**********Bank Loan Case Study**********

https://docs.google.com/spreadsheets/d/16BBUsBpeBC9GToZmsaA5APyHtLe2PWX2/edit?usp=drive_web&ouid=116000631563963995249&rtpof=true

Project Description:-

This project aims to analyse the risk appetite of banks when deciding whether to approve loan applications based on applicant profiles. There are two primary risks associated with these decisions:

1. **Risk of Not Approving a Loan to a Creditworthy Applicant**:
   If a bank rejects a loan application from an applicant who would have repaid the loan successfully, the bank loses potential business.
2. **Risk of Approving a Loan to a Default-Prone Applicant**:
   If a bank approves a loan for an applicant who is likely to default, it could lead to financial losses for the bank.

The data provided includes information about loan applications at the time of submission. This data is divided into two scenarios:

1. **Clients with Payment Difficulties**:
   These applicants have demonstrated late payments exceeding X days on at least one of the first Y loan instalments.
2. **All Other Cases**:
   This category covers applicants who made timely payments throughout.

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

1. Approved: The company has approved the loan application.
2. Cancelled: The customer cancelled the application during the approval process.
3. Refused: The company rejected the loan.
4. Unused Offer: The loan was approved but the customer did not use it.

The objective is to conduct a comprehensive analysis of this data to identify patterns and derive insights. These insights will assist the bank in taking appropriate actions, such as:

- Denying loans to applicants identified as high-risk.
- Reducing loan amounts for certain applicants.
- Offering loans to risky applicants at higher interest rates to mitigate potential losses.

By leveraging these insights, banks can optimize their loan approval processes to ensure that creditworthy consumers are not unfairly denied loans, while effectively managing the risks associated with lending to less creditworthy individuals.

Tech Stack Used:-

Microsoft Excel 365, the purpose is to perform comprehensive analysis and create graphical representations of the results to enhance understanding of the dataset.

Excel is utilized for:

1. **Data Analysis**:
   - Performing calculations, statistical analysis, and data manipulations using built-in functions and formulas.
   - Utilizing features like pivot tables to summarize and analyze large datasets.
   - Implementing conditional formatting to highlight important trends or outliers.
2. **Graphical Representation**:
   - Creating various types of charts (such as bar charts, line graphs, scatter plots, etc.) to visually represent data trends and relationships.
   - Customizing charts with titles, labels, and formatting options to enhance clarity and presentation.
   - Using dynamic charts that update automatically based on changes in underlying data.
3. **Result Interpretation**:
   - Interpreting analysis outcomes through visualizations, making it easier to identify patterns, correlations, and insights.
   - Using Excel's data visualization tools to communicate findings effectively to stakeholders.
   - Extracting actionable insights that inform decision-making processes related to loan approvals and risk assessment.

Approach:-

1. **Data Preprocessing**:
   - Counted total rows in each column using **COUNTA** function.
   - Calculated null value percentages for each column.
   - Removed columns with null value percentages exceeding 35%.

2. **Handling Missing Values**:
   - Imputed missing values (less than 35% null) using mean, median, or mode based on column characteristics.

3. **Outlier Detection**:
   - Identified outliers using interquartile range (IQR) method and BOX_PLOT for relevant columns.

4. **Data Transformation**:
   - Converted columns with day values into years by dividing days by 365.

The analysis aimed to clean and prepare the data for insights, focusing on meaningful columns while addressing missing values and outliers where necessary. The results can be found in the linked Excel file.

Recommendations:

1. **Risk Assessment Enhancements:-**
    - Utilize identified consumer attributes (from EDA) to better assess loan default risk.
    - Implement stricter criteria or adjustments in approval processes based on risk factors.

2. **Improved Screening Criteria:-**
    - Refine applicant screening criteria based on influential loan attributes identified in the analysis.
    - Adjust loan terms or conditions to mitigate potential default risks.

3. **Data-Driven Decision Making:-**
    - Encourage data-driven decision-making by leveraging insights from EDA to optimize lending strategies.

4. **Customer Segmentation:-**
    - Segment customers based on risk profiles identified in the analysis.
    - Tailor loan offerings and terms to different risk segments to optimize portfolio performance.

5. **Continuous Improvement:-**
    - Foster a culture of continuous improvement by incorporating EDA findings into risk management practices.

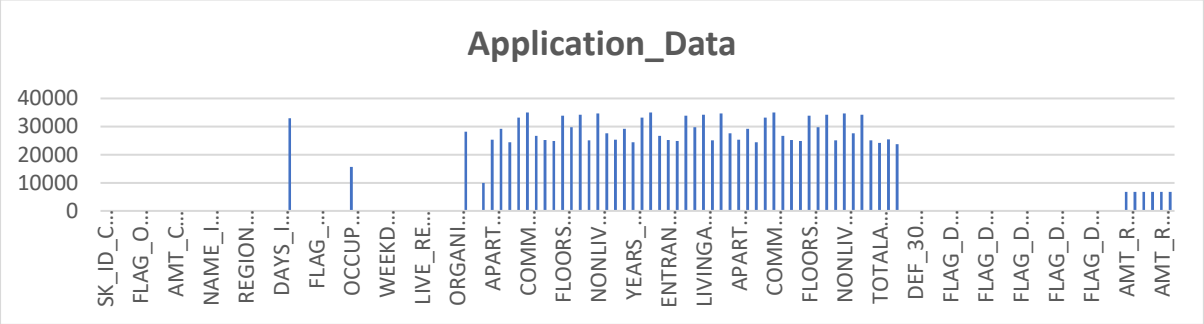
These recommendations aim to assist loan providers in making informed decisions, reducing default risks, and enhancing overall loan portfolio management based on the insights derived from the analysis.

Task A:

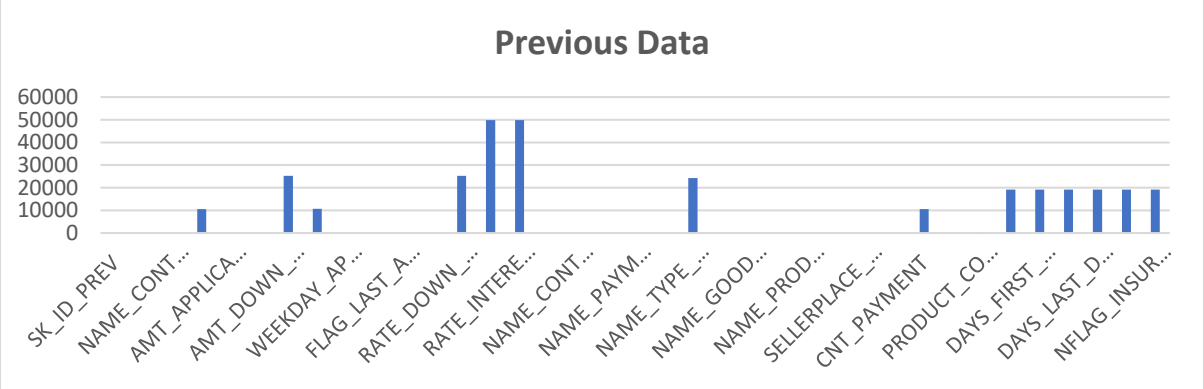
Identify Missing Data and Deal with it Appropriately:

- Prepared an EDA sheet where I see that SK_ID column has 100% rows and no Blanks, taking it as a base calculated Blank rows in rest of the columns.
- All the columns that I see having more than 35% data Blank are removed. As they can't be filled with any formula.
- Name_type_suit had some blank cells, but the column itself has no impact on the loan approval, so dropping the column.
- Sk_ID 148602 had a black cell in CNT_FAM_Members, which is filled with the mean of the count of column. Blanks in categorial data important for Analysis were replaced by Mode values.
- Finally, the column or Features which had no impact on Target Or Label Values were also deleted/dropped
- All columns in the Previous application which have blank rows more than 35% are conditionally formatted in red colour and removed to prepare the final data.
- Occupation column had 31% blank which were replaced by unknown occupation to interpret the univariate analysis.
- Row K47537 and Column AMT_Annuity has a blank cell which is filled with the average of the column AMT_ANNUITY.
- Amount_Goods_Price had some blank columns which were filled using the most frequently occurring number , received by using the formula Mode
- Occupation_Type is Categorial data which can't be treated with mean or Median or Mode but left as it is to Analyse the data

Application_Data:



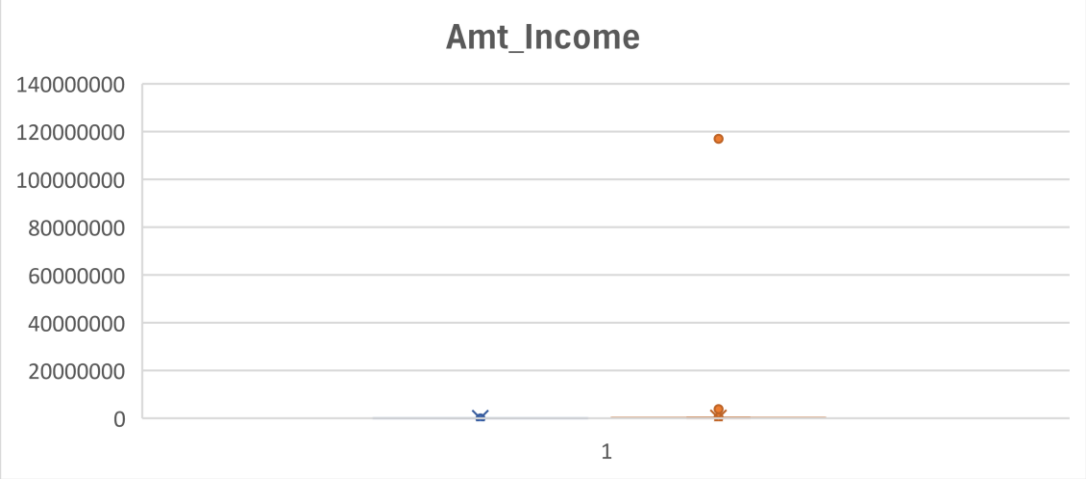
Previous_Data:

