Bank Loan Case Study

Project Description:

The main aim of this project is to identify patterns that indicate if a customer will have difficulty paying their instalments. This information can be used to make decisions such as denying the loan, reducing the amount of loan, or lending at a higher interest rate to risky applicants. The company wants to understand the key factors behind loan default so it can make better decisions about loan approval.

So our task is to use Exploratory Data Analysis (EDA) to analyze patterns in the data and ensure that capable applicants are not rejected.

Problem Statements:

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

- If the applicant can repay the loan but is not approved, the company loses business.
- If the applicant cannot repay the loan and is approved, the company faces a financial loss.

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

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

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

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

Approved: The company has approved the loan application.

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

Refused: The company rejected the loan.

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

Approach:

My first approach in this project was to understand the Risk Analytics in Banking & Finance. Where I need to go through the Risks factors per below to better understand the problem statements.

Credit Risk: The risk of borrowers defaulting on their loans.

Market Risk: The risk of losses due to changes in market conditions, such as interest rates, exchange rates, and stock prices.

Operational Risk: The risk of losses due to internal processes, systems, or human error.

Liquidity Risk: The risk of not being able to meet short-term financial obligations.

Reputation Risk: The risk of damage to an institution's reputation due to negative events or public perception.

Compliance Risk: The risk of failing to comply with regulatory requirements.

Then I went to extract & transform methodology on the dataset provided and followed the below steps.

- The data needs to be cleaned, formatted and to check if there are any missing values in the dataset. If there are, decide on the best strategy to handle them.
- Check and identify the outliers and remove any outliers that may have significant impact on the analysis.
- Apply the best strategy to handle outliers. This could be removing them, replacing them, or leaving them as is, depending on the situation.
- Then Analyse the Data Imbalance in the dataset by checking the distribution of each variable as well as Imbalance ratio.
- After cleaning Perform relevant descriptive statistic calculations to gain a general understanding of dataset
- This could involve calculating averages, medians, or other statistical measures. It could also involve creating visualizations to better understand the data

Tech-Stack Used:

The project was implemented using Python programming language in JupyterLab environment. Python libraries such as Pandas, NumPy, and Matplotlib were utilized for data manipulation, analysis, and visualization. These powerful tools provided the necessary functionality to preprocess the dataset, perform exploratory data analysis, and generate informative visualizations. The combination of Python and its associated libraries enabled efficient and effective analysis of the bank loan data.

Insights:

I performed Exploratory Data Analysis on dataset and found some insights by analysing below tasks.

Datasets:

1) Application_Data- New applications data.

2)Previous_Application- The data of previously applied loans.

3) Columns_description – Describe the column definitions.

A. Identify Missing Data and Deal with it Appropriately:

In this task, we focused on cleaning the dataset and handling missing values in order to ensure the integrity and quality of the data for further analysis.

A. There are total of 49 columns in Application_Data and 11 columns in Previous_Application which have missing values greater than 30%.
B. On checking the relation of 'FLAG_DOCUMENT_X' with loan repayment status, we found that the clients applying for loans only submitted the 'FLAG_DOCUMENT_3'.
C. There is almost no correlation of 'FLAG_MOBIL', 'FLAG_EMP_PHONE', 'FLAG_WORK_PHONE', 'FLAG_CONT_MOBILE', 'FLAG_PHONE', 'FLAG_EMAIL' with the "TARGET" column.
D. 'WEEKDAY_APPR_PROCESS_START', 'HOUR_APPR_PROCESS_START', 'FLAG_LAST_APPL_PER_CONTRACT', 'NFLAG_LAST_APPL_IN_DAY' are the column in the Previous_Application which are not needed for the analysis.
E. Dropping all the above mention columns which will total 76 in Application_Data and 15 in Previous_Application.
F. By identifying we found some other columns that contained irrelevant or redundant information as well as some contains "XNA" & "XAP". so they would not contribute significantly to the analysis. These columns were dropped from the dataset.
G. OCCUPATION_TYPE has large number of missing values so imputed it with mode
H. "AMT_GOODS_PRICE,", "AMT_ANNUITY" we noticed a high standard deviation and a significant number of outliers. Imputing the missing values with the mean would introduce bias. So we imputed these missing values with median.
I. The final cleaned dataset contains 49999 rows and 27 columns.

B. Identify Outliers in the Dataset:

In this task, we focused on transforming the data to improve readability and analysis. We converted the 'DAYS_BIRTH' column into the 'AGE' column for better interpretation and converted the 'DAYS_EMPLOYED' column into the 'YEARS_EMPLOYED' column. Additionally, we addressed outliers in the dataset using the 1.5 IQR (Interquartile Range) rule. First, we performed the following transformations:

1. Converting 'DAYS_BIRTH' into 'AGE' column:

    I. We divided the 'DAYS_BIRTH' values by 365 to convert them into years. o The absolute value was taken to ensure positive values.
    II. The resulting values were assigned to the newly created 'AGE' column.

2. Converting 'DAYS_EMPLOYED' into 'YEARS_EMPLOYED' column:

    I. Similar to the previous step, we divided the 'DAYS_EMPLOYED' values by 365 to convert them into years.
    II. The absolute value was taken to ensure positive values.
    III. The resulting values were assigned to the newly created 'YEARS_EMPLOYED' column.

After the transformations, the 'DAYS_BIRTH' and 'DAYS_EMPLOYED' columns were dropped from the dataset using the `drop` function. The resulting dataset, `new_app`, now contains additional columns 'AGE' and 'YEARS_EMPLOYED', which provide more meaningful information for analysis.

Next, we addressed outliers in both the dataset using the 1.5 IQR rule. The following steps were taken:

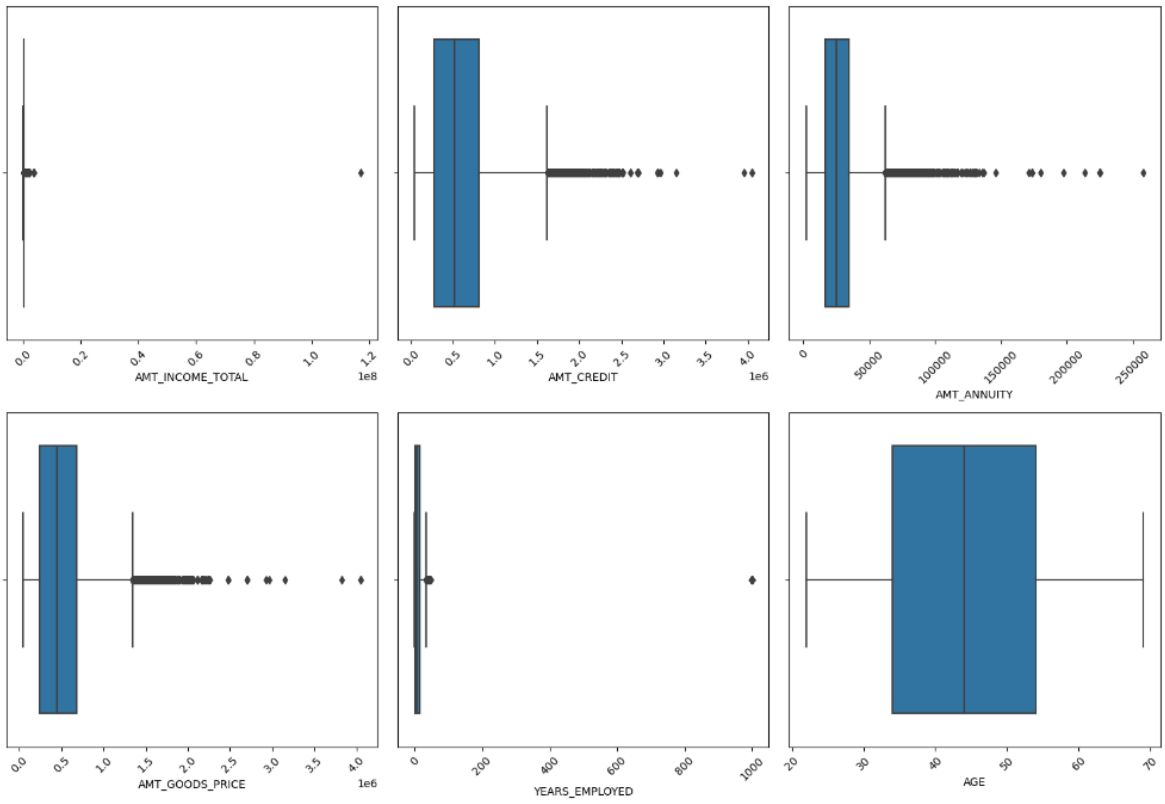
1. Defining a list, ` columns_to_check`, containing the numerical columns for both dataset: 'AMT_INCOME_TOTAL','AMT_CREDIT','AMT_ANNUITY','AMT_GOODS_PRICE','YEARS_EMPLOYED' 'AGE', 'CNT_PAYMENT'.

