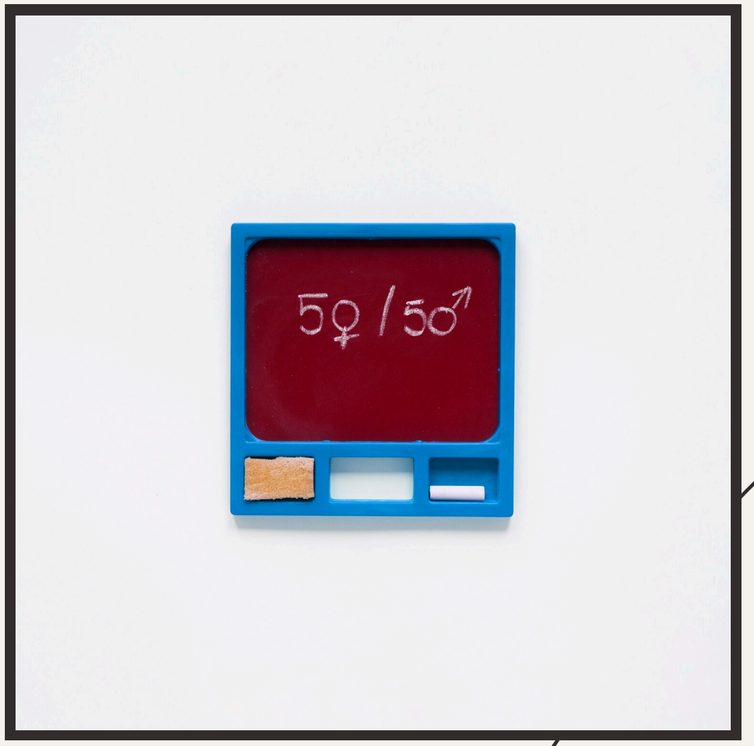
Analyzing Credit EDA: A Case Study by Abhishek Singh Chauhan

Introduction

In this **case study**, we will analyze the Credit EDA dataset to understand the **financial trends** and factors affecting credit risk. We will explore the **data insights** and draw meaningful conclusions.

Understanding Credit EDA

The Credit EDA dataset provides a comprehensive view of **credit-related data**, including customer information, credit history, and risk factors. We will delve into the **data attributes** and their implications on credit analysis.



Exploratory Data Analysis

Through **exploratory data analysis**, we will uncover patterns, anomalies, and correlations in the dataset. We will utilize **statistical techniques** and visualizations to gain insights into credit behavior.

Risk Assessment and Prediction

Using **machine learning models**, we will assess credit risk and predict customer behavior. We will evaluate **predictive features** and their impact on credit outcomes.

Interpreting Results

We will interpret the **model outcomes** and evaluate their implications on credit decision-making. We will discuss the **key findings** and their significance in the financial domain.

Conclusion

In conclusion, our analysis of the Credit EDA dataset has provided valuable insights into credit behavior and risk assessment. We have identified **critical factors** influencing credit outcomes and highlighted the significance of **data-driven analysis** in financial decision-making.

Lets Start with Analyzing Data sets

We have two dataSets. First One is current application Data and the second one is previous Application Data.

Load the Data Frame.

First we need to load the dataset and check for how many rows and Columns are in the data Frame.

Application DataSet.

Check for Number of Rows and Columns in Data Set

```
[8]: application_data.shape

[8]: (307511, 122)
```

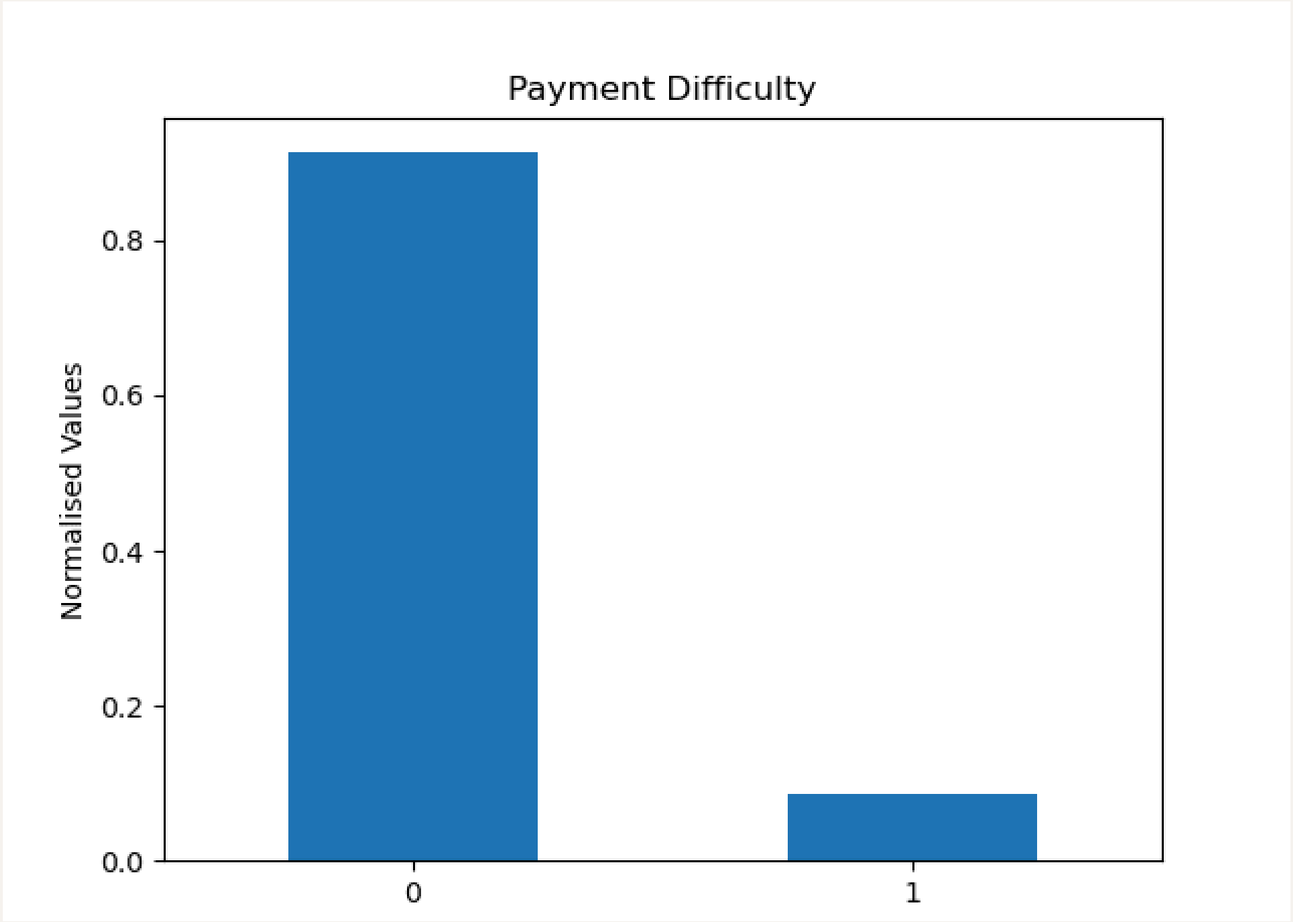
There are 307511 rows and 122 columns in the application dataframe

Computing & Imputing Missing Values.