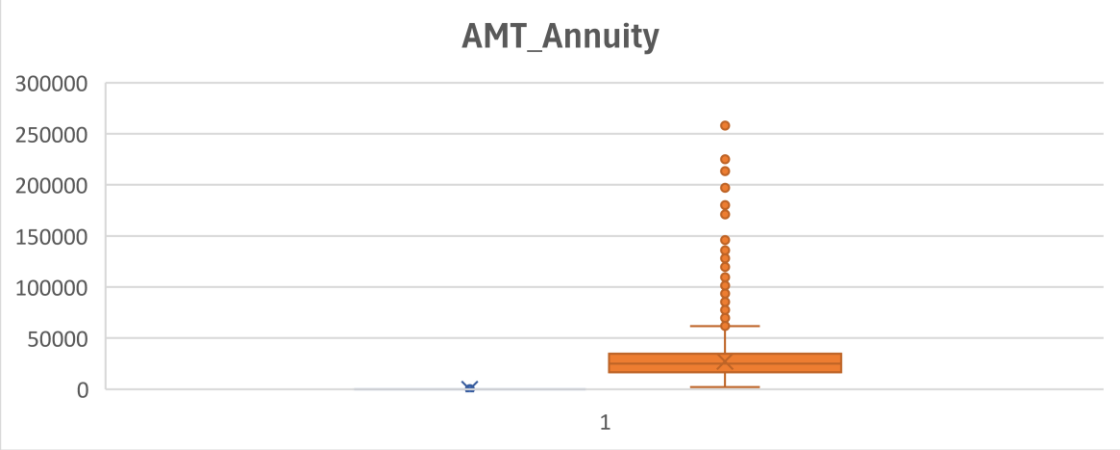


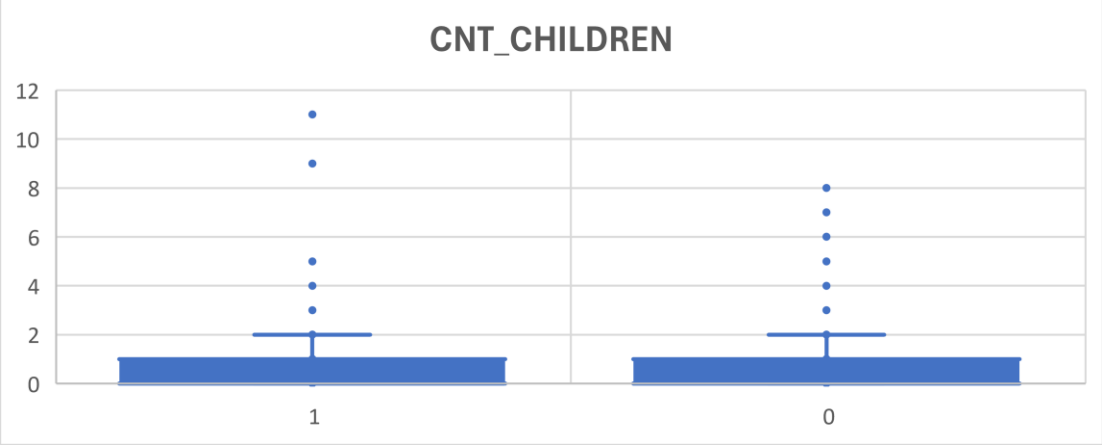
Task B:

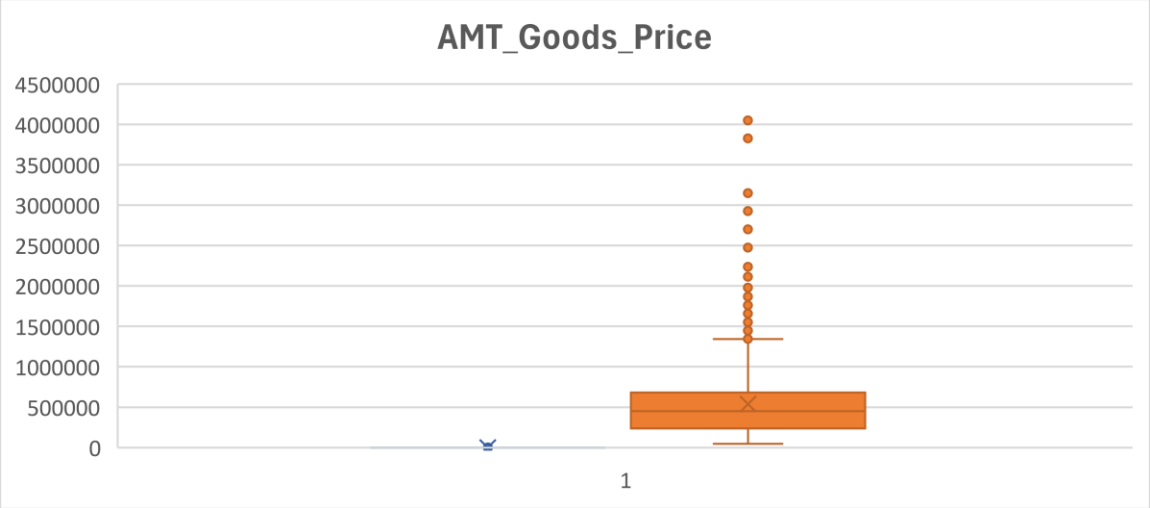
Identify Outliers in the Dataset:

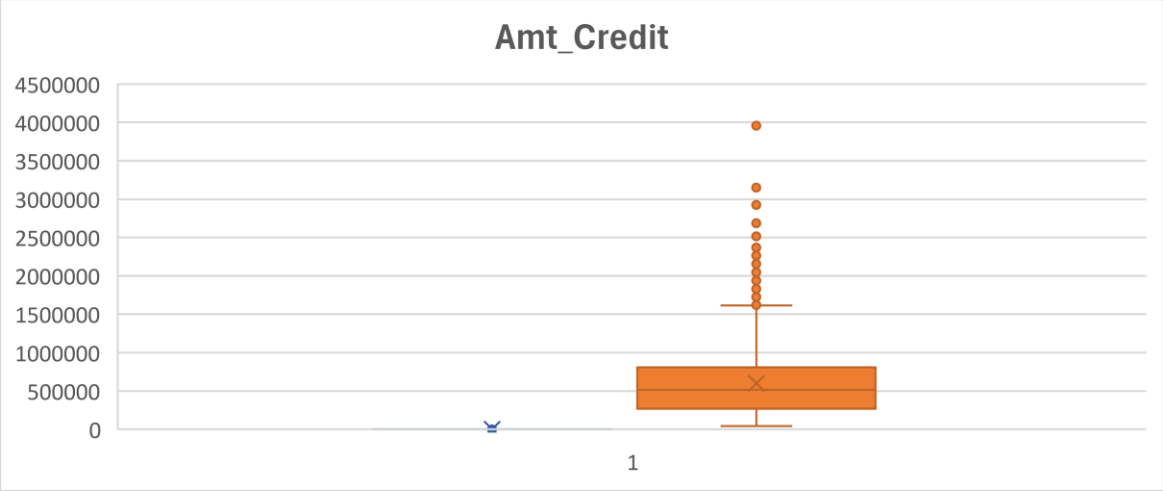
It is observed that Amt_Income has an outlier with Income of 11,70,00,000 and Days_employeed has Outliers with 1001 year employed which is impossible. But, CNT_Children has outliers both in target 0 and 1, whereas Amt_Annuity , Amt_Goods_Price, Amt_Credit and CNT_FAM_Members has outliers present in Target 1.