2. For each column in ` columns_to_check `, we plotted Boxplot to check the outliers for these variables as per 1.5 IQR rule.

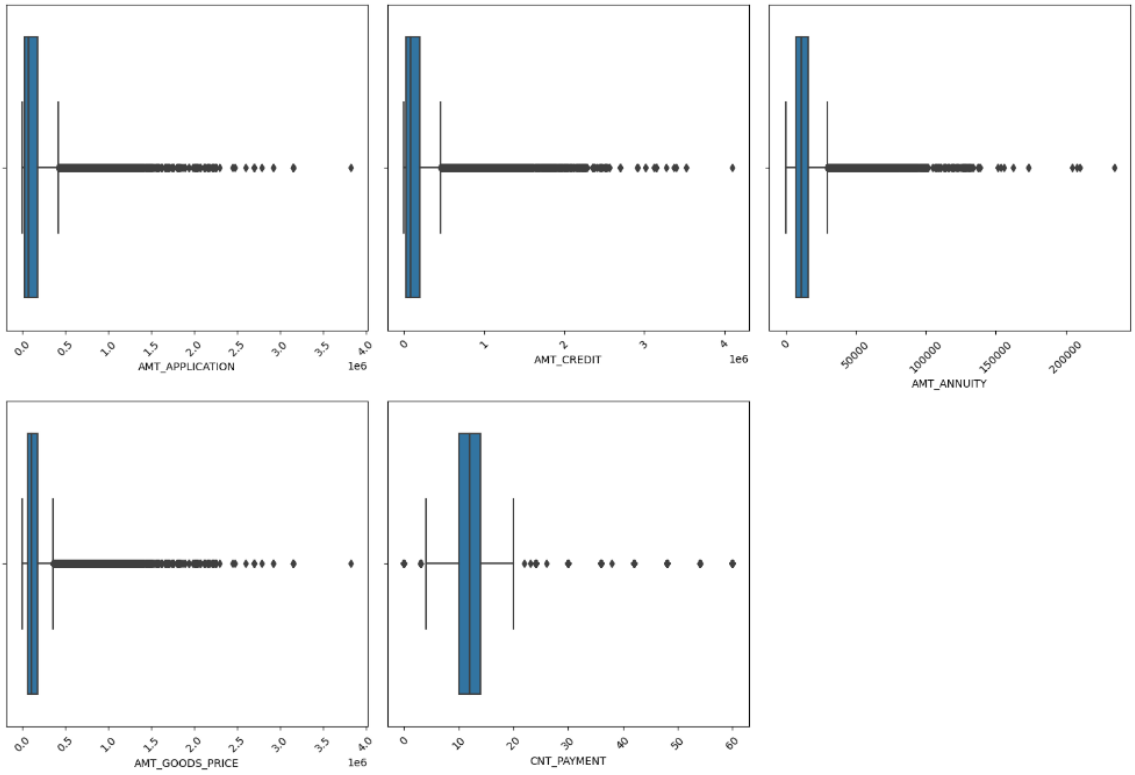
3. As we can see in below Boxplots in application data AGE column don't have any outliers. While all other columns have outliers.

4. In previous application data all the variables have outliers present.

Application_Data:



Previous_Application:



C. Analyze Data Imbalance:

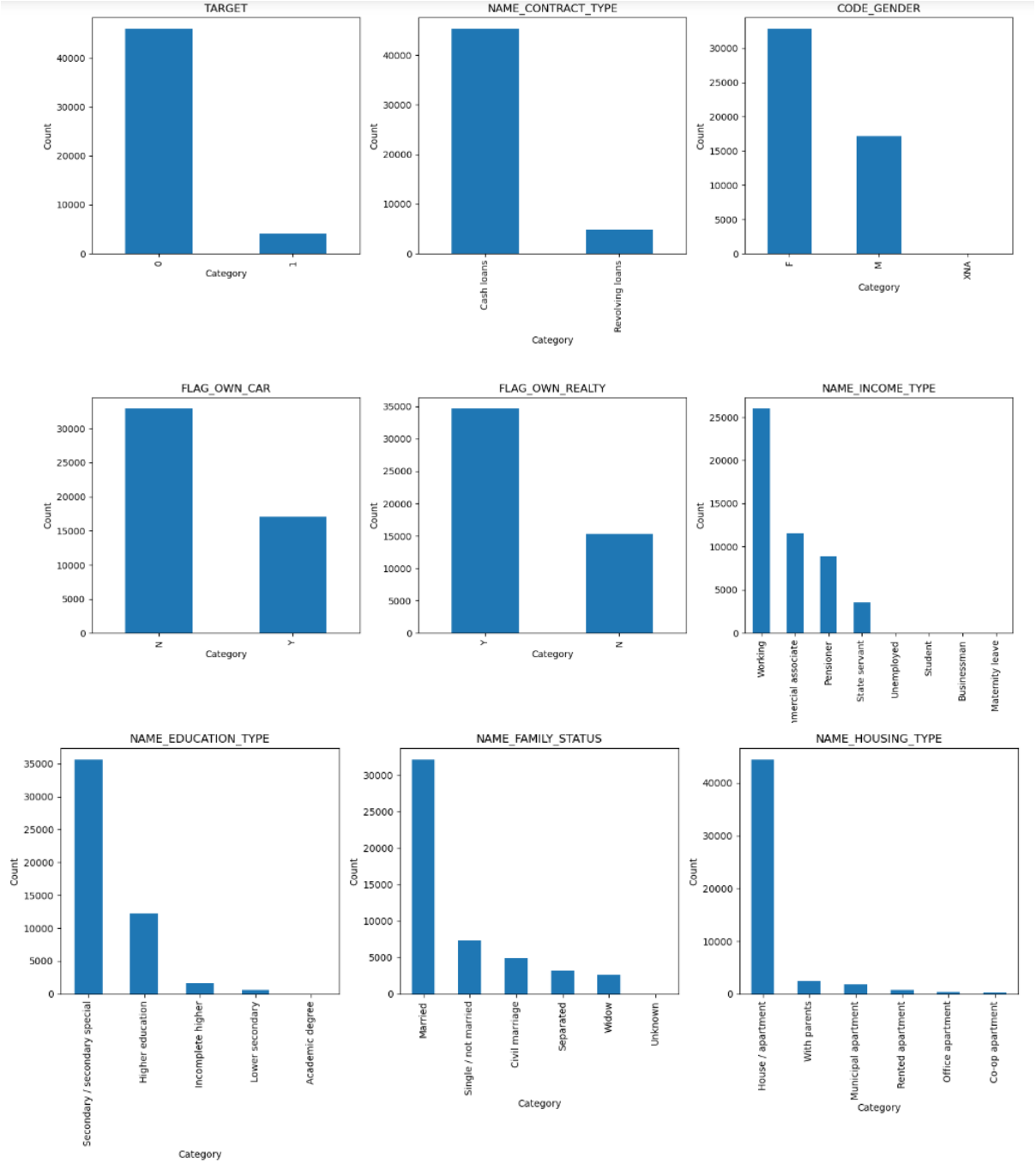
Data imbalance can affect the accuracy of the analysis, especially for binary classification problems. Understanding the data distribution is crucial for building reliable models.

In this task we checked for data imbalance in the following columns: 'TARGET', 'NAME_CONTRACT_TYPE', 'CODE_GENDER', 'FLAG_OWN_CAR', 'FLAG_OWN_REALTY', 'NAME_INCOME_TYPE', 'NAME_EDUCATION_TYPE', 'NAME_FAMILY_STATUS', and 'NAME_HOUSING_TYPE'.

To visualize the data imbalance, we plotted bar charts for each of these columns using the `value_counts()` function and the `plot` function from the matplotlib library. Each subplot in the figure represents one column, and the bar chart shows the counts for each category within the column.