In the first step of data analysis, first we need to clean the dataset by computing and imputing null and empty values by default or mean, median and mode.

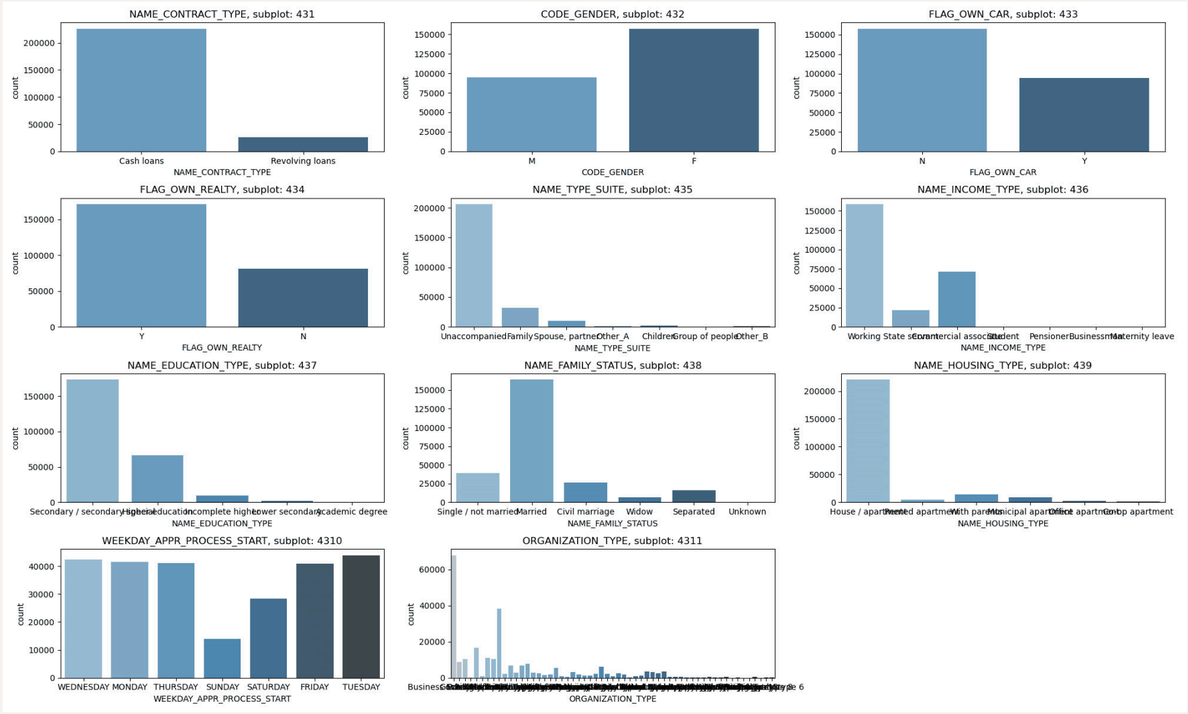
Check for Imbalance Ratio

After cleaning and imputing the dataset, first we need to check for Imbalance Ratio. So that we can know ratio of the person having payment difficulty.



1 in every 10 applicant has payment difficulty.

Check For Some Important Variables and Categories.



Cash loans offered are more than revolving loans, at 90%. 65% of females have taken loans in comparison to 34% males. This is very interesting and needs to be studied further. 65% of applicants don't own cars. 69% of applicants own living quarters. 81% of applicants came accompanied for loan application. While most applicants are working class, 18% are pensioners. 71% have secondary education. 63% are married. 31% have not mentioned their occupation type.

Check for Number of People own Car.

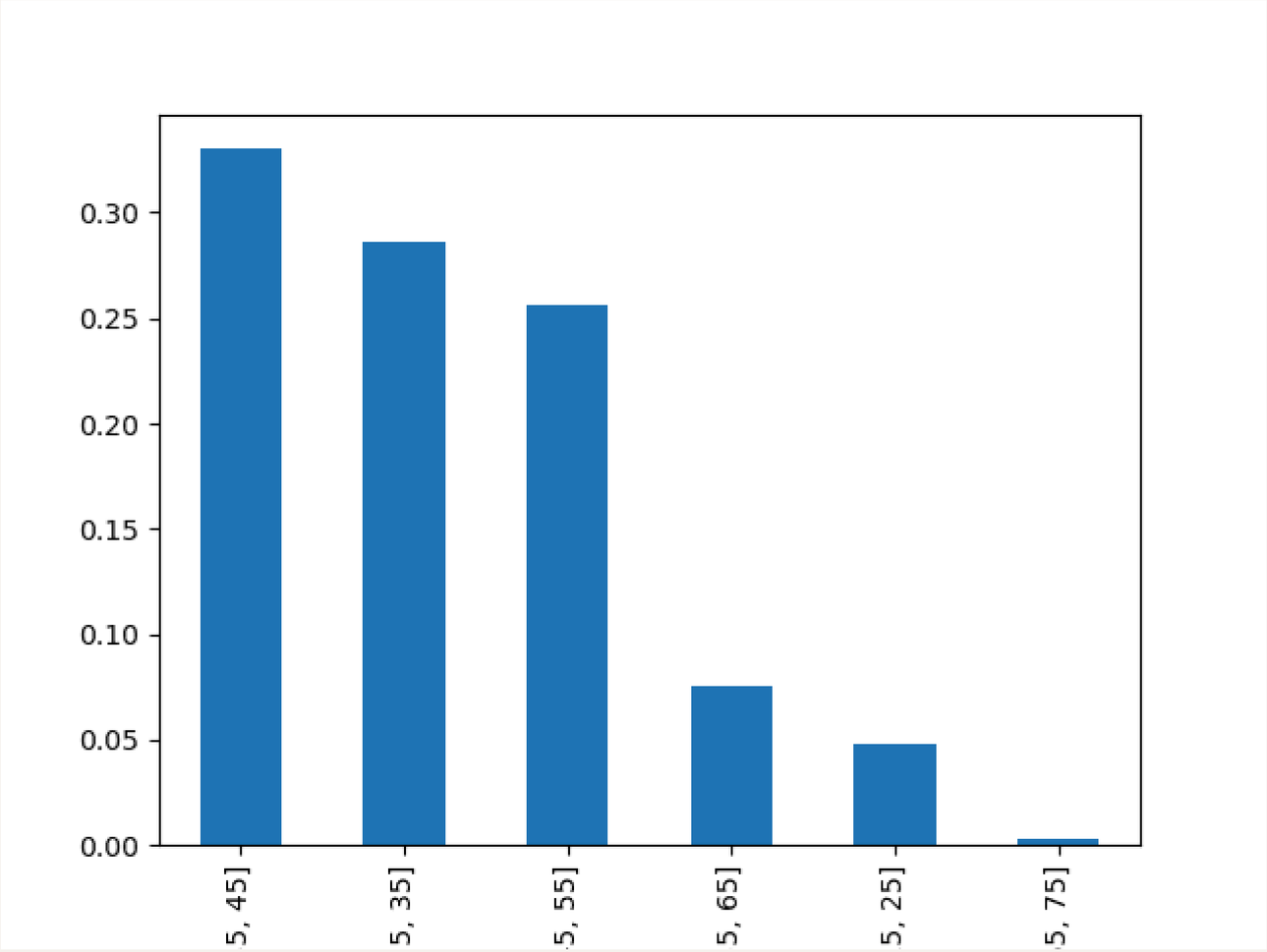
```
]: #Count of people who have own car
   application_data['FLAG_OWN_CAR'].value_counts()
```

```
]: FLAG_OWN_CAR
   N    157719
   Y     94418
   Name: count, dtype: int64
```

