UPGRAD EDA ASSIGNMENT

BINOD KUMAR NAHAK

INTRODUCTION

This case study is to understand real business scenario .

Certainly, understanding risk analytics in the banking and financial services industry is crucial for minimizing the risk of financial loss while lending money to customers. By employing various techniques and leveraging data, organizations can make informed decisions and mitigate potential risks.

By combining techniques and leveraging data effectively, banks and financial institutions can make more informed lending decisions, minimize the risk of financial loss, and improve overall risk management practices.

BUSINESS UNDERSTANDING

Understanding:

- Certainly, in the loan approval process, loan providers face challenges when dealing with applicants who have insufficient or non-existent credit history. They need to assess the applicant's profile to make a decision, considering two types of risks:

- a. The risk of not approving a loan for an applicant who is likely to repay it, resulting in a loss of potential business for the company.

- b. The risk of approving a loan for an applicant who is likely to default, leading to a financial loss for the company.

- To address these risks and make informed decisions, exploratory data analysis (EDA) can be conducted on the given dataset. EDA helps identify patterns and insights related to client behavior, attributes, and payment history, which can be utilized to distinguish between defaulters and non-defaulters.

- By analyzing the data, patterns and relationships can be uncovered, allowing loan providers to better understand the factors that contribute to defaulting behavior. These factors may include demographic information, income levels, loan amounts, credit history, payment patterns, and other relevant variables.

- EDA will provide valuable insights into the dataset, enabling loan providers to identify key variables that significantly impact the likelihood of defaulting. This knowledge can then be utilized in the loan approval process, along with other relevant variables, to assess the creditworthiness and risk associated with each applicant accurately.
- By leveraging EDA and considering multiple variables during the loan approval process, loan providers can make more informed decisions, mitigate the risk of defaults, and minimize financial losses.

OBJECTIVE

To identify patterns indicating if a client has difficulty paying their installments, we can perform Exploratory Data Analysis (EDA) on our dataset. EDA involves examining and visualizing the data to gain insights and discover patterns, relationships, and trends that can help you make informed decisions. Here's a step-by-step approach you can follow:

1. Data Collection
2. Data Cleaning
3. Univariate Analysis
4. Bivariate Analysis
5. Multivariate Analysis
6. Identify Patterns
7. Feature Selection

By following these above steps, we can perform EDA to identify driving factors behind loan repayment difficulties, enabling bank to take appropriate actions for loan applicants based on their profiles and risk assessment. Remember to ensure ethical considerations and comply with legal requirements throughout the process.

STEPS INVOLVED FOR DATA ANALYTICS

- Understanding the domain and variables from the data dictionary is an important step in data analysis. It helps in gaining insights into the meaning and significance of each variable. Here are the subsequent steps involved in the data analysis process:
- Importing Data: The data can be imported into a Python environment using the Pandas library. Pandas provides functions and data structures to efficiently handle and manipulate data.
- Data Visualization: Visualization libraries like seaborn and matplotlib.pyplot can be used to create visual representations of the data. This helps in understanding patterns, distributions, identifying missing values, outliers, and gaining an overall understanding of the dataset.
- Checking Data Structure: It is essential to examine the structure or metadata of the data. This includes inspecting the number of rows and columns, data types of variables, and any initial observations about the dataset.
- Missing Value Handling: Missing values need to be addressed in the dataset. Columns with a high percentage of missing values (e.g., more than 40%) may be dropped based on the specific requirements of the analysis. For the remaining columns, missing values can be imputed using methods like median or mode, depending on the nature of the data.
- Outlier Treatment: Outliers, if present, can be handled by capping or trimming the extreme values. This ensures that extreme values do not significantly impact the analysis or skew the results.
- Categorizing Variables: The remaining variables can be grouped into different categories such as numerical, categorical (object), or date/time variables. This categorization helps in selecting appropriate analysis techniques for each variable type.
- Exploratory Data Analysis (EDA): Univariate, bivariate, and multivariate analyses can be performed to explore the relationships, distributions, and patterns within the data. This includes summary statistics, distribution plots, scatter plots, box plots, etc.
- Correlation Analysis: For numerical columns, a correlation matrix can be computed to identify the relationships between variables. This helps in understanding the strength and direction of associations, which can be useful in further analysis and modeling.
- By following these steps, analysts gain a comprehensive understanding of the dataset, identify patterns and relationships, handle missing values and outliers appropriately, and perform exploratory analysis to extract meaningful insights.

ACTION PERFORMED

To perform the mentioned tasks on the given datasets, we can use the following steps:

- Reading Data: Read the "Application Data" and "Previous Application Data" datasets from external files using a read command such as pd.read_csv() for CSV files or appropriate functions for other file formats.
- Checking Shape and Distribution: Use commands like shape to check the dimensions of the datasets and head and tail to view the sample records. This helps in understanding the structure and initial data distribution.
- Handling Missing Values: Drop the columns with more than 40% missing values using the drop function or similar methods provided by the Pandas library. This step reduces the dimensionality of the dataset and focuses on the important columns for analysis.
- Identifying Important Columns: Identify the columns that are more important for analysis based on the specific requirements or business objectives. These important columns can be selected for further analysis, while less important columns can be ignored or dropped using the drop function.
- Handling Outliers: Use descriptive statistics such as the describe function and visualization techniques like box plots to identify outlier values in the selected columns. Apply capping or trimming methods to handle outliers by replacing extreme values with appropriate thresholds.
- Binning Numerical Continuous Variables: Apply binning techniques, such as creating equal-width or equal-frequency bins, to convert numerical continuous variables into categorical variables. This helps in analyzing the data based on grouped intervals or ranges.
- Checking Data Imbalance: Examine the distribution of the "TARGET" column in the "Application Data" to determine data imbalance. Use functions like value_counts to count the occurrences of each class (defaulters and non-defaulters) and visualize the class distribution using plots like bar charts or pie charts.
- After performing these steps, the data will be ready for further analysis and modeling, taking into account the important columns, handling missing values and outliers, and addressing data imbalance if present.

