

CREDIT EDA ASSIGNMENT

Prepared by: Gurjot Singh Gill

AGENDA

- Introduction
- Business Objective
- Problem Statement
- Analysis and Steps Taken
- Summary

INTRODUCTION

This assignment aims to give us an idea of applying EDA in a real business scenario. In this assignment, apart from applying the techniques that we have learnt in the EDA module, we will also develop a basic understanding of risk analytics in banking and financial services and understand how data is used to minimize the risk of losing money while lending to customers.



BUSINESS OBJECTIVE

This case study aims to identify patterns that indicate if a client has difficulty paying their installments which may be used for taking actions such as denying the loan, reducing the amount of the loan, lending (to risky applicants) at a higher interest rate, etc. This will ensure that the consumers capable of repaying the loan are not rejected. Identification of such applicants using EDA is the aim of this case study.

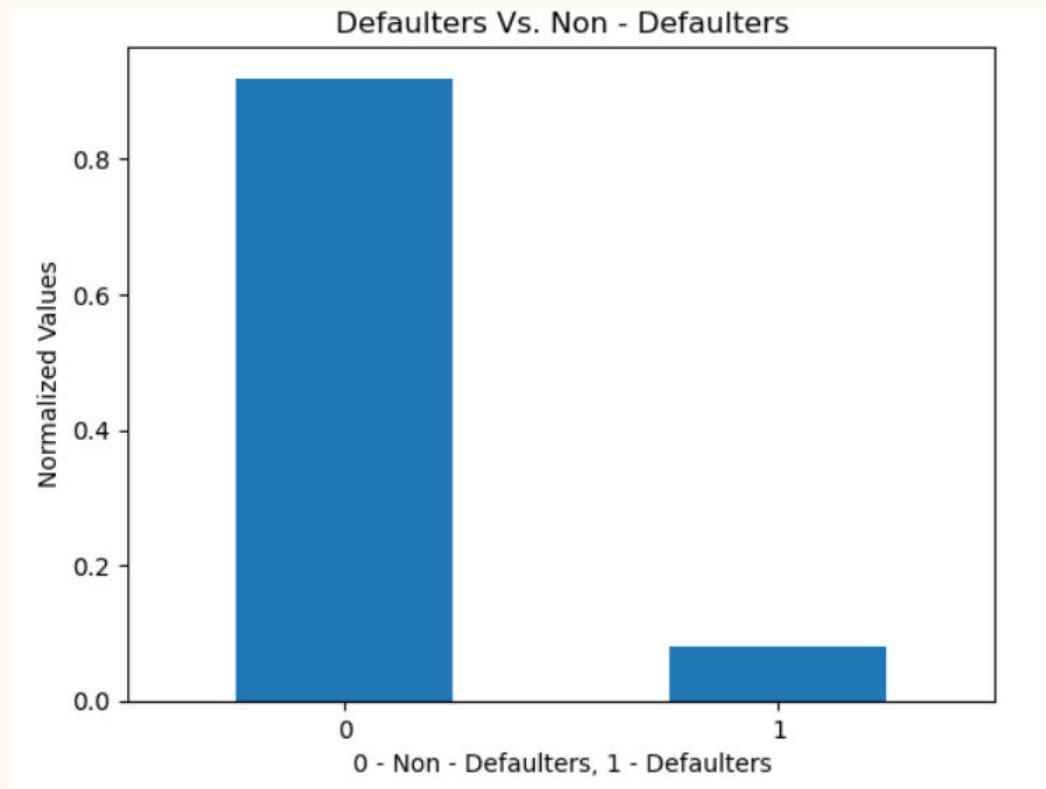
PROBLEM STATEMENT

- Understanding the driving variables behind loan defaults and finding out which variables are strong enough indicators of a future default.
- The data provided to us contains two datasets namely 'application_data' which has the information of the applicant's current applications information.
- The other dataset called 'Previous_application' contains the data pertaining to the previous applications of an applicant.
- The first data set 'application_data' contains two types of scenarios :
 - Client with payment difficulties: he/she had late payment more than X days on at least one of the first Y installments of the loan in the sample.
 - All other cases: Where all of the payment are made on time without any discrepancies.

ANALYSIS AND STEPS TAKEN

- Importing the dataset
- Cleaning the data:
 - Checking for missing/null values in the dataset.
 - Handling missing values (Imputing/removing them).
 - Dropping the columns that did not pose a lot of significance.
- Univariate Analysis:
 - Analysis on Categorical Ordered data
 - Analysis on Categorical Unordered data
 - Analysis on Numerical Data
- Bivariate Analysis (on the same categories as above)
- Multivariate Analysis

