



EDA - Case Study on Loan Application

Submitted by Nikhil M. Kanchan

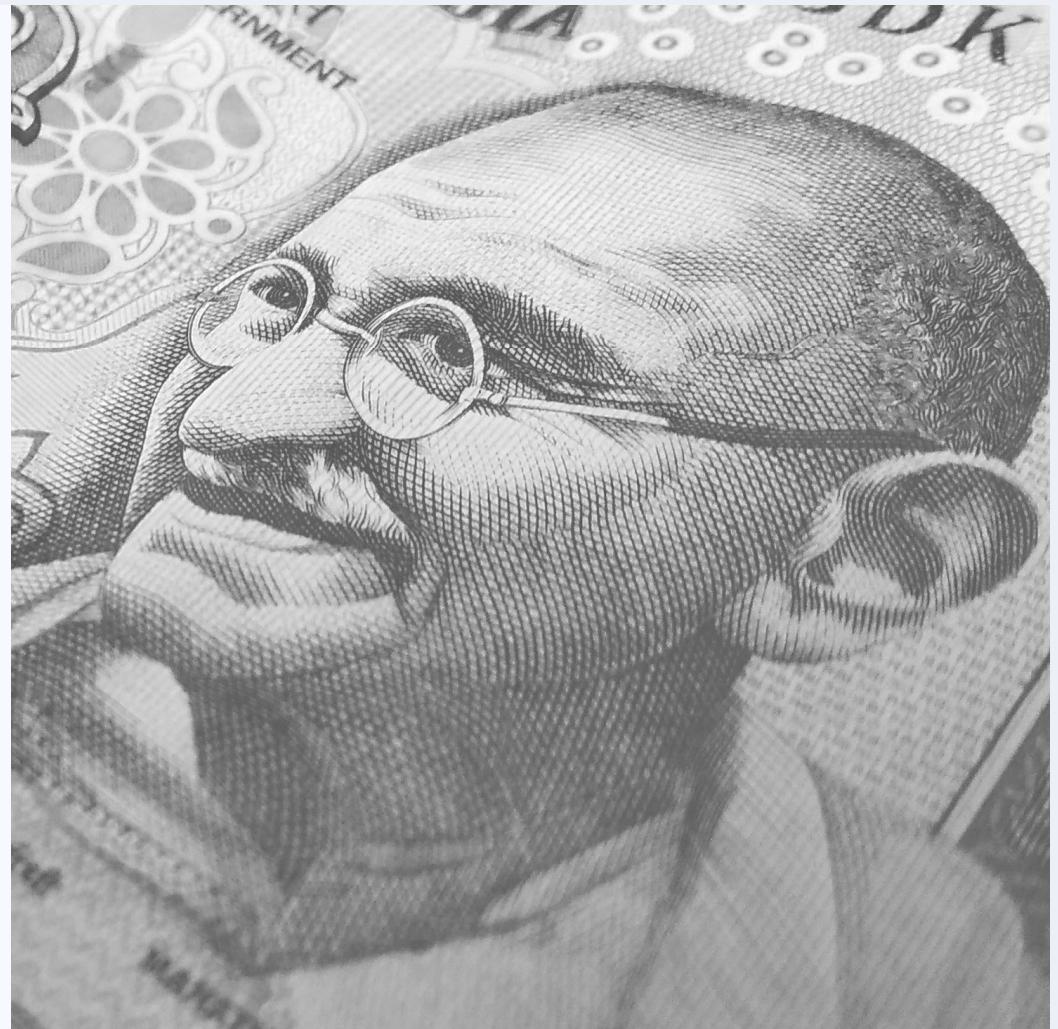
Date 5th May 2021

OUTLINE

1. Objective
2. Problem Statement
3. Exploratory Data Analysis
4. Process Followed
5. Major Inferences
6. Preferred Customers
7. Alert
8. Other Stats
9. Important Variables

OBJECTIVE

As part of a team studying lending process for various types of loans to urban customers; the objective here is to understand various factors that drive customers to get different type of loans and provide recommendation to Business to drive the approval process for the loan application.