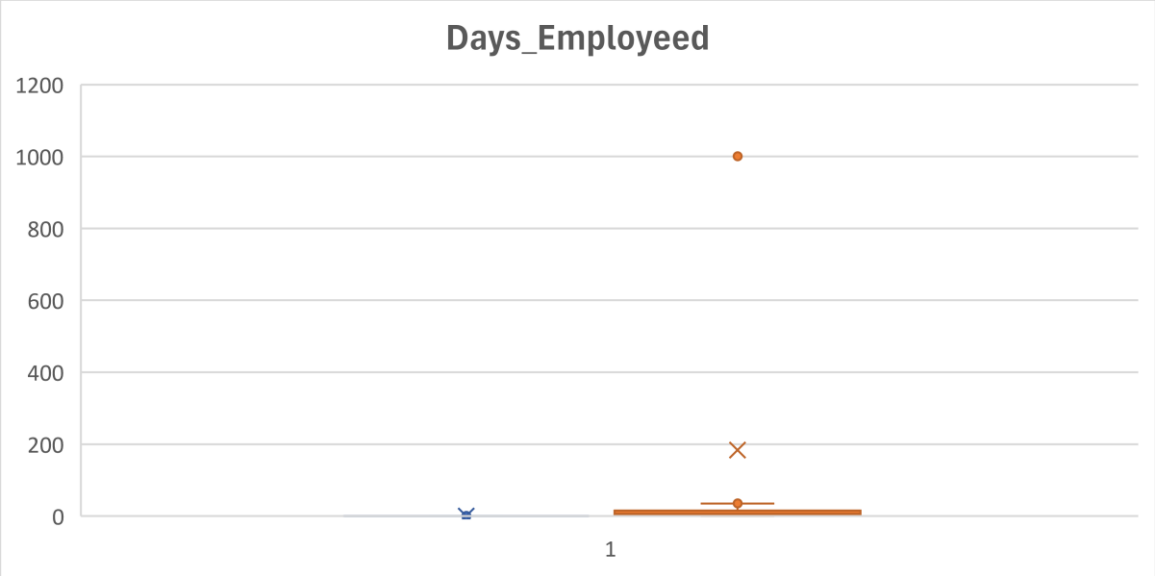


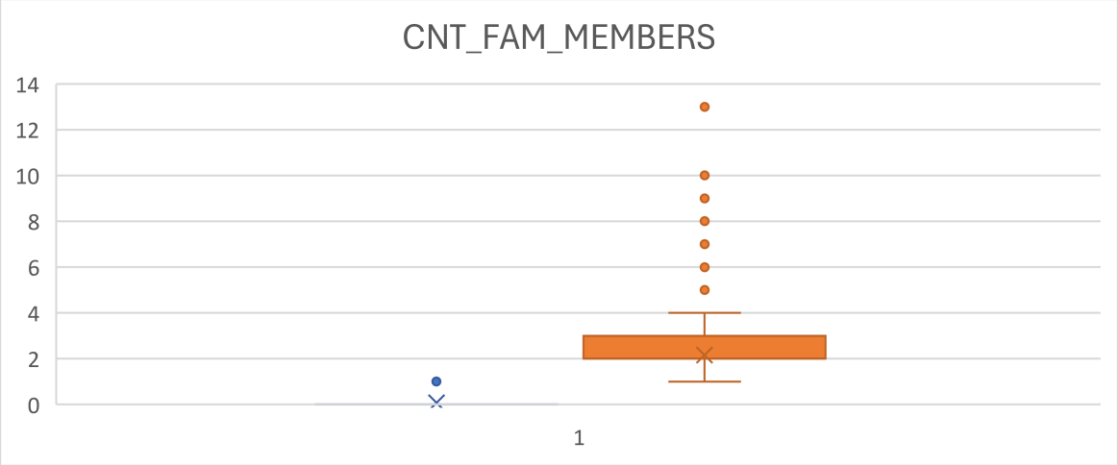





AMT_Goods_Price
4500000
4000000
3500000
3000000
2500000
2000000
1500000
1000000
500000
0
1


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


Days_Employeed
1200
1000
800
600
400
200
0
1



Task C:

Analyse 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.

Target	Count of Males	Count of females	Count of XNA	Total
1	1762	2264	0	4026
0	15412	30559	2	45973
Total	17174	32823	2	49999

- Count of Female Candidates is more than the male candidates

Target	Flag own cars(Y)	Flag own Cars(N)	Total
1	1253	2773	4026
0	15797	30176	45973
Total	17050	32949	49999

- Clients applied for loans do not Own Cars.

Target	Flag Own Realty(Y)	Flag Own Realty(N)	Total
1	2752	1274	4026
0	31939	14034	45973
Total	34691	15308	49999

- We observe that Clients owning Properties/Realty/Estate has applied for Loans

Target	Cash Loans	Revolving Loans	Total
Target	Cash Loans	Revolving Loans	Total

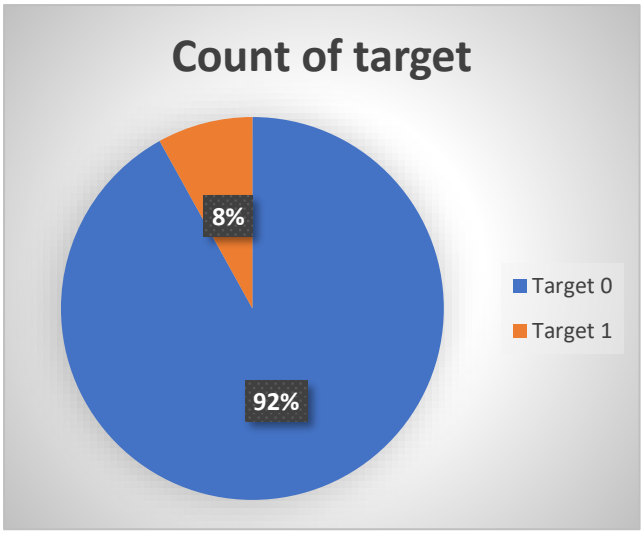
1	3792	234	4026
0	41484	4489	45973
Total	45276	4723	49999

- Clients have applied for Cash loans more than revolving loans

Target	Count of Children
1	4026
0	45973
Total	49999

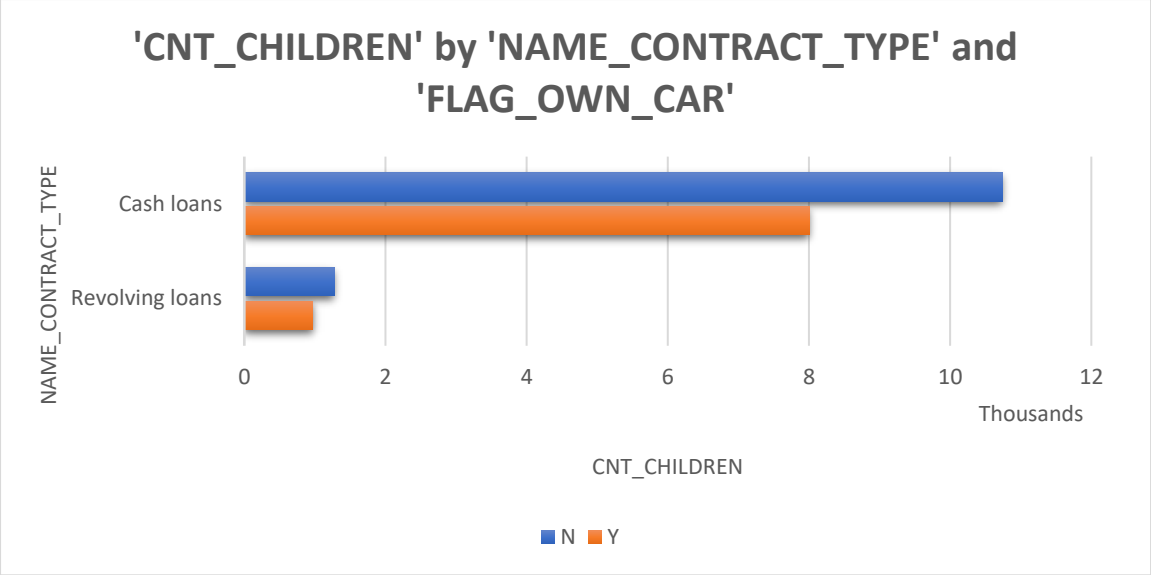
- Count of children of Clients with payment difficulties are lesser than normal Payers.

Target	Count of target
Target 0	45973
Target 1	4026
Total	49999



Ratio of Data Imbalance = $45973:4026 \approx 11.42$

Sum of CNT_CHILDREN	FLAG_OWN_CAR		
NAME_CONTRACT_TYPE	N	Y	Grand Total
Cash loans	10740	8006	18746
Revolving loans	1274	972	2246
Grand Total	**12014**	**8978**	**20992**



- The Data seems to be very Imbalanced and is more inclined towards the target Variable 0. Cash Loans are disbursed more and More in demand than Revolving loans.

Task D:

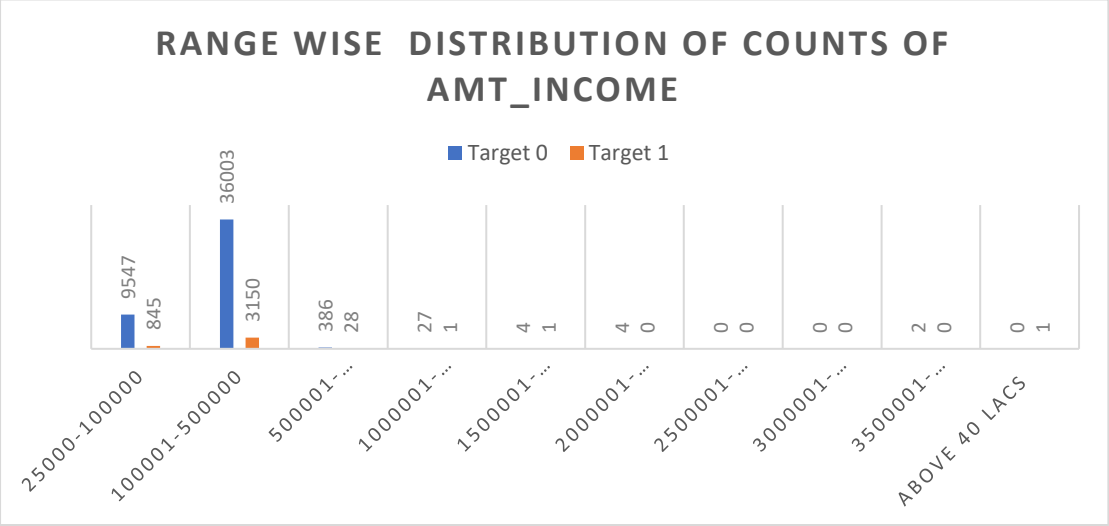
Perform Univariate, Segmented Univariate, and Bivariate Analysis

To gain insights into the driving factors of loan default, it is important to conduct various analyses on consumer and loan attributes.
Using Pivot table, Slicer, Range, Charts, Segmentation Analysis has been done.