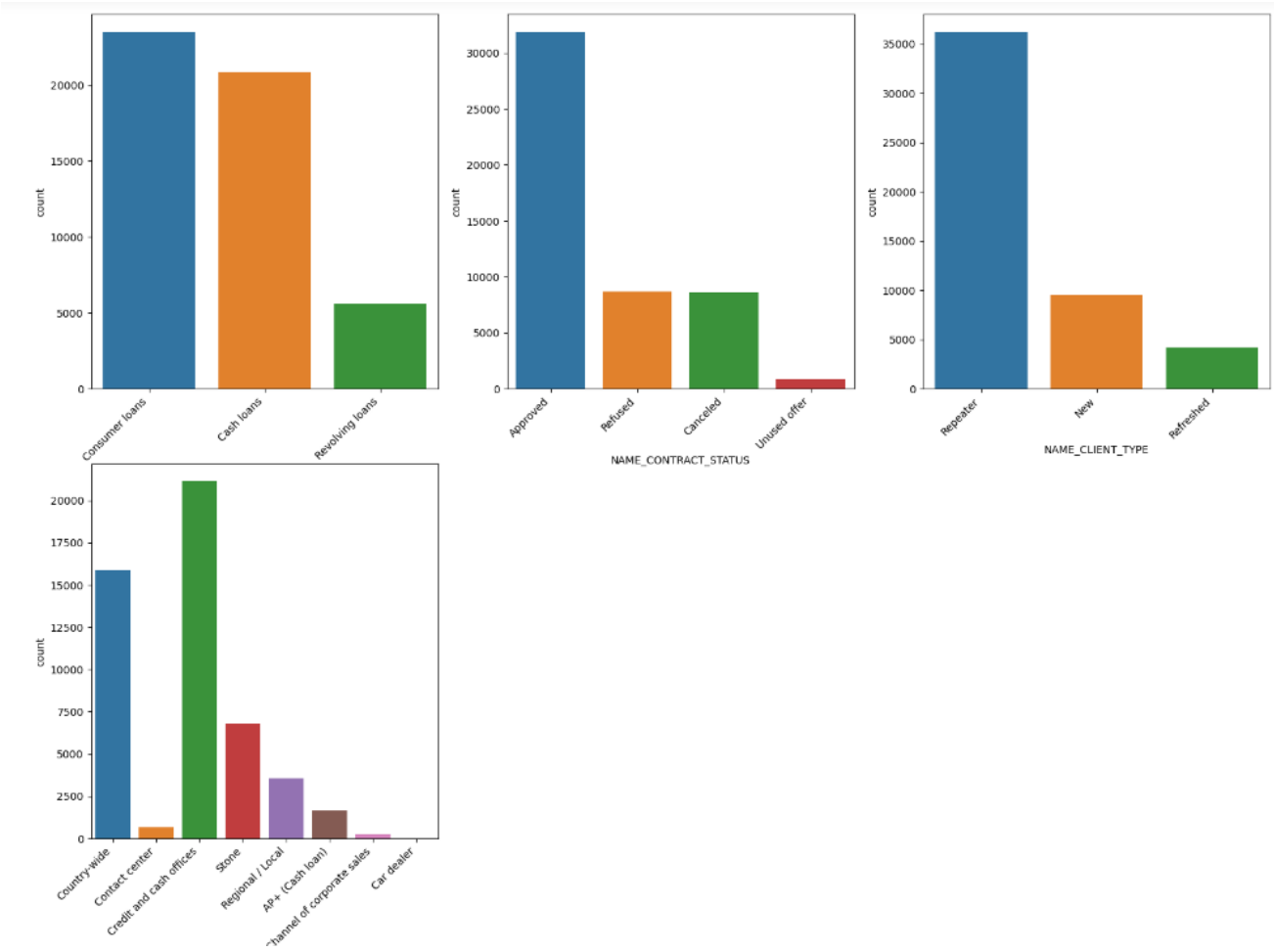
Based on the bar charts, we observed the following data imbalances in Application Data:

- TARGET: There are very few defaulters (1) compared to non-defaulters (0).
- NAME_CONTRACT_TYPE: There are very few revolving loans compared to cash loans.
- NAME_EDUCATION_TYPE: Most of the loans were applied by individuals with a secondary/secondary special education.
- NAME_FAMILY_STATUS: Most of the loans were applied by married individuals.
- NAME_HOUSING_TYPE: Most of the loan applications came from homeowners or individuals living in apartments



Data imbalances in Previous Application:

- NAME_CONTRACT_TYPE: The majority of loans are consumer loans, while the number of revolving loans is relatively low.
- NAME_CONTRACT_STATUS: The majority of loan were approved in status.
- NAME_CLIENT_TYPE: Repeater clients are high in numbers.
- CHANNEL_TYPE: Credit & Cash offers are high in numbers followed by country-wide



D. Perform Univariate, Segmented Univariate, and Bivariate Analysis:

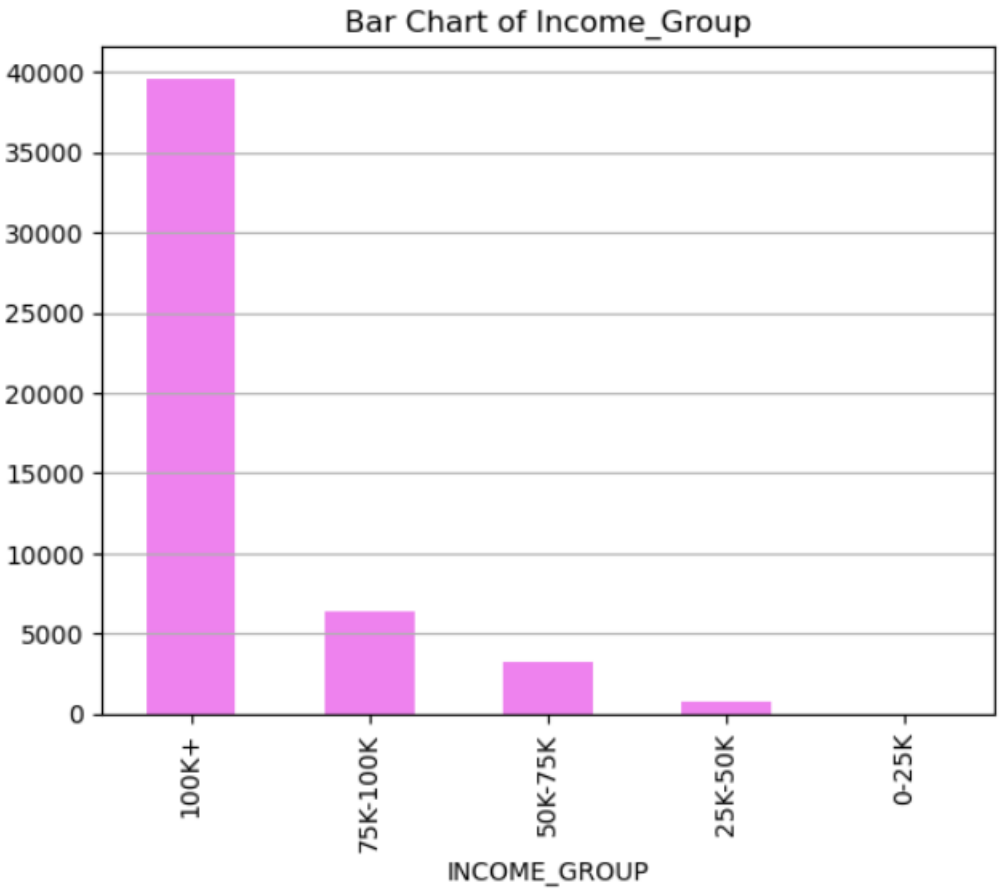
Univariate Analysis:

- NAME_CONTRACT_TYPE: Cash loans were taken by 90.55% of the people and only 9.44% people took Revolving Loans. While in previous application Consumer loans were the highest applied for loan, followed by Cash loans and then Revolving loans
- NAME_INCOME_TYPE: The majority of the clients who applied for loans were working class professionals with 26010 applications. Followed by Commercial associate 11543 applications, which is followed by pensioners and State servants.
  Unemployed, Students, Bussinessman, Maternity leave people were the least amount of people who applied for loan.
- NAME_FAMILY_STATUS: Married people were the higher percentage of loan applicants with 64.18%. Followed by Single /Not Married people having 14.61%
  Civil marriage, Separate, Widow were among the lower percentage category who applied for loans.
- NAME_HOUSING_TYPE: 88.73% of the applicants were living in a House / Apartment
  4.79% were living with their parents followed by 3.69% living in Muncipal Apartment
  Rented apartment accounted for 1.538031%, Office apartment accounted for 0.854017% and Co-op apartment accounted for 0.382008%.

