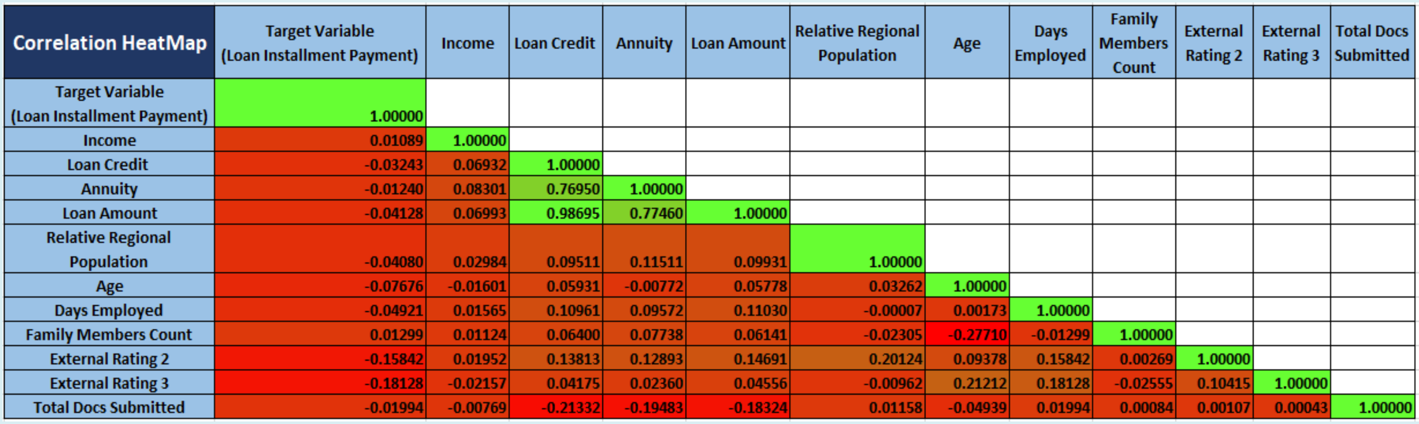


E) Correlation between Variables:-

Bivariate Analysis is done on Target Variable with other variable. Similarly we can find insights between other variables and there are multiple combinations possible. Based on Stakeholders or Executive questionaries' analysis can be performed.

One quick way to find the trend between parameters is correlation coefficient. To find degree of changes of one variable with respect to other correlation is best statistical way to find. However the limitations are that it deals only with linearity and only on numerical data.

Upon selecting some relevant variable a heatmap of correlations is created to understand relation between variables.



Correlation HeatMap	Target Variable (Loan Installment Payment)	Income	Loan Credit	Annuity	Loan Amount	Relative Regional Population	Age	Days Employed	Family Members Count	External Rating 2	External Rating 3	Total Docs Submitted
Target Variable (Loan Installment Payment)	1.00000											
Income	0.01089	1.00000										
Loan Credit	-0.03243	0.06932	1.00000									
Annuity	-0.01240	0.08301	0.76950	1.00000								
Loan Amount	-0.04128	0.06993	0.98695	0.77460	1.00000							
Relative Regional Population	-0.04080	0.02984	0.09511	0.11511	0.09931	1.00000						
Age	-0.07676	-0.01601	0.05931	-0.00772	0.05778	0.03262	1.00000					
Days Employed	-0.04921	0.01565	0.10961	0.09572	0.11030	-0.00007	0.00173	1.00000				
Family Members Count	0.01299	0.01124	0.06400	0.07738	0.06141	-0.02305	-0.27710	-0.01299	1.00000			
External Rating 2	-0.15842	0.01952	0.13813	0.12893	0.14691	0.20124	0.09378	0.15842	0.00269	1.00000		
External Rating 3	-0.18128	-0.02157	0.04175	0.02360	0.04556	-0.00962	0.21212	0.18128	-0.02555	0.10415	1.00000	
Total Docs Submitted	-0.01994	-0.00769	-0.21332	-0.19483	-0.18324	0.01158	-0.04939	0.01994	0.00084	0.00107	0.00043	1.00000

- Focusing on Target Variable none of other variable shows positive variation with changes. Income however shows slightest of positive change with it.
- Although all the other variables have weak strength and no linear relationship can be concluded, external rating 2 and 3 are weakest among all. So it is recommended not to go with point to point analysis but to categorise and study data carefully.
- Loan Credited and Loan Amount have strongest linear relationship among all. While most of the other are positive yet close to zero or negative. Among all Age and Family Member Count have no clear relation between them?
- Clearly Linear Relationship does not seems to be established between variables Correlation coefficient deals with linearity. If any other relation is there, we might have to scatter plot the variable, look for Trend line and its strength.

Conclusions….

Conclusions:

- Certainly based on a lot of factors a loan amount and duration should be decided. But we have to understand interlink between all other data as well. We saw how Income Locality and occupation may collectively give a solution, yet a single variable can give totally unexpected result.
- A large number of clients have less than 5LPA Income yet they are able to pay the instalments. This has to be looked by banks to decide their Interest rates, their plans and schemes to attract more and more customers.
- Locality makes no difference. Even people living in remote areas are able to repay the loan. Yes it is a factor to be considered but the weighted should be low.
- Now a days everything is digitalised and having a strong reminder system shall help to pay instalments on time. Services like Email, text massaging or even linking with UPI ID can ease process for both bank and customer. Using OTP we can confirm email and mobile number so that client is always reachable and under radar.
- Therefore it's important to interview the client and have a panel discussion. A weightage ratio can be given to parameters to ease the process.

THANK YOU!!

(Looking Forward for Valuable Suggestions and Feedbacks)