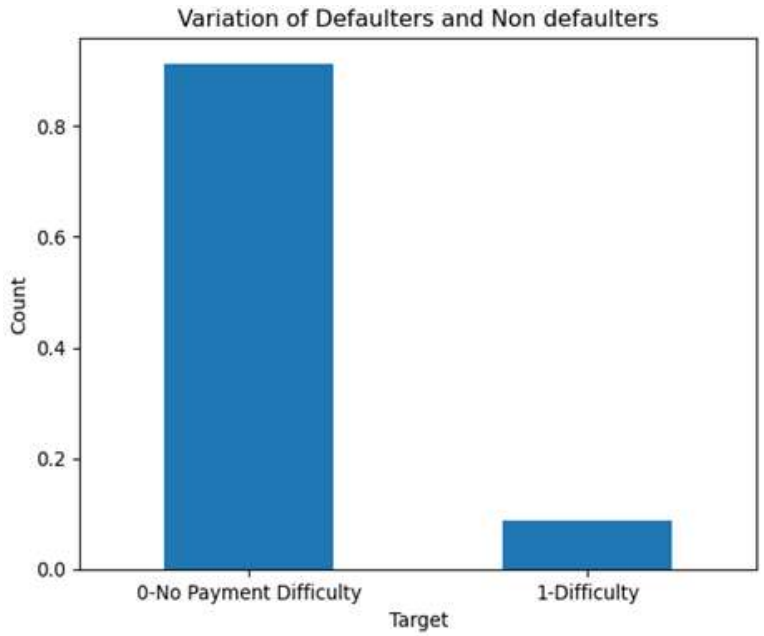
DATA IMBALANCE

To calculate the ratio of Target 0 to Target 1 and assess the data imbalance, we can use the following steps:

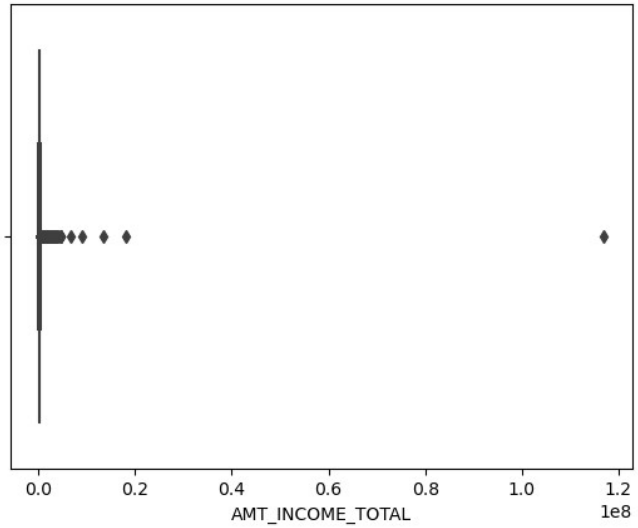
Count the number of occurrences of Target 0 and Target 1 in the "TARGET" column of the "Application Data" dataset.

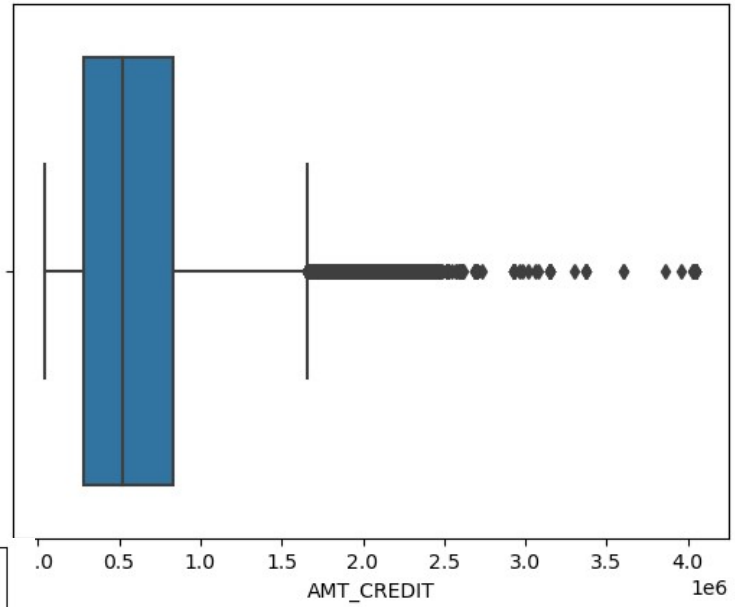
Calculate the ratio by dividing the count of Target 0 by the count of Target 1.

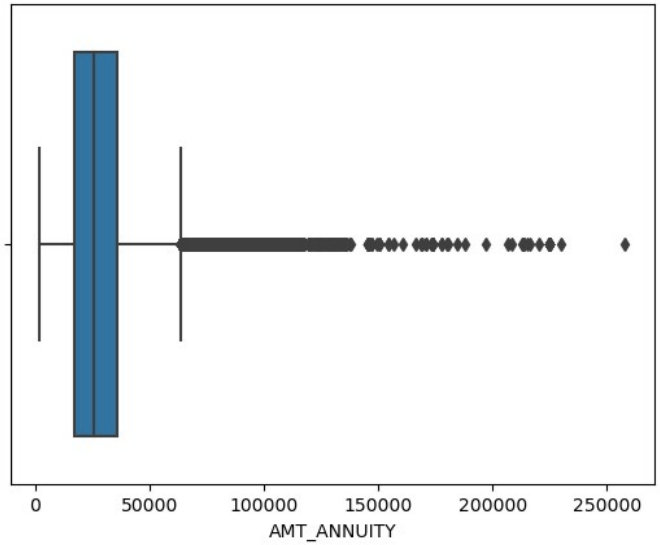
The resulting ratio represents the imbalance between the two classes, indicating whether one class is significantly more prevalent than the other.



OUTLIER ANALYSIS







Observation-

Considering the presence of outliers in the "income total", "credit amount" and "annuity amount" variables, it is advisable to employ a technique such as capping and flooring to achieve the desired variance in these values. By setting appropriate upper and lower thresholds, extreme values can be truncated, ensuring that the data remains within a reasonable range and allowing for more meaningful analysis and interpretation.