There are 1557719 people who dont own car and 94418 people own car.

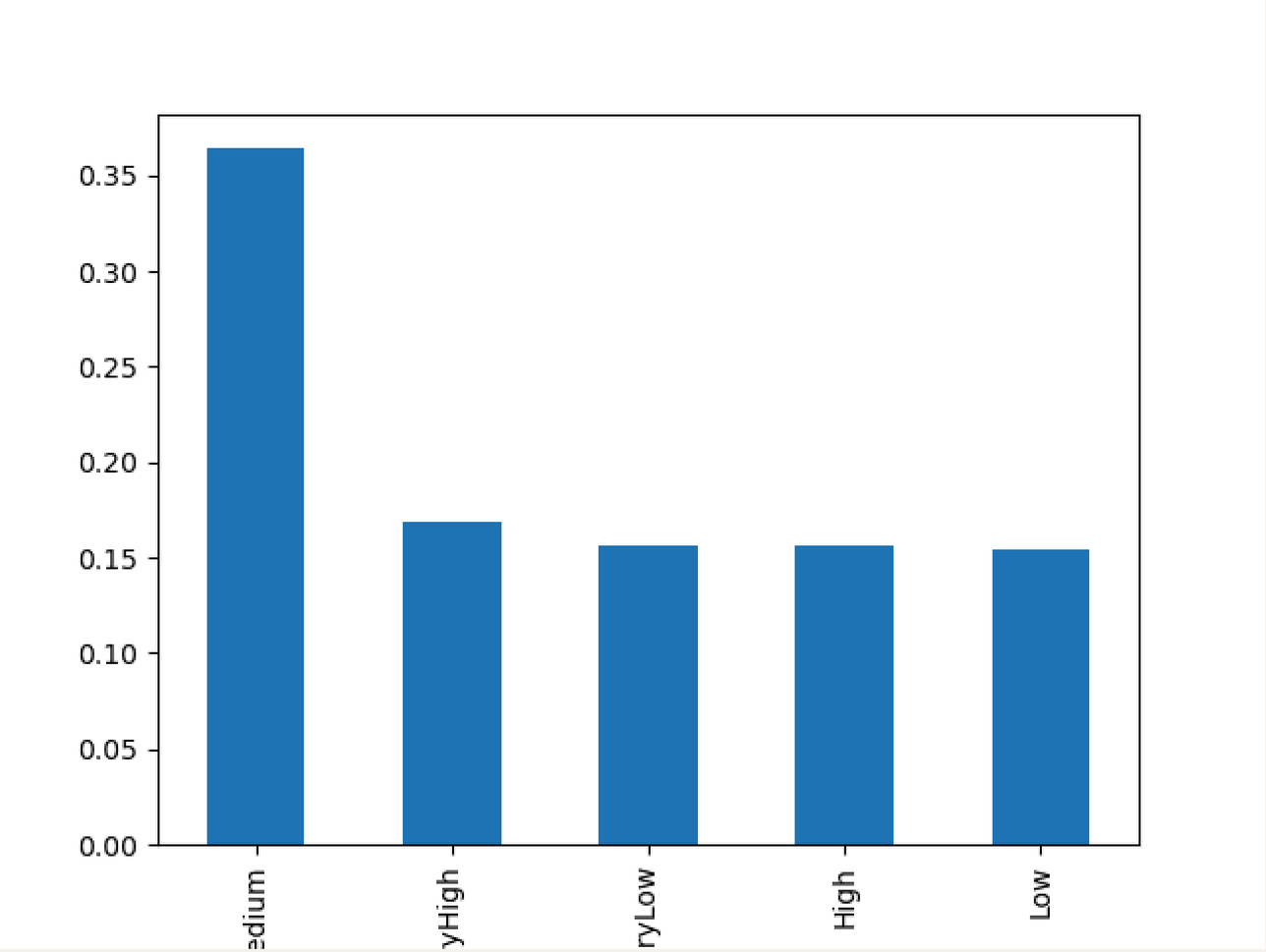
- N means No.
- Y means Yes

Analyzing for different Age Groups.



35-45 Age group is the largest Group of Age applying for loans. This may be attributed to consumerism aspect at that age.