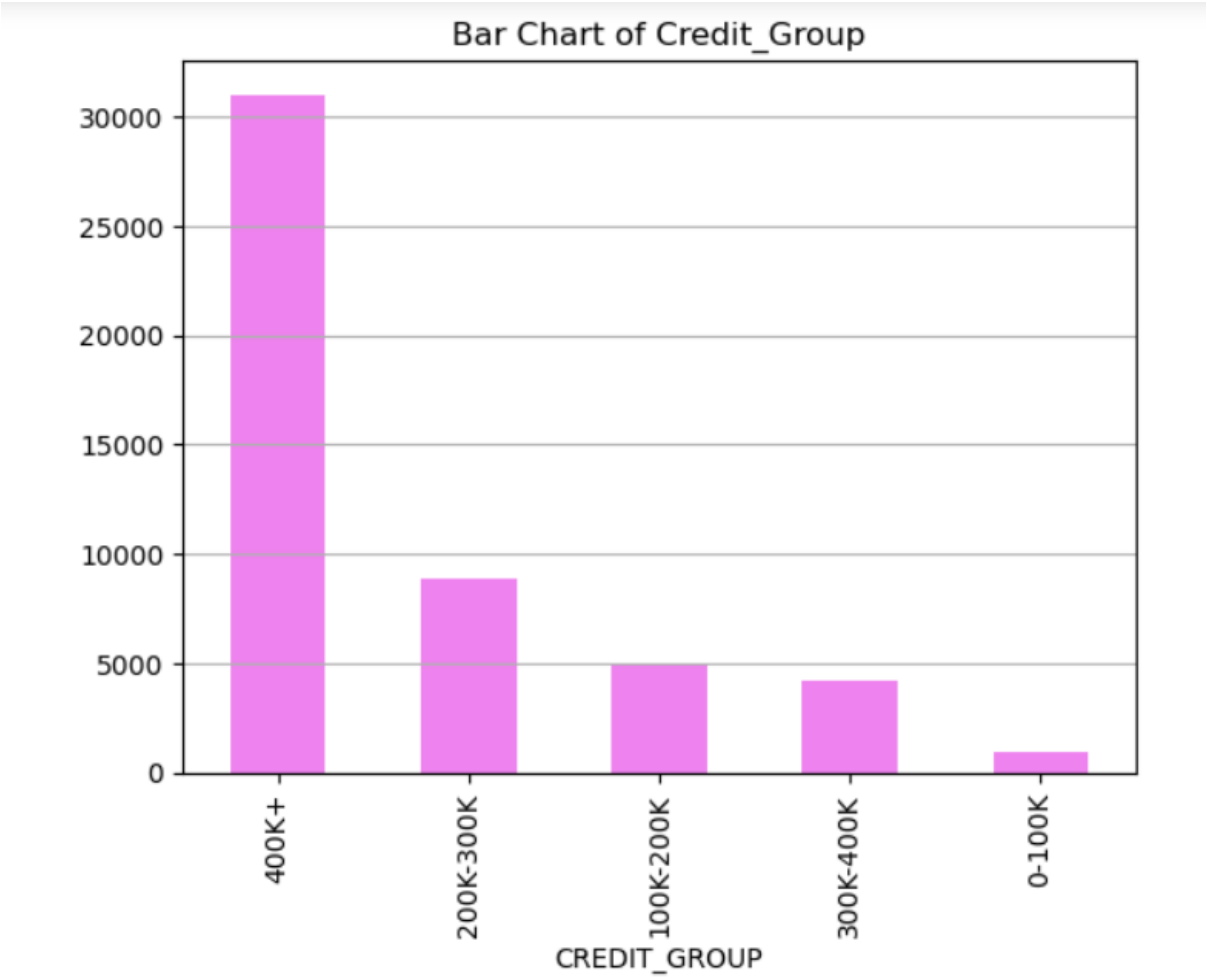
- OCCUPATION_TYPE: Laborers were the highest percentage of people who applied for a loan having the value of 49.21%, Followed by Sales staff (10.32%), Core staff (8.86%), Managers (6.97%) etc.
- ORGANIZATION_TYPE: Business Entity Type3, Self employed, Medicine, Government were the major organizations the loan applicants worked for. However 17.84% people have not disclosed their Organizations.
- CODE_GENDER: 65.64% of Females have applied for loan while remaining 34.34% being males.
- FLAG_OWN_CAR : In Flag Own Car 65.89% of the people who do not own a car and they were the majority who applied for loans. 34.10% of the people who applied owns a car.
- FLAG_OWN_REALTY: 69.38% people own a Realty(House or flat) While 30.61% people don't have own any realty.
- NAME_CONTRACT_STATUS: Majority of the applicants got approved of getting the loans. Cancelled and Refused are near about 17% each.
- NAME_CLIENT_TYPE: Repeater clients are more in number who aaplied for loan, while New & Refreshed are very low.
- CHANNEL_TYPE: Received majority of the loan applications from Credit and cash offices.

Segmented Univariate:

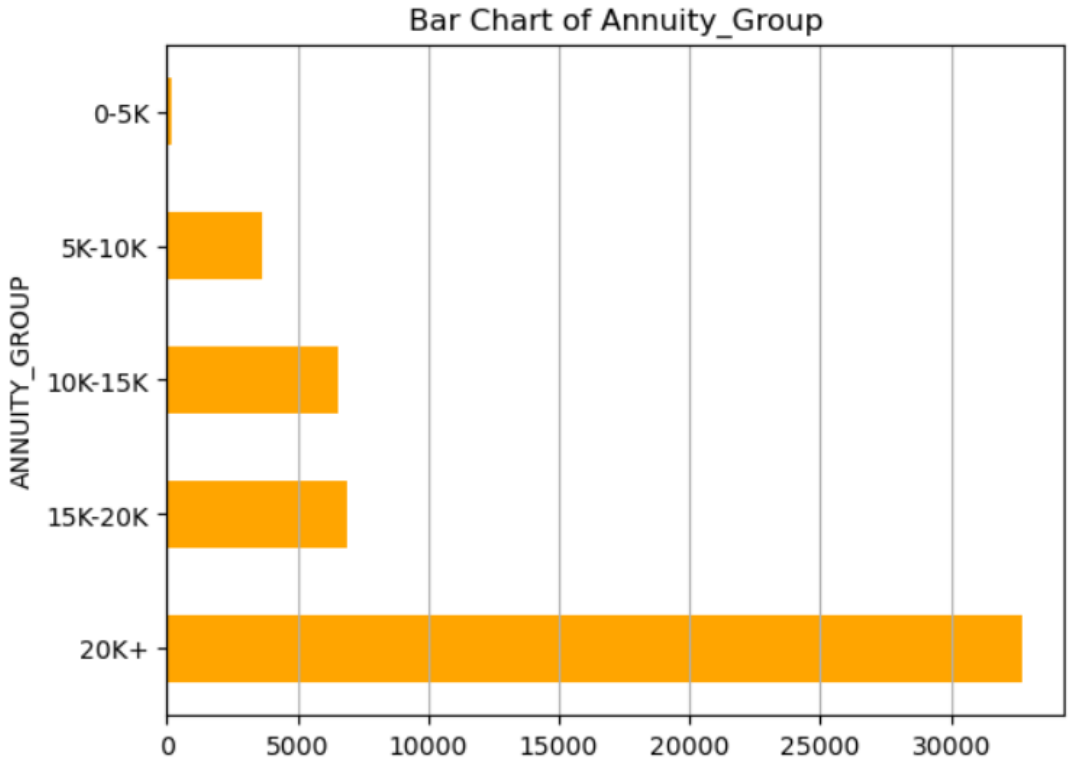
- **Income Group:** Majority of the loan appliers were from 100k+ (Very High) wage Income being 79.21% total followed by 75k-100k (High class) income being 26.73% & 50k-75k (Medium Class) income which is 6.45% Low Class income wage people were 1.60% applied for loan. It shows loan applied by Very High and High wage people can repay the loan