ASSUMPTION

```
NaN                    96391
Laborers               55186
Sales staff            32102
Core staff             27570
Managers               21371
Drivers                18603
High skill tech staff  11380
Accountants             9813
Medicine staff          8537
Security staff          6721
Cooking staff           5946
Cleaning staff          4653
Private service staff   2652
Low-skill Laborers      2093
Waiters/barmen staff    1348
Secretaries             1305
Realty agents            751
HR staff                 563
IT staff                 526
Name: OCCUPATION_TYPE, dtype: int64
```

```
Cleaning staff    2
Laborers          1
Medicine staff    1
Sales staff       1
Name: OCCUPATION_TYPE, dtype: int64
```

'OCCUPATION_TYPE' value counts where 'NAME_INCOME_TYPE' has value "Pensioner"

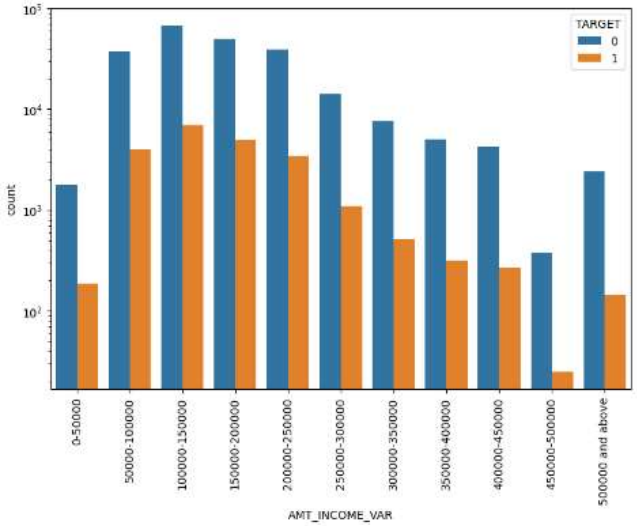
Income type value counts for null values of occupation type.

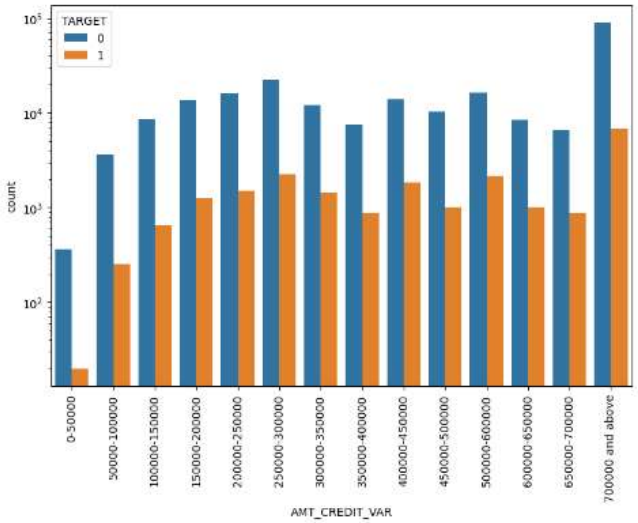
```
Pensioner               55357
Working                 24920
Commercial associate    12297
State servant            3787
Unemployed                 22
Student                     5
Businessman                 2
Maternity leave             1
Name: NAME_INCOME_TYPE, dtype: int64
```

Observation-

Upon careful observation, it has come to light that over 50% of the instances where the occupation data is missing pertain to individuals classified as pensioners. Given this notable trend, a viable course of action would be to disregard the "occupation" column and instead rely on the "NAME_INCOME_TYPE" column for the purpose of analysis and decision-making.

UNIVARIATE ANALYSIS FOR CATEGORICAL ORDERED VALUE





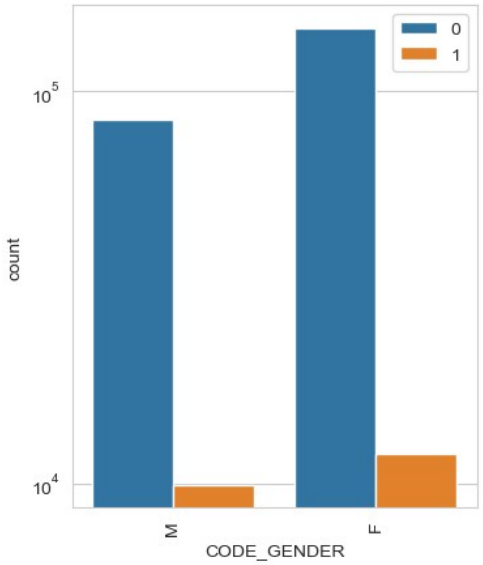
Observation-

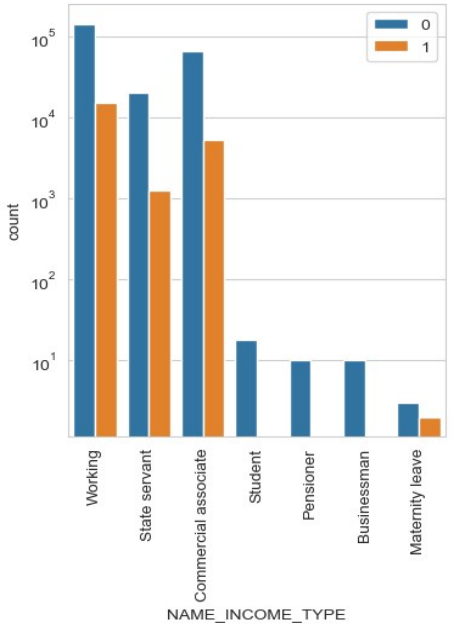
a. Regarding the "AMT_INCOME_VAR" it is observed that customers with an annual income ranging from 1,00,000 to 2,00,000 exhibit a higher number of loans, along with a relatively high percentage of defaulters. However, this trend gradually diminishes as the annual income increases.