Check for Income Group

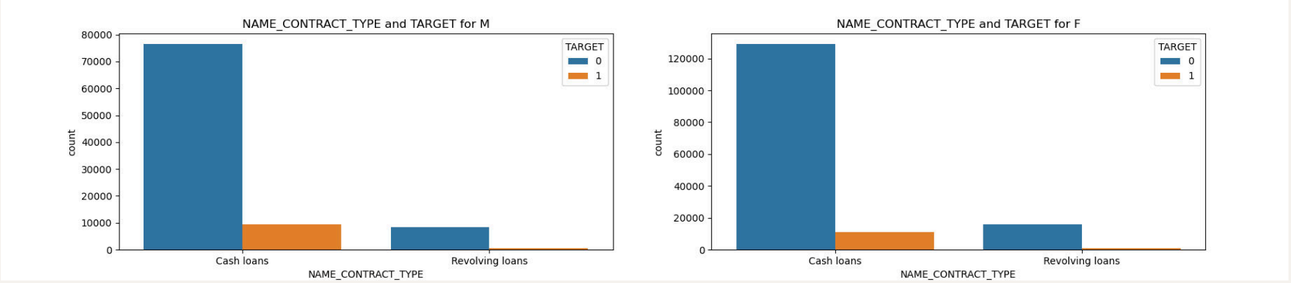


Medium Income group is the largest Group applying for loans.

Univariate Analysis on Categorical Nominal to analyse both data frames

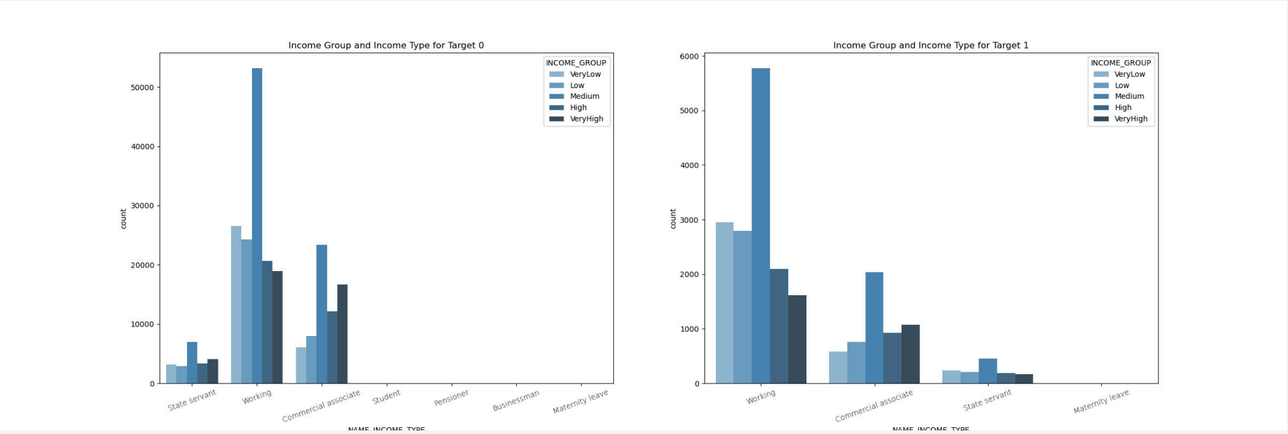
- NAME_CONTRACT TYPE- Cash Loans are large part of the company's portfolio. For Target 0 - 85% and almost 95% for Target-1
- NAME_TYPE_SUIT - 80-90% in Target 0 and Target 1 are applying for loan Unaccompanied. Indicating, this is not a parameter that can influence payment default.
- NAME_INCOME_TYPE - 50% working in case of Target 0 and 60% in case of Target 1 are working income types.
- NAME_EDUCATION_TYPE - In both Target 0 and 1, applicants with Secondary Education has applied for loans more than others.90% of defaulting payments are from applicants with secondary income. Needs further analysis
- NAME_FAMILY_STATUS - Married applicants - almost 60% have defaulted on payments
- NAME_HOSUING_TYPE -85-90% in Target 0 and Target 1 applicants are staying in "House/apartment". Indicating, this is not a parameter that can influence payment default.
- OCCUPATION_TYPE - Labourers, sales staff, core staff, drivers constitute of 50% of defaulters. Labourers is the highest percentage of applicants too.
- ORGANIZATION_TYPE - Business ENTITY TYPE 3 AND SELF EMPLOYED add upto 40% defaulters. The highest % of loan takers are also this category.

Gender wise Analization



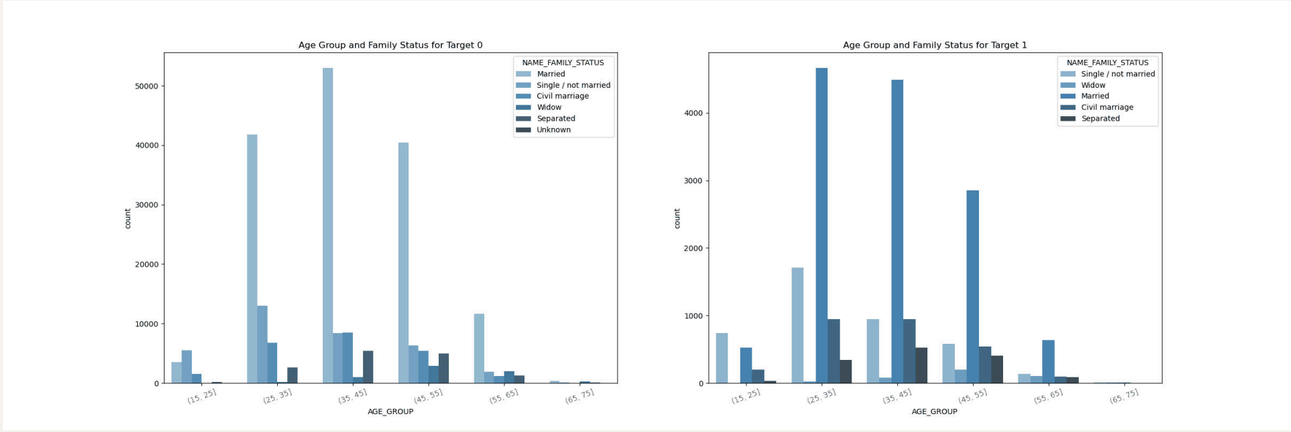
Male applicants are defaulting more that female applicants

Income Group And Income Type



Medium income group with income type has almost 1 in 12 defaults. Higher than the avg 1 in 11 defaults.