ANALYSIS ON THE TARGET VARIABLE

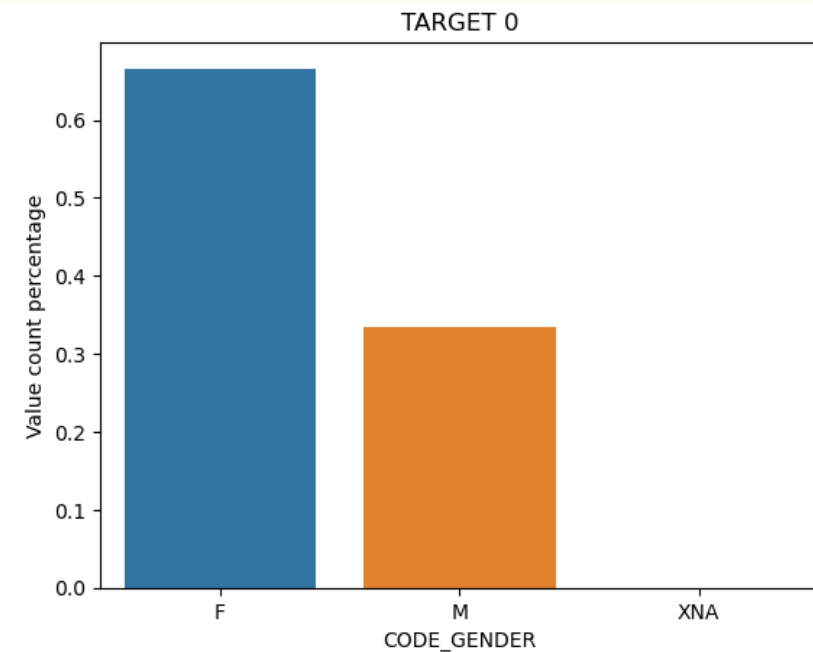
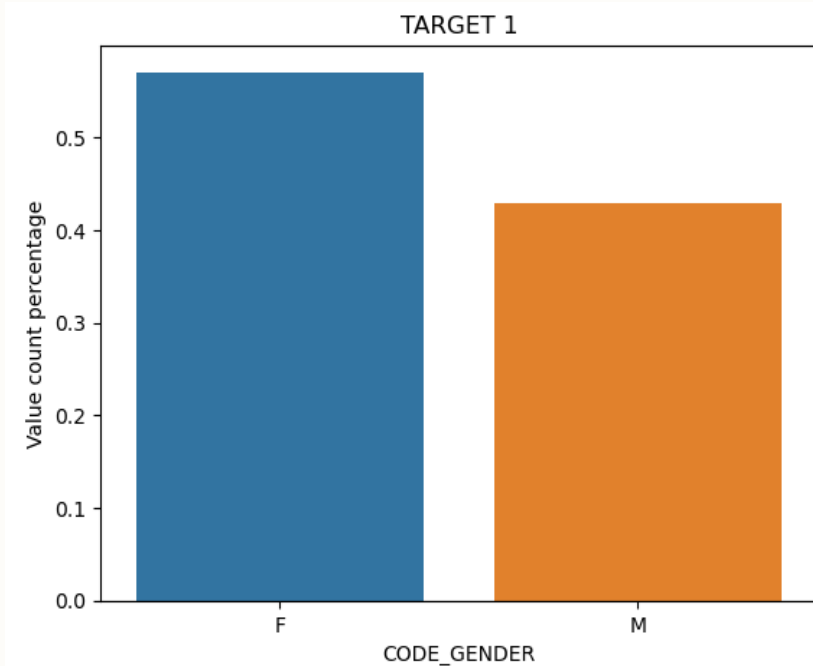


Inference: As we see in the graph, 90% of the applicants are non defaulters and 10% are defaulters.

UNIVARIATE ANALYSIS

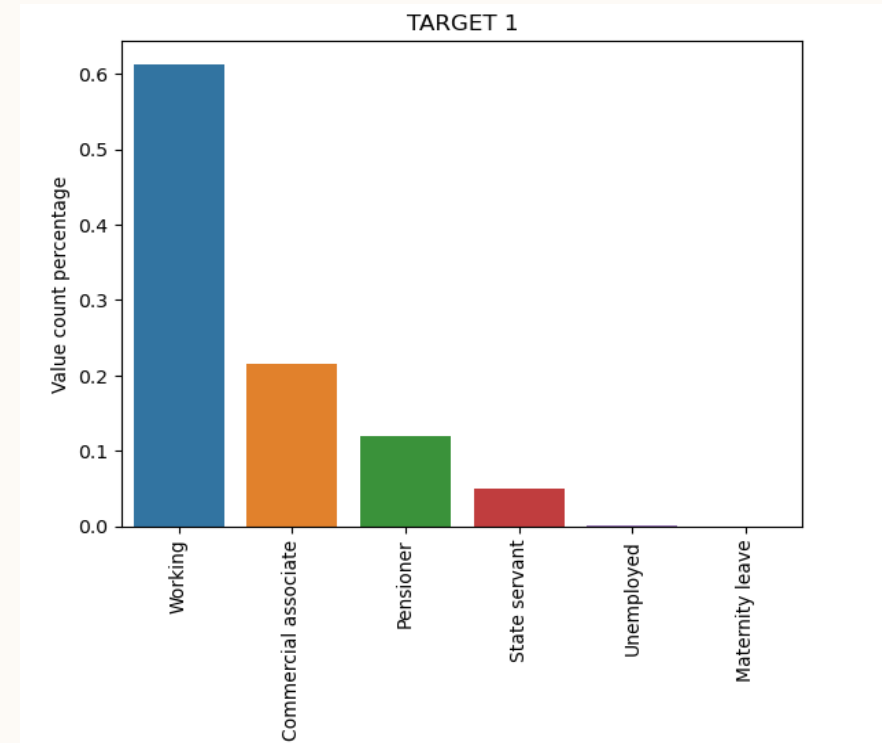
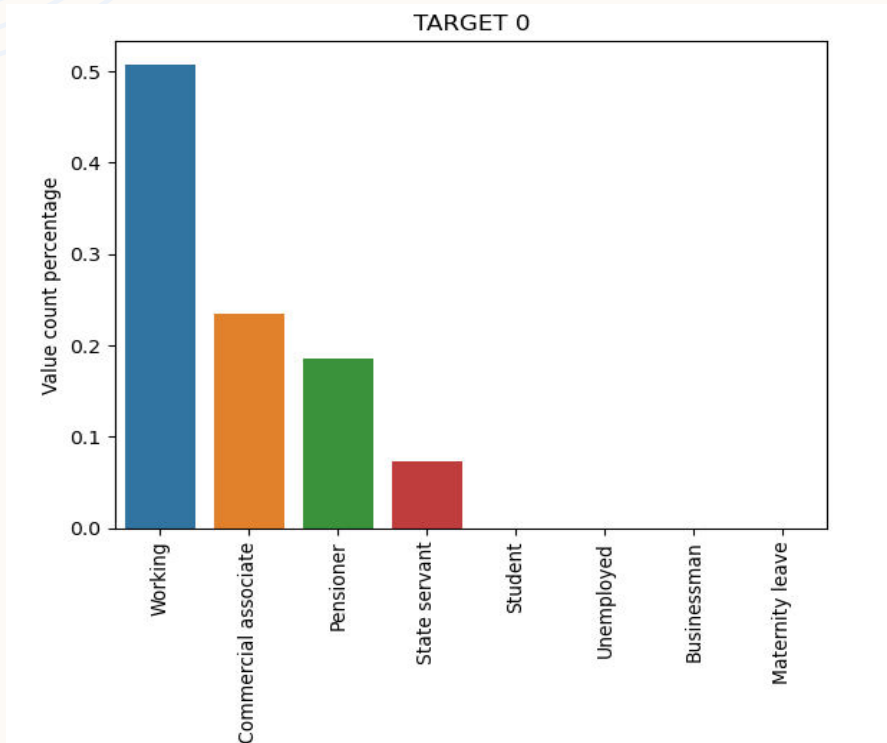
ON CATEGORICAL DATA (UNORDERED)