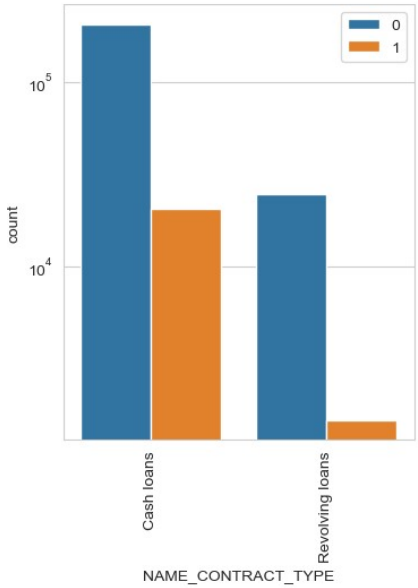
b. In relation to the "AMT_CREDIT_VAR" customers who acquire credit amounts around 50,000 display a significantly lower quantity of defaulters. However, there is a notable upsurge in the number of defaulters when the credit amount surpasses 50,000. Interestingly, there is a decline in the quantity of defaulters when the credit amount lies within the range of 3,00,000 to 4,00,000. Furthermore, a higher number of defaulters is observed for credit values exceeding 700,000.

c. As for the "AMT_ANNUITY_VAR" customers with an annual annuity amount ranging from 10,000 to 50,000 demonstrate both a higher quantity of loans and a correspondingly higher number of defaulters.

CATEGORICAL UNORDERED VALUE







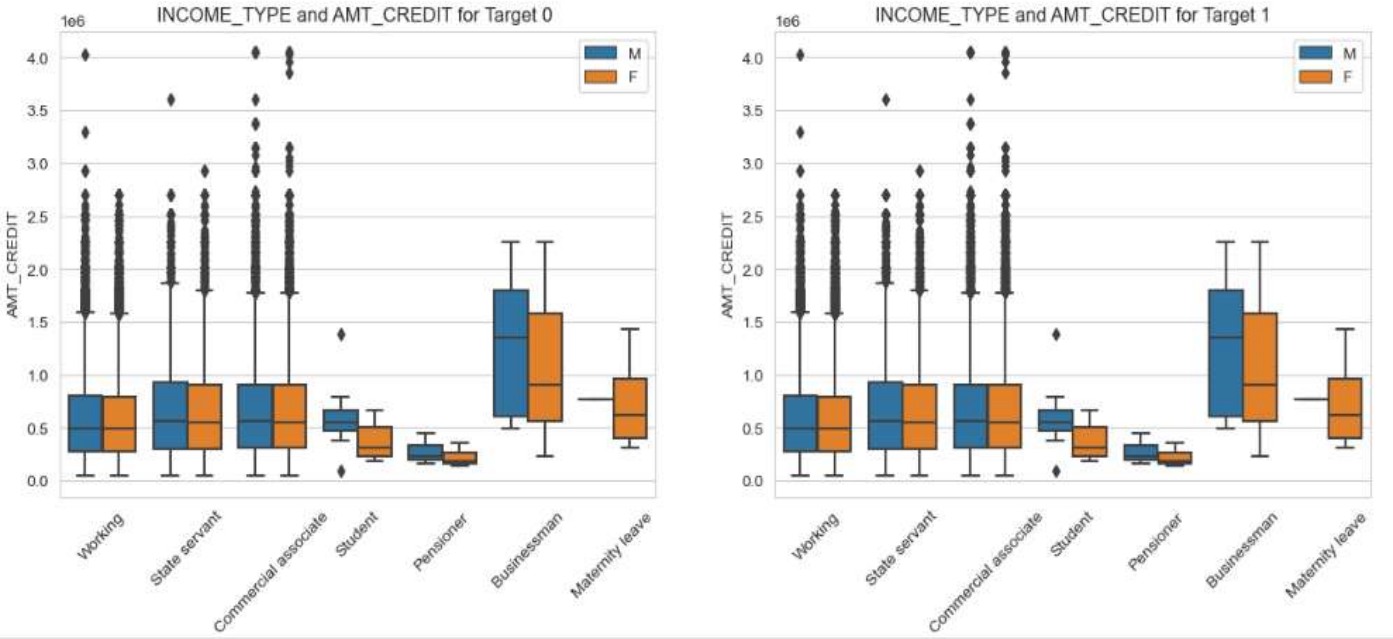
Observation-

a. Regarding the "CODE_GENDER" variable, there is a higher quantity of female customers compared to male customers. Additionally, the distribution of loans is significantly higher among female applicants compared to their male counterparts.

b. When considering the "NAME_INCOME_TYPE" variable, it is observed that students and businessmen exhibit a notably lower percentage of defaulters. On the other hand, working professionals, state servants, and commercial associates have a significantly higher default rate.

c. Analyzing the "NAME_CONTRACT_TYPE" variable, it becomes evident that the frequency of credits for the "Cash loans" contract type is higher than that of the "Revolving loans" contract type. Furthermore, the "Cash loans" category demonstrates a considerably higher number of defaulters compared to the 'Revolving loans' category.

BIVARIATE ANALYSIS FOR CATEGORICAL VALUES



Observation:

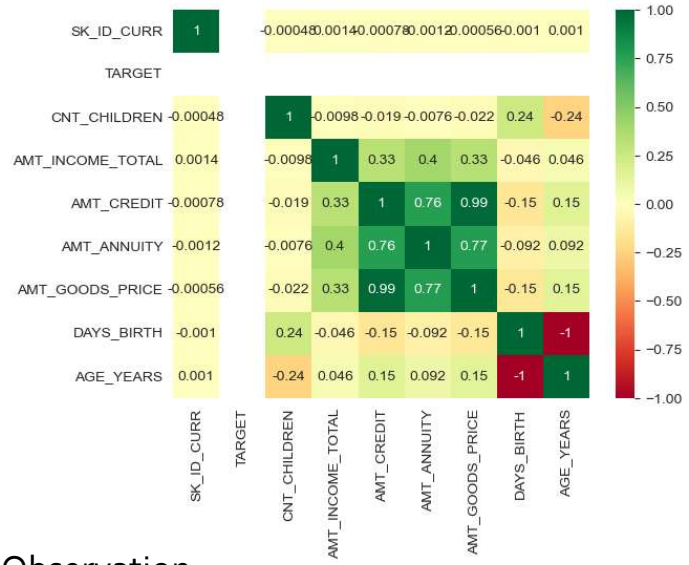
a. The plot clearly indicates that businessmen have the highest credit amounts compared to individuals in other professions, regardless of gender.

b. Commercial associates, working professionals, and state servants exhibit a greater number of outliers in terms of income, both among defaulters and non-defaulters, when compared to other income types.