Age Group & Family Status



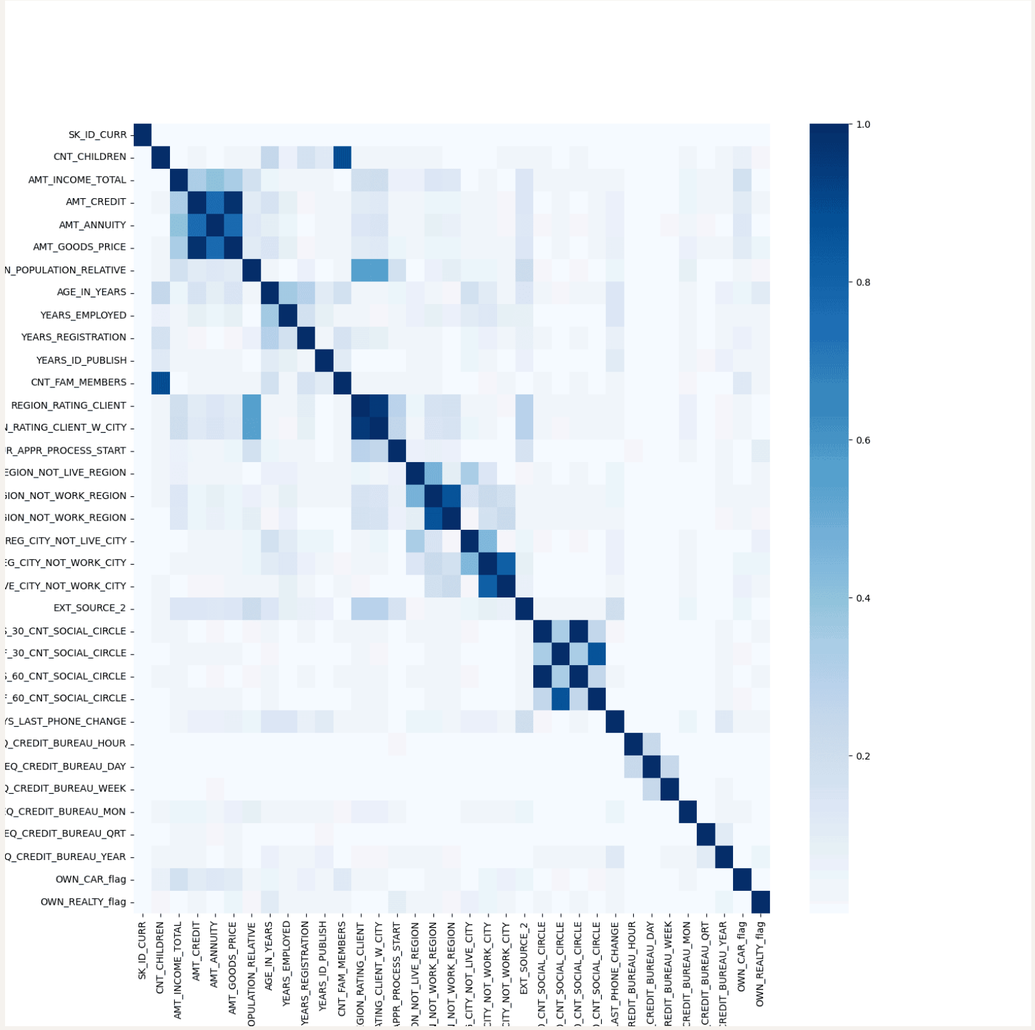
Married applicant in the age group 25-35 and 35-45 is the largest group of applicant with payment difficulties.

Top Correlations.

	Column1	Column2	Correlation
862	OBS_60_CNT_SOCIAL_CIRCLE	OBS_30_CNT_SOCIAL_CIRCLE	0.998289
178	AMT_GOODS_PRICE	AMT_CREDIT	0.982536
467	REGION_RATING_CLIENT_W_CITY	REGION_RATING_CLIENT	0.956531
386	CNT_FAM_MEMBERS	CNT_CHILDREN	0.893829
898	DEF_60_CNT_SOCIAL_CIRCLE	DEF_30_CNT_SOCIAL_CIRCLE	0.867983
611	LIVE_REGION_NOT_WORK_REGION	REG_REGION_NOT_WORK_REGION	0.846872
719	LIVE_CITY_NOT_WORK_CITY	REG_CITY_NOT_WORK_CITY	0.768247
179	AMT_GOODS_PRICE	AMT_ANNUITY	0.749591
143	AMT_ANNUITY	AMT_CREDIT	0.748708
575	REG_REGION_NOT_WORK_REGION	REG_REGION_NOT_LIVE_REGION	0.506747

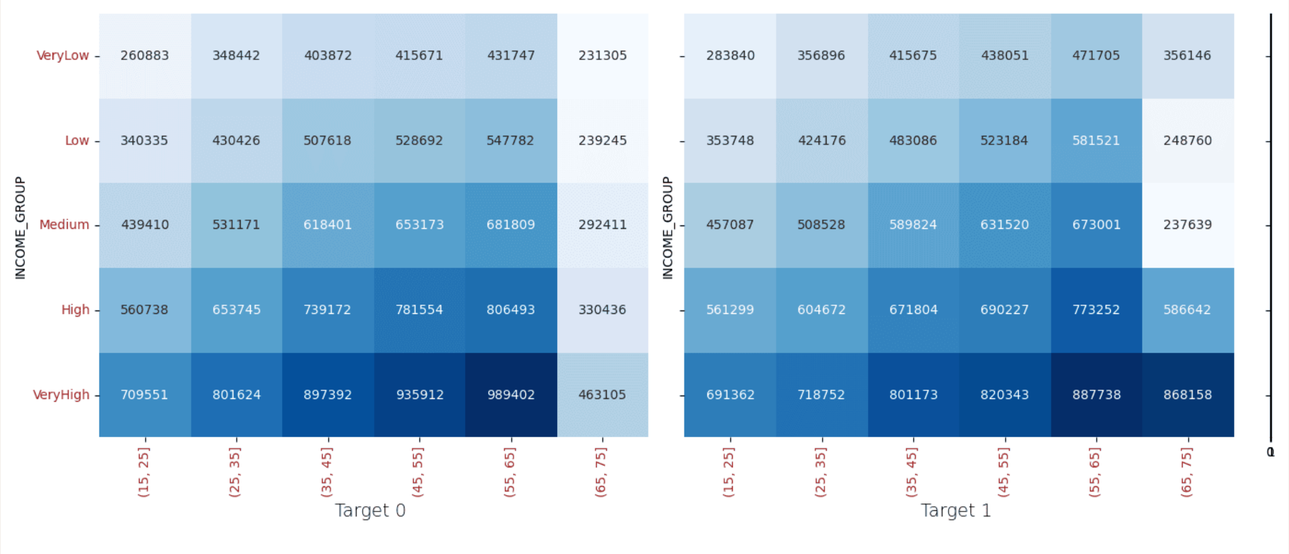
We obesrve that the TOP 10 correlation columns are same for Target 0 and Target 1 dataframes

Correlation by Heatmap Plot



HeatMap says the same story of correlation in pictoral manner

Analysing relationship of AMT_CREDIT with AGE GROUP and income group



Age Group 55-65 in Very High income group has high amount credit. As explained above, this could result as loss in loan book

Summary on Application Dataframe

The comprehensive analysis of the application DataFrame involved examining various categorical and continuous variables to understand their relationships and impacts on loan defaults. Below is a detailed summary of the findings from the analysis:

Age Group and Income Group Analysis

- **Age Group:** The 35-45 age group is the most represented among non-defaulters (Target 0). However, the 25-35 age group has a higher proportion of defaulters (Target 1), indicating that younger individuals might have a higher default risk.
- **Income Group:** Both non-defaulters and defaulters are predominantly from the Medium income group. This suggests that income group alone may not be a strong indicator of default risk without considering other factors.