GENDER



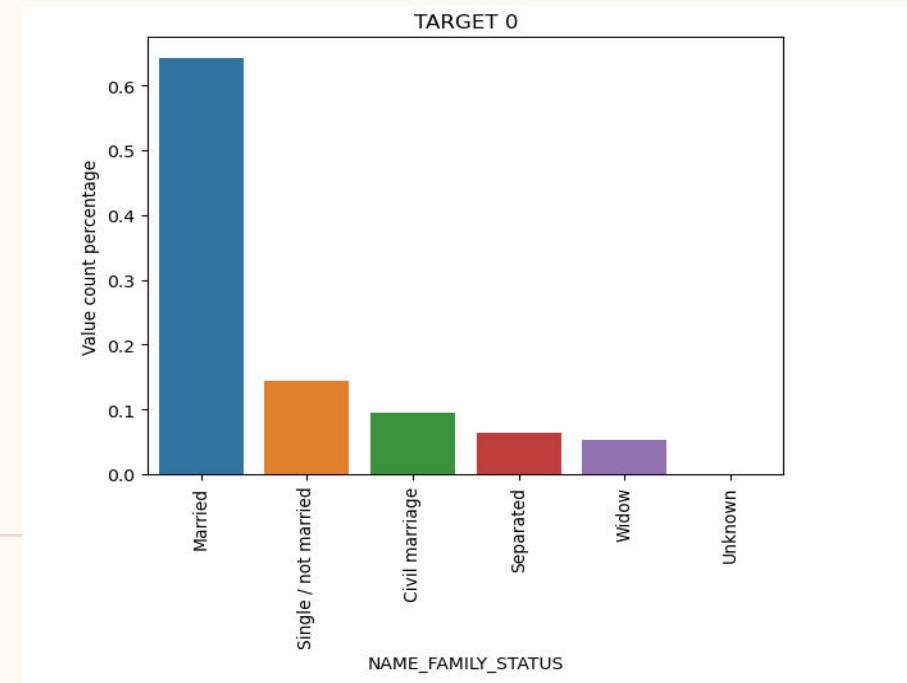
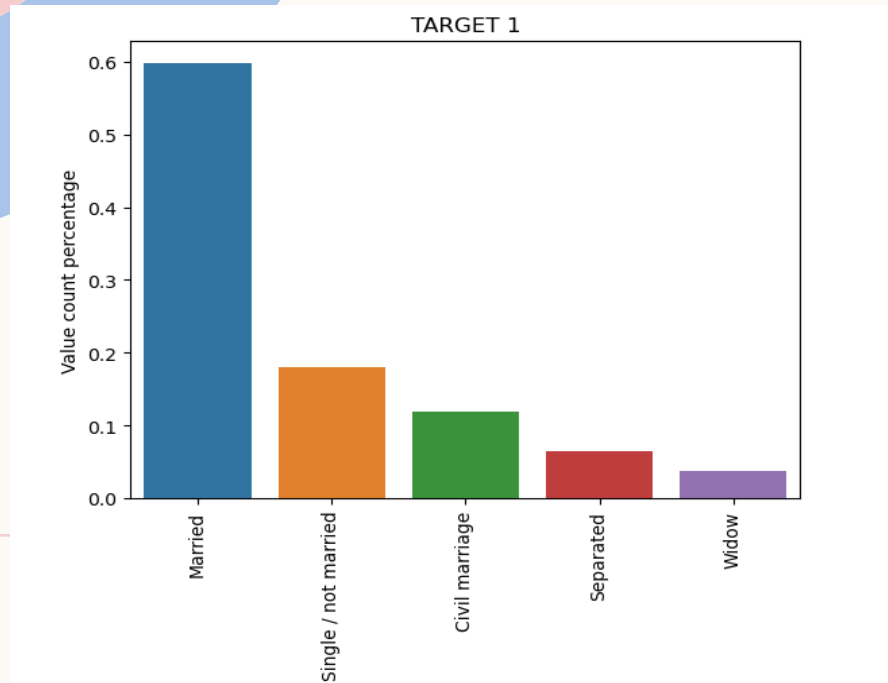
Both the cases have a Gender majority of Female population but, there is an increase in the Male population in the TARGET 1 category (Defaulters).

INCOME TYPE



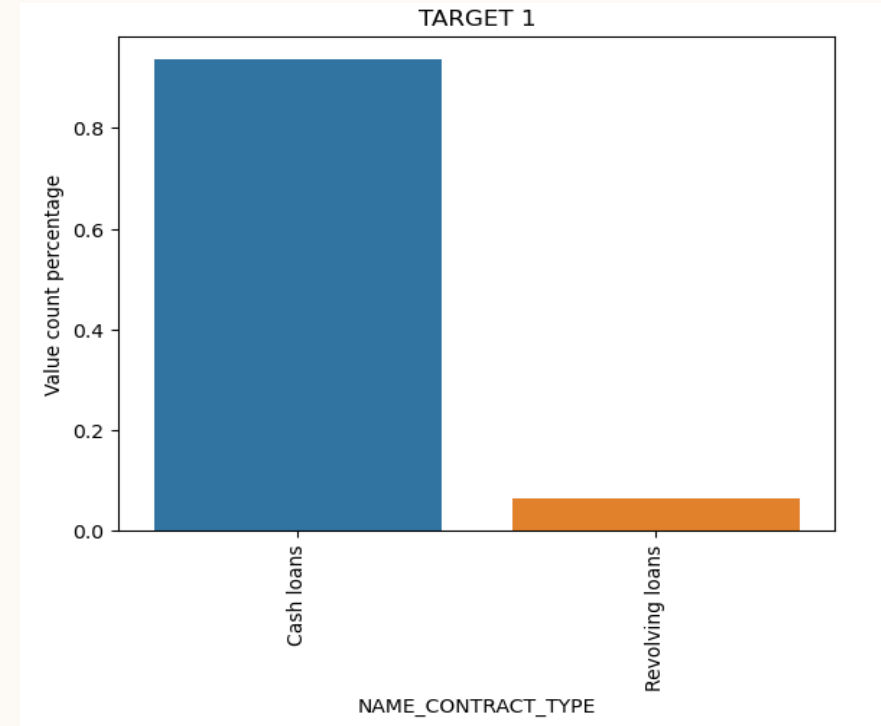
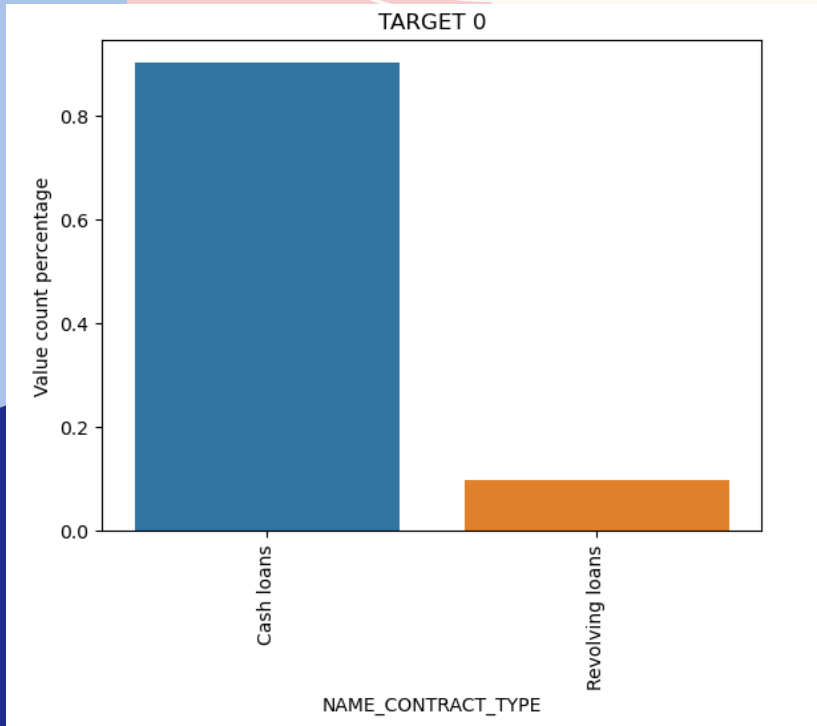
- Both the TARGET values have the same highest Income type.
- If you take a look at the other Income types, you will notice that the 'Pensioner' is higher in case of TARGET 0 (Defaulters).