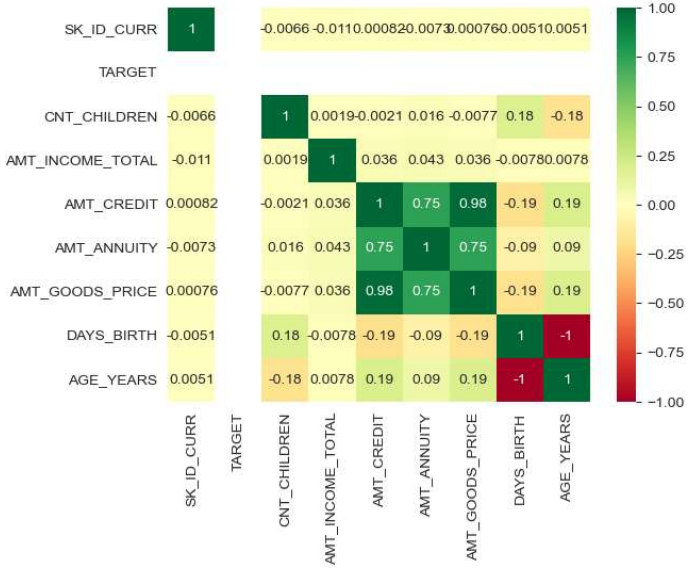
c. Specifically, female businessmen tend to have a higher number of credits in the third quartile, suggesting that they obtain larger loan amounts compared to other female applicants in different professions.

CORRELATION MATRIX

For Non Defaulter



For Defaulter



Observation-

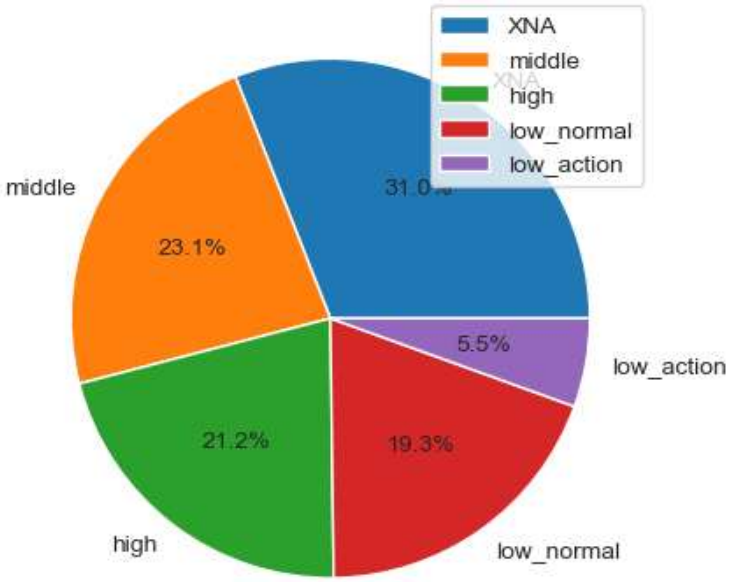
Based on the provided correlation table and heatmap, it is evident that there are several noteworthy correlations. Among them, the top three correlations are as follows:

Guarantor good price and credit amount: These two variables exhibit a strong positive correlation, indicating that as the guarantor's good price increases, the credit amount also tends to increase.

Credit amount and annuity amount: There exists a positive correlation between credit amount and annuity amount, suggesting that higher credit amounts are associated with larger annuity payments.

Guarantor good price and annuity amount: There is a correlation between the guarantor’s good price and the annuity amount, although the strength of this correlation may be lower compared to the first two mentioned. However, it is important to note that correlation does not imply causation, and further analysis is required to establish any causal relationships between these variables.

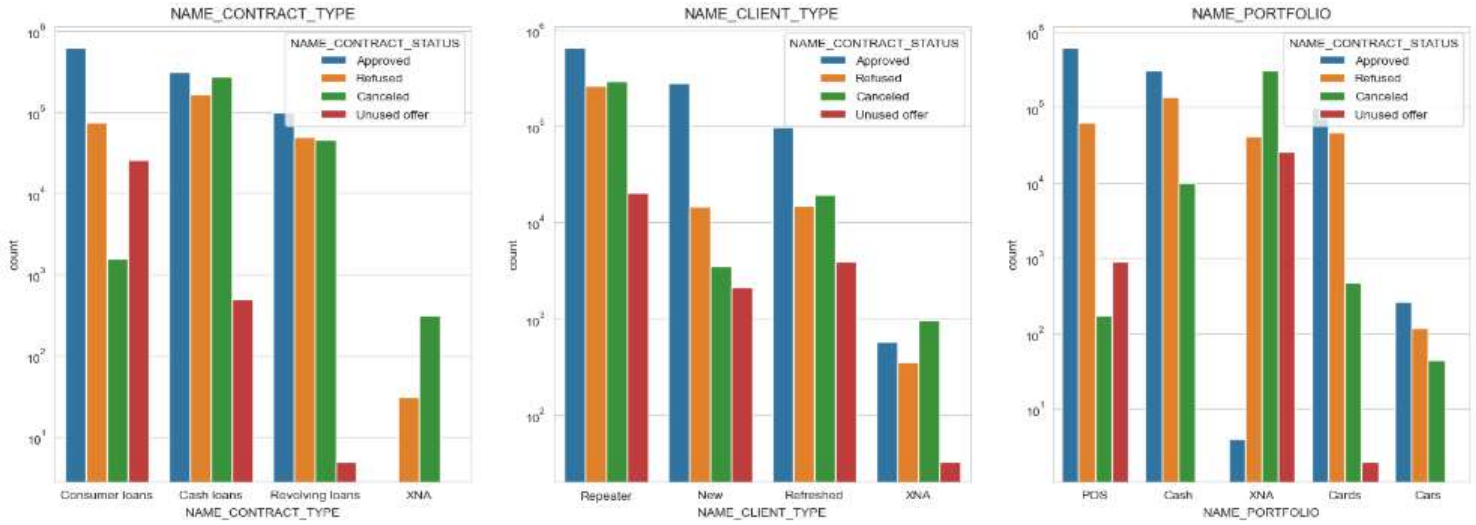
PERFORMING UNIVARIATE ANALYSIS FOR PREVIOUS APPLICATION DATA



Observation-

Based on the given information, it is stated that the middle rate of interest for loans is the highest among other types. This suggests that, in the context being discussed, the middle rate of interest stands out as the most elevated compared to alternative interest rates associated with loans.