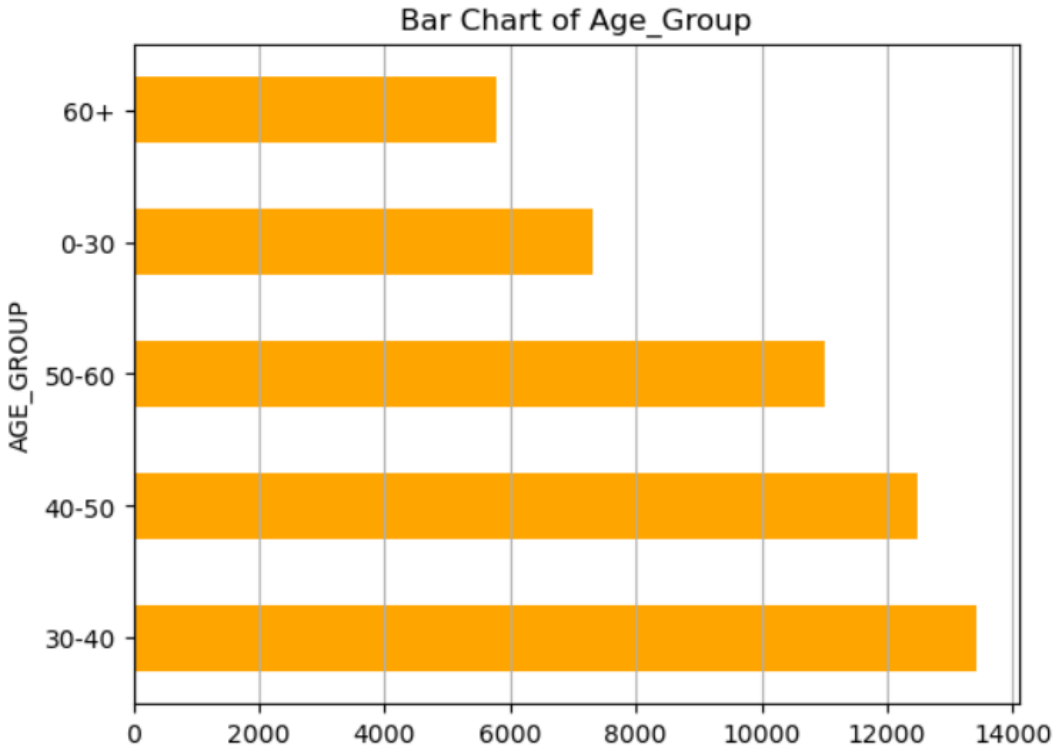
- **Credit Group:** Mainly the credit amount of loan ranged from 400k+ followed by 200k-300k and 100k-200k



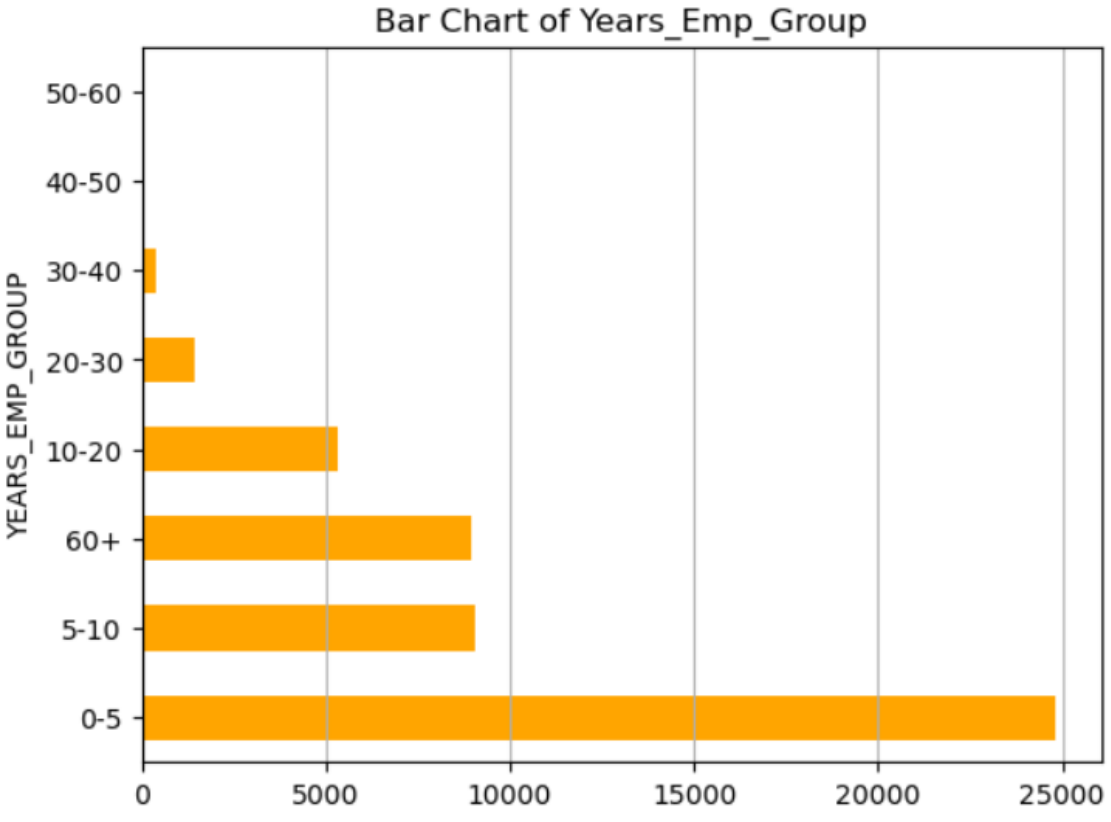
❖ **ANNUITY_GROUP:** Applicants took Annuities ranging in 20K which topped the graph with 65.41% Followed by 15k-20k (13.79%), 10k-15k (13.12%) etc.



❖ **AGE_GROUP:** Majority of the loans applied by younger age people ranging in 30-40 age group which is 26.84% followed by mid senior people from 40-50 age group & 50-60 senior age group.



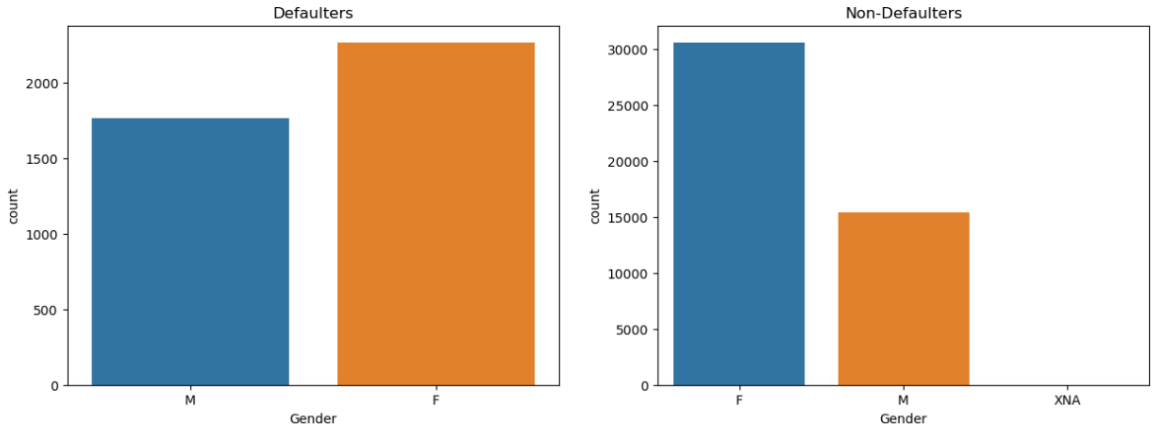
- **YEARS_EMP_GROUP:** 0-5 years of experience people topped the chart with 49.63% with none being from 50-60 years , and some being from 40-50 years.