FAMILY STATUS



- Single and Civil marriages increase in case of the Defaulters variable when compared to Non Defaulters. This indicate that Single and Civil marriages applicants are prone to default.
- However, there is not a lot of difference between Married applicants. Hence, it is safe to sanction loans to married applicants.

CASH LOANS VS REVOLVING LOANS

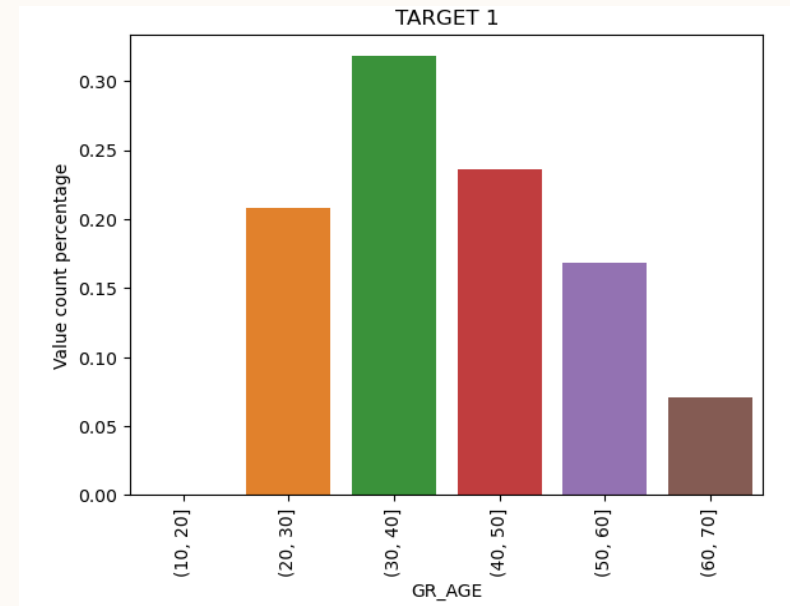
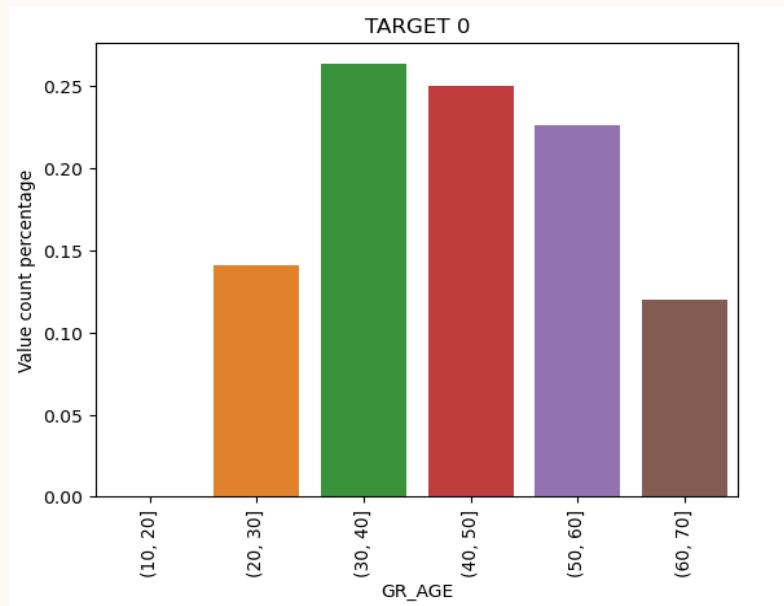


- The majority of the people have chosen Cash loans over Revolving loans. 85% in Target 0 and 95% in Target 1

