Bank Loan Case Study

LINK TO EXCEL SHEET

Project Description

The purpose of this project is to conduct Exploratory Data Analysis (EDA) to address the challenges of managing loan default risks among customers who don't have sufficient credit history. The dataset includes different outcomes like approved loans, cancelled application, unused loans and rejected loans.

By cleaning and examining the dataset, the main aim is to point out the factors which affects customers having trouble paying loan. This analysis will help in deriving actionable insights which will help the finance company in making decisions like denying the loan, reducing the amount of loan or lending at a higher interest rate to risky applicants, all while aiming to grow the business safely.

Data Analytics Task

- Identify Missing Data and Deal with it Appropriately: Handling missing data in the loan application dataset is essential to ensure the accuracy of analysis.

- Identify Outliers in the Dataset: Outliers can significantly impact the analysis and distort the results.

- 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.

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

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

Approach

To accomplish the necessary tasks and to finalize the project, the approach to this project involves a structured methodology to identify patterns in the data to make informed loan approval decisions.

- First of all to better understand the project, research was done about risk analytics in banking and financial services.

- Then, the provided files were imported into Excel to understand and observe the structure of the data. Three files had been provided for our analysis and reference. The dataset contains information about loan applications like income of applicant, credit amount, etc.

previous_application.csv: Contains information about previous loan applications.

application_data.csv: Provides details about the current loan applications.

columns_description.csv: Describes the columns present in the above datasets, explaining what each column represents.

Approach

- Then, data pre-processing was performed which is one of the data analytics task like data cleaning by handling missing columns, duplicates and errors and detecting potential outliers.

- After cleaning and preparing the data, Exploratory Data Analysis (EDA) was performed to analyze the patterns in the data and to derive key insights.

- Lastly, after performing all the necessary analysis and deriving key insights, a detailed report was made and it was handed over to the required department to make informed decisions on loan approval.

Tech-Stack Used

Microsoft Excel 2019 has been used to clean and prepare the dataset for analysis. It was also used for basic statistical calculations. Pivot tables were created, Excel charts were utilized for visualization and potential outliers were identified.

The reason behind using excel was that is widely used and easy to understand. It is great tool for analysing data for large datasets and works well with other tools like Word and PowerPoint, so it is easy to create reports and presentation based on the analysis. It is cost effective as it is often installed in our computers.

Insights

A.) Identify Missing Data and Deal with it Appropriately

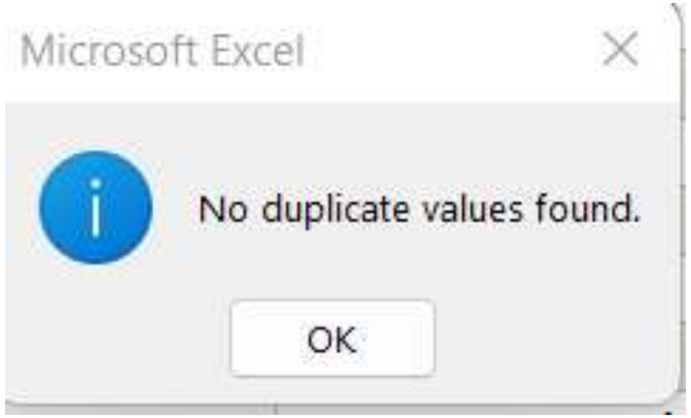
- The task is to identify the missing data in the dataset and decide on an appropriate method to deal with it using Excel built-in functions and features.

- Create a bar chart or column chart to visualize the proportion of missing values for each variable.

Checking for Duplicate Data

In the application_data, column SK_ID_CURR was checked for duplicate values because it is ID of loan in the dataset provided. It should be unique because duplicate values will cause discrepancies in the analysis.

After checking for duplicates it was observed that there are no duplicate values in the column.

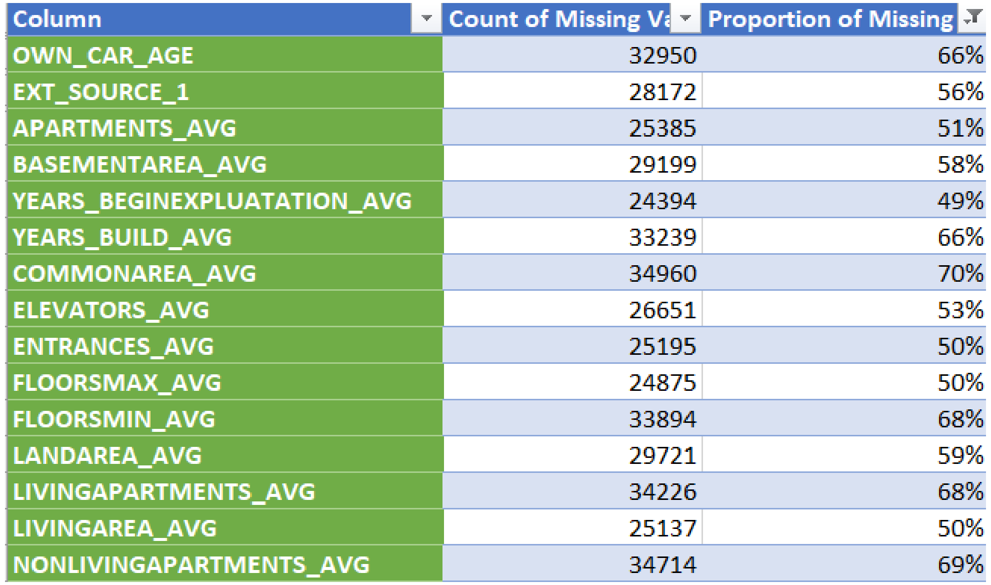

Microsoft Excel

No duplicate values found.

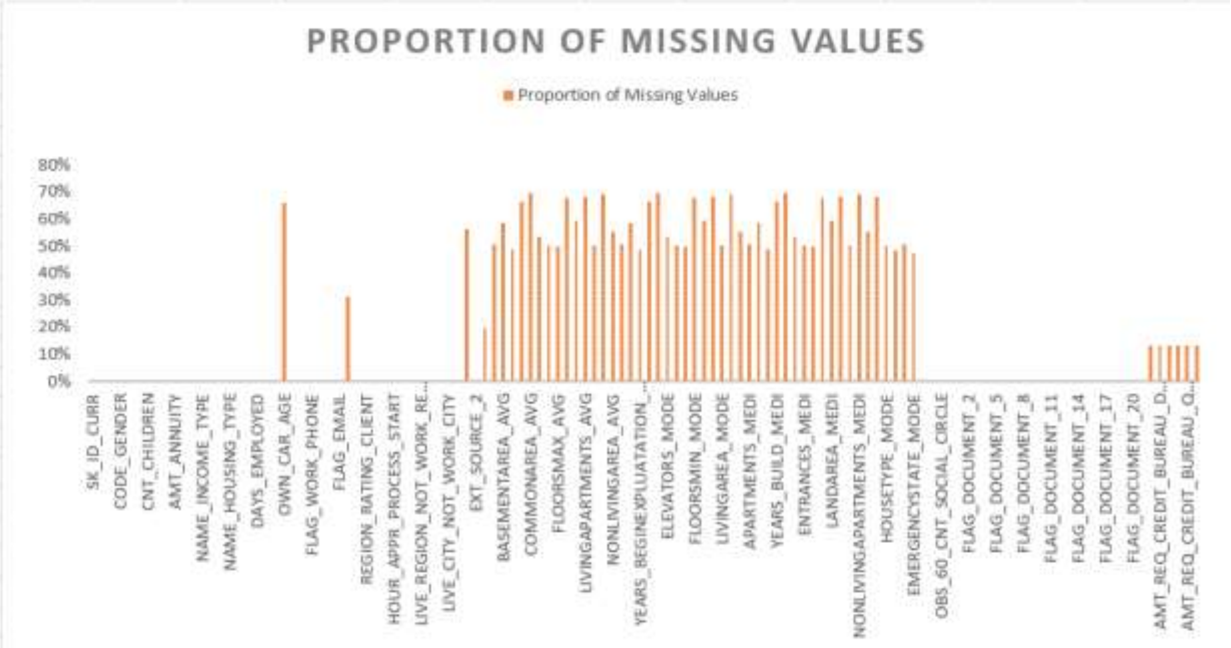
Checking for Missing Data

- In the dataset, there are total 122 columns, each column was checked for the number of missing values as well as the proportion of missing values.
- It was observed that there are 49 columns having more than 40% missing values. The best way to deal with them is to drop them.
- As for the columns having less than 40% missing values, performing imputation is the best way to deal with them.
- Also, columns which are irrelevant for our analysis have also been removed.
- After dropping all unnecessary columns, the dataset has total 32 columns which will be used for further analysis.

Checking for Missing Data



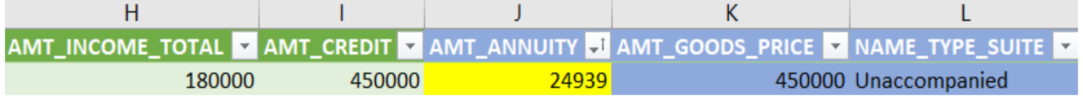
Column	Count of Missing Va	Proportion of Missing
OWN_CAR_AGE	32950	66%
EXT_SOURCE_1	28172	56%
APARTMENTS_AVG	25385	51%
BASEMENTAREA_AVG	29199	58%
YEARS_BEGINEXPLUATATION_AVG	24394	49%
YEARS_BUILD_AVG	33239	66%
COMMONAREA_AVG	34960	70%
ELEVATORS_AVG	26651	53%
ENTRANCES_AVG	25195	50%
FLOORSMAX_AVG	24875	50%
FLOORSMIN_AVG	33894	68%
LANDAREA_AVG	29721	59%
LIVINGAPARTMENTS_AVG	34226	68%
LIVINGAREA_AVG	25137	50%
NONLIVINGAPARTMENTS_AVG	34714	69%



Above data shows table highlighting few columns having missing values greater than 40% and visualisation of proportion of missing values.

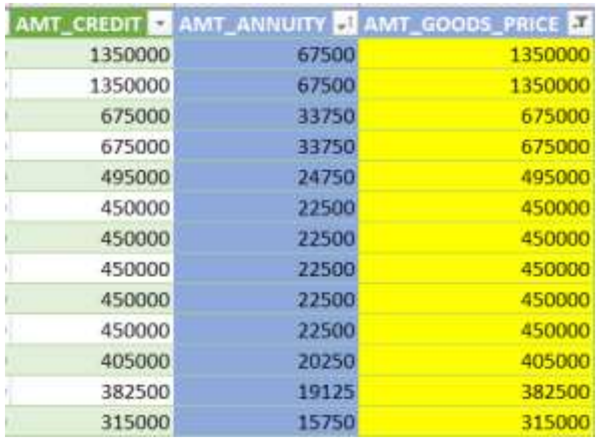
Performing Imputation

- In column AMT_ANNUITY, there is 1 missing value, it has been replaced by median value of AMT_ANNUITY



H	I	J	K	L
AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRICE	NAME_TYPE_SUITE
180000	450000	24939	450000	Unaccompanied

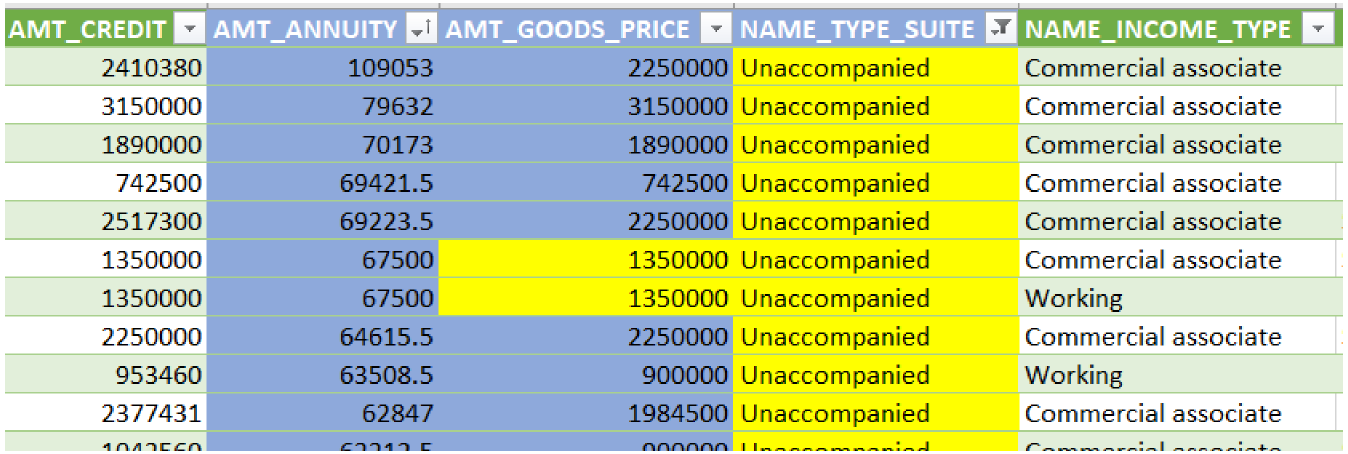
- In column AMT_GOODS_PRICE, there are 38 missing values, they have been replaced by the corresponding row value of AMT_CREDIT column after observing that maximum values in the AMT_GOODS_PRICE column do correspond to the respective value of AMT_CREDIT column.



AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRICE
1350000	67500	1350000
1350000	67500	1350000
675000	33750	675000
675000	33750	675000
495000	24750	495000
450000	22500	450000
450000	22500	450000
450000	22500	450000
450000	22500	450000
450000	22500	450000
405000	20250	405000
382500	19125	382500
315000	15750	315000

Performing Imputation

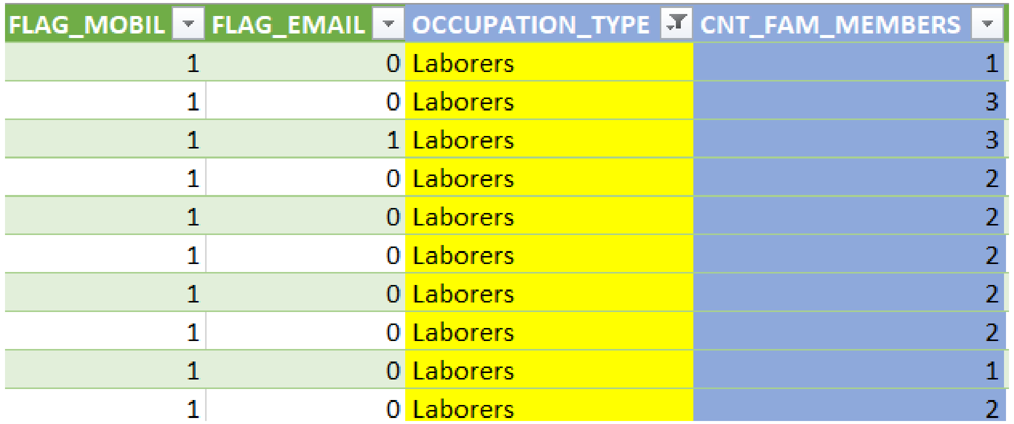
- In column NAME_TYPE_SUITE, there are 192 missing values, they have been replaced by most common value of this column which is Unaccompanied after performing some filtration in corresponding NAME_INCOME_TYPE and NAME_FAMILY_STATUS column.



AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRICE	NAME_TYPE_SUITE	NAME_INCOME_TYPE
2410380	109053	2250000	Unaccompanied	Commercial associate
3150000	79632	3150000	Unaccompanied	Commercial associate
1890000	70173	1890000	Unaccompanied	Commercial associate
742500	69421.5	742500	Unaccompanied	Commercial associate
2517300	69223.5	2250000	Unaccompanied	Commercial associate
1350000	67500	1350000	Unaccompanied	Commercial associate
1350000	67500	1350000	Unaccompanied	Working
2250000	64615.5	2250000	Unaccompanied	Commercial associate
953460	63508.5	900000	Unaccompanied	Working
2377431	62847	1984500	Unaccompanied	Commercial associate
1042560	62212.5	900000	Unaccompanied	Commercial associate

Performing Imputation

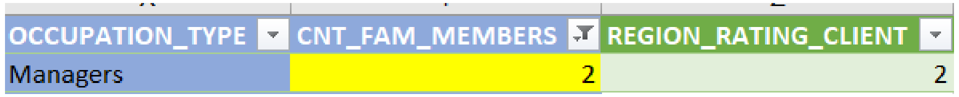
- In column OCCUPATION_TYPE, there are 15654 missing values, they have been replaced by most common value of this column which is Laborers after performing some filtration in corresponding NAME_INCOME_TYPE and NAME_EDUCATION_TYPE column.



FLAG_MOBIL	FLAG_EMAIL	OCCUPATION_TYPE	CNT_FAM_MEMBERS
1	0	Laborers	1
1	0	Laborers	3
1	1	Laborers	3
1	0	Laborers	2
1	0	Laborers	2
1	0	Laborers	2
1	0	Laborers	2
1	0	Laborers	2
1	0	Laborers	1
1	0	Laborers	2

Performing Imputation

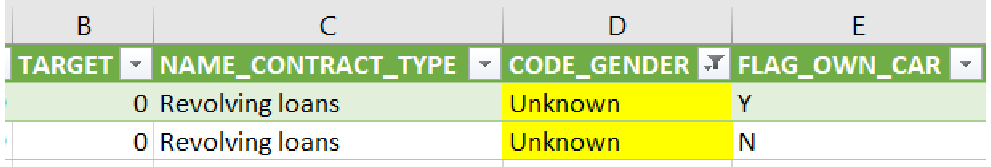
- In column CNT_FAM_MEMBERS, there is 1 missing value, it has been replaced by the median value of CNT_FAM_MEMBERS.



OCCUPATION_TYPE	CNT_FAM_MEMBERS	REGION_RATING_CLIENT
Managers	2	2

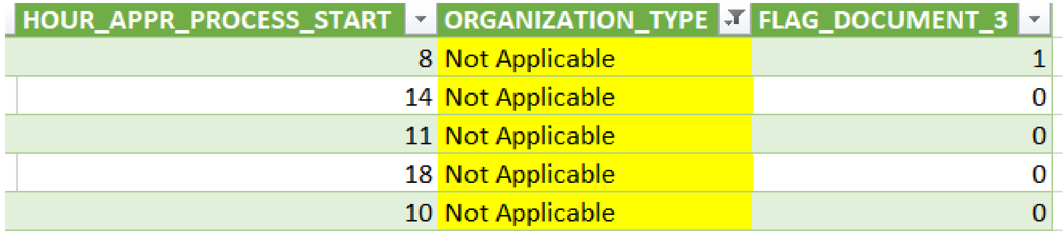
Error Rectification

- In column CODE_GENDER, there are some error values, they have been replaced with Unknown.



B	C	D	E
TARGET	NAME_CONTRACT_TYPE	CODE_GENDER	FLAG_OWN_CAR
0	Revolving loans	Unknown	Y
0	Revolving loans	Unknown	N

- In column ORGANIZATION_TYPE, there are some error values, they have been replaced with Not Applicable as the corresponding column NAME_INCOME_TYPE contains only Pensioner and Unemployed.



HOUR_APPR_PROCESS_START	ORGANIZATION_TYPE	FLAG_DOCUMENT_3
8	Not Applicable	1
14	Not Applicable	0
11	Not Applicable	0
18	Not Applicable	0
10	Not Applicable	0

Insights

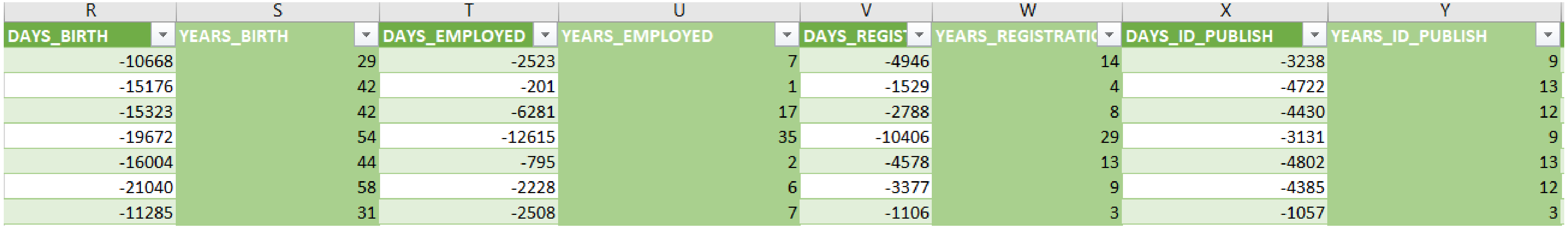
- Performing the given task helped in identifying variables which had missing value and focusing where data is incomplete.
- Understanding the missing data helped in deciding whether to fix them or remove certain data points like variable having more than 40% missing values were removed which was necessary. Also certain irrelevant columns which were not useful for analysis were also removed.
- Visualization of missing data through chart made it easier to understand and explain the data’s condition.
- Overall, it was found that there was high missing rate in the dataset indicating issues with how data is being collected or entered which suggests a need of better data practices.
- Handling the missing data is necessary because it can affect accuracy which will make it harder to assess loan risks properly. Therefore, all the missing data were dealt with properly which will lead to informed decision making in the further analysis.

B.) Identify Outliers in the Dataset

- The task is to detect and identify outliers in the dataset using statistical functions and features, focusing on numerical variables.
- Create box plots or scatter plots to visualize the distribution of numerical variables and highlight the outliers.

Standardizing values

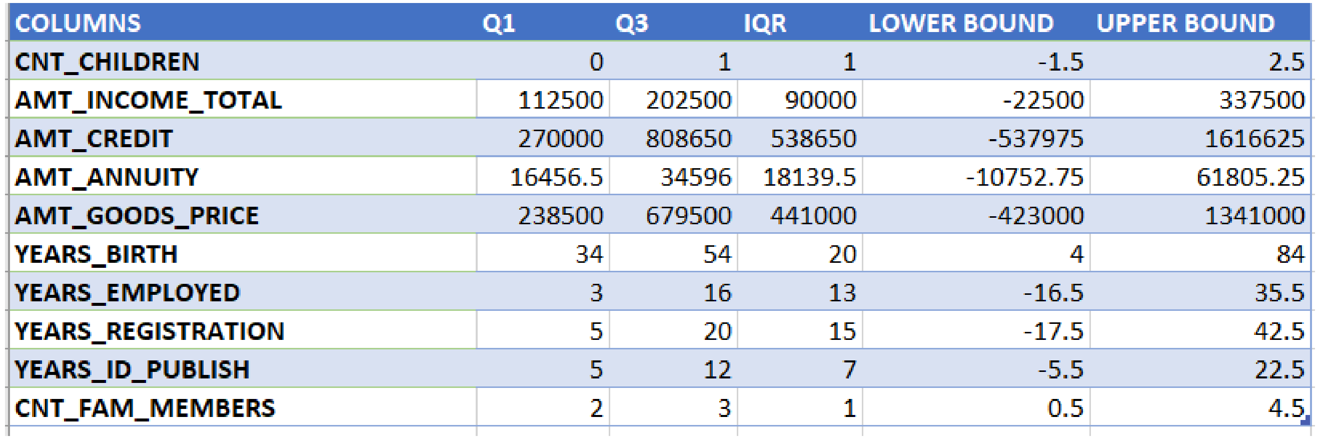
Columns DAYS_BIRTH, DAYS_EMPLOYED, DAYS_REGISTRATION,DAYS_ID_PUBLISH were converted into respective years in different columns for clear analysis in detecting outliers.



R	S	T	U	V	W	X	Y
DAYS_BIRTH	YEARS_BIRTH	DAYS_EMPLOYED	YEARS_EMPLOYED	DAYS_REGIST	YEARS_REGISTRATI	DAYS_ID_PUBLISH	YEARS_ID_PUBLISH
-10668	29	-2523	7	-4946	14	-3238	9
-15176	42	-201	1	-1529	4	-4722	13
-15323	42	-6281	17	-2788	8	-4430	12
-19672	54	-12615	35	-10406	29	-3131	9
-16004	44	-795	2	-4578	13	-4802	13
-21040	58	-2228	6	-3377	9	-4385	12
-11285	31	-2508	7	-1106	3	-1057	3

Statistical Calculations

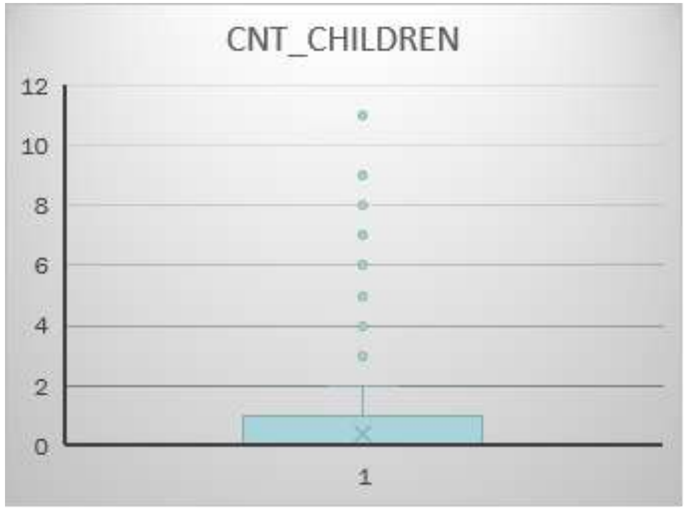
The table shows calculation of Quartiles, Interquartile Range (IQR) and Outlier Boundaries.

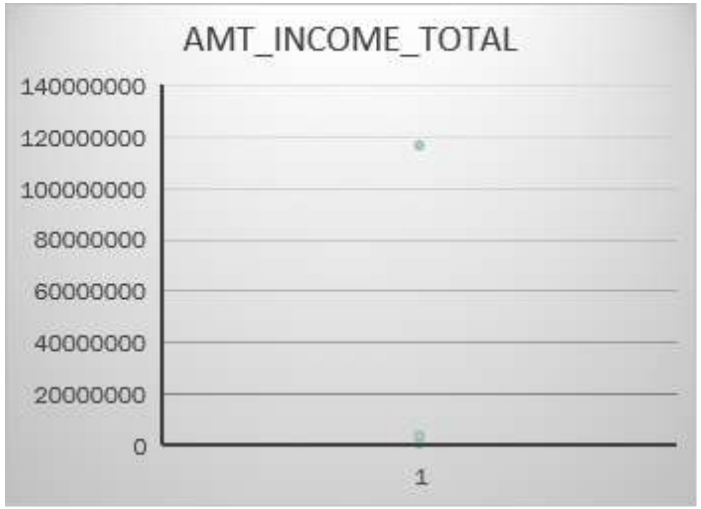


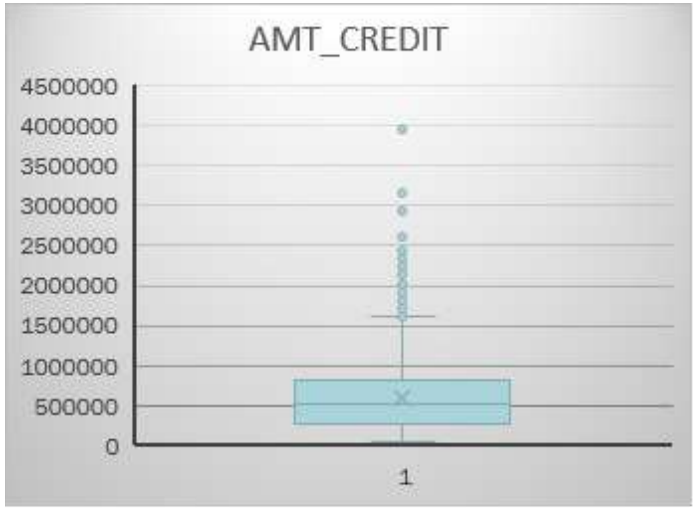
COLUMNS	Q1	Q3	IQR	LOWER BOUND	UPPER BOUND
CNT_CHILDREN	0	1	1	-1.5	2.5
AMT_INCOME_TOTAL	112500	202500	90000	-22500	337500
AMT_CREDIT	270000	808650	538650	-537975	1616625
AMT_ANNUITY	16456.5	34596	18139.5	-10752.75	61805.25
AMT_GOODS_PRICE	238500	679500	441000	-423000	1341000
YEARS_BIRTH	34	54	20	4	84
YEARS_EMPLOYED	3	16	13	-16.5	35.5
YEARS_REGISTRATION	5	20	15	-17.5	42.5
YEARS_ID_PUBLISH	5	12	7	-5.5	22.5
CNT_FAM_MEMBERS	2	3	1	0.5	4.5

From the above table it can be observed that IQR of AMT_INCOME_TOTAL is very less. Third quartile of AMT_CREDIT and AMT_GOODS_PRICE is larger as compared to the first quartile. AMT_ANNUITY third quartile is bit larger than first quartile.