BIVARIATE ANALYSIS



Observation-

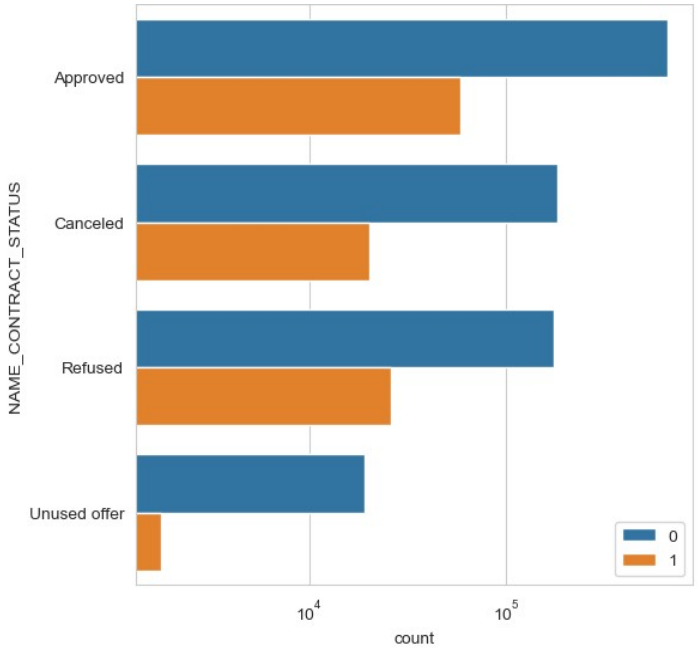
a. It can be observed that consumer loans and cash loans are among the loan types with the highest approval rates. This implies that a significant number of applications for consumer loans and cash loans have been successfully approved.

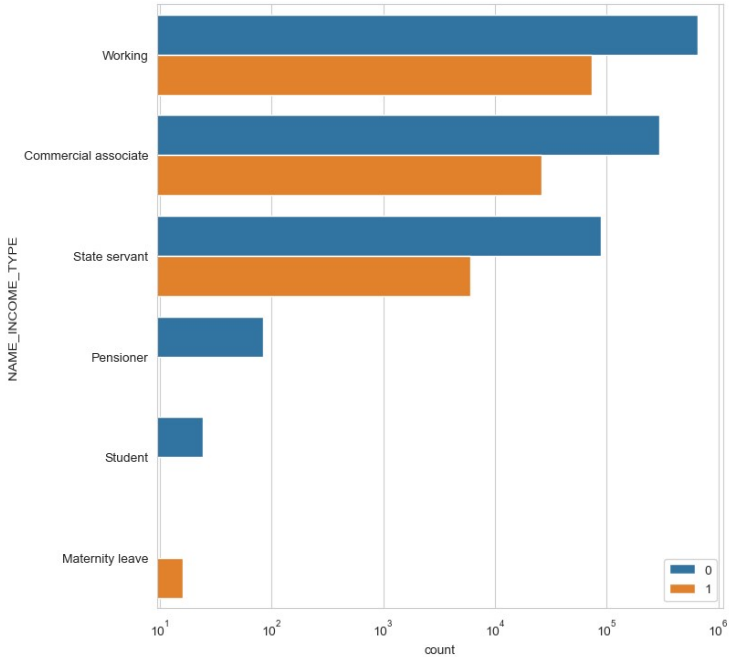
b. On the other hand, there is a greater number of loan refusals recorded for cash loans compared to consumer loans. This suggests that cash loans have a higher likelihood of being declined when compared to consumer loans.

c. Based on the provided graph, it is evident that there is a higher count of repeated customers compared to new customers. This indicates that the number of customers who have availed loans in the past and are returning for additional loans is higher than the count of new customers.

d. Additionally, the number of loan refusals for cash loans exceeds the number of refusals for POS (Point of Sale) loans. This indicates that a larger proportion of cash loan applications have been declined compared to POS loan applications.