UNIVARIATE ANALYSIS

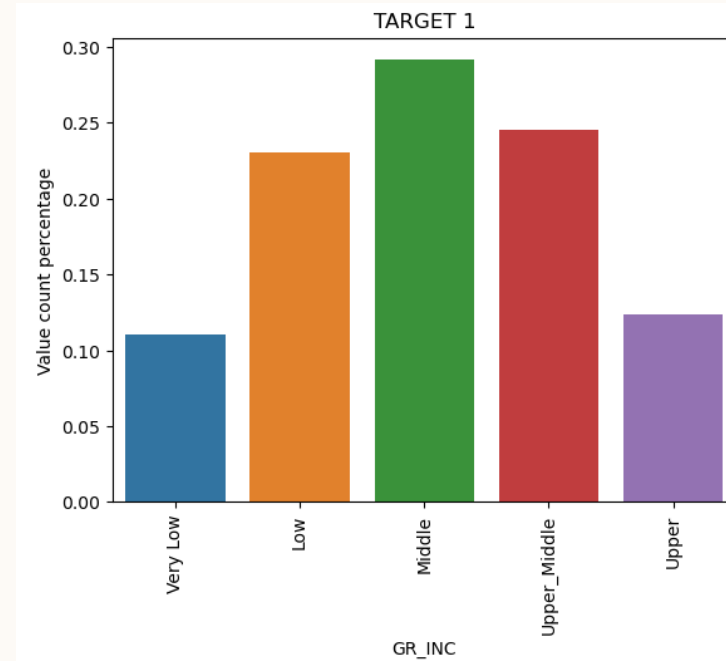
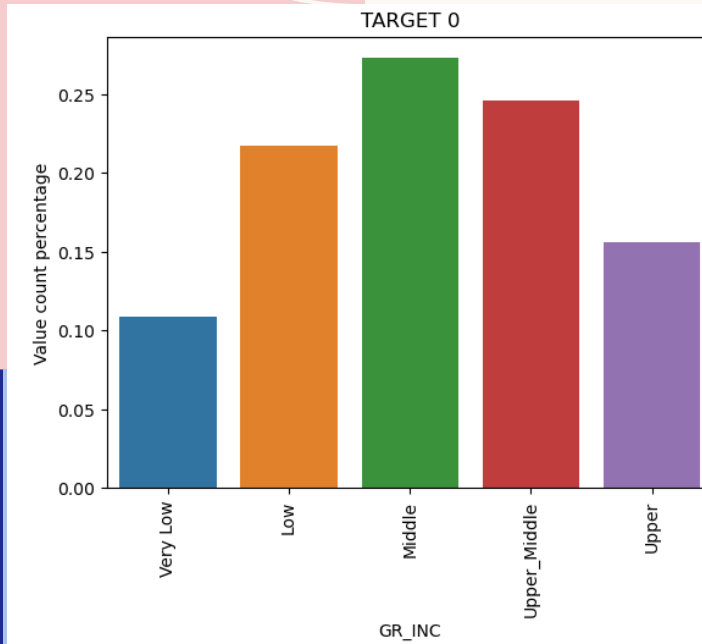
ON CATEGORICAL DATA (ORDERED)

AGE GROUP



- The age group 30 to 40 have applied for the most number of loans but have difficulties paying them back.
- The count of the age group 40 to 50 is more in case of defaulters than non defaulters.

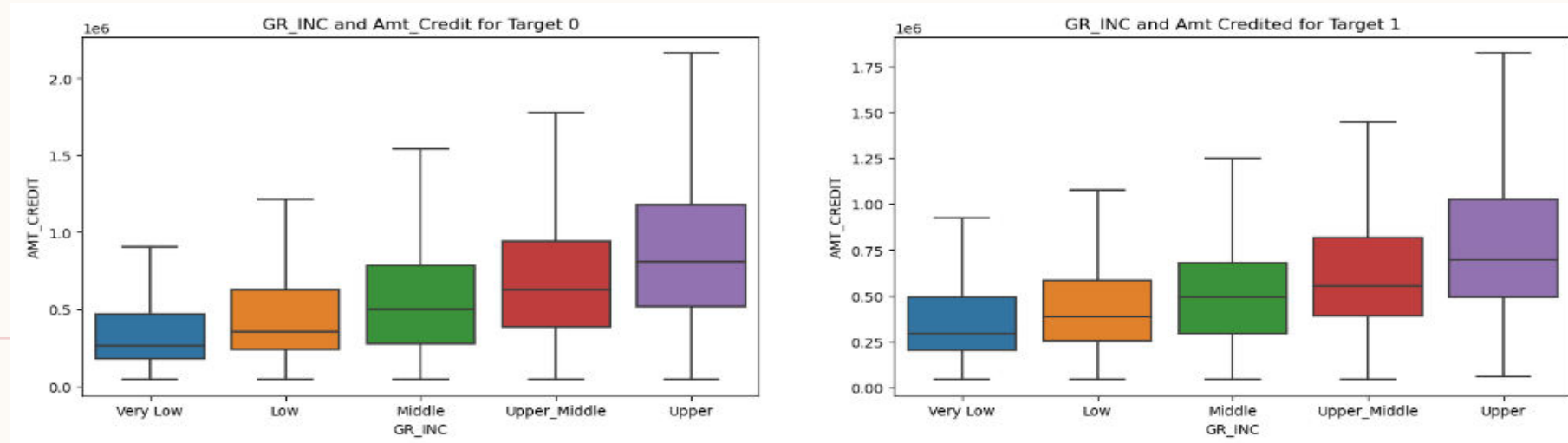
INCOME GROUP



- Middle income group is the largest category in both the Target variable.

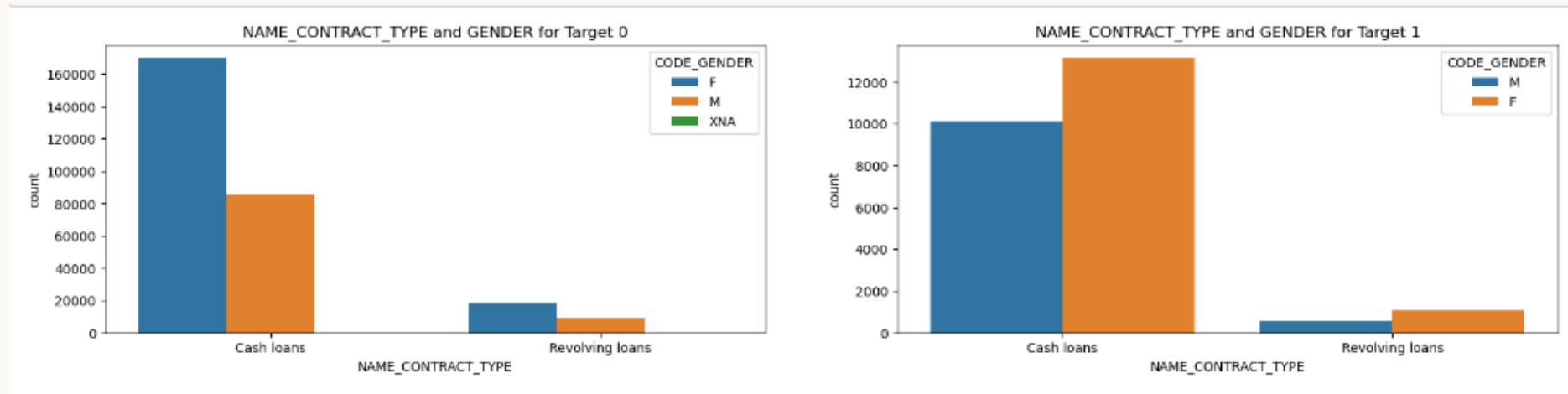
BIVARIATE ANALYSIS

CATEGORICAL AND NUMERICAL DATA



- For both the targets, the IQR increases with the income group
- - The outliers are more extreme and frequent for the Upper group, suggesting that some borrowers in this group have exceptionally high or low credit amounts.
- - The median credit increases with the increase in income group for both the targets