Highlighting Outliers

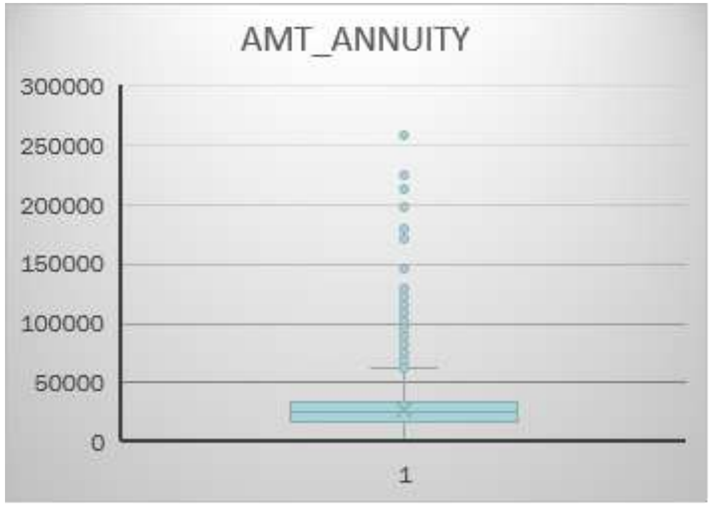


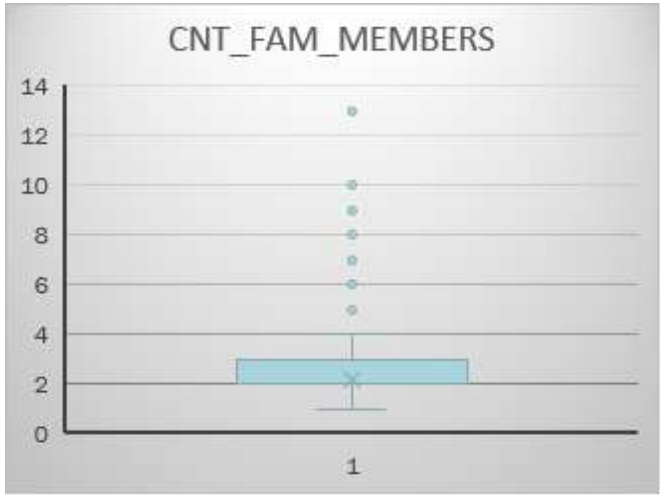


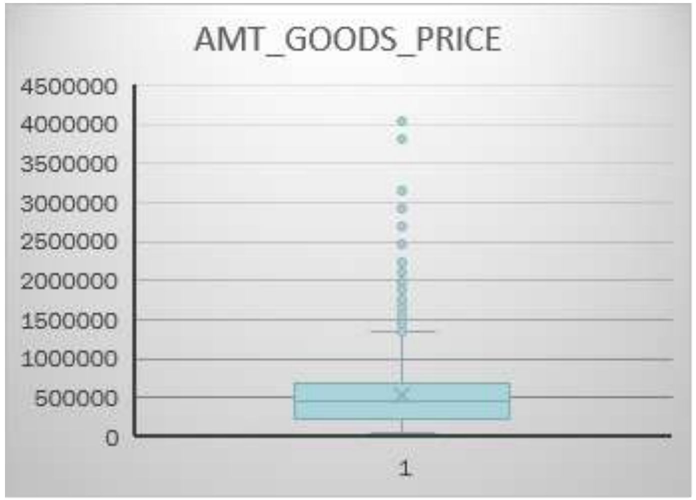


From the above box plot it can be observed that CNT_CHILDREN and AMT_CREDIT have some outliers with outlier values till 11 and 4000000 respectively whereas AMT_INCOME_TOTAL has huge number of outliers with outlier values going up to 120000000.

Highlighting Outliers

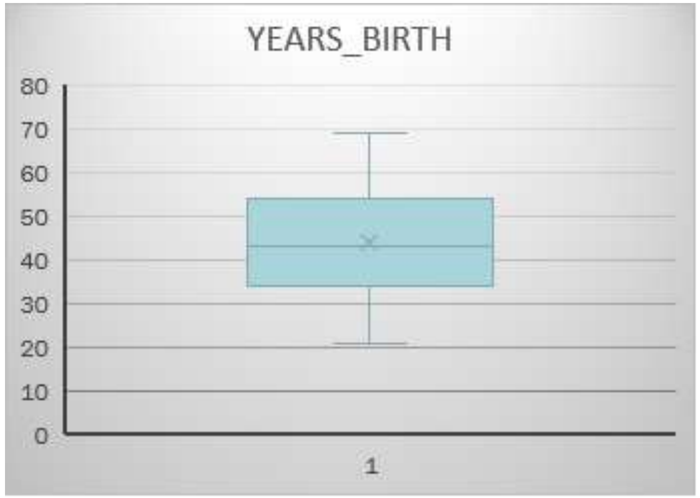


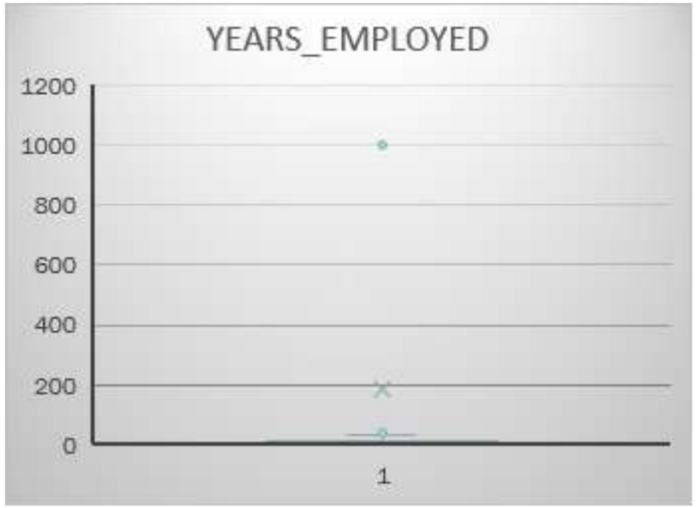


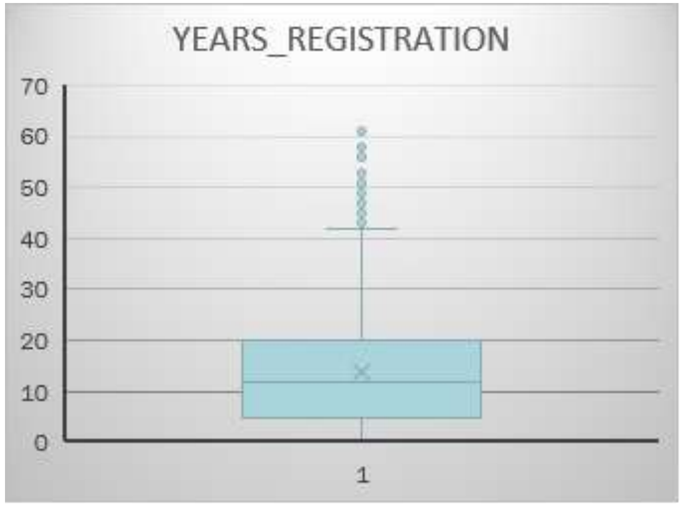


From the above box plot it can be observed that AMT_ANNUITY, CNT_FAM_MEMBERS and AMT_GOODS_PRICE have some amount of outliers with outlier values going above 250000, 12, 4000000 respectively.

Highlighting Outliers

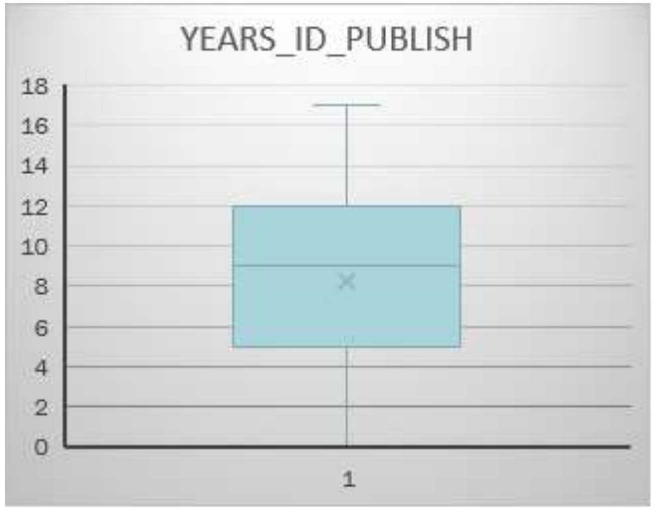






From the above box plot it can be observed that YEARS_BIRTH has no outliers, YEARS_EMPLOYED has significant amount of outliers with value going up to 1000 and YEARS_REGISTRATION has some outliers with value going above 60.

Highlighting Outliers



From the above box plot it can be observed that YEARS_ID_PUBLISH has no outliers.

Insights

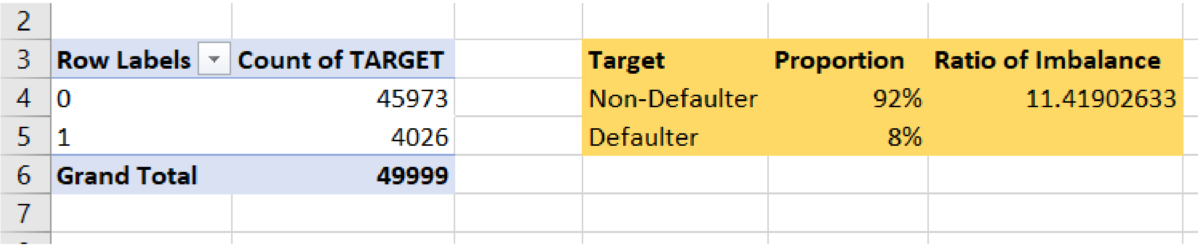
- YEARS_BIRTH and YEARS_ID_PUBLISH has no outliers which suggests that these data are reliable and there are no unusual patterns or errors that could impact the analysis.
- AMT_INCOME_TOTAL has significant amount of outliers suggesting that some individuals have high incomes applying for loan, they can be businessmen or entrepreneur who might have non-traditional source of income like investments leading to unusual values.
- AMT_ANNUITY also has significant amount of outliers suggesting that there might be loans with large amounts like individuals having hight net worth or business.
- AMT_CREDIT and AMT_GOODS_PRICE has significant amount of outliers. Observing their quartile calculation, it indicates that most credit amount and price of goods of customers are in third quartile.
- YEARS_EMPLOYED has outlier value above 1000 years which is not possible suggesting incorrect data entry which needs to be rectified.

B.) Analyse Data Imbalance

- The task is to determine if there is data imbalance in the loan application dataset and calculate the ratio of data imbalance using Excel functions.

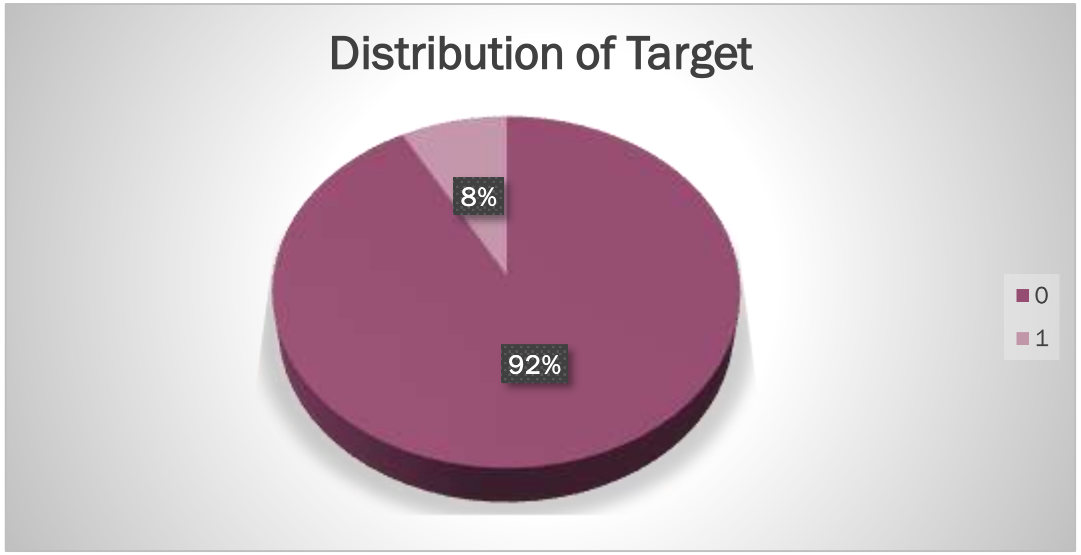
- Create a pie chart or bar chart to visualise the distribution of the target variable and highlight the class imbalance.

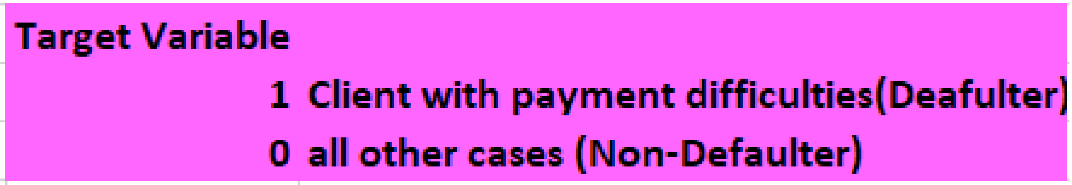
Data Imbalance



Row Labels	Count of TARGET
0	45973
1	4026
Grand Total	**49999**

Target	Proportion	Ratio of Imbalance
Non-Defaulter	92%	11.41902633
Defaulter	8%	





Target Variable

1 Client with payment difficulties(Deafulter)

0 all other cases (Non-Defaulter)

Insights

- From the data it can be observed that there is significant imbalance between the Defaulter and Non-Defaulter classes.
- The ratio of imbalance shows a large disparity and is 11.42 approximately which means that Defaulter class is 11.42 times less frequent than the Non-Defaulter class.
- This imbalance will be a problem while modelling as the model might struggle to identify Defaulter class correctly which is a big issue because Defaulter class is important.
- Strategies like oversampling or under sampling where more examples of Defaulter class are added or the number of examples of Non-Defaulter class is reduced can be used to improve the model’s performance.
- Feature Engineering technique can also be used where new features can be created that might help the model distinguish between the classes better.