Problem Statement

- ▶ The loan providing companies find it hard to give loans to the people due to their insufficient or non-existent credit history. Because of that, some consumers use it as their advantage by becoming a defaulter.
- ▶ As a Data Analyst, the requirement for me is to use EDA process to analyze the patterns present in the data and make recommendations.
- ▶ When the company receives a loan application, it has to decide for loan approval based on the applicant's profile drilled through various recommendations made in this document.
- ▶ The recommendation made should ensure that the following points are always adhered to:-
 - ▶ If the applicant is likely to repay the loan, then not approving the loan results in a loss of business to the company
 - ▶ If the applicant is not likely to repay the loan, i.e. he/she is likely to default, then approving the loan may lead to a financial loss for the company.

EXPLORATORY DATA ANALYSIS

Exploratory Data Analysis or the EDA is an approach of analyzing data sets to summarize their main characteristics, often using statistical graphics and other data visualization methods. A statistical model can be used or not, but primarily EDA is for seeing what the data can tell us beyond the formal modeling or hypothesis testing task.

The Purpose of EDA is to:-

- Maximize the insight in the data set
- Detect outliers and anomalies
- Test underlying assumptions

Based on the results of EDA, companies also make business decisions, which can have repercussions later. Hence, we observe the following:

- If not performed properly, EDA can hamper the further steps in the machine learning model building process.
- If done well, it may improve the efficacy of all we do in the next steps.



Process Followed

- Data sourcing
- Data cleaning
- Univariate analysis
- Bivariate and multivariate analysis

MAJOR INFERENCES

- People prefer the client mainly for Cash Loan.
- Current customer base of Client consists of following category of people:-

Female	People not owning Cars
People with secondary education	Mid income range people
House/ Apartment Owners	Existing Clients
Married People	Working professionals

- New Targets –
 - People with Academics background or Higher Education are generally the one with higher Income.
 - Males have higher salary but are not applying for loans in Clients Bank.

MAJOR INFERENCES

- More scrutiny is recommended for the applicants with following characteristics:-
 - Young people with Age < 35
 - People with less number of months employed (Low factor)
 - People living in less populated areas (Low factor)
- Younger people or the one below the age of 35 are more likely to face payment difficulty as compared to people with higher age.
- People in early 30's, Mid 40's and late 50's are more likely to apply for loan - this can be linked to important stages of a person's life - initial settlement, buying a house and education expenses for kids.
- Payment difficulty is faced less by applicants with higher number of Months Employed.
- Majority clients for the bank lives in area where population is less however people living in higher population regions may face less payment difficulties.

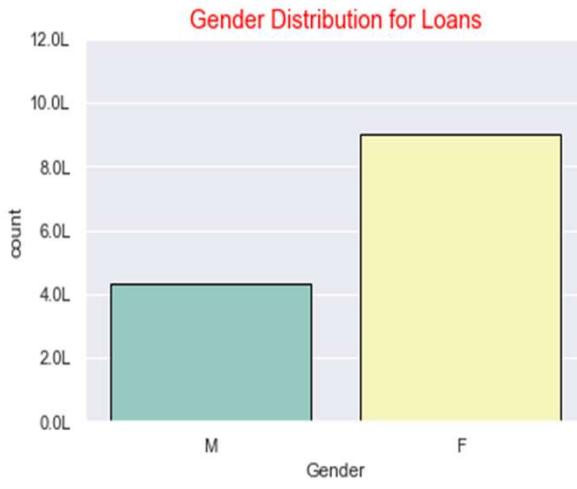
MAJOR INFERENCES

- Decisions should never be made in haste. It is noticed that applications with higher processing time have less chances of payment difficulty.
- Bank is already doing good in terms of handling applicants with chances of payment difficulty with applicants with payment difficulties less than 9%.
- Very less application processing is done on Sundays & Saturdays. It looks like the bank used to be aggressive on Saturdays in the past.
- The major source for the application is Cash & Credit offices of the client. Other channels are not that effective.
- Credit Amount, Annuity Amount & Goods Price have high positive correlation.
- Social surrounding has very high correlation for 30 DPD and 60 DPD. However, there is no such observation of increase in payment difficulty for people with social surrounding where the 30/ 60 DPD is high.