Bivariate Analysis

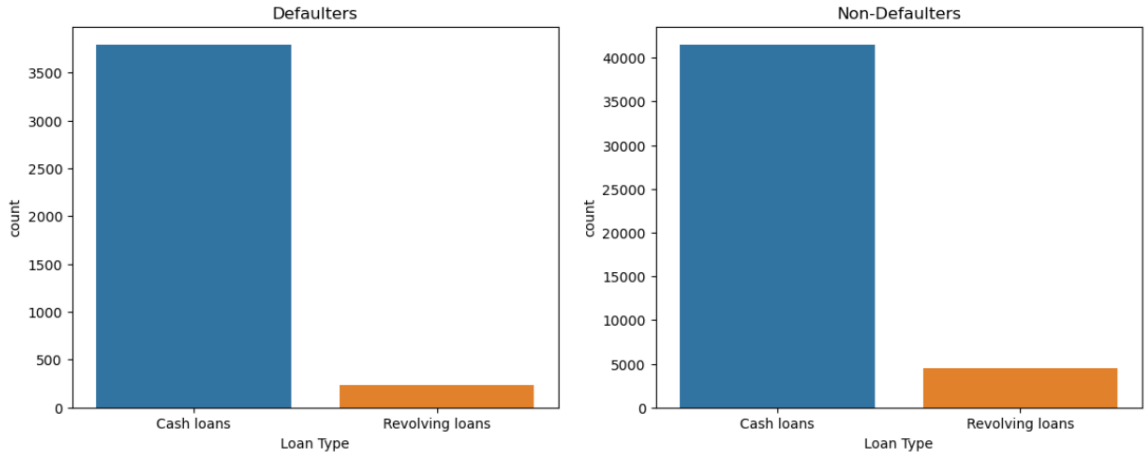
a) Gender Analysis:

- Defaulter Analysis: We observe a slight gender imbalance among defaulters, with a slightly higher number of females compared to males. This suggests that there may be certain factors or variables that contribute to a higher likelihood of defaulting for females.
- Non-Defaulter Analysis: Similarly, among non-defaulters, we continue to observe a higher representation of females compared to males. This finding indicates that gender may not be a significant differentiating factor in determining loan repayment behavior, as both genders have a similar proportion of non-defaulters.



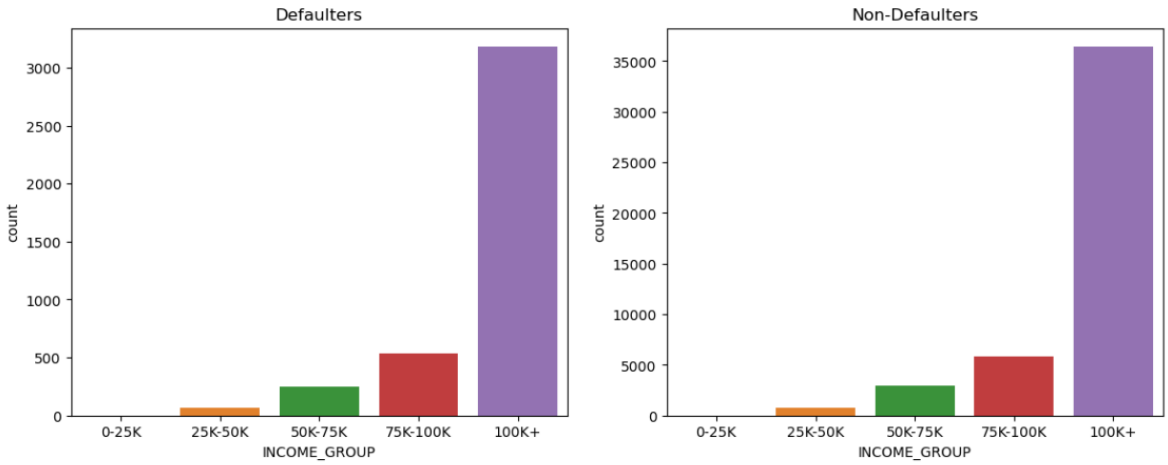
b) Contract type Analysis:

- In both the cases of defaulters and non-defaulters, we observe a significant difference in the number of Revolving loans compared to Cash loans. Revolving loans are considerably less prevalent among both defaulters and non-defaulters. This indicates that the majority of loan applications, regardless of the repayment status, are for Cash loans rather than Revolving loans.



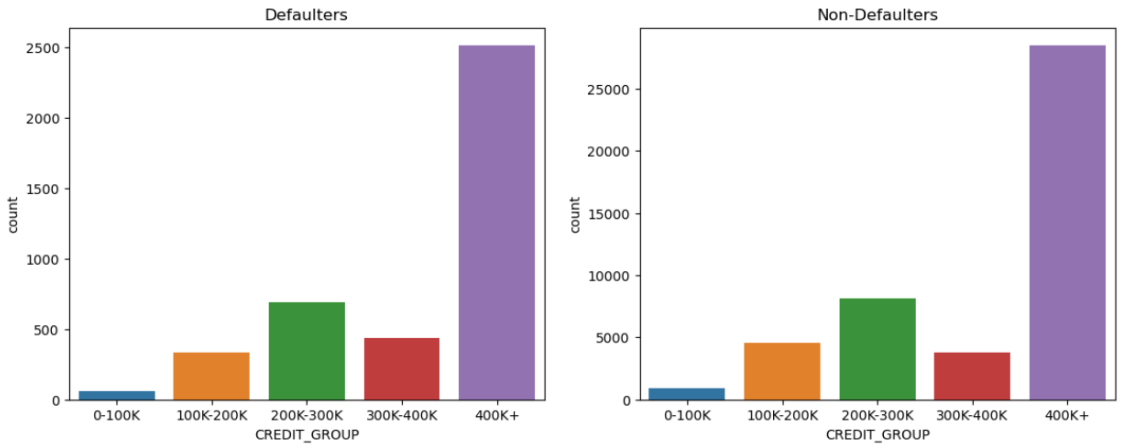
c) Income Group Analysis:

- Defaulters: Higher-income individuals surprisingly have a higher default rate compared to medium and low-income groups. The low-income group has a relatively lower count of defaulters.
- Non-defaulters: non-defaulters are more common in the high-income group and less common in the low-income group. - Insight: Income level alone does not determine loan default risk. Other factors, such as financial management habits and personal circumstances, play a significant role.



d) Credit Group Analysis:

- Upon analyzing the data, we observe interesting patterns related to the credit amount and the likelihood of defaulting. For defaulters, there is a higher concentration of individuals in the higher credit amount groups, indicating that those with higher credited amounts are more likely to default on their loans. On the other hand, non-defaulters also show a similar trend, with a larger proportion of individuals falling into the higher credit amount groups.

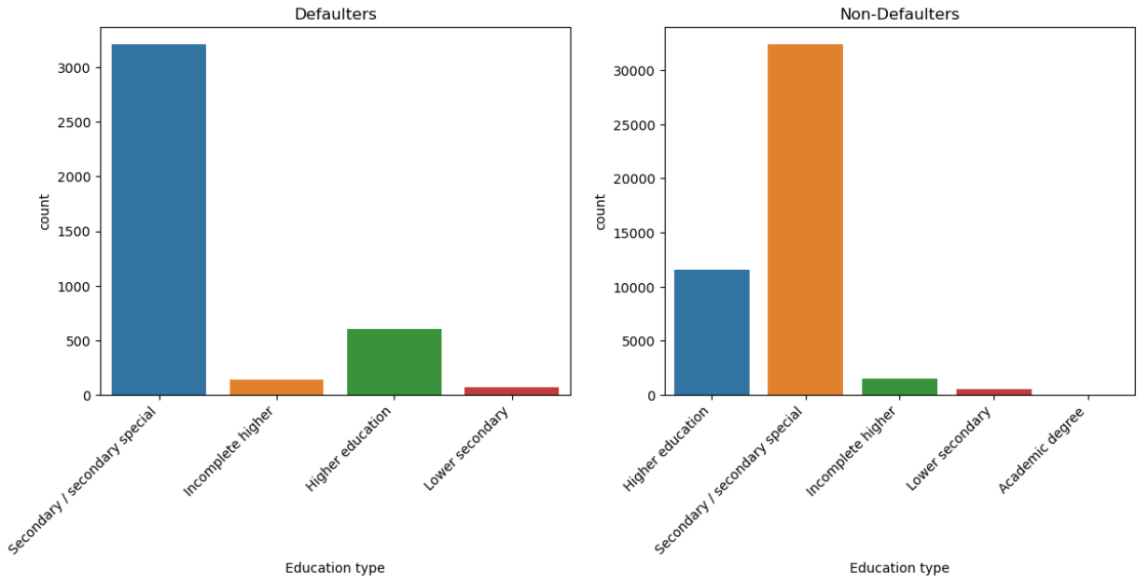


e*) Education Analysis:*

- Defaulters: Among the defaulters, individuals with a secondary or secondary special level of education have the highest representation. This indicates that individuals with lower levels