B.) Perform Univariate, Segmented Univariate and Bivariate Analysis

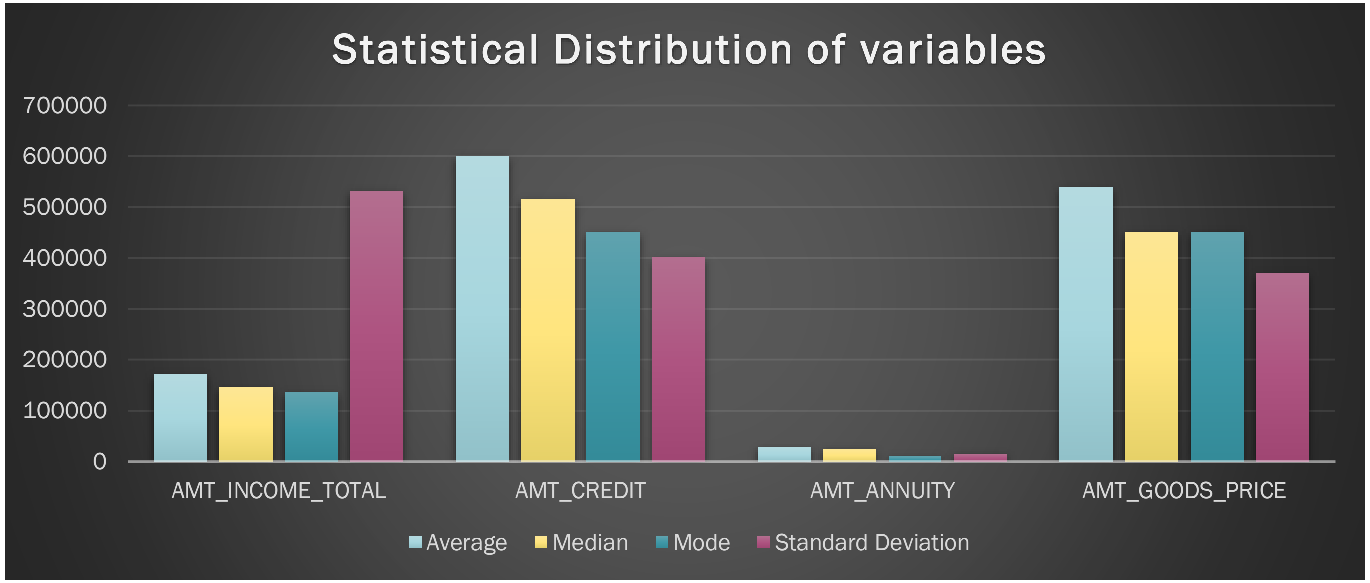
- The task is to perform univariate analysis to understand the distribution of individual variables, segmented univariate analysis to compare variable distributions for different scenarios, and bivariate analysis to explore relationships between variables and the target variable using Excel functions and features.

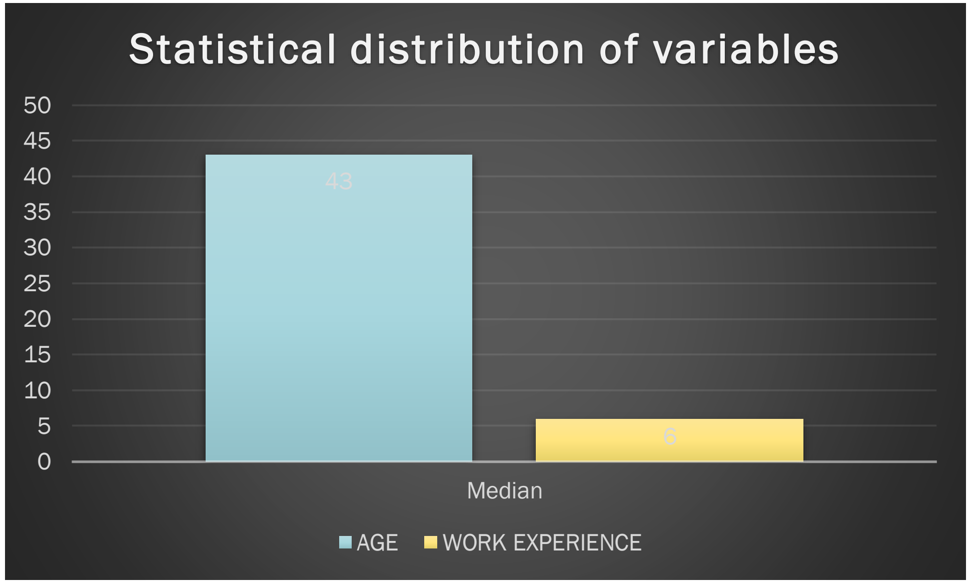
- Create histograms, bar charts, or box plots to visualize the distributions of variables. Create stacked bar charts or grouped bar charts to compare variable distributions across different scenarios. Create scatter plots or heatmaps to visualize the relationships between variables and the target variable.

Univariate Analysis on Numerical variables

	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRICE
Average	170768	599701	27107	538803
Median	145800	514778	24939	450000
Mode	135000	450000	9000	450000
Standard Deviation	531814	402411	14563	369922

	AGE	WORK EXPERIENCE
Median	43	6





Insights

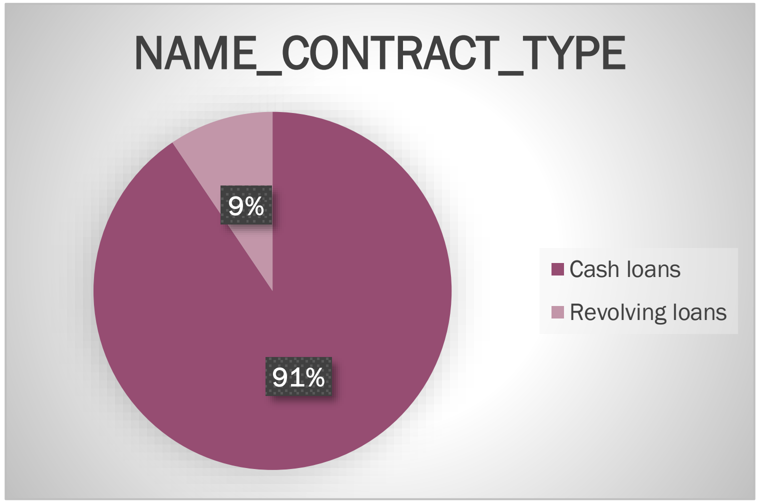
- From the univariate analysis on numerical variables, it can be observed that most applicants income is less than 400000 but the amount credited for loan is greater than 400000 indicating that most of the applicants applied for loan with amount greater than their income which further suggests potential financial strain.

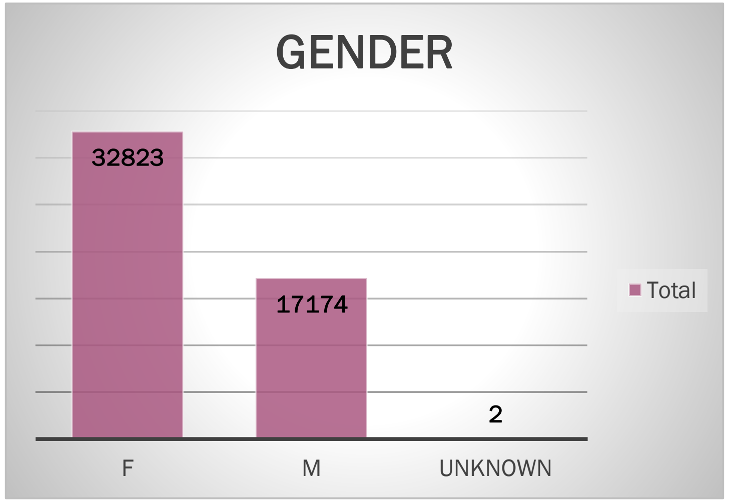
- Most of the applicants annuity amount is greater than 50000 which is approximately 12% of income of most applicant. It can be considered a manageable debt-to-income ratio, as many financial advisors recommend keeping total debt payments below 30-40% of income.

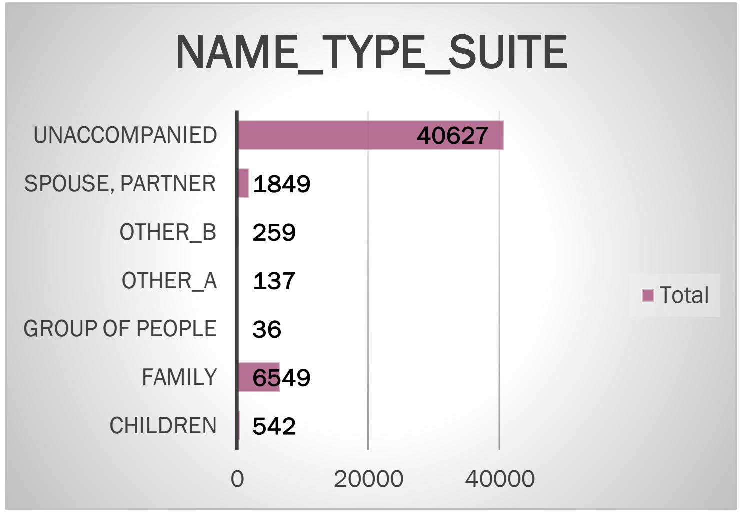
- It can also be observed that AMT_GOODS_PRICE closely follows the distribution of AMT_CREDIT. It can have significant implication for credit risk assessments. It suggests that the applicant might be taking too much debt and might struggle to pay it back. The bank needs to check how much people are borrowing and adjust their limits accordingly.

- The typical loan applicant is 43 years of age and has 6 years of work experience suggesting they are established in their career and have stable income. They might be facing mid-life financial responsibility like mortgages or family expenses.

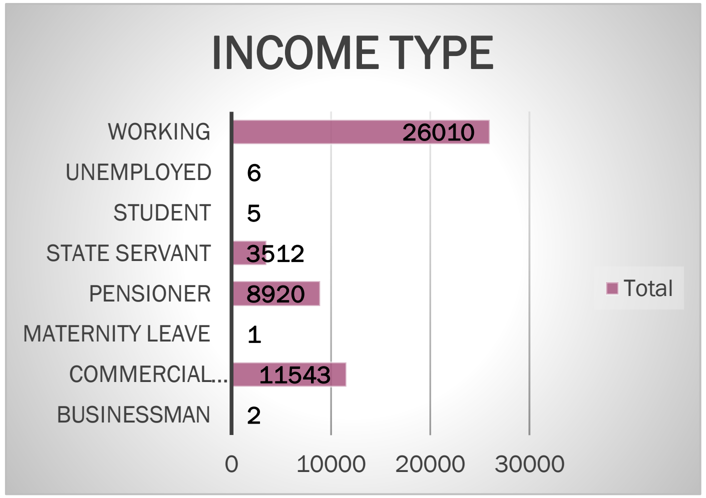
Univariate Analysis on Categorical variables

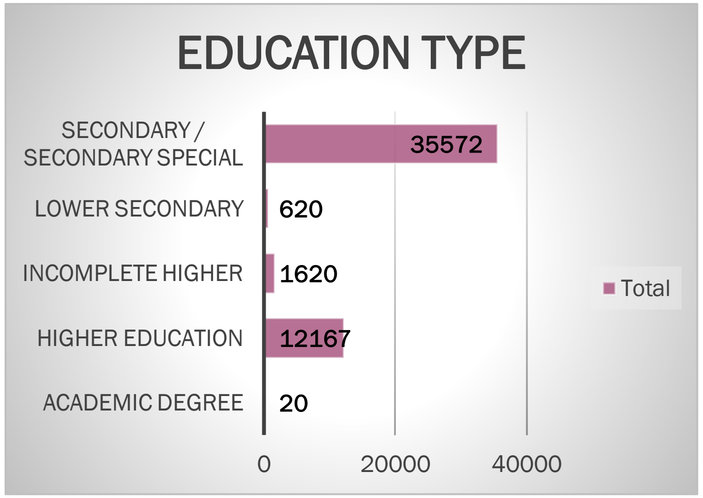


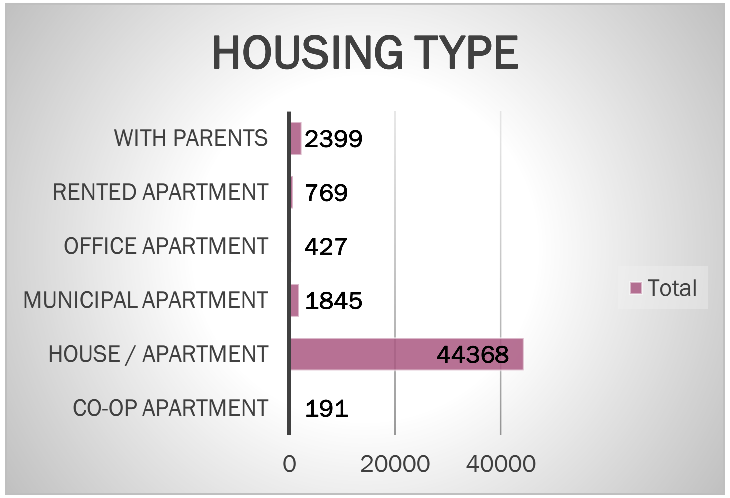




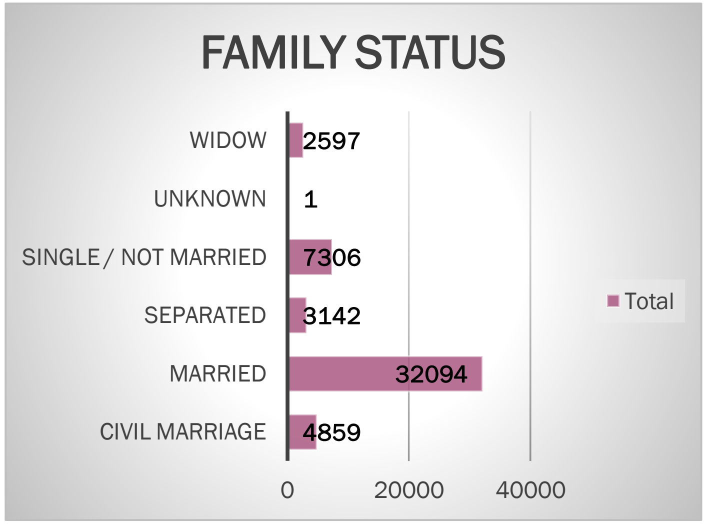
Univariate Analysis on Categorical variables