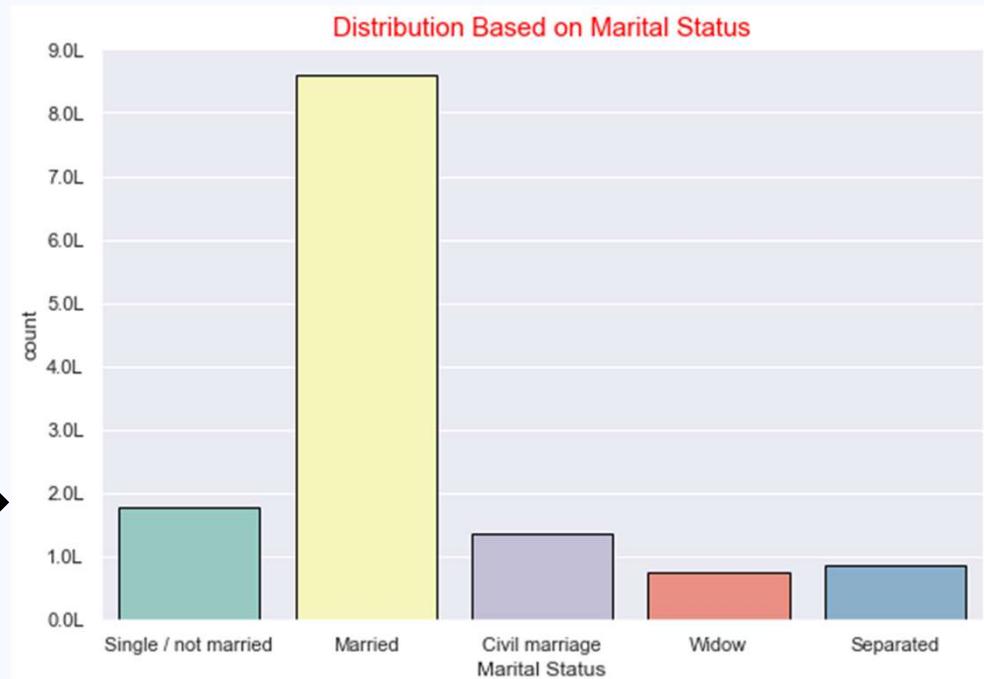
MAJOR INFERENCES

- There is a very high correlation between the historic data for Amount application & Amount Credited. This means that bank normally approves the loan for the requested amount.
- Payment Difficulty is not dependent or very less driven by Income, Amount Annuity or Amount Credit.