Application_data

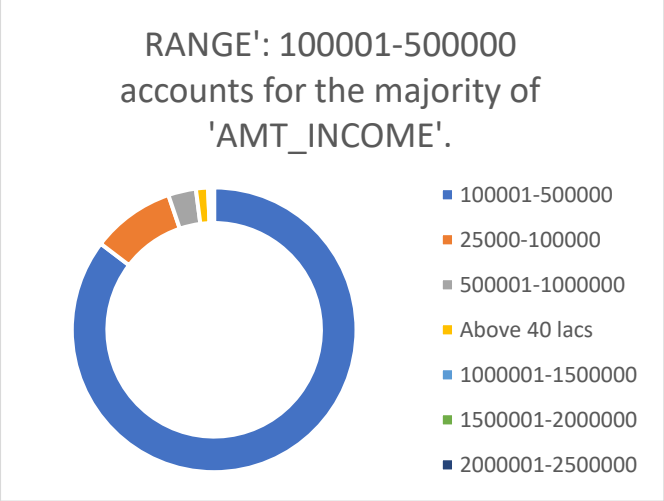
Count of target 0 and Target 1 as per the Income.

Amount_Income_Total	Target 0	Target 1
25000-100000	9547	845
100001-500000	36003	3150
500001-1000000	386	28
1000001-1500000	27	1
1500001-2000000	4	1
2000001-2500000	4	0
2500001-3000000	0	0
3000001-3500000	0	0
3500001-4000000	2	0
Above 40 lacs	0	1

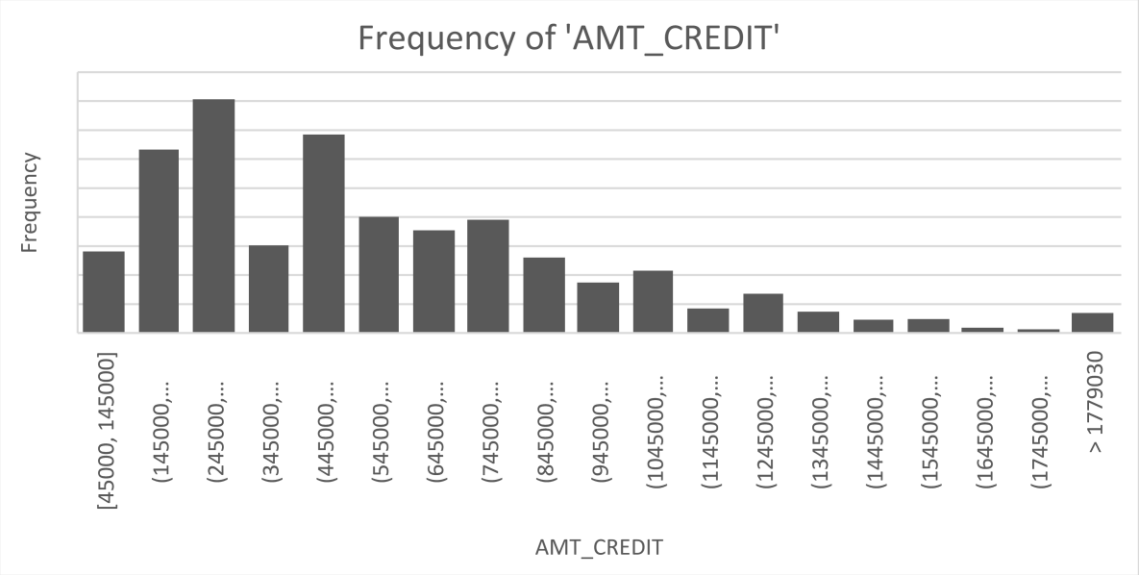


We can see that Clients who apply for Loans fall into the category of 100001-500000, and their loans get approval with target 0.

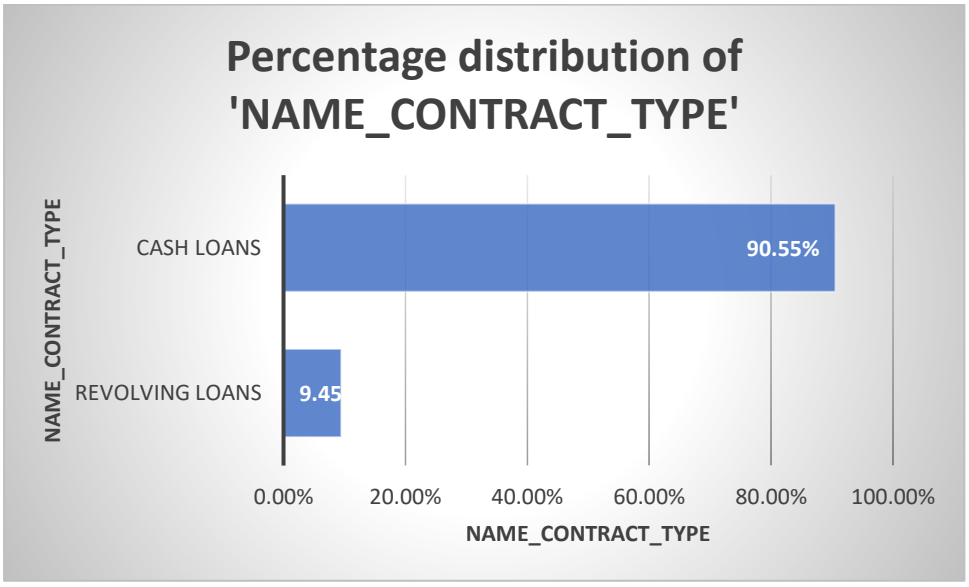
RANGE	Sum of AMT_INCOME
100001-500000	7286100131
25000-100000	807168449.7
500001-1000000	269327677.5
Above 40 lacs	117000000
1000001-1500000	33682500
1500001-2000000	8955000
2000001-2500000	8550000
3500001-4000000	7425000
Grand Total	**8538208758**



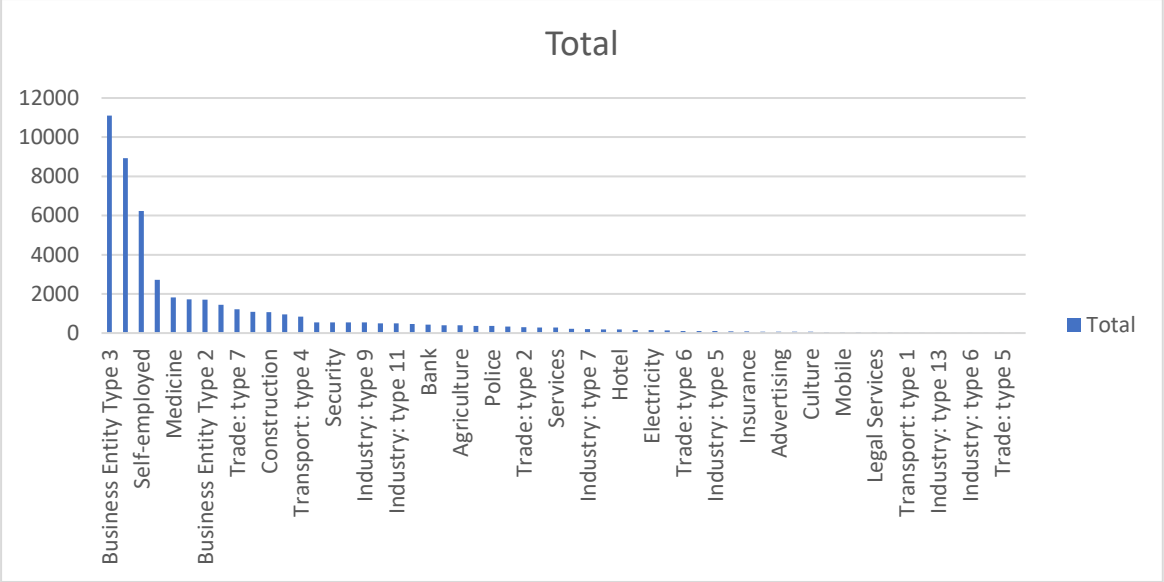
- Clients having Income Between 100001-500000 has applied for Loans. But only 36003 got the Loans approved.



- Clients have received the max amount credit between 245000-345000. As their AMT_Income falls in criteria 100001-500000.