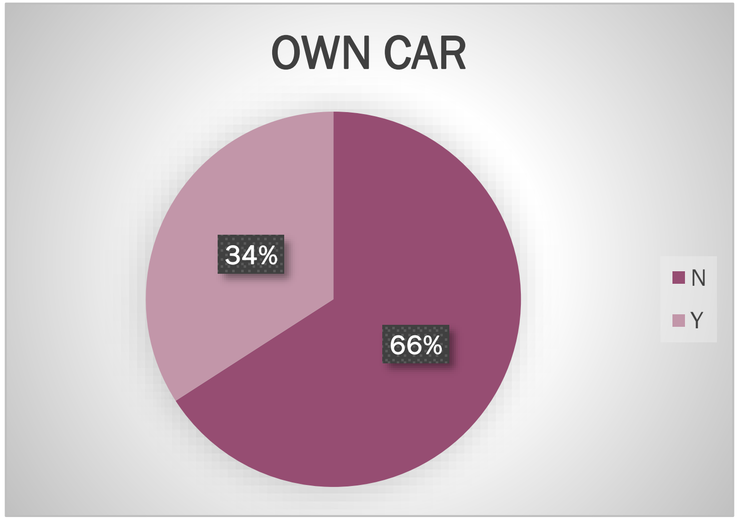


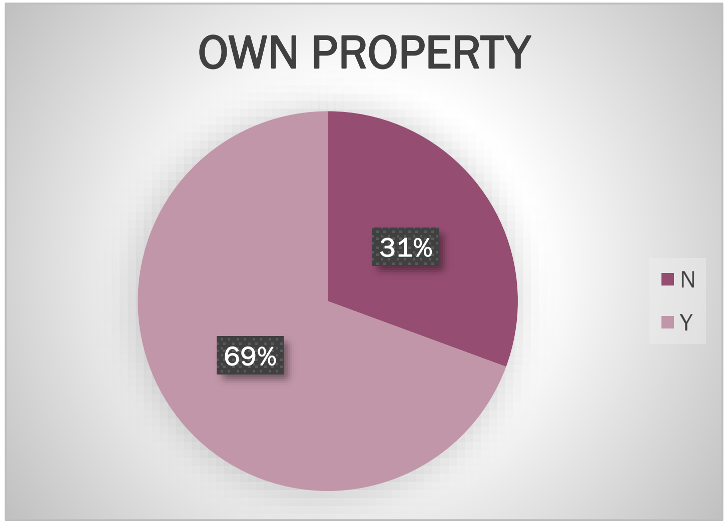




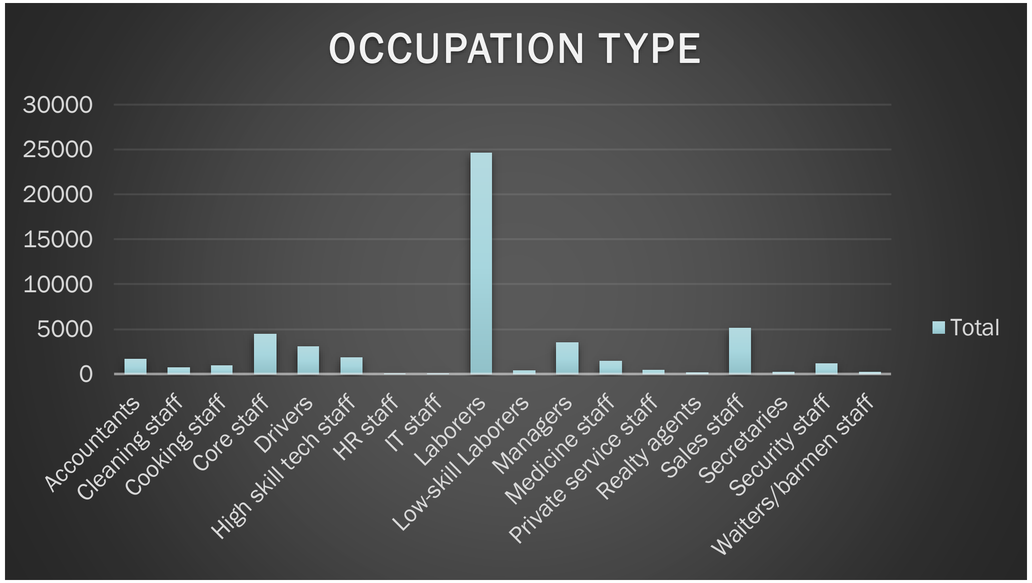
Univariate Analysis on Categorical variables

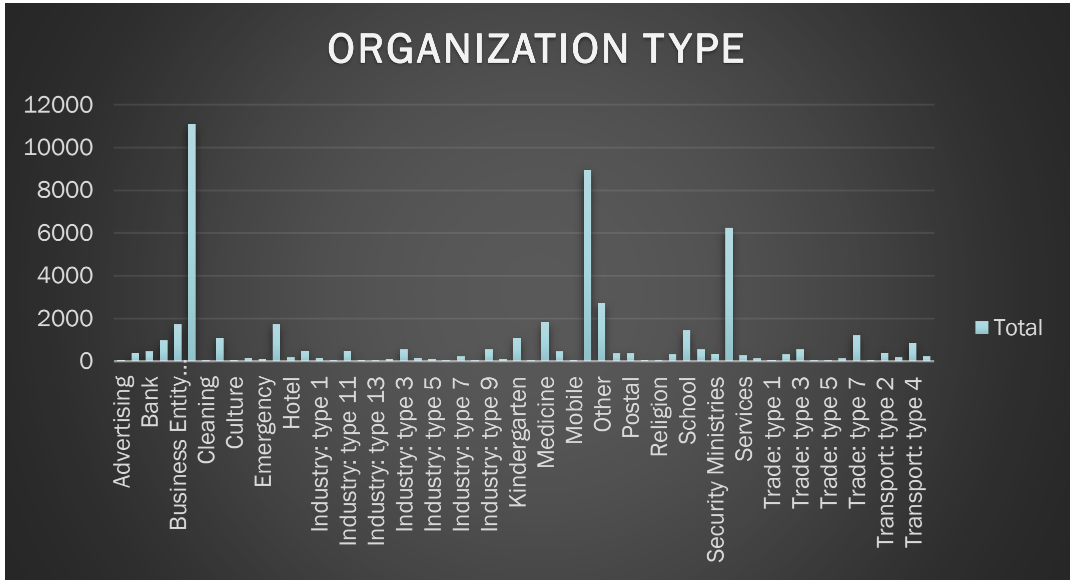






Univariate Analysis on Categorical variables





Insights

- From the univariate analysis of categorical variables, it can be observed that most applicants are for cash loans and very less for revolving loans. Cash loans are more popular because fixed payment make it easier to budget.

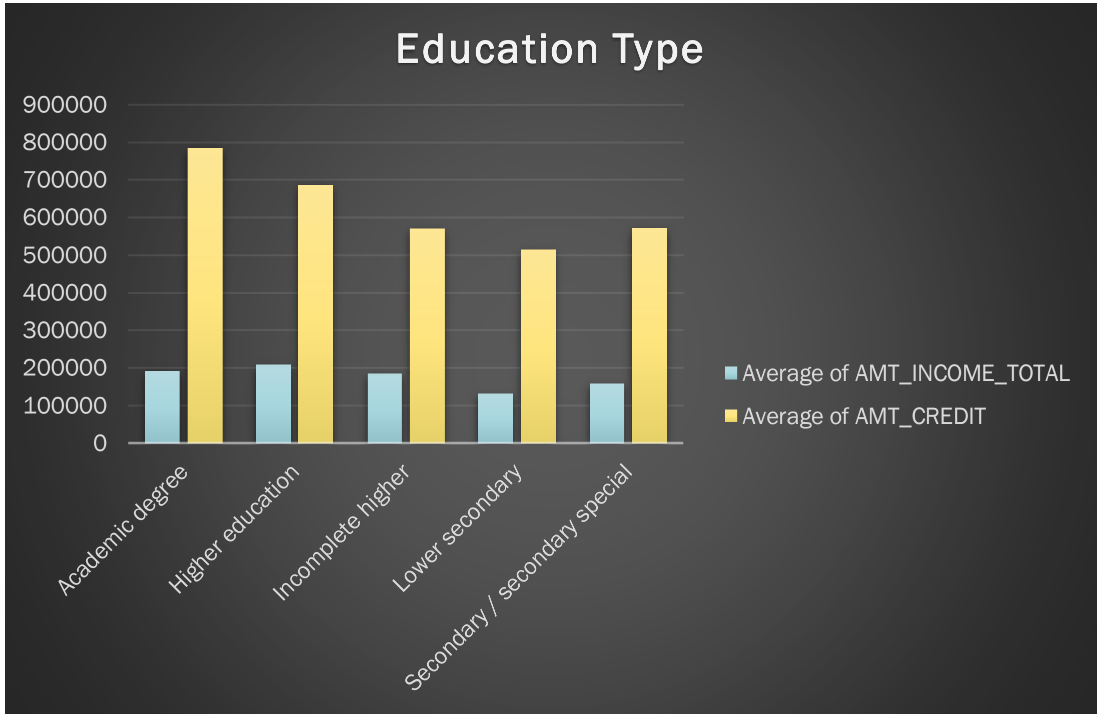
- Most of the loan applicants are females compared to males and most applicants were not accompanied by anyone else followed by people who were accompanied by family members.

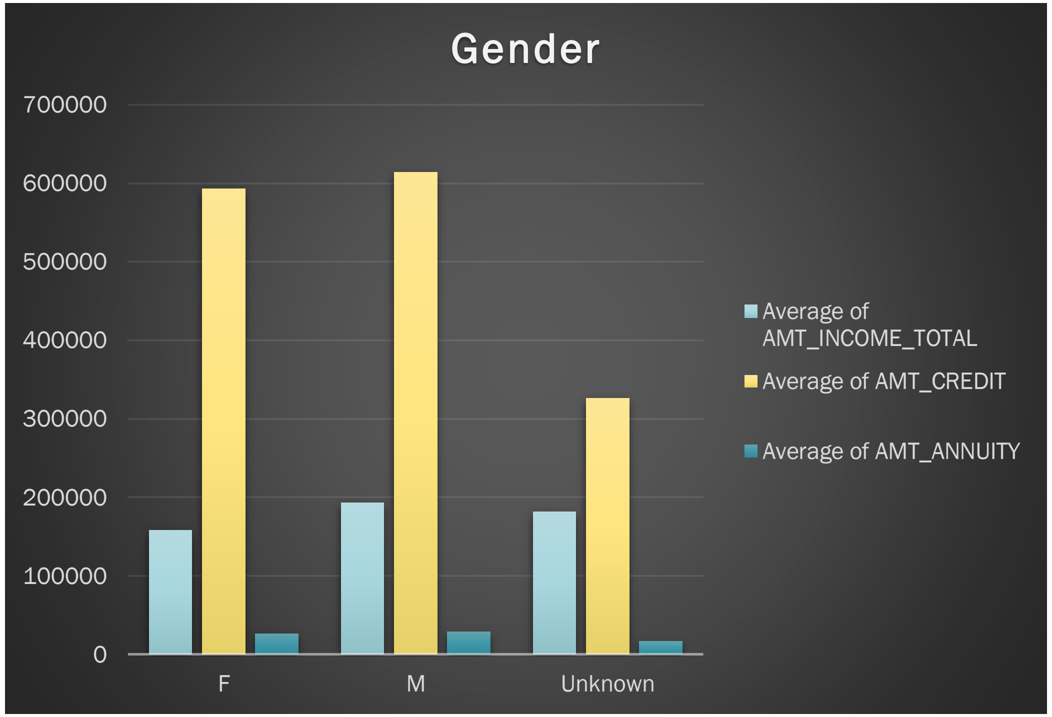
- Most of the applicants are working class followed by commercial associate. This means most applicants have regular income and may need loan to cover financial gaps whereas people who work in business may require loan for business purposes.

- Most of the applicants have education up to Secondary and were normally married.

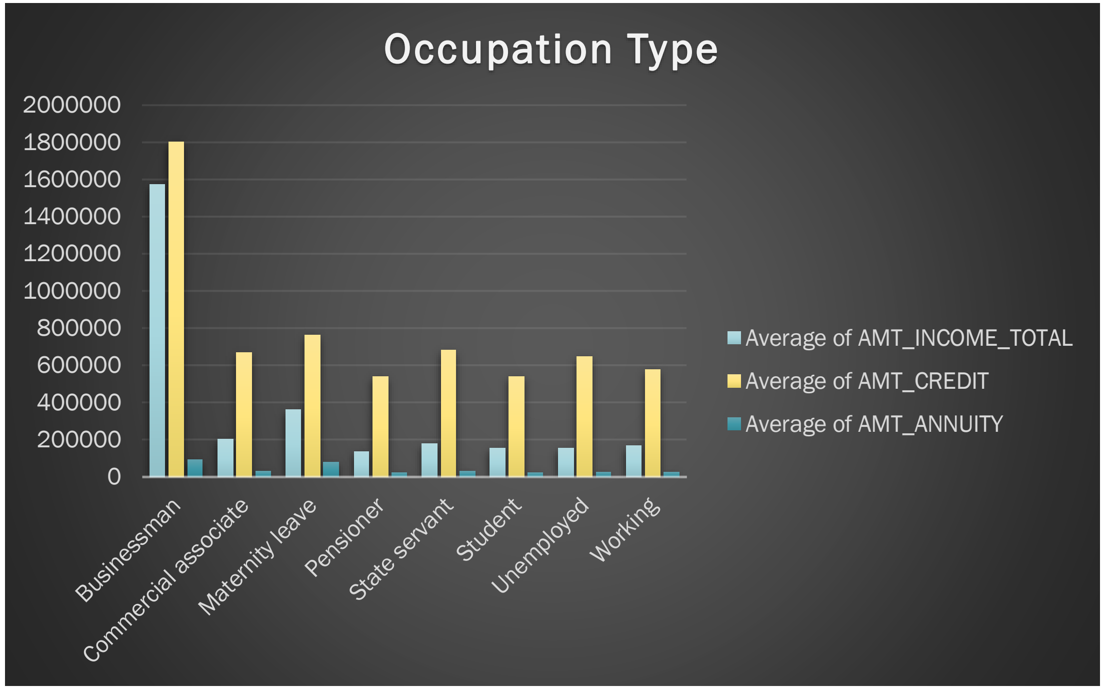
- Most of the applicants didn’t own a car but had their own property. This could indicate that they have some level of financial stability and security having invested in real estate and might need loan to cover the expenses.