Preferred Customers - I

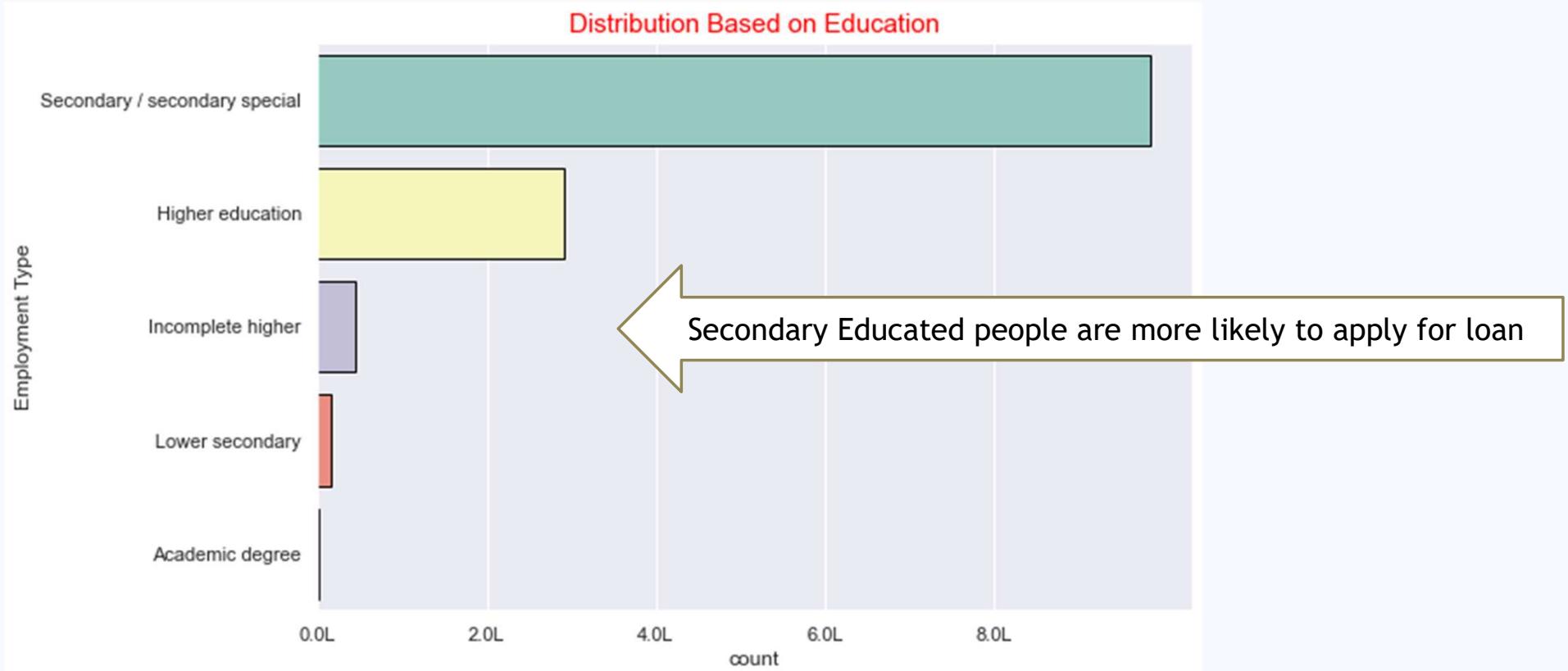


Females are more likely to apply for loan

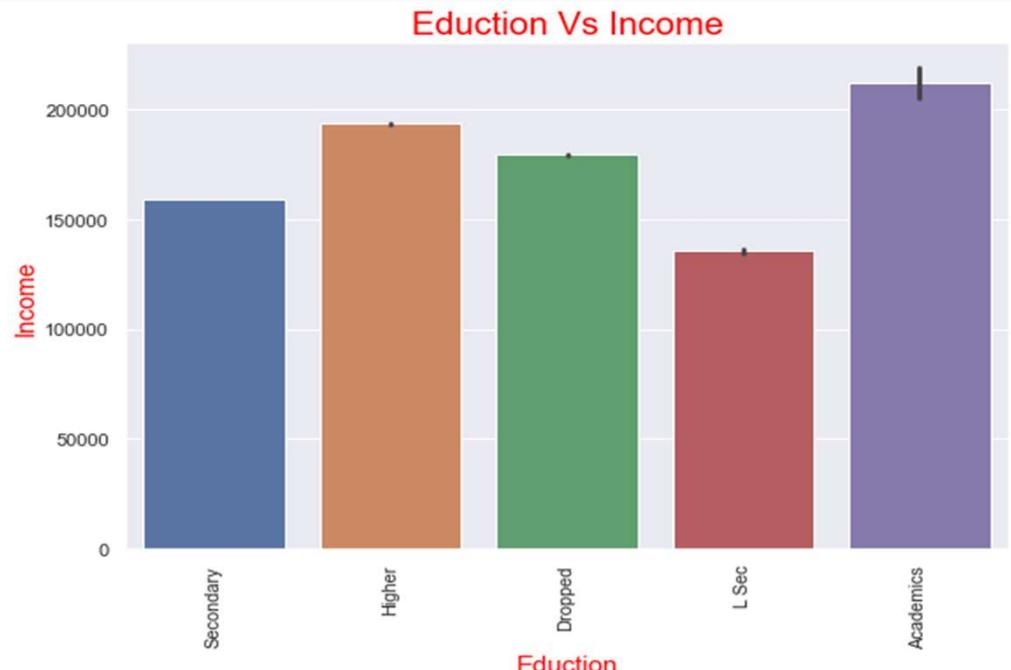
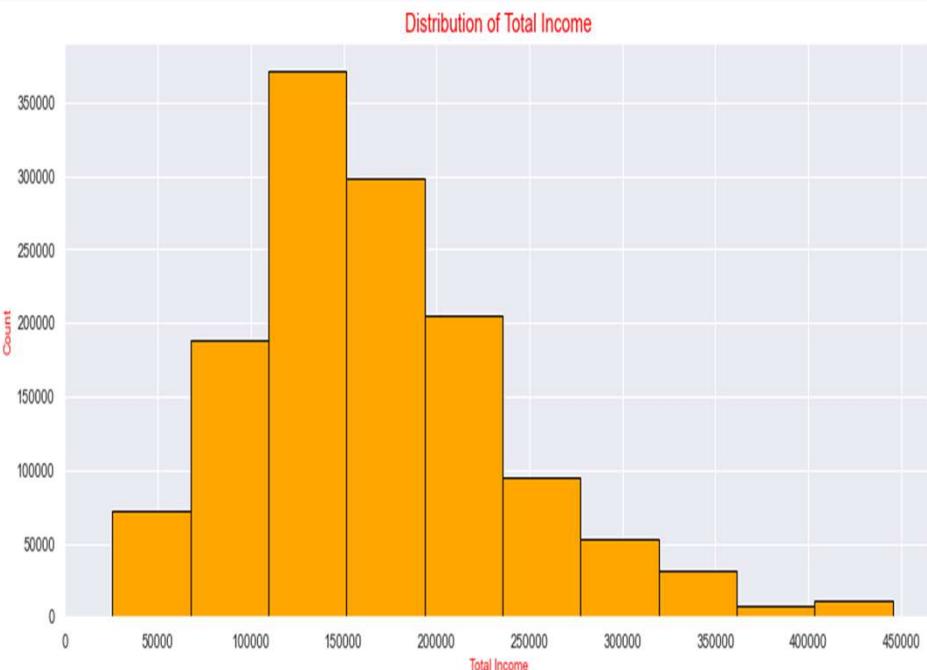