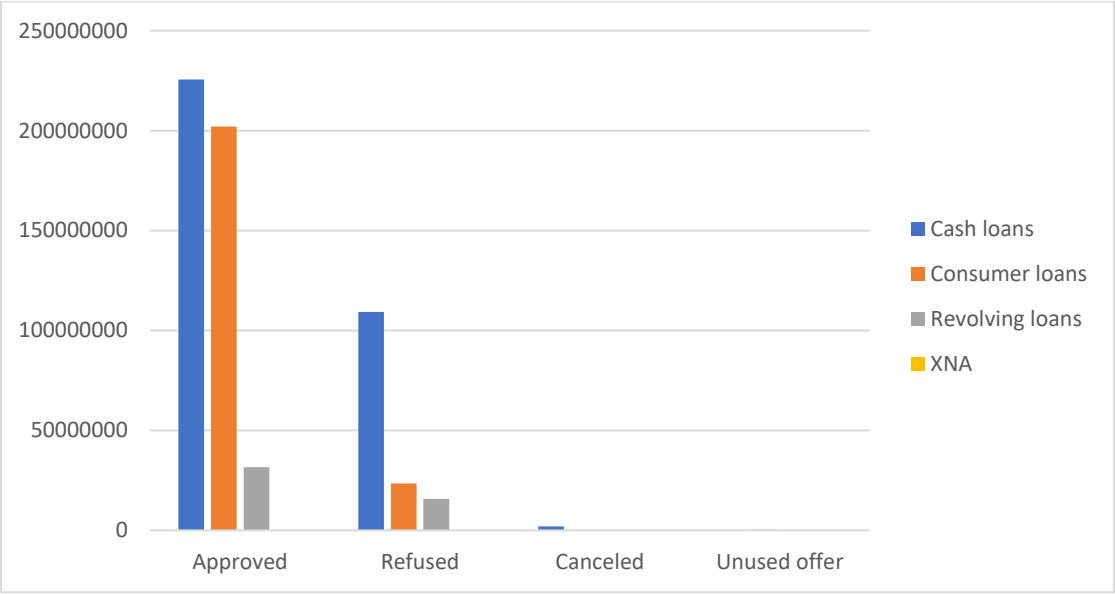


- 90.55% Loans applied for are Cash loans.

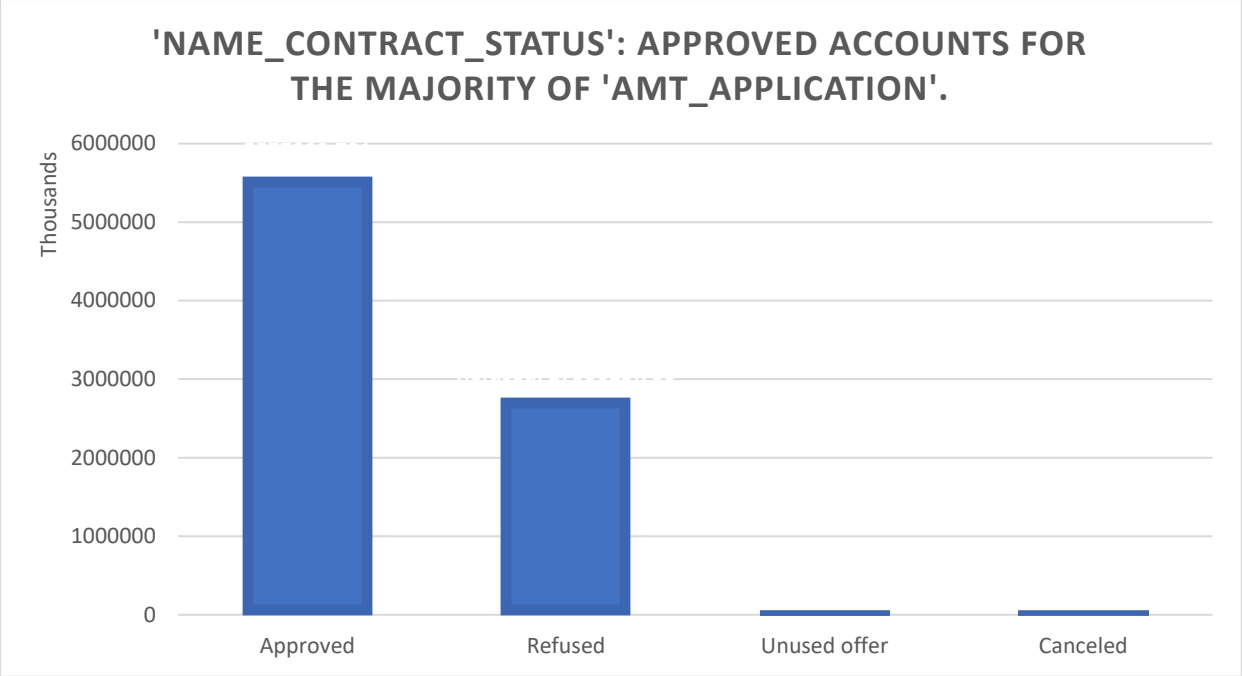


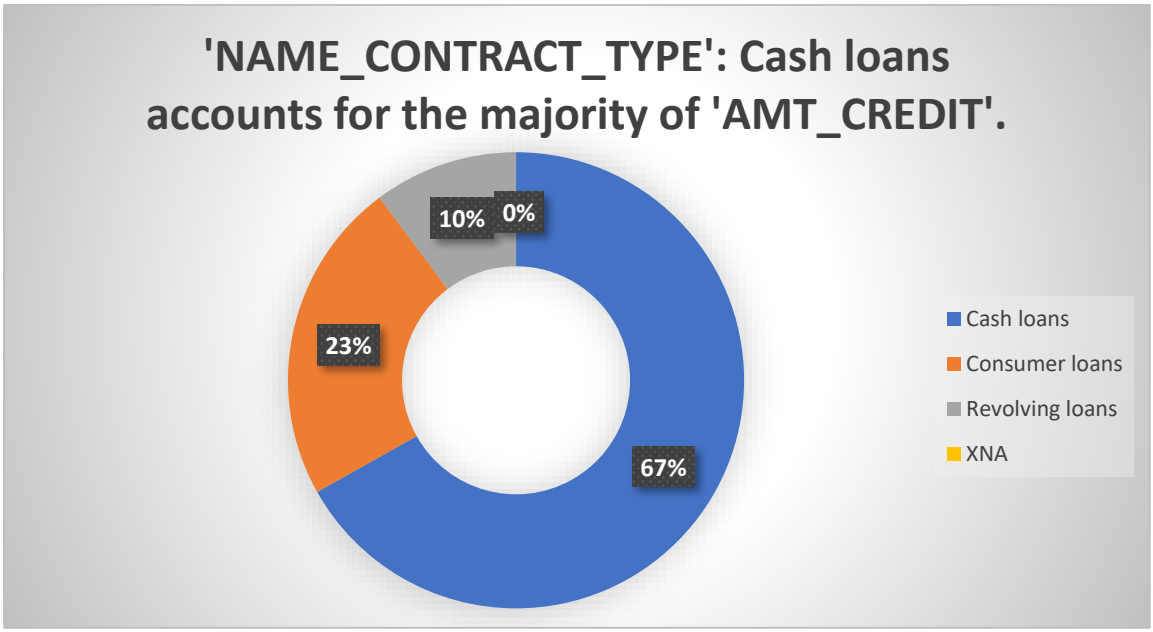
- Clients applying more for Loans belong to business Entity Type 3.
-

Previous data:-

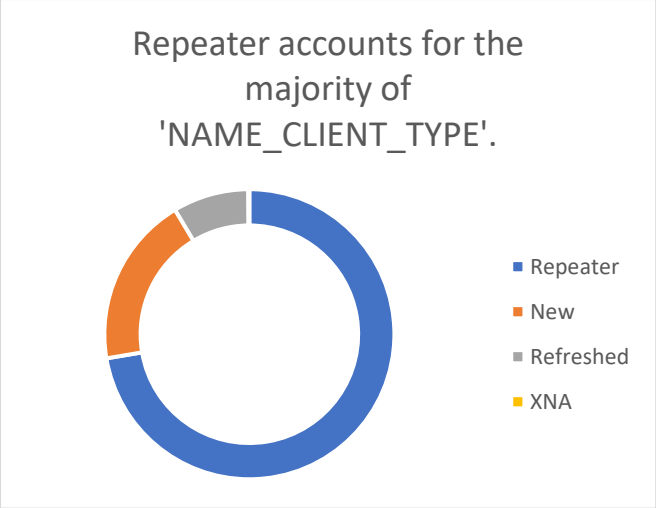


- The amount of loans which is approved is a Cash Loan and Consumer Loans.