AFTER MERGING ANALYSIS OF PREVIOUS APPLIED CUSTOMER WITH TARGET VALUE





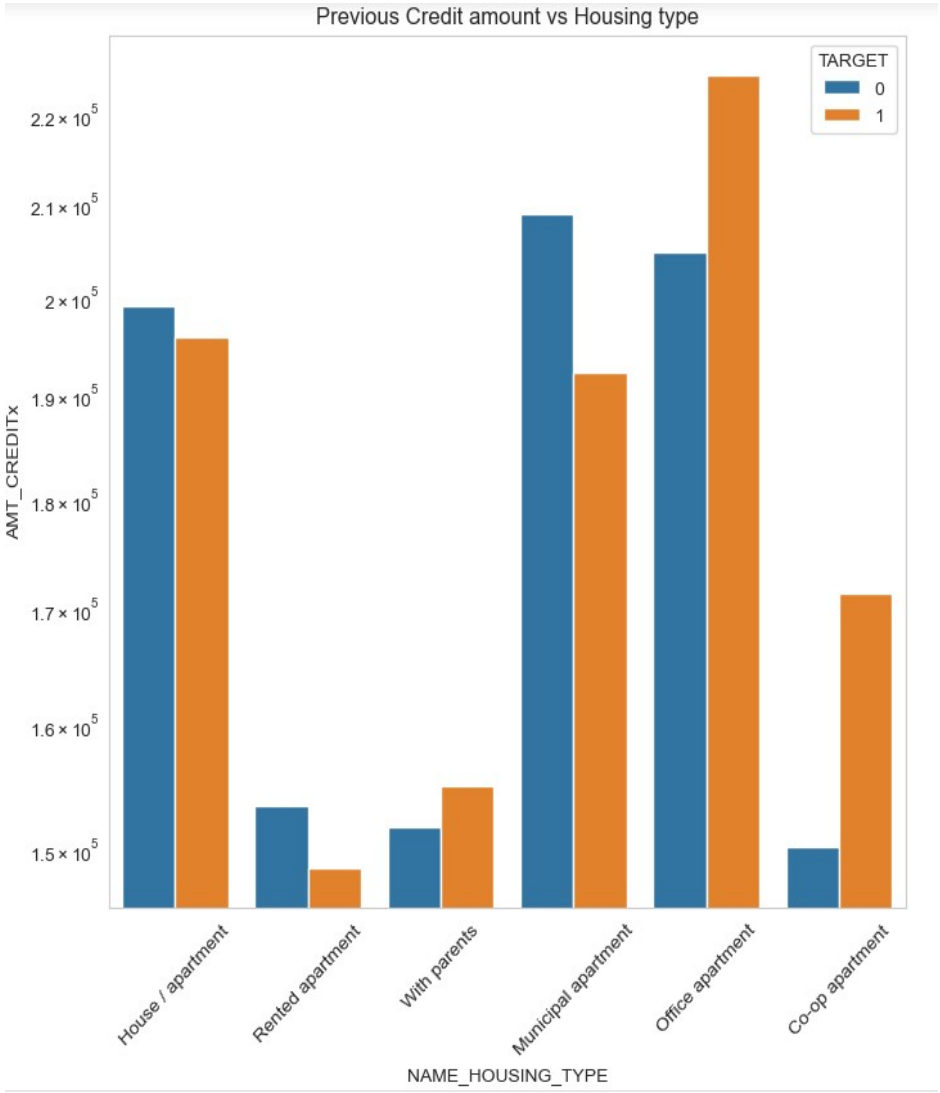
Observations-

a. It is evident that customers who reside in housing apartments or those who stay with their parents exhibit a higher proportion of non-defaulters. This suggests that individuals in these living arrangements have a lower likelihood of defaulting on their loans compared to customers with different housing situations.

b. Students and pensioners demonstrate a notably lower rate of defaulting on their loans. This indicates that individuals categorized as students or pensioners are less likely to default compared to other groups.

c. On the other hand, applicants who are on maternity leave show a higher incidence of defaulting on their loans. This suggests that individuals on maternity leave have a higher likelihood of encountering difficulties in meeting their loan obligations, leading to a higher default rate.

MERGED DATASET PREVIOUS CREDIT AMOUNT VS HOUSING TYPE



Observation-

a. Municipal and office apartments are associated with higher credit amounts for non-defaulters. This suggests that individuals residing in municipal and office apartments tend to have larger loan amounts when they have a good repayment history.

b. Among the different types of apartments, office apartments have the highest proportion of defaulters. This indicates that individuals living in office apartments are more likely to default on their loans compared to other types of apartments.

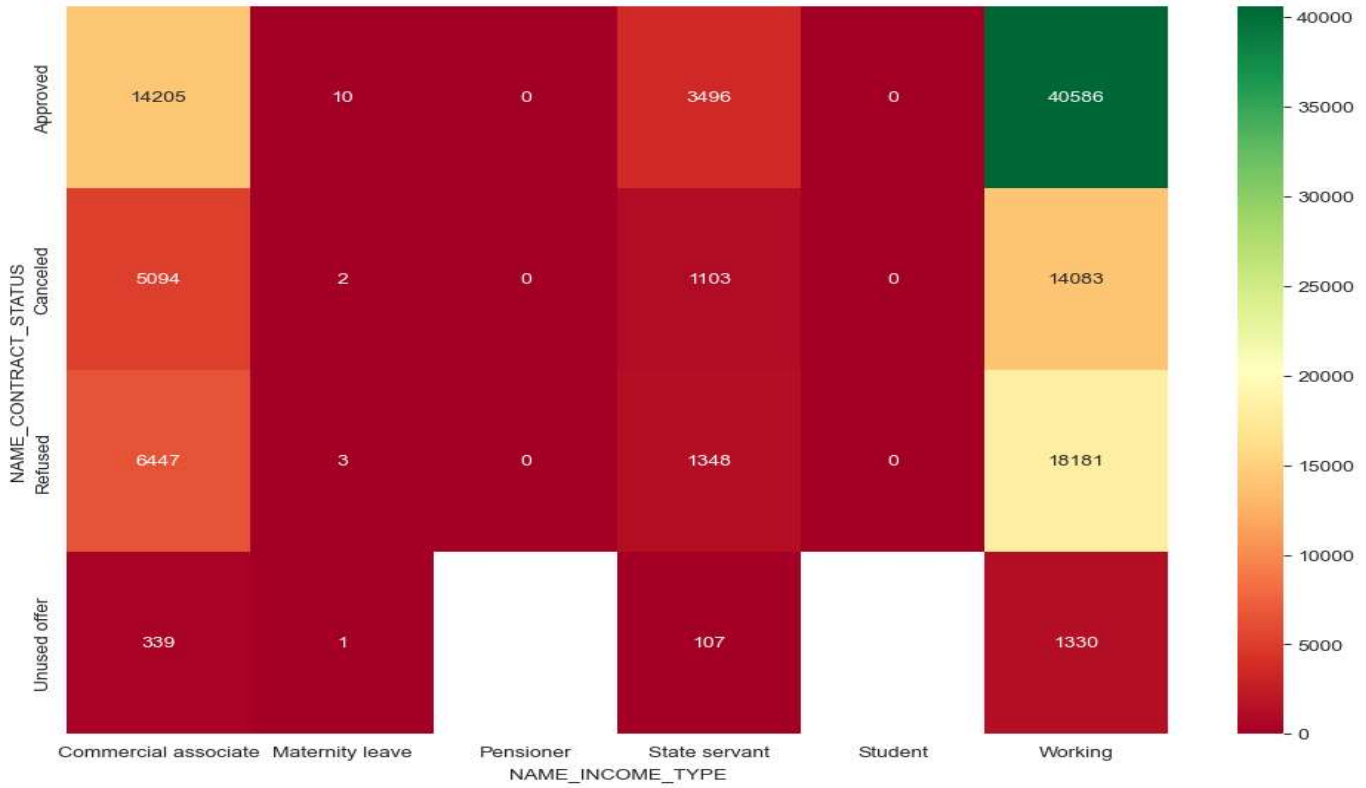
c. The default rate is significantly higher for co-op apartments compared to the non-defaulter rate. This implies that individuals residing in co-op apartments have a higher likelihood of defaulting on their loans.

d. Given the higher default rates observed for office apartments and co-op apartments, it is advisable for the bank to exercise caution and implement additional safety measures when granting loans to applicants from these categories.

e. To mitigate risks, the bank should prioritize its focus on house apartments and rented apartments. This suggests that individuals residing in these types of apartments are associated with lower default rates, making them potentially safer candidates for loan approvals.