CATEGORICAL AND CATEGORICAL DATA



- For both the targets, there are significantly more cash loans than revolving loans.
- For both the targets, females have more loans than males
- The difference between males and females is almost similar signifying that males have difficulty in paying back the loans
- Females have no difficulties paying back the revolving loans



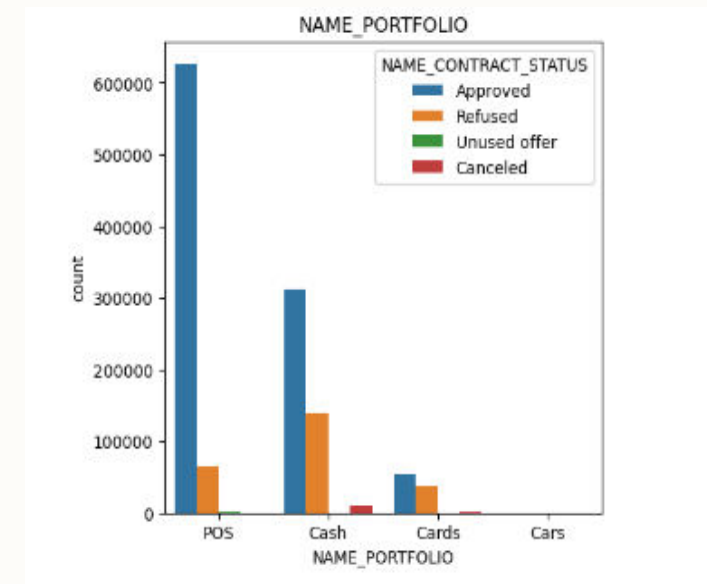
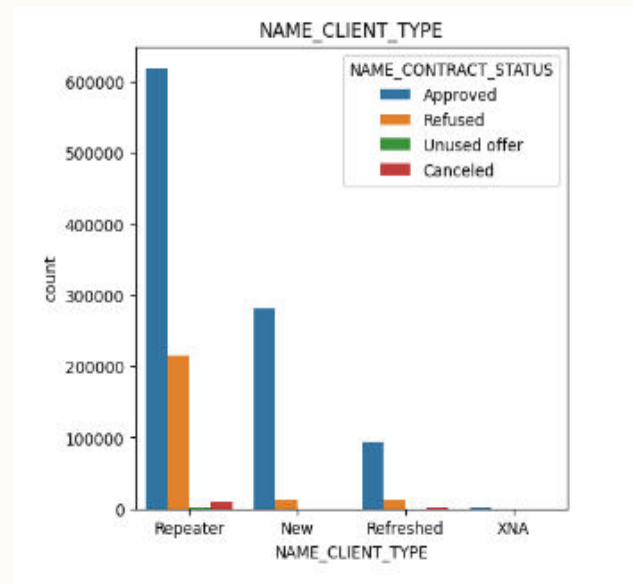
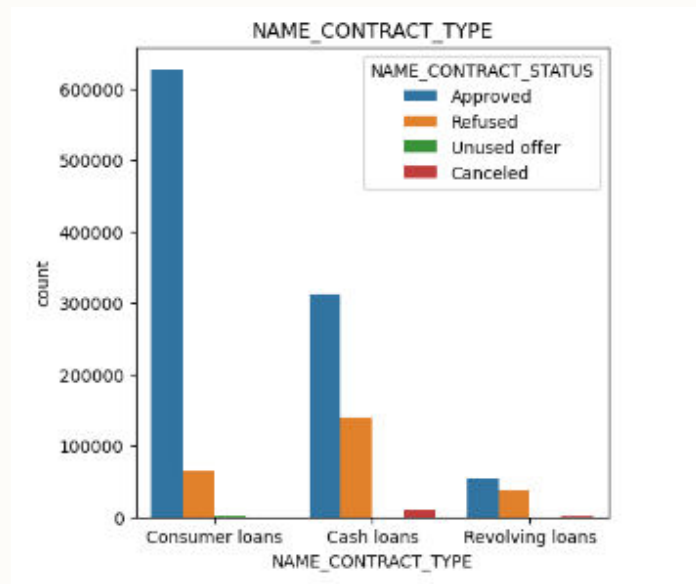
PREVIOUS_DATA DATASET

ANALYSIS AND STEPS TAKEN

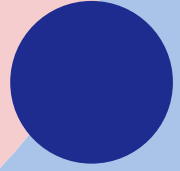
- Importing the dataset
- Cleaning the data:
 - Checking for missing/null values in the dataset.
 - Handling missing values (Imputing/removing them).
 - Dropping the columns that did not pose a lot of significance.
- Univariate Analysis:
 - Analysis on Categorical Ordered data
 - Analysis on Categorical Unordered data
 - Analysis on Numerical Data
- Bivariate Analysis (on the same categories as above)
- Multivariate Analysis

BIVARIATE ANALYSIS

THE STATUS OF THE APPLICATION



- Consumer loans are most popular and common types of loans availed by applicants
- Consumer loans have the highest rate of approval
- Cash loans have the highest rate of refusals
- Repeaters have a high chance of getting their loans approved
- While repeaters have a highest chance of refusal as well among others, New applicants have a good chance of their loan getting approved too
- POS is the most dominant portfolio among all the others.
- It has the highest approval rate followed by Cash portfolio. However, Cash portfolio has the highest rate of refusals as well among all the other portfolios.