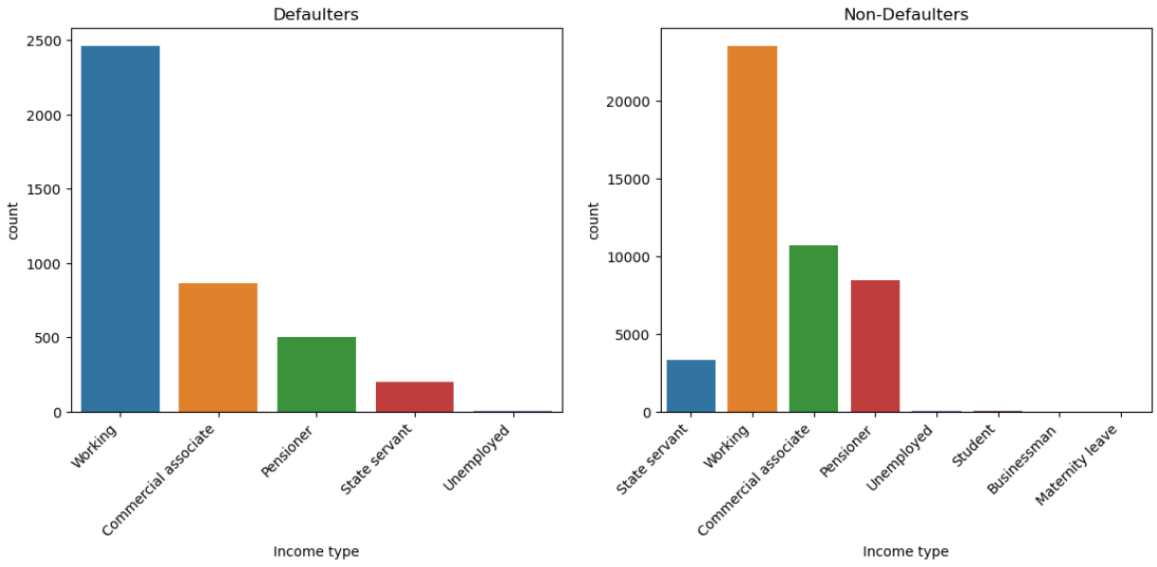
of education are more likely to default on their loans compared to those with higher levels of education.

- Non-defaulters: Similarly, among the non-defaulters, the majority are individuals with a secondary or secondary special level of education. This suggests that individuals with lower levels of education are also more likely to fulfil their loan repayment obligations and demonstrate responsible borrowing behaviour.



f) Income Type Analysis:

- Defaulters: Among the defaulters, we observe that individuals belonging to the "Working" profession have the highest number. This suggests that working individuals are more prone to defaulting on their loans compared to individuals in other professions.
- Non-defaulters: Similarly, among the non-defaulters, the majority are individuals from the "Working" profession. This indicates that individuals in the working category are more likely to meet their loan repayment obligations and are considered reliable borrowers.

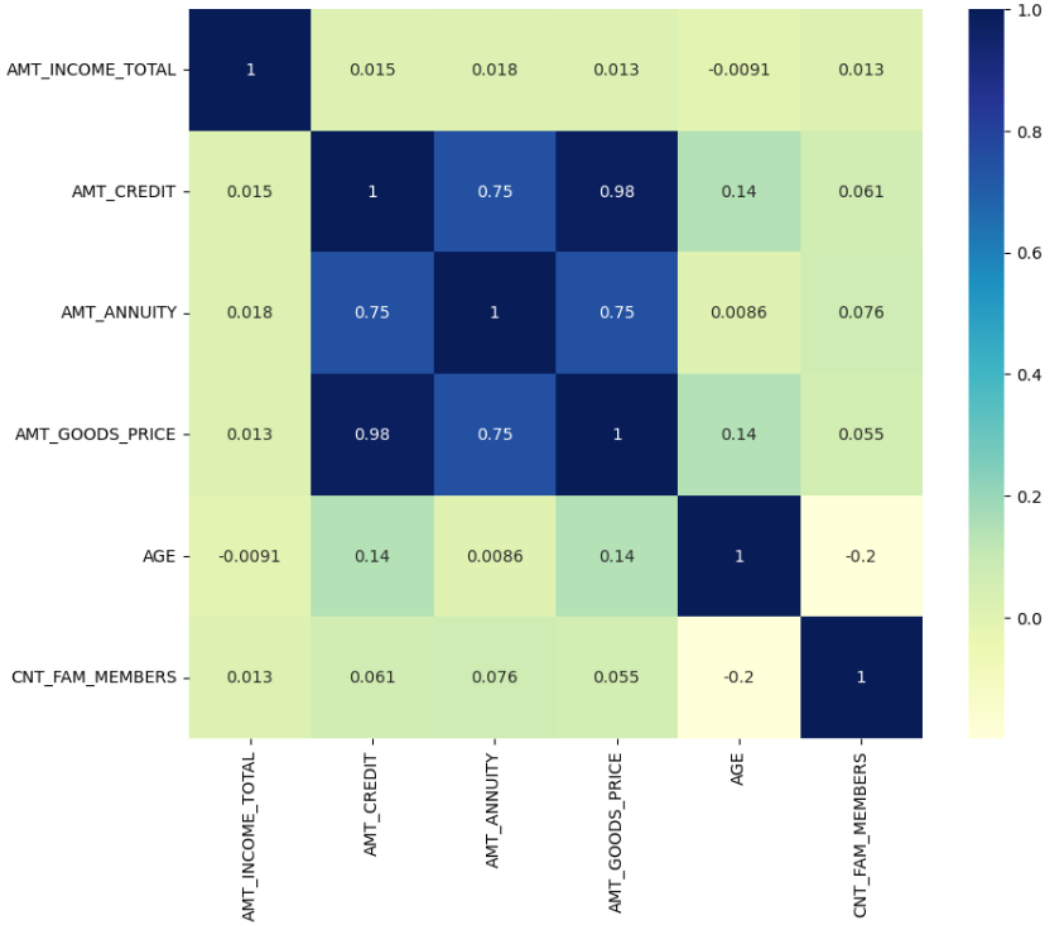


E. Identify Top Correlations for Different Scenarios:

Corelation of defaulters(Clients with payment difficulties):

i. There is a strong positive correlation between the "AMT_CREDIT" variable (amount of credit requested) and the "AMT_ANNUITY" variable (loan annuity).

ii. A moderate positive correlation is observed between the "AMT_CREDIT" variable and the "AMT_GOODS_PRICE" variable (price of the goods for which the loan is requested).