Income Group and AMT_CREDIT Analysis

- **Boxplot Analysis:** The amount credited (AMT_CREDIT) varies across income groups for both non-defaulters and defaulters. High and Very High income groups have higher credit amounts. For defaulters, the average credit amount is lower in higher income groups compared to non-defaulters, indicating that higher-income individuals might manage larger credits better, but when defaults occur, they are substantial.
- **Pivot Table Insights:** The average credit amount (AMT_CREDIT) is higher for higher income groups, especially for non-defaulters. However, defaulters in the Very High income group have significantly high credit amounts, which could impact the financial institution's loan book adversely if these defaults are not managed effectively.

Education Type and AMT_CREDIT Analysis

Boxplot Analysis: Applicants with academic degrees tend to have higher median loan values, though their numbers are few. No strong inferences can be drawn from this due to the small sample size of academic degree holders.

Gender and Income Type Analysis

- **Pivot Table Analysis**: The relationship between AMT_CREDIT, gender, and income type shows that males in the Businessman category have higher average credit amounts compared to females. For defaulters, females on maternity leave have high credit amounts, highlighting a potential risk area.
- **Countplot Analysis**: Male applicants, though fewer in number, have a higher default ratio compared to female applicants. This suggests a gender-based differential in default risk that requires further investigation.

Age Group and Family Status Analysis

Countplot Analysis: The relationship between age group and family status shows different patterns for non-defaulters and defaulters. For instance, older age groups (55-65) in Very High income groups have higher credit amounts, which could lead to significant losses if defaults occur.

Correlation Analysis

Top Correlations: Key correlations include:

- OBS_30_CNT_SOCIAL_CIRCLE and OBS_60_CNT_SOCIAL_CIRCLE (very high correlation), indicating that a client's social surroundings are closely related.
- AMT_GOODS_PRICE and AMT_CREDIT (high correlation), suggesting that the goods price is strongly tied to the credit amount.
- CNT_FAM_MEMBERS and CNT_CHILDREN (high correlation), showing a logical connection between family size and number of children.

Heatmap Analysis: The heatmap visually confirmed these correlations, highlighting significant relationships among various numerical variables.

Bivariate Analysis on Continuous Variables

- Scatterplots: Scatterplots of the top correlated variables revealed trends and potential risk areas:
- OBS_30_CNT_SOCIAL_CIRCLE and OBS_60_CNT_SOCIAL_CIRCLE are strongly correlated for both targets, with steeper trends for defaulters.
- AMT_CREDIT and AMT_GOODS_PRICE showed proportional increases for non-defaulters, while defaulters had less proportional increases, indicating potential repayment issues.
- DEF_30_CNT_SOCIAL_CIRCLE trends showed up, but the data was sparse for defaulters.

Implications and Recommendations

Enhanced Risk Assessment:

- The higher default rates in younger age groups (25-35) and high credit amounts in older high-income groups (55-65) suggest the need for targeted risk assessment strategies.
- Implementing stricter assessment criteria for high credit amounts, especially for older and higher-income applicants, could mitigate default risks.

Gender-Specific Policies:

- The higher default ratio among male applicants indicates the necessity for gender-specific credit policies and risk management strategies.
- Female applicants on maternity leave with high credit amounts should be monitored closely to prevent defaults.

Social Circle Impact:

The strong correlation between social circle observations (OBS_30_CNT_SOCIAL_CIRCLE and OBS_60_CNT_SOCIAL_CIRCLE) and defaults suggests that social surroundings should be a key factor in credit risk assessment.

Tailored Loan Products:

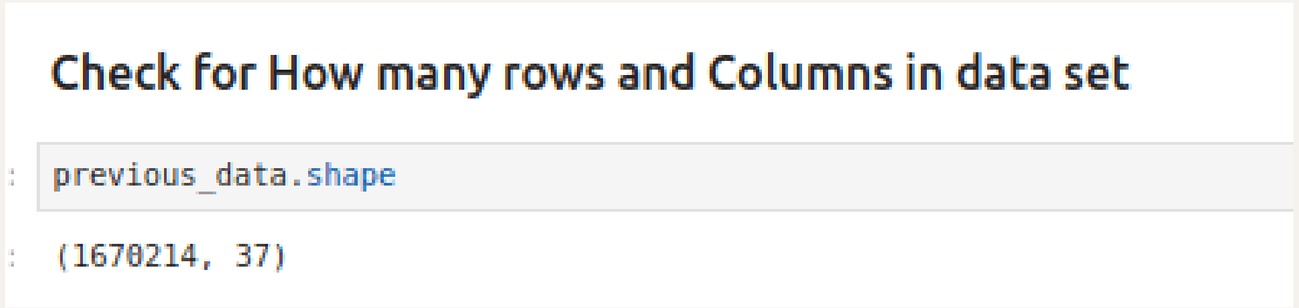
Based on the varied credit amounts across different education types and income groups, tailored loan products that cater to specific segments (e.g., academic degree holders, high-income older individuals) could be developed to better manage risks.

Continuous Monitoring and Early Intervention:

Continuous monitoring of high-risk segments, such as the 55-65 age group with high credit amounts, and early intervention strategies (e.g., financial counseling, adjusted repayment plans) could help reduce default rates and protect the financial institution’s loan book.