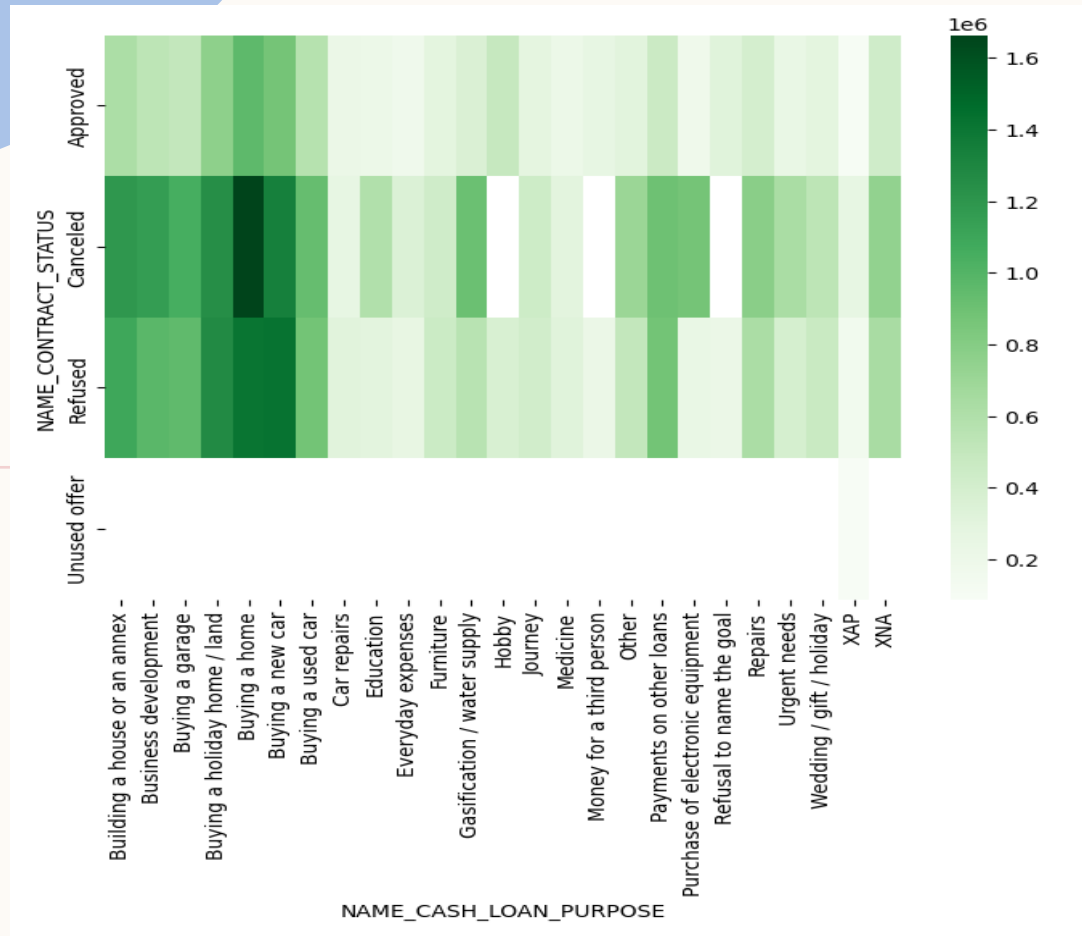
MULTIVARIATE ANALYSIS

NAME_CONTRACT_STATUS, NAME_CLIENT_TYPE AND AMT_CREDIT



- For some unknown reason XNA has the highest credit amount in the 'cancelled' contract
- Unused offer has the minimum credit amount
- Cancelled and refused have the highest and the second highest credit amount whereas approved should be the one having more credit amount

NAME_CONTRACT_STATUS, NAME_CASH_LOAN_PURPOSE AND AMT_CREDIT



- The purposes 'Hobby', 'Money for a third person' and 'Refusal to name the goal' have low to none credit amount. That could be the reason of the cancellation of their application
- Not sure why Buying a house with good credit amount is Cancelled/Refused
- Building a house or an annex has the highest number of approved loans
- Buying a new car or buying a used car are not very influential in making a credit decision.