CORRELATION MATRIX

CONTRACT_STATUS AND INCOME_STATUS WITH AGGREGATE FUNCTION ON TARGET



Observation-

a. The aforementioned matrix specifically pertains to instances where the TARGET variable is equal to 1, indicating cases of default.

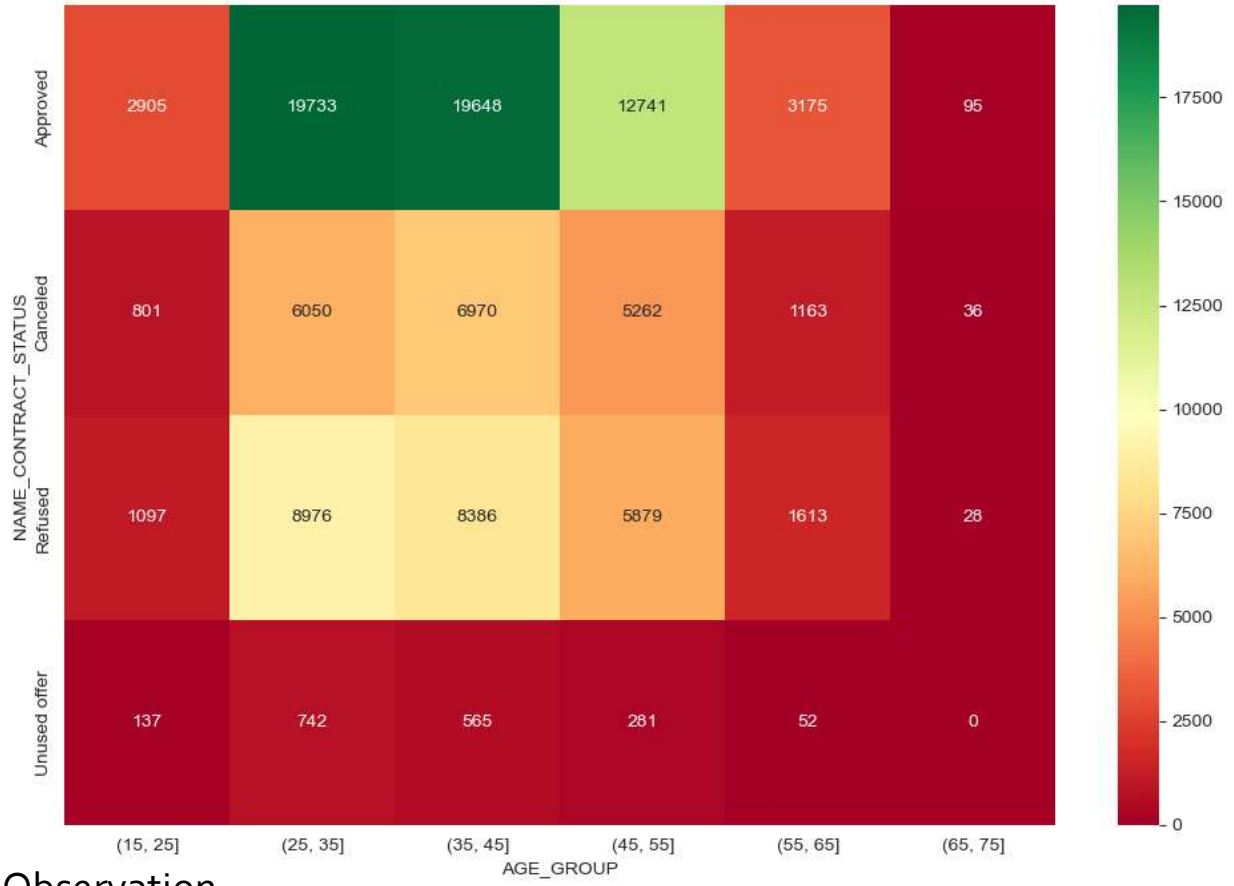
b. Among working applicants with an approved status, a considerable number have defaulted on their loans. This suggests that even among those with an approved loan status, working applicants exhibit a higher propensity for default.

c. Previous loan applications that were either refused, cancelled, or left unused also exhibit instances of default. This indicates that the history of previous loan applications, regardless of their outcome, can be indicative of potential default risk.

d. It is observed that customers who were previously part of the working class and had their loan applications refused are now more likely to be among the defaulters.

CORRELATION MATRIX

CONTRACT_STATUS AND AGE GROUP WITH AGGREGATE FUNCTION ON TARGET



Observation-

a. Approved loans belonging to the age groups of 25-35 and 35-45 demonstrate a higher incidence of default. This implies that individuals within these age ranges, despite having their loans approved initially, have a greater tendency to default on their loan obligations.

RECOMMENDATION

- The loan amount is observed to be lower for unused applications, and it becomes imperative to ascertain the underlying cause for this phenomenon.
- In view of the lower default rates associated with female applicants, it is advisable to accord them additional weightage during the evaluation process.
- A significant proportion of defaulters comprises individuals who are employed. However, this fact should not automatically warrant the rejection of working applicants; instead, a comprehensive assessment of other pertinent parameters is necessary prior to proceeding with their loan applications.
- In instances where a previous application exhibited refused or cancelled loans, it is essential to investigate the rationale behind the sanctioning of the current loan application with a defaulter option.
- Notably, previous applications featuring instances of refused, cancelled, or unused loans also encompass cases belonging to the non-defaulter category, where timely payments are being made in the current application. This observation strongly suggests the possibility of erroneous decisions having been made in those particular cases.
- Particular attention ought to be directed towards clients who reside in rented apartments or live with their parents, as these groups exhibit a markedly lower incidence of defaults compared to non-defaulters.
- Based on the aforementioned analysis, it is evident that the number of repeated customers surpasses that of new customers, underscoring the need for increased focus on nurturing and retaining existing clientele.

THANK YOU