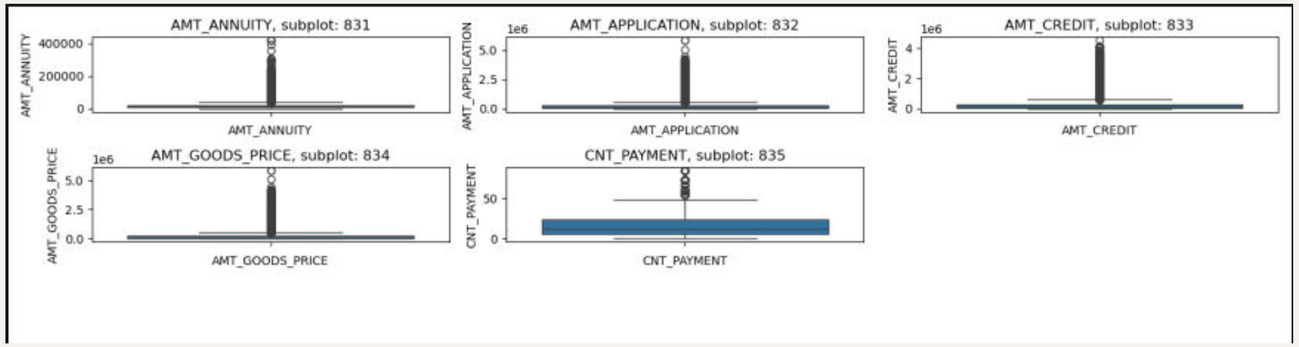
Now Analyse the Previous Data

Check for Rows and Columns.



There are 1670214 rows and 37 columns in this dataset.

Analysing Continuos Variable



Continuous Variables seem to have high percentage of outliers. Checking distribution

Univariate Analysis of Categorical Variables

1. This dataframe has a differnt type of loan called Consumer Loan, which was not there in Application data frame. 55% of loans are consumer loans. 37% cash loans and rest revolving
2. Approved loans are 79% and refused, cancelled, unused - rest. **IMBALANCE IN DATA**
3. 67% are repeaters. NAME_CLIENT_TYPE also has some null values showing as XNA
4. 55% of the applicants have taken loan for POS purchase.
5. Name seller industry has 37% XNA values, Consumer electronics is next highest category at 30%

Univariate Analysis on Numerical Variable

Overview:

The univariate analysis of the numerical variables in the dataset reveals significant insights, particularly highlighting the presence of outliers in the continuous variables. Below is a summary of the descriptive statistics and graphical analysis.

SK_ID_PREV and SK_ID_CURR

Unique identifiers for previous and current applications, respectively. Both have a large range of values with a mean of approximately 1.92 million for SK_ID_PREV and 278,357 for SK_ID_CURR.

Loan Amounts:

- AMT_ANNUITY: Mean of 15,955 with a standard deviation of 14,782
- AMT_APPLICATION: Mean of 175,233 with a standard deviation of 292,780.
- AMT_CREDIT: Mean of 196,114 with a standard deviation of 318,575.
- AMT_DOWN_PAYMENT: Mean of 6,697 with a standard deviation of 20,921.
- AMT_GOODS_PRICE: Mean of 227,847 with a standard deviation of 315,397.

Interest Rates:

- RATE_DOWN_PAYMENT: Mean of 0.0796 with a standard deviation of 0.1078.
- RATE_INTEREST_PRIMARY: Mean of 0.1884 with a standard deviation of 0.0877.
- RATE_INTEREST_PRIVILEGED: Mean of 0.7735 with a standard deviation of 0.1009.

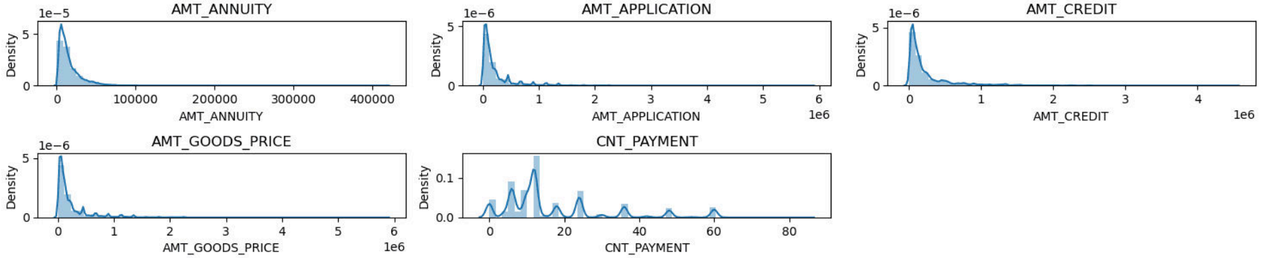
Payment Counts and Days:

- CNT_PAYMENT: Mean of 16.05 with a standard deviation of 14.57.
- DAYS_FIRST_DRAWING, DAYS_FIRST_DUE, DAYS_LAST_DUE_1ST_VERSION, DAYS_LAST_DUE, DAYS_TERMINATION: All show a wide range of values with significant standard deviations, reflecting variability in loan repayment schedules.

Insurance Flag:

NFLAG_INSURED_ON_APPROVAL: Mean of 0.3326, indicating that approximately 33.26% of loans were insured upon approval.

Check for outliers.



Continuous Variables seem to have high percentage of outliers. Box plot and distribution both signify the same.

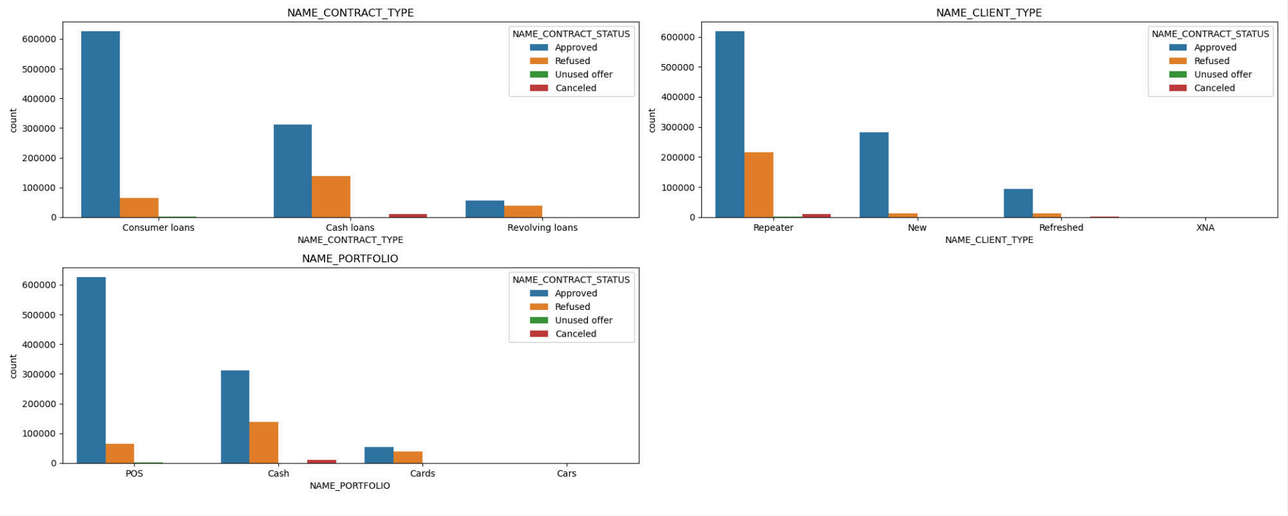
Treatment of Outliers.

```
AMT_ANNUITY : (61959, 16) Percentage of Outlier rows:  4.97
AMT_APPLICATION : (75698, 16) Percentage of Outlier rows:  6.07
AMT_CREDIT : (69957, 16) Percentage of Outlier rows:  5.61
AMT_GOODS_PRICE : (75714, 16) Percentage of Outlier rows:  6.08
CNT_PAYMENT : (103224, 16) Percentage of Outlier rows:  8.28
```