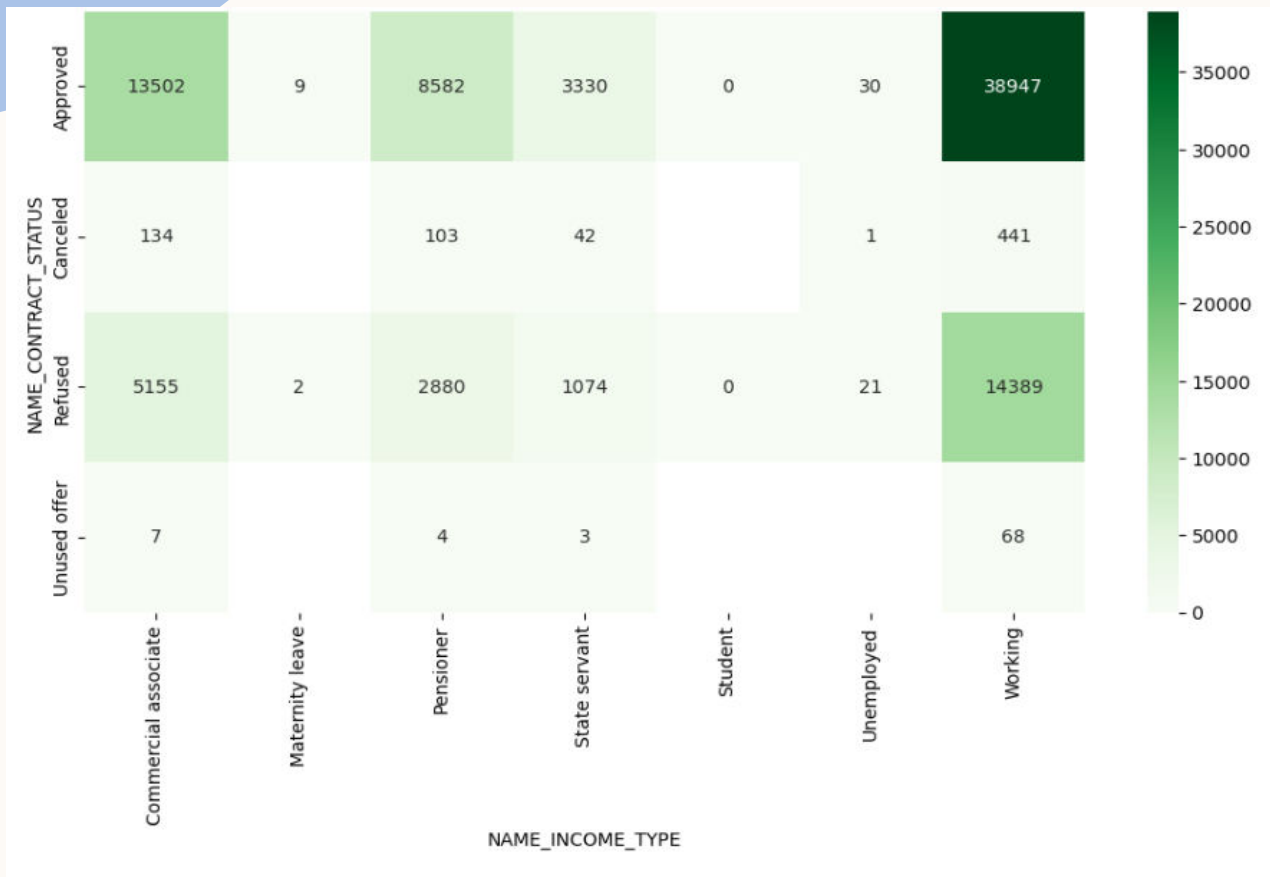
MERGING THE DATASET

ANALYSIS AND STEPS TAKEN

- Importing the dataset
- Cleaning the data:
 - Checking for missing/null values in the dataset.
 - Handling missing values (Imputing/removing them).
 - Dropping the columns that did not pose a lot of significance.
- Multivariate Analysis

MULTIVARIATE ANALYSIS

NAME_CONTRACT_STATUS, NAME_INCOME_TYPE AND THE TARGET COLUMNS USING A HEATMAP



- The Working class people have the maximum loans approved and are the highest defaulters among other Income types
- 14,389 working class people were refused before and now have defaulted
- Commercial associates come second in the highest defaulters list.

NAME_CONTRACT_STATUS, GR_AGE AND THE TARGET COLUMNS

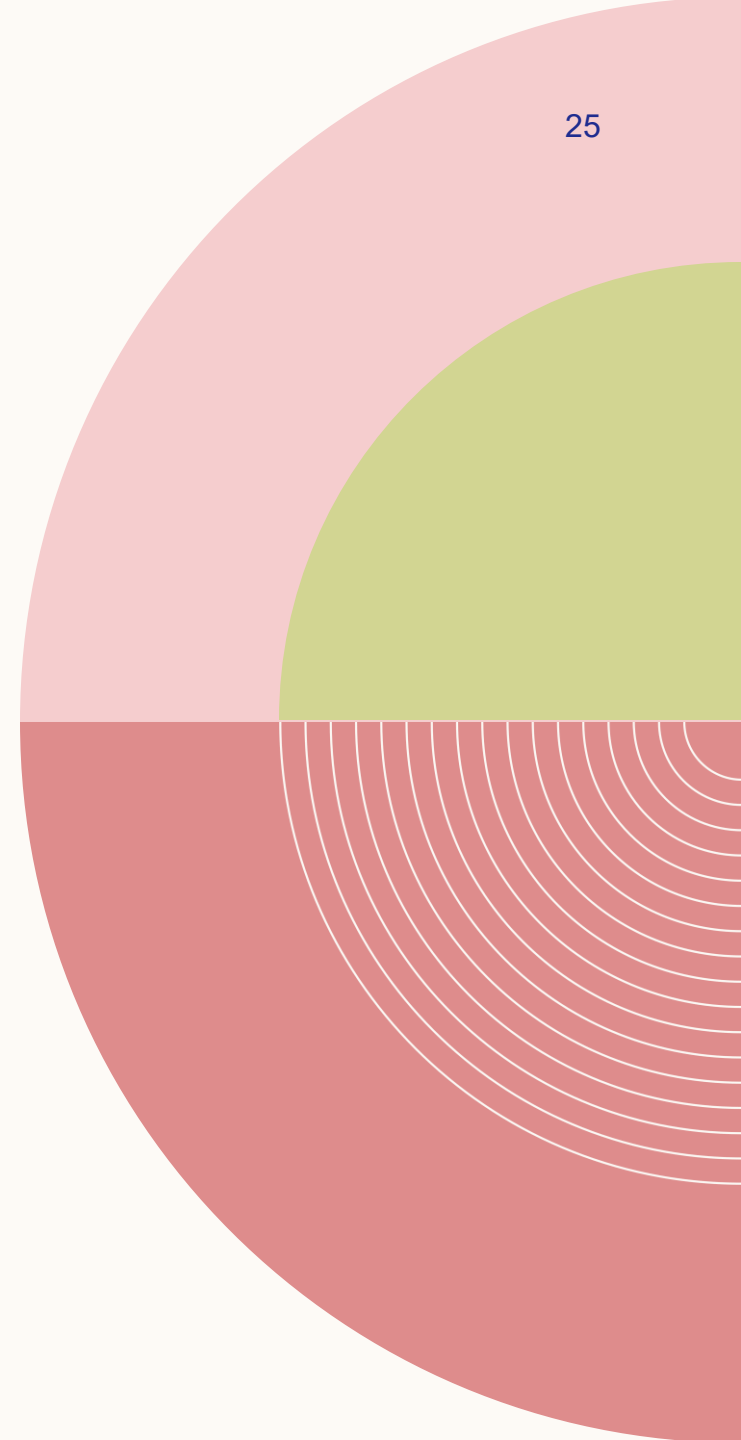
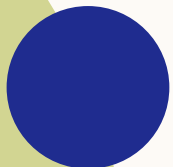


- The age group of 20 to 30 and 30 to 40 have the highest approved loans and have defaulted the highest as well.
- Similarly 30 to 40 and 40 to 50 age group has been refused of a loan and have defaulted once their loan got approved.
- Surprisingly, 11,766 applicants from 20 to 30 age group have got their loans approved and was the least age group to have their application refused.

SUMMARY

A FEW POINTERS TO KEEP IN MIND BEFORE LENDING OUT A LOAN

- Defaulters usually belong to the 'Middle' income group.
- Defaulters belong to the 'Working' income type. That does not mean they are not capable of paying back the loan, a proper scrutiny is needed.
- The age group that defaults usually belong to is between 30 to 50 years of age.
- People who don't have a strong educational background usually land as defaulters in the future.
- Marital status does effect the individual's financial situation. According to the data, Married individuals are the highest defaulters.





THANK YOU