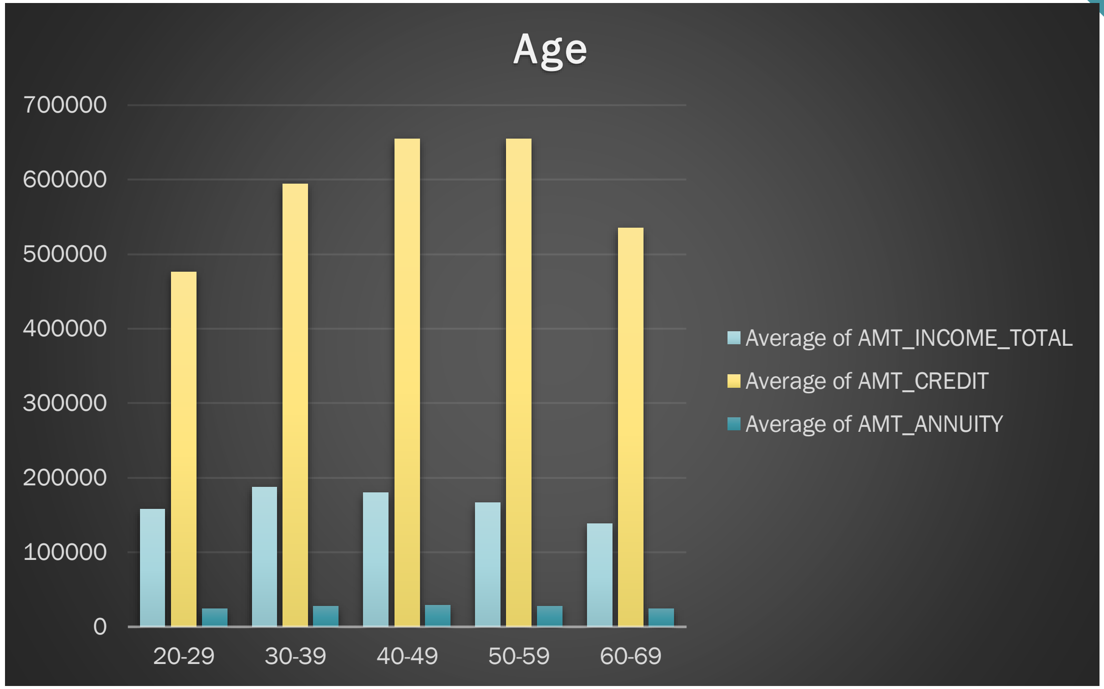
Segmented Univariate Analysis





Segmented Univariate Analysis

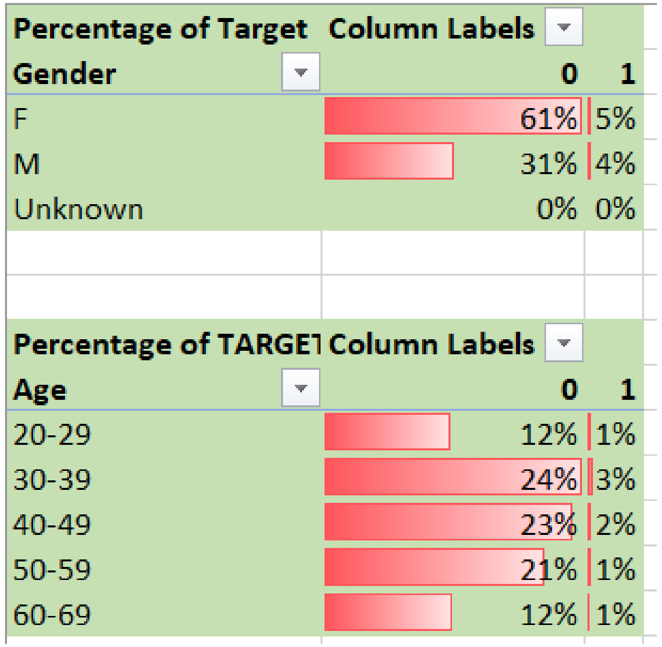




Insights

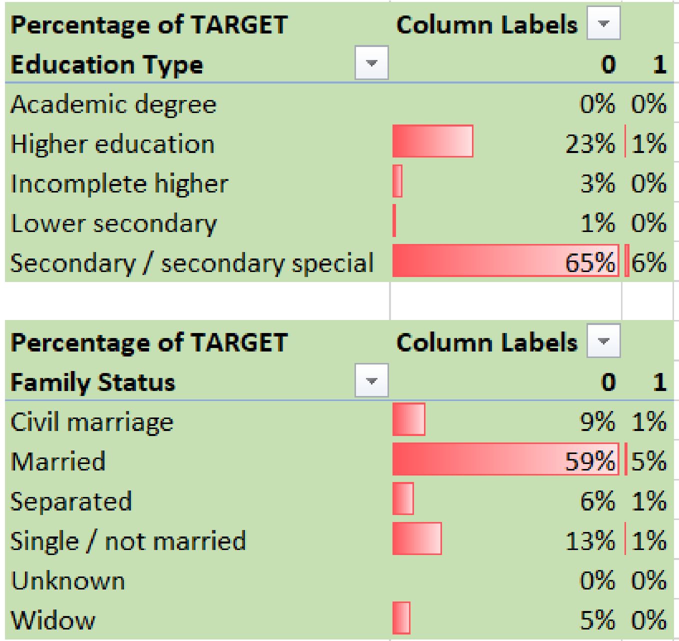
- From segmented univariate analysis it can be observed that, in terms of education applicants having higher education have highest average income. These applicants are less likely to default loans because higher incomes enable them to comfortably repay their loans.
- It can also be observed that applicants having academic degree have highest average credit amount. This suggests that bank view degree holders as more creditworthy which further suggests that they are less likely to default.
- The average income of male applicants is more than that of a female. The analysis remains the same for average credit amount and annuity amount. This suggests women often earn less than men indicating payment gap.
- Applicants who are businessmen have highest average amount income and highest average amount credit. This suggests that businessmen have more financial resources and use loan amount to invest in their business which in turns increases their credit amount.
- Applicants whose age are between 40 to 60 have highest average income and highest credit amount. This suggests that applicants in this age group have their careers established leading to higher incomes. These applicants might have had the time to build credit history which lead to higher credit amount. Applicants falling in this age group are more likely to take debt for investments, or personal finances.

Bivariate Analysis



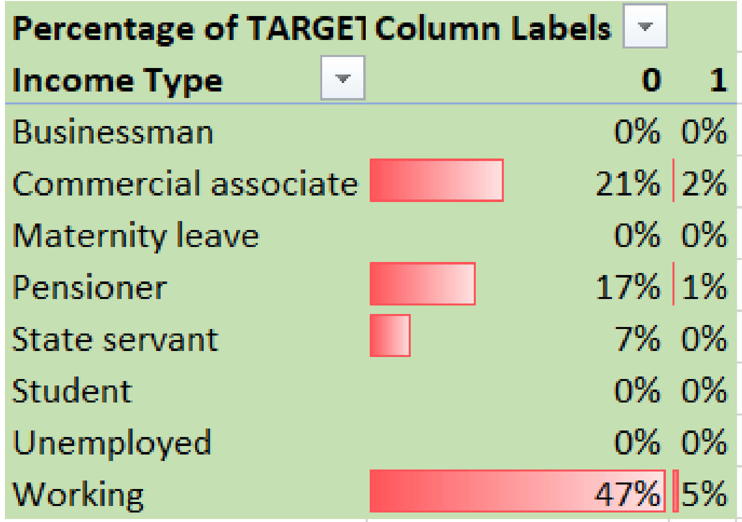
Percentage of Target	Column Labels	
Gender	0	1
F	61%	5%
M	31%	4%
Unknown	0%	0%

Percentage of TARGET	Column Labels	
Age	0	1
20-29	12%	1%
30-39	24%	3%
40-49	23%	2%
50-59	21%	1%
60-69	12%	1%

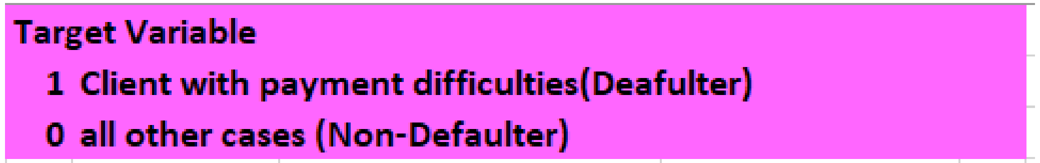


Percentage of TARGET	Column Labels	
Education Type	0	1
Academic degree	0%	0%
Higher education	23%	1%
Incomplete higher	3%	0%
Lower secondary	1%	0%
Secondary / secondary special	65%	6%

Percentage of TARGET	Column Labels	
Family Status	0	1
Civil marriage	9%	1%
Married	59%	5%
Separated	6%	1%
Single / not married	13%	1%
Unknown	0%	0%
Widow	5%	0%



Percentage of TARGET	Column Labels	
Income Type	0	1
Businessman	0%	0%
Commercial associate	21%	2%
Maternity leave	0%	0%
Pensioner	17%	1%
State servant	7%	0%
Student	0%	0%
Unemployed	0%	0%
Working	47%	5%

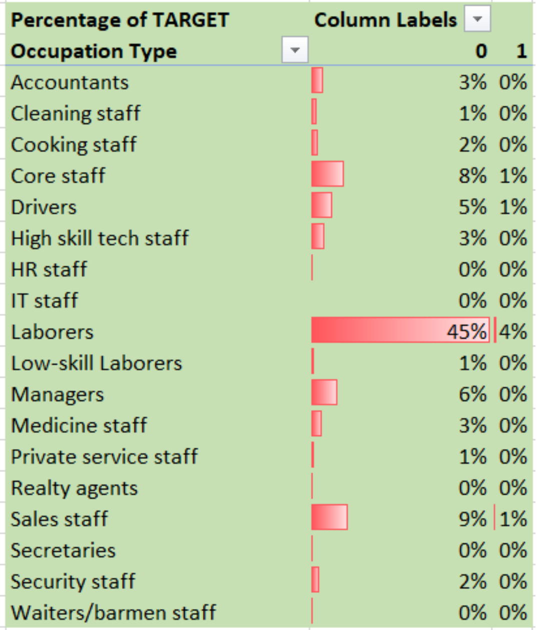


Target Variable

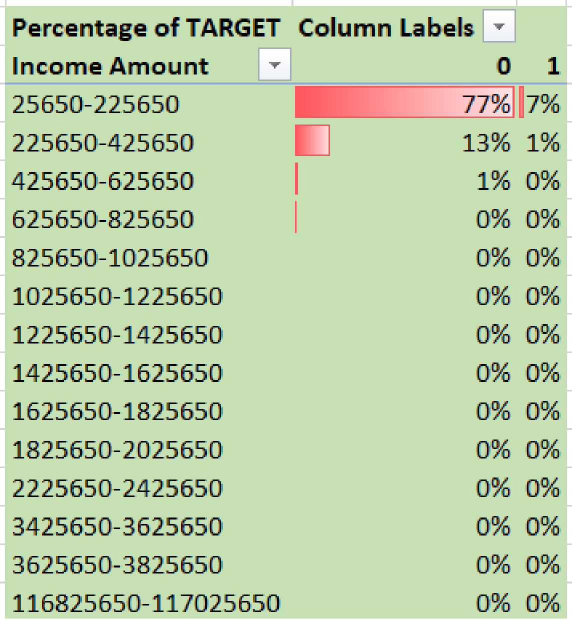
1 Client with payment difficulties(Deafulter)

0 all other cases (Non-Defaulter)

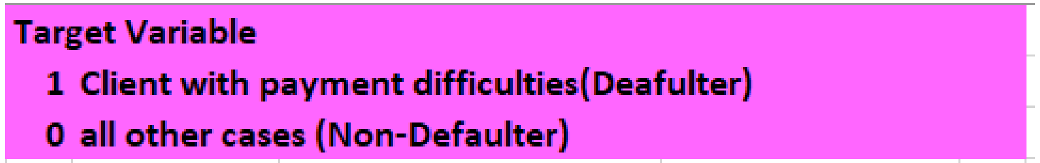
Bivariate Analysis



Percentage of TARGET	Column Labels	
Occupation Type	0	1
Accountants	3%	0%
Cleaning staff	1%	0%
Cooking staff	2%	0%
Core staff	8%	1%
Drivers	5%	1%
High skill tech staff	3%	0%
HR staff	0%	0%
IT staff	0%	0%
Laborers	45%	4%
Low-skill Laborers	1%	0%
Managers	6%	0%
Medicine staff	3%	0%
Private service staff	1%	0%
Realty agents	0%	0%
Sales staff	9%	1%
Secretaries	0%	0%
Security staff	2%	0%
Waiters/barmen staff	0%	0%



Percentage of TARGET	Column Labels	
Income Amount	0	1
25650-225650	77%	7%
225650-425650	13%	1%
425650-625650	1%	0%
625650-825650	0%	0%
825650-1025650	0%	0%
1025650-1225650	0%	0%
1225650-1425650	0%	0%
1425650-1625650	0%	0%
1625650-1825650	0%	0%
1825650-2025650	0%	0%
2225650-2425650	0%	0%
3425650-3625650	0%	0%
3625650-3825650	0%	0%
116825650-117025650	0%	0%



Target Variable	
1	Client with payment difficulties(Deafulter)
0	all other cases (Non-Defaulter)

Insights

- From bivariate analysis, it can be observed that females have high percentage of difficulty with payment (5%) comparing to male (4%). Considering only gender to determine the default risk can be misleading, so other factors like income, age, education, etc should be considered for more accurate analysis.