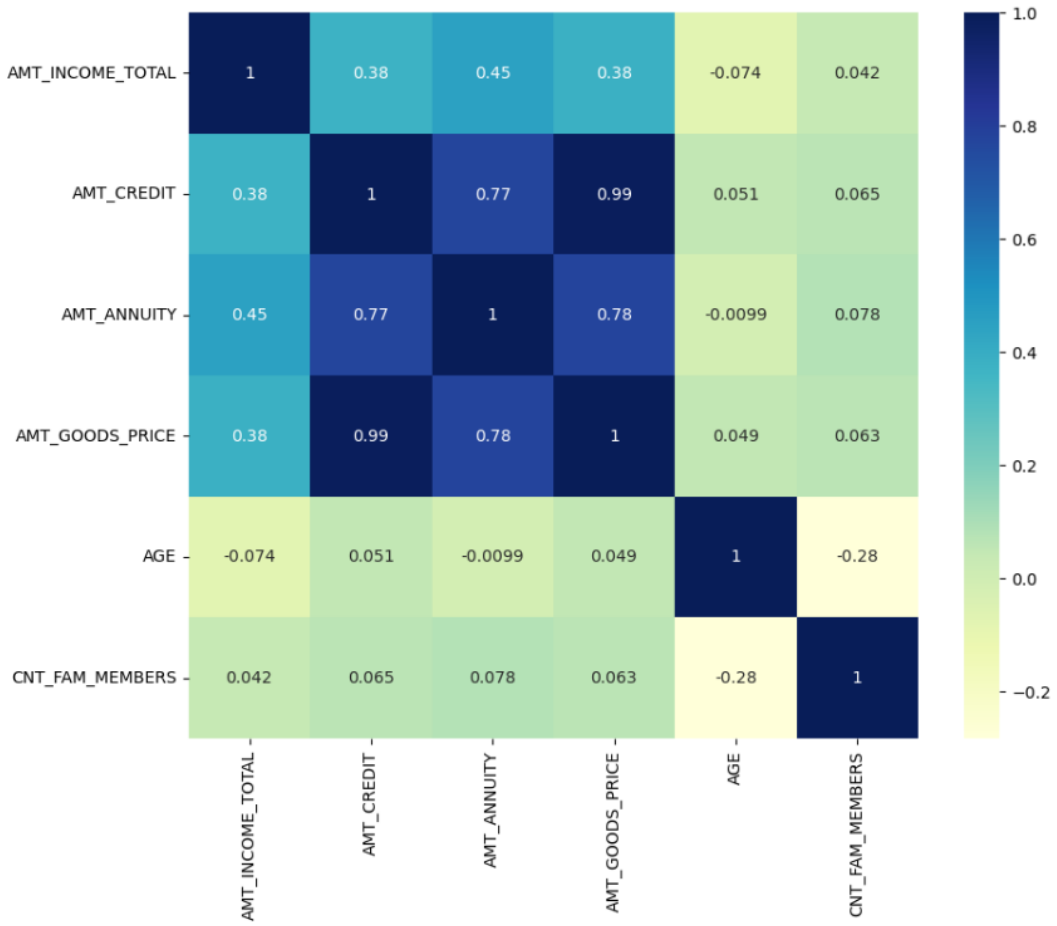


Corelation of non defaulters:

iii. AMT_CREDIT and AMT_ANNUITY: Among non-defaulters, there is a strong positive correlation of 0.77 between the loan amount (AMT_CREDIT) and the corresponding annuity payments (AMT_ANNUITY). This suggests that as the loan amount increases, the associated annuity payments also tend to increase.

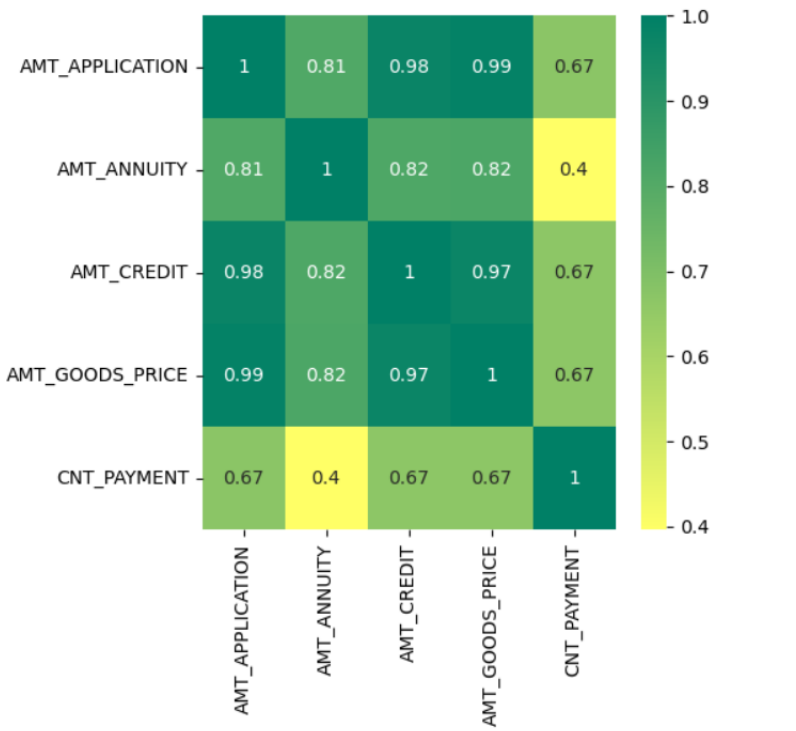
iv. AMT_CREDIT and AMT_GOODS_PRICE: Non-defaulters exhibit a high correlation of 0.99 between the loan amount (AMT_CREDIT) and the price of the financed goods (AMT_GOODS_PRICE). This indicates that the loan amount is closely related to the value of the purchased goods.

v. AMT_ANNUITY and AMT_GOODS_PRICE: There is a moderate positive correlation of 0.78 between the annuity payments (AMT_ANNUITY) and the price of the goods (AMT_GOODS_PRICE) among non-defaulters. As the price of the goods increases, the corresponding annuity payments also tend to increase.



Correlation Insights from previous application:

vi. A strong positive correlation is observed between the "AMT_APPLICATION" variable (amount of loan application) and the "AMT_CREDIT" variable across all cases, irrespective of the target variable. This indicates that the amount of credit requested is strongly related to the loan application amount.



Conclusion:

In this analysis of defaulters and non-defaulters, we observed that the same pairs of columns exhibit high correlation for both groups. Specifically, the columns AMT_CREDIT and AMT_ANNUITY, AMT_CREDIT and AMT_GOODS_PRICE, and AMT_ANNUITY and AMT_GOODS_PRICE showed significant correlations in both cases.

The strong positive correlations between AMT_CREDIT and AMT_ANNUITY, as well as AMT_CREDIT and AMT_GOODS_PRICE, suggest that higher loan amounts are associated with higher annuity payments and the purchase of more expensive goods. This finding indicates that defaulters tend to borrow larger amounts to finance costly purchases, potentially contributing to their higher default risk. Lenders should carefully evaluate the borrower's ability to handle larger loan amounts and associated annuity payments, as well as the relationship between the loan amount and the value of the purchased goods, to mitigate default risks.

Key Insights

Key Insights for Non-Default:

- NAME_EDUCATION_TYPE: Academic degree has less defaults.
- NAME_INCOME_TYPE: Student and Businessmen have no defaults.
- ORGANIZATION_TYPE: Clients with Trade Type 4 and Industry type 8 have no defaults while Trade Type 5 has less than 3% defaults.
- DAYS_BIRTH: People above age of 50 have low probability of defaulting.
- DAYS_EMPLOYED: Clients with range 40-50 & 50-60 year experience having less than 1% default rate
- AMT_INCOME_TOTAL: Applicant with Income range 25k-50k & 50k-75k are less likely to default.
- CNT_CHILDREN: People with zero to two children tend to repay the loans.
- CHANNEL_TYPE: Channel of Corporate sales have no defaults.

Key Insights for Default:

- CODE_GENDER: Men are at relatively higher default rate.
- NAME_FAMILY_STATUS : People who have civil marriage or who are single default a lot.
- NAME_EDUCATION_TYPE: People with Lower Secondary & Secondary education.
- NAME_INCOME_TYPE: Clients who are Unemployed default a lot.
- OCCUPATION_TYPE: Avoid Low-skill Laborers, Drivers and Waiters/barmen staff, Security staff, Laborers and Cooking staff as the default rate is huge.
- ORGANIZATION_TYPE: Organizations with highest percent of loans not repaid are Transport: type 3 (16%), Industry: type 13 (13.5%) and Restaurant (less than 12%). Self-employed people have relative high defaulting rate, and thus should be avoided to be approved for loan or provide loan with higher interest rate to mitigate the risk of defaulting.
- DAYS_BIRTH: Avoid young people who are in age group of 20-40 as they have higher probability of defaulting
- DAYS_EMPLOYED: People who have less than 5 years of employment have high default rate.
- CNT_CHILDREN & CNT_FAM_MEMBERS: Client who have children equal to or more than 9 default 100% and hence their applications are to be rejected.
- AMT_GOODS_PRICE: When the credit amount goes beyond 3M, there is an increase in defaulters.

Result:

The project helps me to understand how to perform Data Analysis based on all the features of the Movies. As if we stick with one feature then we might have to lose the other important factors. Movie success depends on different factors to get attention of every categories audience, some like Horror movies, others like Drama depending on these categories audience. Also the time duration of movies, their Language if having Multi language might have get more audience, so considering all these factors a Data Analyst have to extract meaningful insights from the data to help producers, make a successful movies.

.