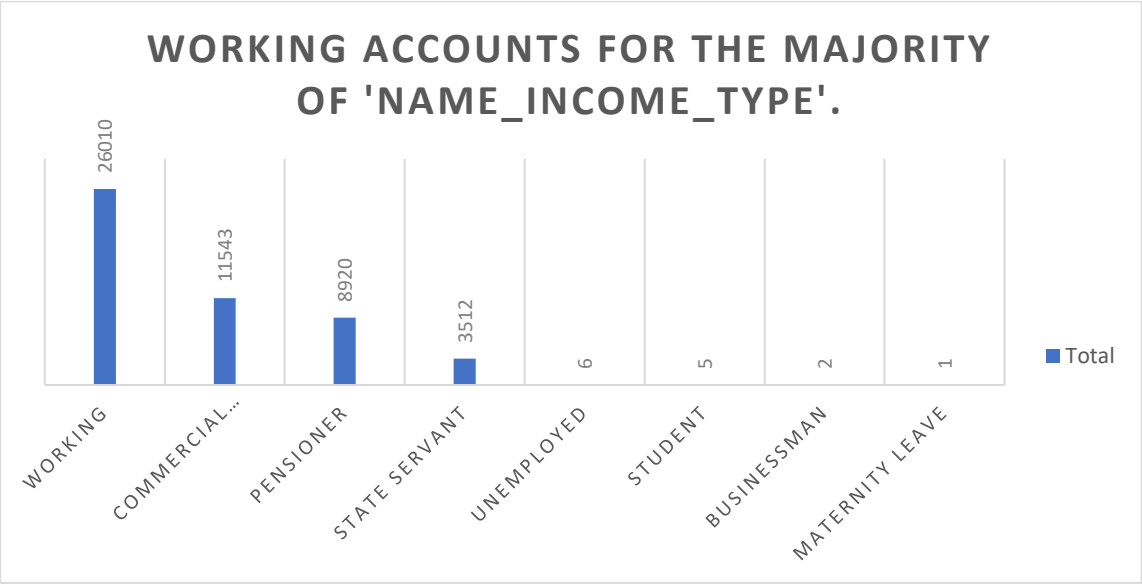


- Cash Loans are majorly approved than other Loans in AMT_CREDIT

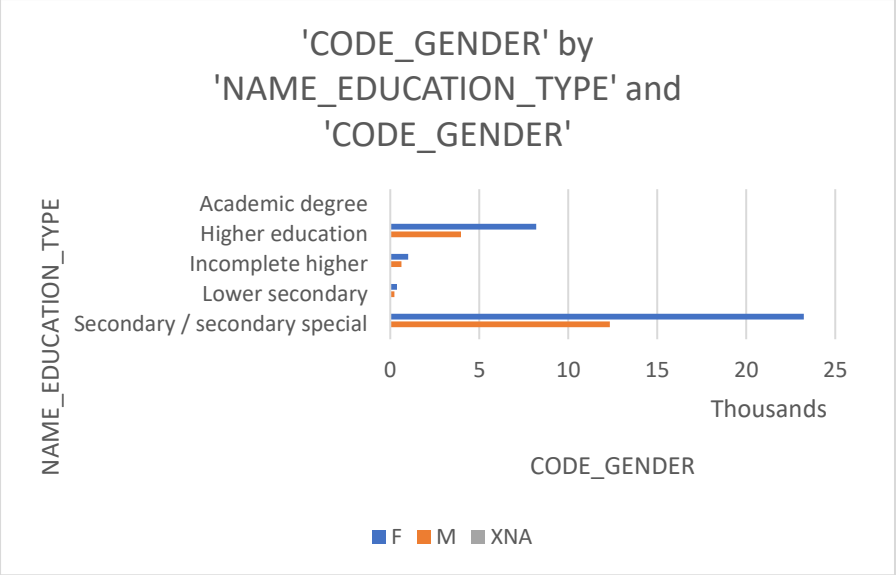


- Majority of the Clients who have already taken loans earlier(Repeater) have again applied for Loans, followed by new ones.

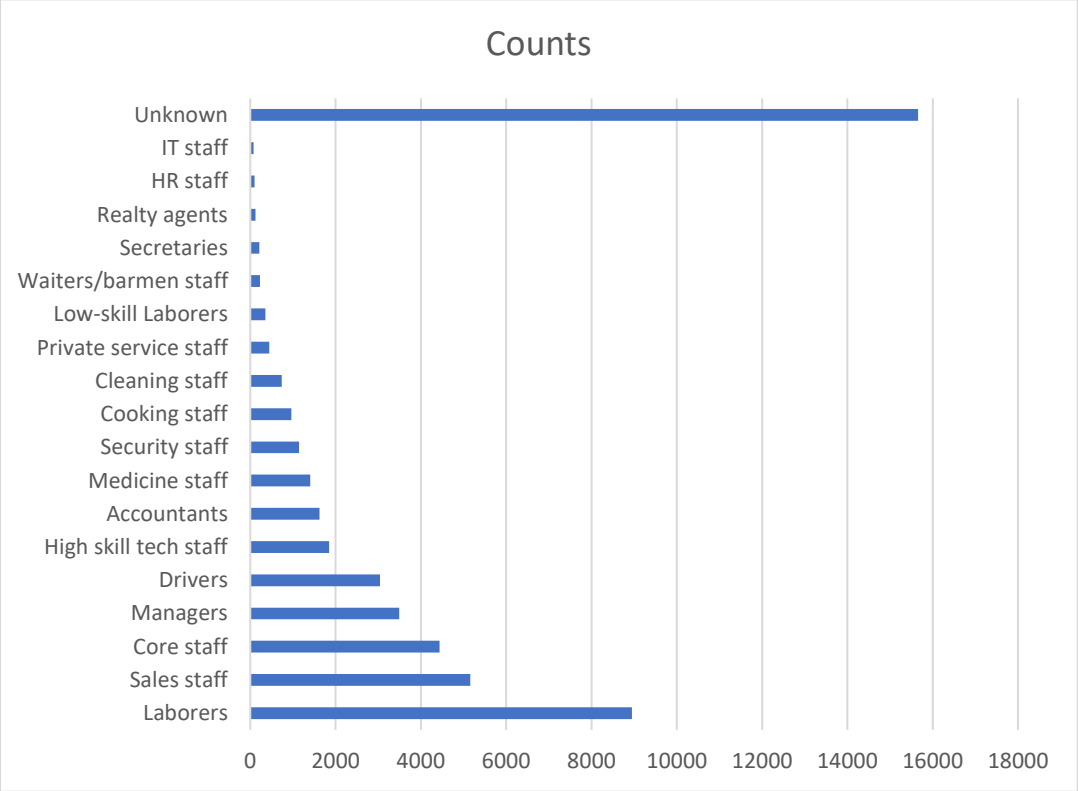
Multi-Variate Analysis



- 26010 Working Professional are the one applying more for cash loans followed by Commercials and then Pensioners.



- Females who have complete their secondary education have applied for most of the cash loans. Followed by males completed their secondary education. Clients completing higher education are on rank 2.



- It is observed that clients whose occupation is unknown have the major counts for applying loans followed by Laborers. So we interpret that the clients who did not mention their Occupation or Laborers are either completed their Secondary or Higher Education.

Descriptive Statistics:-

Descriptive Statistics	AMT_INCOME_TOTAL	AMT_CREDIT	CNT_Children	AMT_ANNUITY	AMT_GOODS_PRICE	Days_Employed
Mean	170767.5905	599700.5815	0.419848397	27107.37739	538992.3491	184.0008887
Median	145800	514777.5	0	24939	450000	6.071232877
Mode	135000	450000	0	9000	450000	1000.665753
Standard Deviation	531813.7768	402411.4096	0.724031307	14562.65317	369717.1252	380.7027603
Variance	282825893198.07	161934942605.41	0.524221333	212070867.3	136690752694.68	144934.5917
Minimum	25650	45000	0	2052	45000	0
Maximum	117000000	4050000	11	258025.5	4050000	1000.665753
Range	116974350	4005000	11	255973.5	4005000	1000.665753
Sum	8538208758	29984429376	20992	1355341762	26949078465	9199860.436
Count	49999	49999	49999	49999	49999	49999
Quartile 1	112500	270000	0	16456.5	238500	2.556164384
Quartile 3	202500	808650	1	34596	679500	15.66575342
IQR	90000	538650	1	18139.5	441000	13.10958904

- It is observed that, in most of the columns Mean is greater than Median which means the data is Right skewed or positively skewed. It is also observed that there is a huge difference between Quartile3(75%) and Maximum, which means there are Outliers are present in the data.

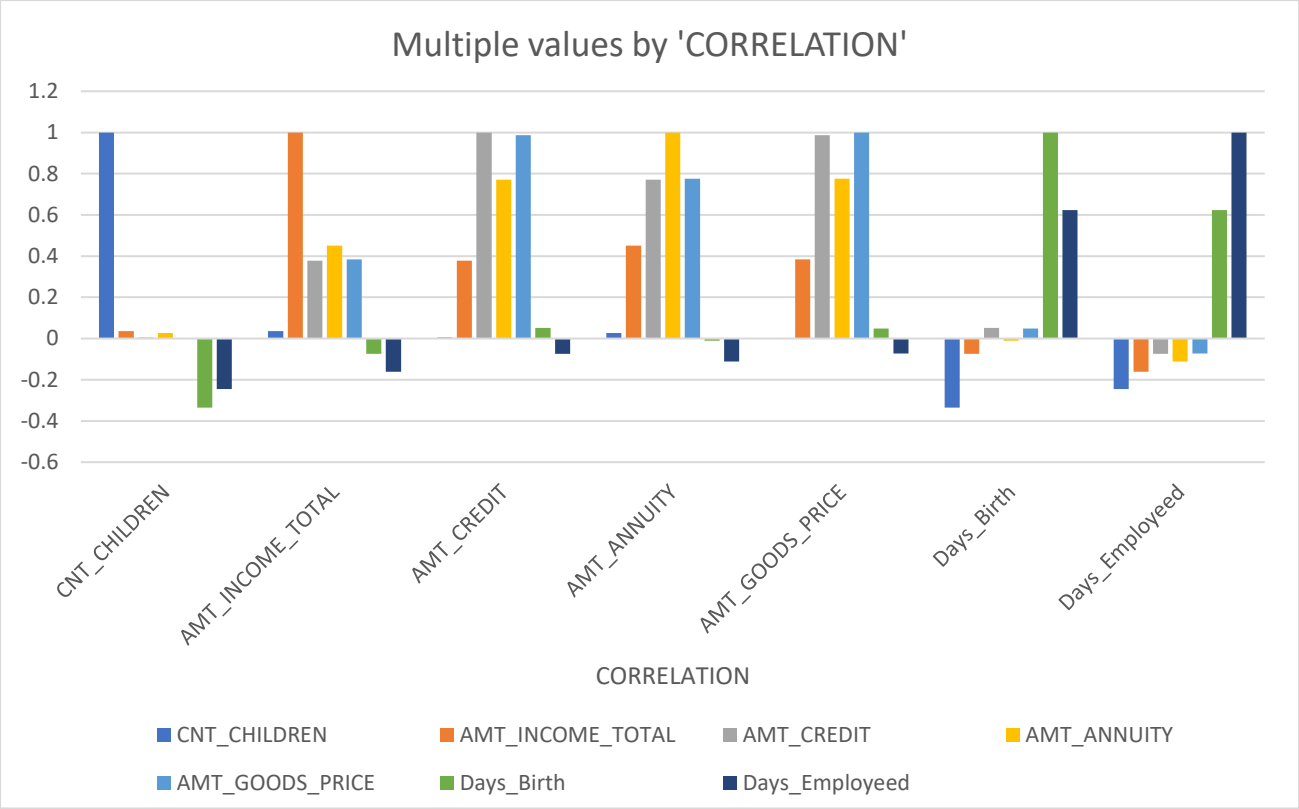
Task E.