These rows can be deleted considering the size of DF. Not doing so, as it is not in scope of the project

Bivariate Analysis

Categorical Variable



- In approved category, consumer loan has largest no of applicants.
- There seem to be no cancelled loans in cash loan category than consumer loan.
- More cash loans have been refused than consumer loans.
- The bank has more repeaters in all approved, refused, unused, cancelled categories
- POS transactions seem to be consumer loans and similar to point 2 - more cash laons have been refused than POS.

Top Correlations.

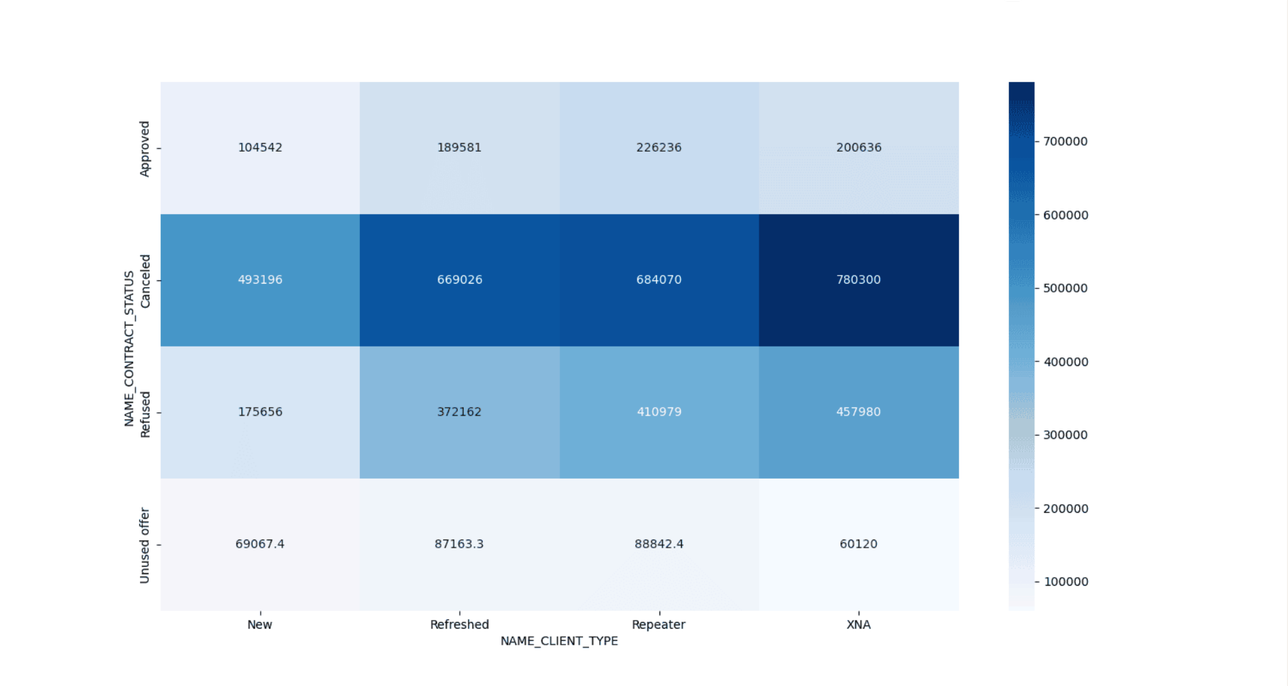
```
print(correlation)
```

```
                AMT_ANNUITY  AMT_APPLICATION  AMT_CREDIT  AMT_GOODS_PRICE  \
AMT_ANNUITY        1.000000         0.820831    0.814884         0.820895
AMT_APPLICATION    0.820831         1.000000    0.992965         0.999883
AMT_CREDIT         0.814884         0.992965    1.000000         0.993028
AMT_GOODS_PRICE    0.820895         0.999883    0.993028         1.000000
CNT_PAYMENT        0.401020         0.672276    0.700323         0.672129

                CNT_PAYMENT
AMT_ANNUITY        0.401020
AMT_APPLICATION    0.672276
AMT_CREDIT         0.700323
AMT_GOODS_PRICE    0.672129
CNT_PAYMENT        1.000000
```

You can see the correlations here.

Multivariate Analysis



- Unused offer application amount is low
- Cancelled application amount is high. The bank may be refusing these possibly as the Debt liabiltiy ratio of consumer must be going high due to the high amount and thus credit defualt risk.
- Repreater's application amount is higeht than the New customers. This may indicate that the bank has more conducive policies/rate of interest etc for repeat applicants

Credit EDA case
Summary

Defaulter Demography.

All the below variables were established in analysis of Application dataframe as leading to default. Checked these against the Approved loans which have defaults, and it proves to be correct -Medium income -25-35 years ols , followed by 35-45 years age group -Male -Unemployed -Labourers, Salesman, Drivers -Business type 3 -Own House - No Other IMPORTANT Factors to be considered -Days last phone number changed - Lower figure points at concern -No of Bureau Hits in last week. Month etc – zero hits is good -Amount income not correspondingly equivalent to Good Bought – Income low and good value high is a concern -Previous applications with Refused, Cancelled, Unused loans also have default which is a matter of concern. This indicates that the financial company had Refused/Cancelled previous application but has approved the current and is facing default on these.

Credible Applications refused

- -Unused applications have lower loan amount. Is this the reason for no usage?
- -Female applicants should be given extra weightage as defaults are lesser.
- -60% of defaulters are Working applicants. This does not mean working applicants must be refused. Proper scrutiny of other parameters needed
- -Previous applications with Refused, Cancelled,Unused loans also have cases where payments are coming on time in current application. This indicates that possibly wrong decisions were done in those cases.

Case Study By Abhishek Singh Chauhan.

For detailed Case study report, refer to the jupytor notebook which is inside the folder.