- When considering family status, married people have highest percentage of difficulty with payment, with Widow being the lowest (Unknown is considered exception).

- Applicants having secondary/secondary special education have the highest percentage of payment difficulty whereas applicants having academic degree have lowest default rate.

- Applicants who are working have highest percentage of payment difficulty followed by commercial associate. Students and businessmen have no default record. These two category can be considered safe for providing loan.

- Applicants in the age group of 30-50 have high percentage of payment difficulty.

- Applicants who are labourers have highest percentage of payment difficulty.

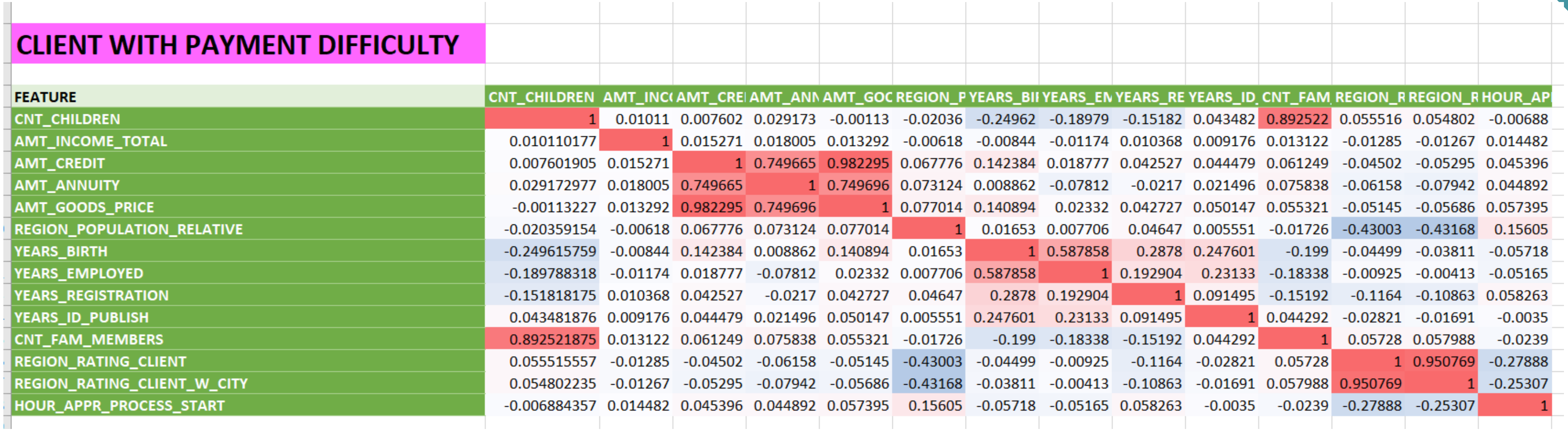
- Applicants having income less than 300000 are more likely to have difficulty in loan repayment whereas those having income more than 500000 are less likely to default.

B.) Identify Top Correlations for Different Scenarios

- The task is to 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.

- Create correlation matrices or heatmaps to visualize the correlations between variables within each segment. Highlight the top correlated variables for each scenario using different colors or shading.

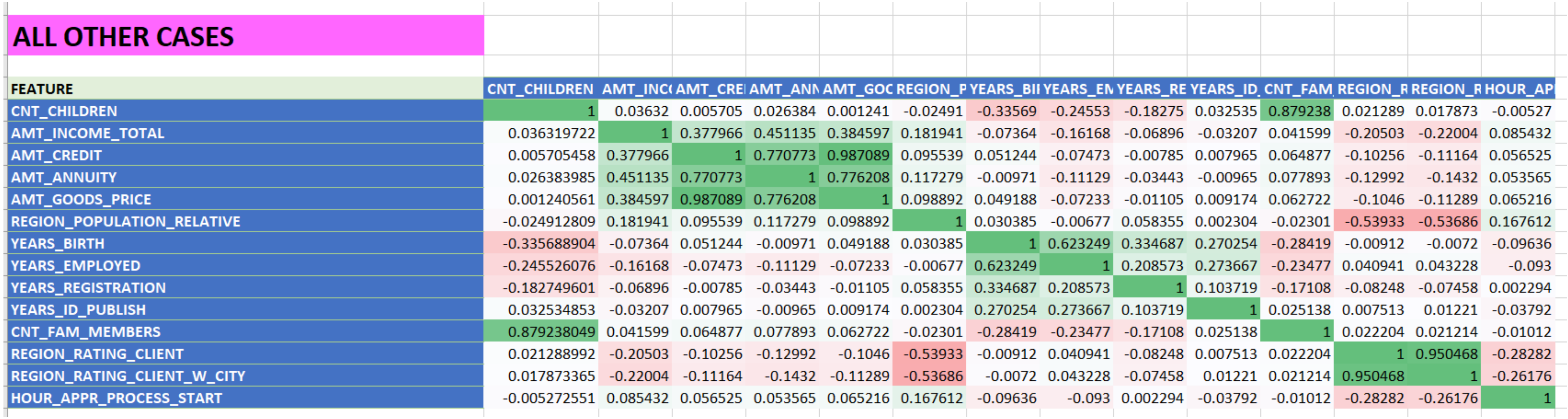
Correlation Matrix for Loan Defaulters



CLIENT WITH PAYMENT DIFFICULTY

FEATURE	CNT_CHILDREN	AMT_INC	AMT_CRE	AMT_ANN	AMT_GOC	REGION_P	YEARS_BI	YEARS_EN	YEARS_RE	YEARS_ID_	CNT_FAM_	REGION_R	REGION_R	HOUR_AP
CNT_CHILDREN	1	0.01011	0.007602	0.029173	-0.00113	-0.02036	-0.24962	-0.18979	-0.15182	0.043482	0.892522	0.055516	0.054802	-0.00688
AMT_INCOME_TOTAL	0.010110177	1	0.015271	0.018005	0.013292	-0.00618	-0.00844	-0.01174	0.010368	0.009176	0.013122	-0.01285	-0.01267	0.014482
AMT_CREDIT	0.007601905	0.015271	1	0.749665	0.982295	0.067776	0.142384	0.018777	0.042527	0.044479	0.061249	-0.04502	-0.05295	0.045396
AMT_ANNUITY	0.029172977	0.018005	0.749665	1	0.749696	0.073124	0.008862	-0.07812	-0.0217	0.021496	0.075838	-0.06158	-0.07942	0.044892
AMT_GOODS_PRICE	-0.00113227	0.013292	0.982295	0.749696	1	0.077014	0.140894	0.02332	0.042727	0.050147	0.055321	-0.05145	-0.05686	0.057395
REGION_POPULATION_RELATIVE	-0.020359154	-0.00618	0.067776	0.073124	0.077014	1	0.01653	0.007706	0.04647	0.005551	-0.01726	-0.43003	-0.43168	0.15605
YEARS_BIRTH	-0.249615759	-0.00844	0.142384	0.008862	0.140894	0.01653	1	0.587858	0.2878	0.247601	-0.199	-0.04499	-0.03811	-0.05718
YEARS_EMPLOYED	-0.189788318	-0.01174	0.018777	-0.07812	0.02332	0.007706	0.587858	1	0.192904	0.23133	-0.18338	-0.00925	-0.00413	-0.05165
YEARS_REGISTRATION	-0.151818175	0.010368	0.042527	-0.0217	0.042727	0.04647	0.2878	0.192904	1	0.091495	-0.15192	-0.1164	-0.10863	0.058263
YEARS_ID_PUBLISH	0.043481876	0.009176	0.044479	0.021496	0.050147	0.005551	0.247601	0.23133	0.091495	1	0.044292	-0.02821	-0.01691	-0.0035
CNT_FAM_MEMBERS	0.892521875	0.013122	0.061249	0.075838	0.055321	-0.01726	-0.199	-0.18338	-0.15192	0.044292	1	0.05728	0.057988	-0.0239
REGION_RATING_CLIENT	0.055515557	-0.01285	-0.04502	-0.06158	-0.05145	-0.43003	-0.04499	-0.00925	-0.1164	-0.02821	0.05728	1	0.950769	-0.27888
REGION_RATING_CLIENT_W_CITY	0.054802235	-0.01267	-0.05295	-0.07942	-0.05686	-0.43168	-0.03811	-0.00413	-0.10863	-0.01691	0.057988	0.950769	1	-0.25307
HOUR_APPR_PROCESS_START	-0.006884357	0.014482	0.045396	0.044892	0.057395	0.15605	-0.05718	-0.05165	0.058263	-0.0035	-0.0239	-0.27888	-0.25307	1

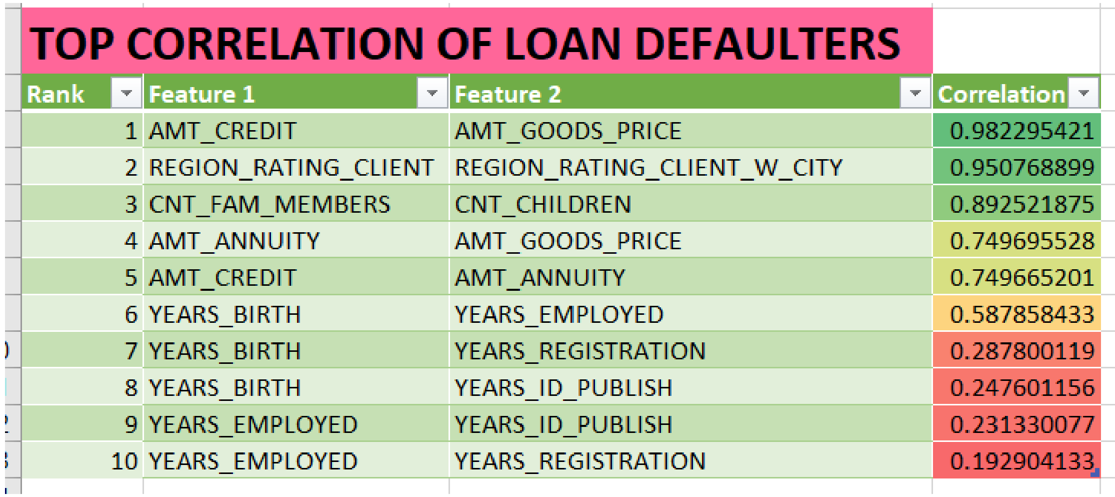
Correlation Matrix for Non-Defaulters



ALL OTHER CASES

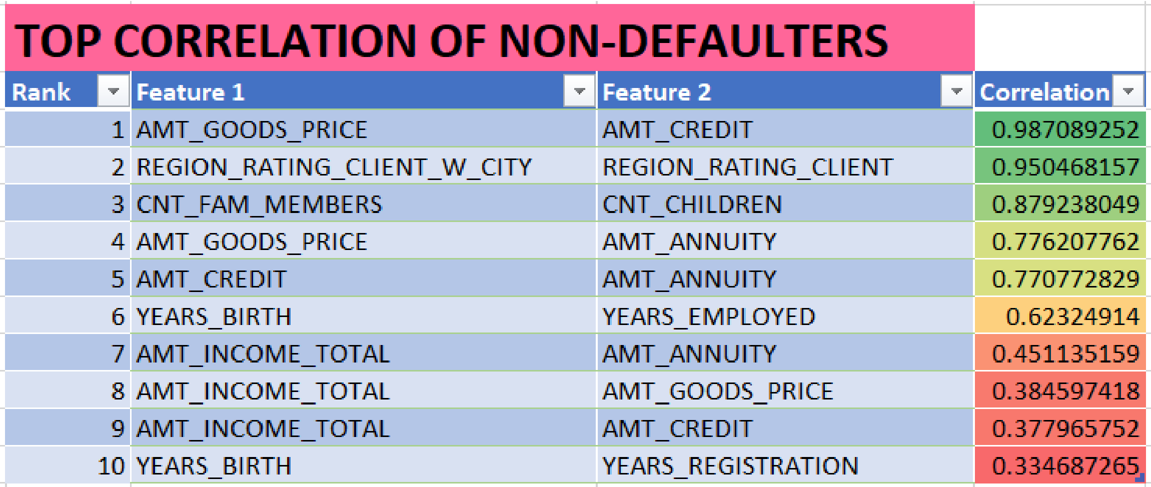
FEATURE	CNT_CHILDREN	AMT_INC	AMT_CRE	AMT_ANN	AMT_GOC	REGION_P	YEARS_BI	YEARS_EN	YEARS_RE	YEARS_ID_	CNT_FAM_	REGION_R	REGION_R	HOUR_AP
CNT_CHILDREN	1	0.03632	0.005705	0.026384	0.001241	-0.02491	-0.33569	-0.24553	-0.18275	0.032535	0.879238	0.021289	0.017873	-0.00527
AMT_INCOME_TOTAL	0.036319722	1	0.377966	0.451135	0.384597	0.181941	-0.07364	-0.16168	-0.06896	-0.03207	0.041599	-0.20503	-0.22004	0.085432
AMT_CREDIT	0.005705458	0.377966	1	0.770773	0.987089	0.095539	0.051244	-0.07473	-0.00785	0.007965	0.064877	-0.10256	-0.11164	0.056525
AMT_ANNUITY	0.026383985	0.451135	0.770773	1	0.776208	0.117279	-0.00971	-0.11129	-0.03443	-0.00965	0.077893	-0.12992	-0.1432	0.053565
AMT_GOODS_PRICE	0.001240561	0.384597	0.987089	0.776208	1	0.098892	0.049188	-0.07233	-0.01105	0.009174	0.062722	-0.1046	-0.11289	0.065216
REGION_POPULATION_RELATIVE	-0.024912809	0.181941	0.095539	0.117279	0.098892	1	0.030385	-0.00677	0.058355	0.002304	-0.02301	-0.53933	-0.53686	0.167612
YEARS_BIRTH	-0.335688904	-0.07364	0.051244	-0.00971	0.049188	0.030385	1	0.623249	0.334687	0.270254	-0.28419	-0.00912	-0.0072	-0.09636
YEARS_EMPLOYED	-0.245526076	-0.16168	-0.07473	-0.11129	-0.07233	-0.00677	0.623249	1	0.208573	0.273667	-0.23477	0.040941	0.043228	-0.093
YEARS_REGISTRATION	-0.182749601	-0.06896	-0.00785	-0.03443	-0.01105	0.058355	0.334687	0.208573	1	0.103719	-0.17108	-0.08248	-0.07458	0.002294
YEARS_ID_PUBLISH	0.032534853	-0.03207	0.007965	-0.00965	0.009174	0.002304	0.270254	0.273667	0.103719	1	0.025138	0.007513	0.01221	-0.03792
CNT_FAM_MEMBERS	0.879238049	0.041599	0.064877	0.077893	0.062722	-0.02301	-0.28419	-0.23477	-0.17108	0.025138	1	0.022204	0.021214	-0.01012
REGION_RATING_CLIENT	0.021288992	-0.20503	-0.10256	-0.12992	-0.1046	-0.53933	-0.00912	0.040941	-0.08248	0.007513	0.022204	1	0.950468	-0.28282
REGION_RATING_CLIENT_W_CITY	0.017873365	-0.22004	-0.11164	-0.1432	-0.11289	-0.53686	-0.0072	0.043228	-0.07458	0.01221	0.021214	0.950468	1	-0.26176
HOUR_APPR_PROCESS_START	-0.005272551	0.085432	0.056525	0.053565	0.065216	0.167612	-0.09636	-0.093	0.002294	-0.03792	-0.01012	-0.28282	-0.26176	1