Identify Top Correlations for Different Scenarios:

Understanding the correlation between variables and the target variable can provide insights into strong indicators of loan default.

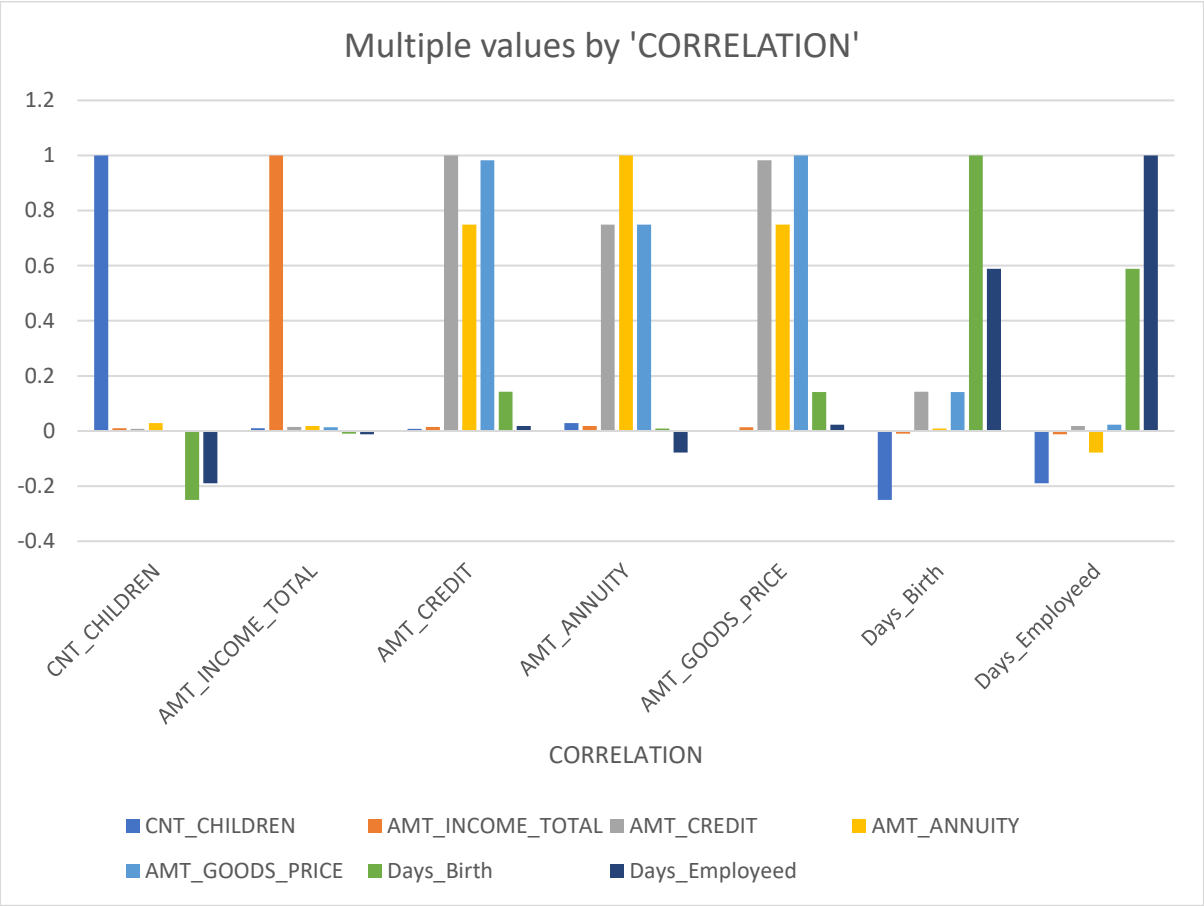
- Correlation of all the features for the Target 0:-

CORRELATION	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRICE	Days_Birth	Days_Employeed
CNT_CHILDREN	1	0.0363197	0.0057055	0.0263821	0.0015181	-0.3358763	-0.2455215
AMT_INCOME_TOTAL	0.0363197	1	0.3779658	0.4511356	0.3845759	-0.0737694	-0.1616809
AMT_CREDIT	0.0057055	0.3779658	1	0.7707718	0.9869998	0.0510842	-0.0747334
AMT_ANNUITY	0.0263821	0.4511356	0.7707718	1	0.7758346	-0.0099154	-0.1112939
AMT_GOODS_PRICE	0.0015181	0.3845759	0.9869998	0.7758346	1	0.0487733	-0.0724465
Days_Birth	-0.3358763	-0.0737694	0.0510842	-0.0099154	0.0487733	1	0.6234747
Days_Employeed	-0.2455215	-0.1616809	-0.0747334	-0.1112939	-0.0724465	0.6234747	1



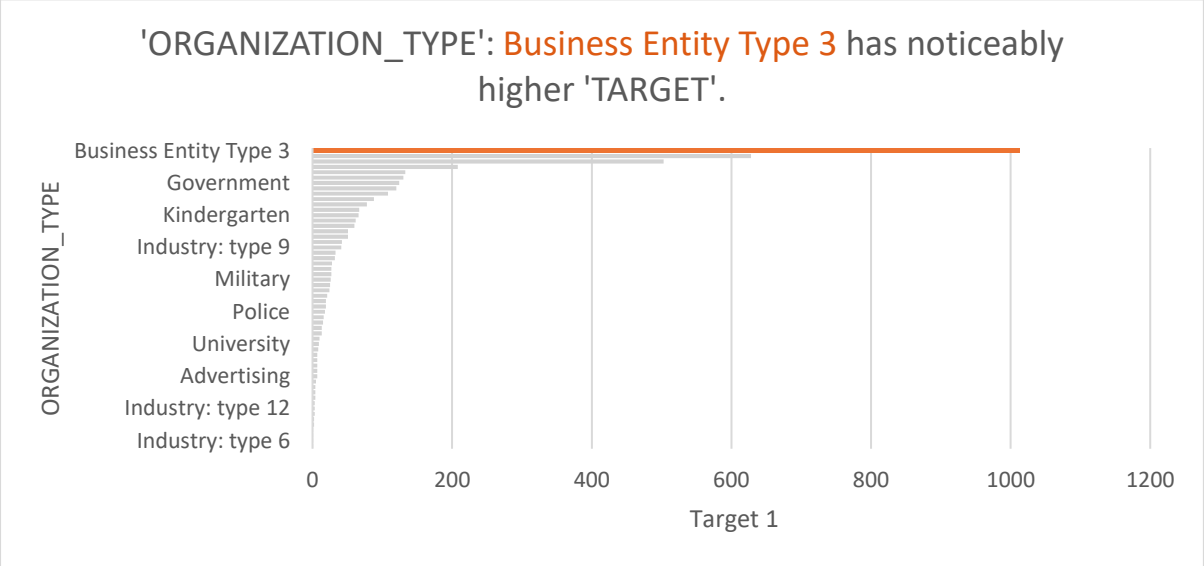
➢ Correlation of all the features for the Target 1:-

CORRELATION	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRICE	Days_Birth	Days_Employeed
CNT_CHILDREN	1	0.0101102	0.0076019	0.029173	-0.0010797	-0.2496732	-0.1897732
AMT_INCOME_TOTAL	0.0101102	1	0.0152714	0.0180046	0.0132695	-0.0090337	-0.0117587
AMT_CREDIT	0.0076019	0.0152714	1	0.7496652	0.982268	0.142506	0.0187822
AMT_ANNUITY	0.029173	0.0180046	0.7496652	1	0.749504	0.0087517	-0.0781139
AMT_GOODS_PRICE	-0.0010797	0.0132695	0.982268	0.749504	1	0.1410059	0.0231816
Days_Birth	-0.2496732	-0.0090337	0.142506	0.0087517	0.1410059	1	0.5882428
Days_Employeed	-0.1897732	-0.0117587	0.0187822	-0.0781139	0.0231816	0.5882428	1