Top Correlation



TOP CORRELATION OF LOAN DEFAULTERS

Rank	Feature 1	Feature 2	Correlation
1	AMT_CREDIT	AMT_GOODS_PRICE	0.982295421
2	REGION_RATING_CLIENT	REGION_RATING_CLIENT_W_CITY	0.950768899
3	CNT_FAM_MEMBERS	CNT_CHILDREN	0.892521875
4	AMT_ANNUITY	AMT_GOODS_PRICE	0.749695528
5	AMT_CREDIT	AMT_ANNUITY	0.749665201
6	YEARS_BIRTH	YEARS_EMPLOYED	0.587858433
7	YEARS_BIRTH	YEARS_REGISTRATION	0.287800119
8	YEARS_BIRTH	YEARS_ID_PUBLISH	0.247601156
9	YEARS_EMPLOYED	YEARS_ID_PUBLISH	0.231330077
10	YEARS_EMPLOYED	YEARS_REGISTRATION	0.192904133



TOP CORRELATION OF NON-DEFAULTERS

Rank	Feature 1	Feature 2	Correlation
1	AMT_GOODS_PRICE	AMT_CREDIT	0.987089252
2	REGION_RATING_CLIENT_W_CITY	REGION_RATING_CLIENT	0.950468157
3	CNT_FAM_MEMBERS	CNT_CHILDREN	0.879238049
4	AMT_GOODS_PRICE	AMT_ANNUITY	0.776207762
5	AMT_CREDIT	AMT_ANNUITY	0.770772829
6	YEARS_BIRTH	YEARS_EMPLOYED	0.62324914
7	AMT_INCOME_TOTAL	AMT_ANNUITY	0.451135159
8	AMT_INCOME_TOTAL	AMT_GOODS_PRICE	0.384597418
9	AMT_INCOME_TOTAL	AMT_CREDIT	0.377965752
10	YEARS_BIRTH	YEARS_REGISTRATION	0.334687265

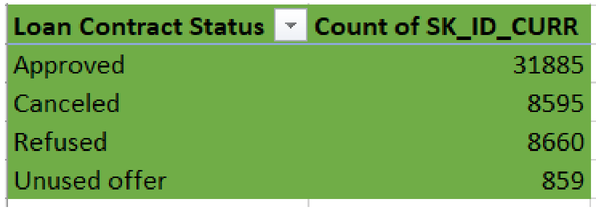
Insights

- From the correlation analysis of different segments, it can be observed that AMT_CREDIT and AMT_GOODS_PRICE has strong correlation which suggests that as the AMT_GOODS_PRICE increases AMT_CREDIT also increases suggesting this correlation is not useful for predicting loan defaults.
- It can also be observed that correlation of AMT-ANNUITY with AMT_CREDIT is slightly reduced in loan defaulters as compared to non-defaulters.
- It can be observed that correlation of AMT_INCOME_TOTAL with AMT_CREDIT, AMT_ANNUITY and AMT_GOODS_PRICE is decent in case of non- defaulter case whereas in loan defaulter case it is not strong. This suggest that high income and low credit or annuity predicts loan default.
- By comparing these correlation the company can better understand the reason for loan default and make informed decision.

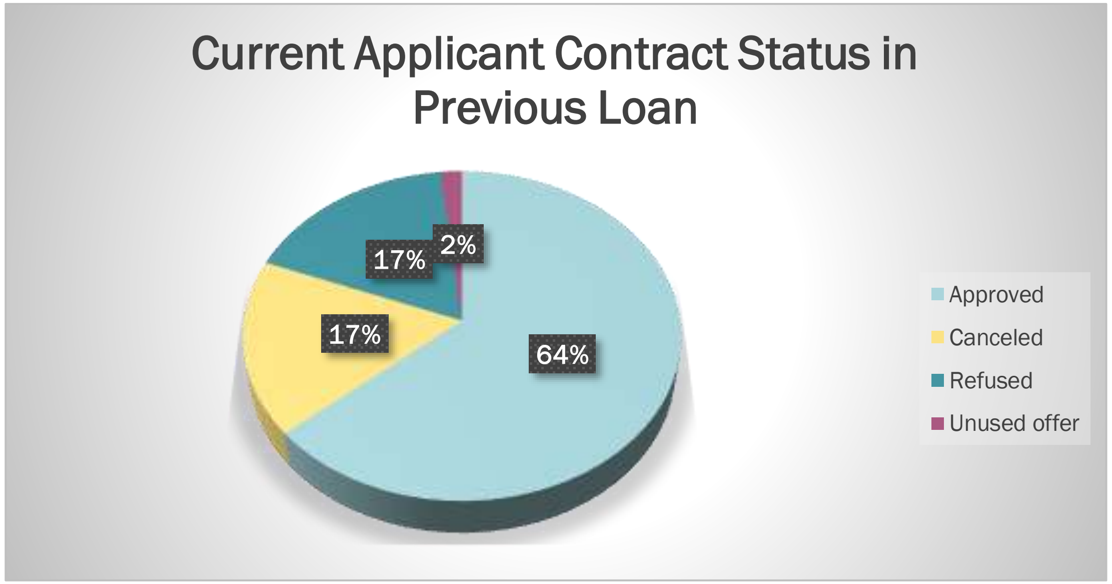
Analysis on Previous Application

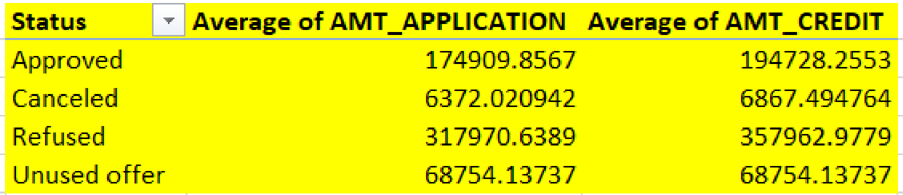
- For a deeper dive or more nuanced insights, some analysis has been done to gain more insights on the patterns of customer being a defaulter or non-defaulter in bank loan.

Analysis

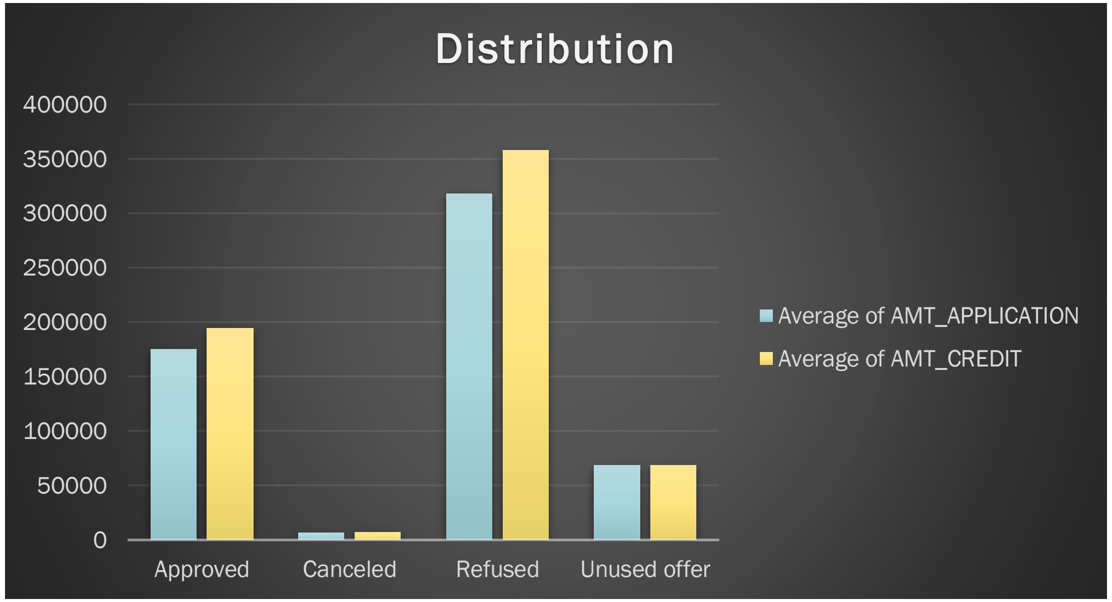


Loan Contract Status	Count of SK_ID_CURR
Approved	31885
Canceled	8595
Refused	8660
Unused offer	859

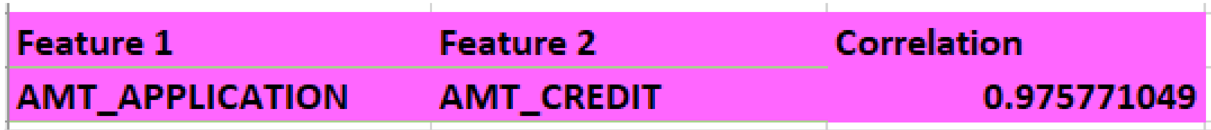




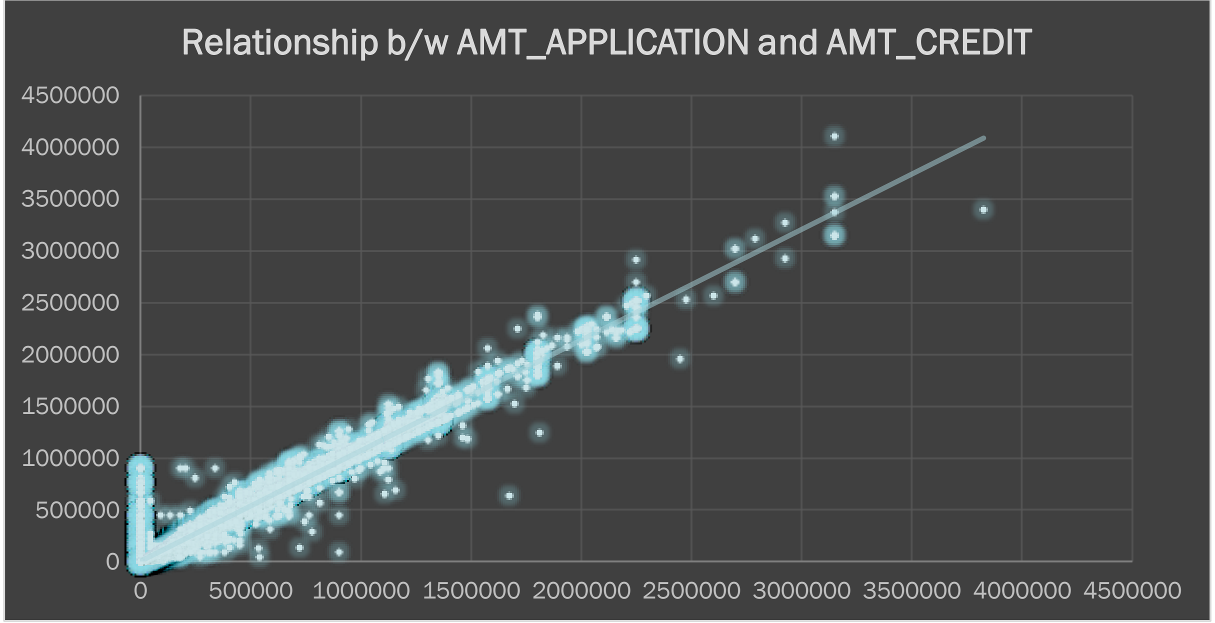
Status	Average of AMT_APPLICATION	Average of AMT_CREDIT
Approved	174909.8567	194728.2553
Canceled	6372.020942	6867.494764
Refused	317970.6389	357962.9779
Unused offer	68754.13737	68754.13737



Analysis



Feature 1	Feature 2	Correlation
AMT_APPLICATION	AMT_CREDIT	0.975771049



Insights

- By analysing some aspects of previous loan dataset, it can be observed that applicants from current application loan who had previously applied for loan, 64% applicants loan application had been approved whereas 17% applicants had cancelled the application during the approval process. 17% applicants have been rejected for loan and the remaining 2% did not use the approved loan.

- It can also be observed that applicants whose loan had been approved had an average application amount of 180000 and average amount credited was around 200000.

- Applicants who cancelled the application during the approval process had average application amount and average credit amount around 7000 approximately.

- Applicants whose loan were rejected had average application amount around 320000.

- Applicants who did not use the loan amount had an average application amount as well as credit amount around 70000.

- By analysing the relationship between Amount application and Amount credit it can be observed that the correlation is positive and is quite close to 1 suggesting higher application amount leads to higher credit amount. The scatter plot has a clear trendline which suggests strong correlation.

Result

- Through this project as a data analyst, I gained significant insights that proved to be valuable. It deepened my knowledge of data analysis in organizations and process of making data driven decisions. This experience helped in improving my analytical skills and MS-Excel skills. I became comfortable in performing various functions of MS-Excel and how to execute them effectively.

- I found and fixed missing data which made the dataset more reliable. I identified potential outliers. I also measured the imbalance of data which helped in interpreting the results better. I also analyzed the data to see what factors affect loan defaults which would help in making decisions on whether to approve loans or not. I also identified important factors that predict loan default which would help assess risk in a more accurate way.

- Overall, this project helped me in understanding loan default risk better which would lead to smarter decisions on loan approach and improve risk management for the finance company. This experience will further help in boosting my career as data analyst in uncovering valuable insights from data and making informed decisions.

Thank you

Garima Thakur

garima18thakur@gmail.com