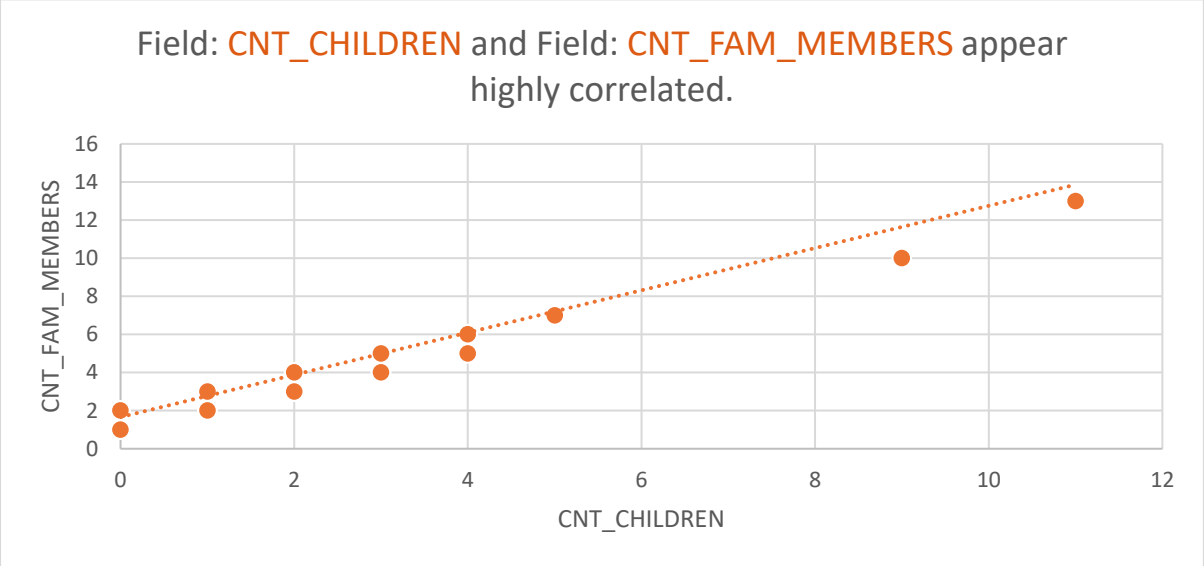


Findings:- AMT_CREDIT is is highly correlated to AMT_ANNUITY, AMT_GOODS.

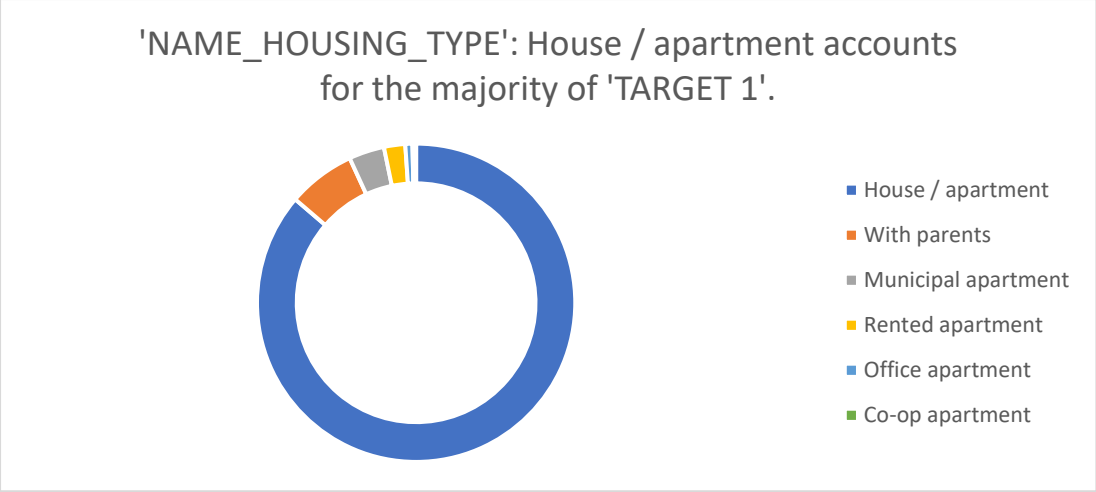
Other Analysis and Interpretations of Defaulters:-



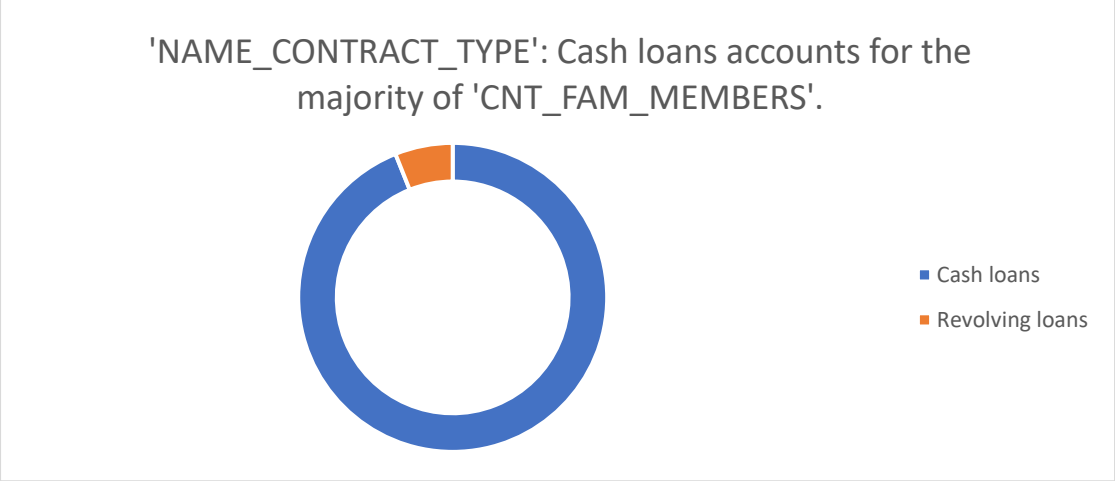
- So the clients belonging to Business Entity Type 3 have more number of children , which can be one of the reason for defaulting.



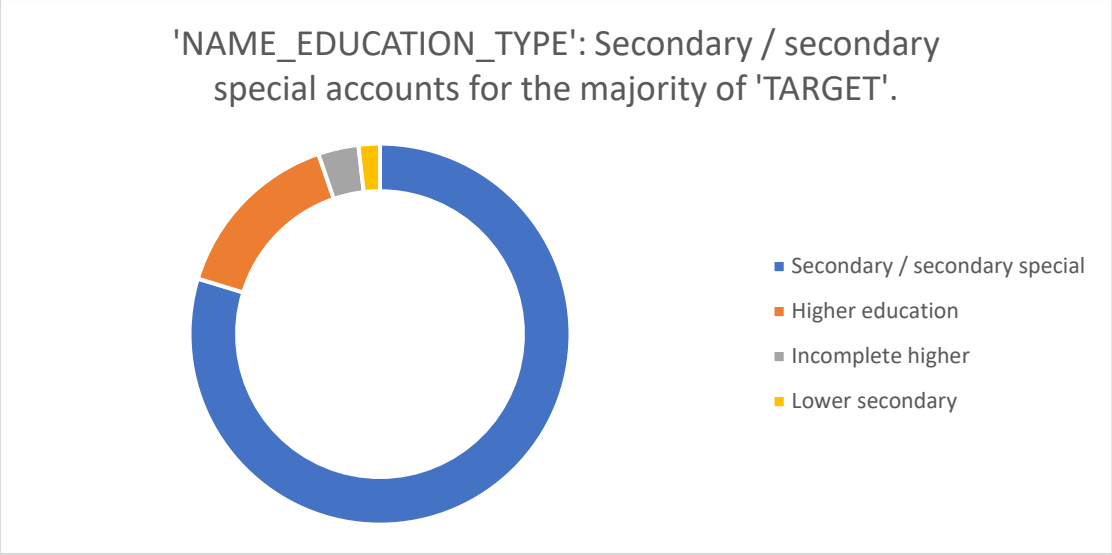
- CNT_Children and CNT_FAM_Members are highly correlated. This means the more number of children or Family members , higher is the chance of defaulting



- It is observed that clients belong to House /Apartments have defaulted a lot as they can stay on rent or have to pay huge EMIs.



- Clients have higher number of of CNT_FAM_Members have taken Cash Loans and defaulted.



- The clients who have completed their secondary Education/special have defaulted than others.
- Majority of clients live in House/Apartment.
- Client with SK_ID 114967 Income 11,70,00,000 is a working professional Labourer but falls in the target 1. Observation says he owns a Realty.
- The lower secondary education category have the largest rate of not retuning the loans.
- Most of the clients stated that the years employed are 1001 , which is not possible.
- Cancelled and unused Loans are very rare.
- Most of the clients applied for Loans are from Business Entity type 3 Organization.
- CNT_FAM_MEMBERS are highly correlated to CNT_CHILDREN. Hence the clients with higher number of children or higher number of family members have high chances of non-repayment of Loans.
- Severe drop in correlation between AMT_INCOME and AMT_CREDIT.
- Avoid young people (20-40 yrs.) as they have higher probability of defaulting.

Conclusion:

This project demonstrates effective techniques for handling large datasets, particularly through the application of exploratory data analysis (EDA). When dealing with extensive datasets, it's crucial to streamline the analysis by selecting only the most relevant columns. Exploring correlations between columns can significantly aid in this process, optimizing time and resources by identifying key variables for analysis. Additionally, this project contributes to a better understanding of terminology commonly used in the banking domain, enhancing knowledge and proficiency in financial data analysis. Overall, leveraging EDA in conjunction with strategic column selection can streamline analysis workflows and yield valuable insights from large datasets, particularly in the context of banking and finance.