And so are married people

Preferred Customers - II

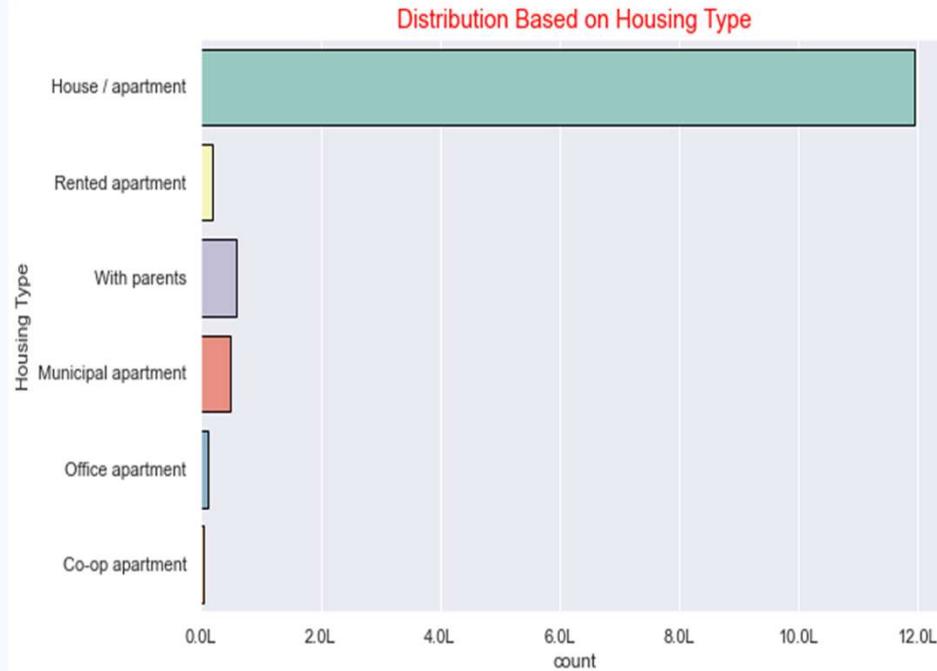


Preferred Customers - III

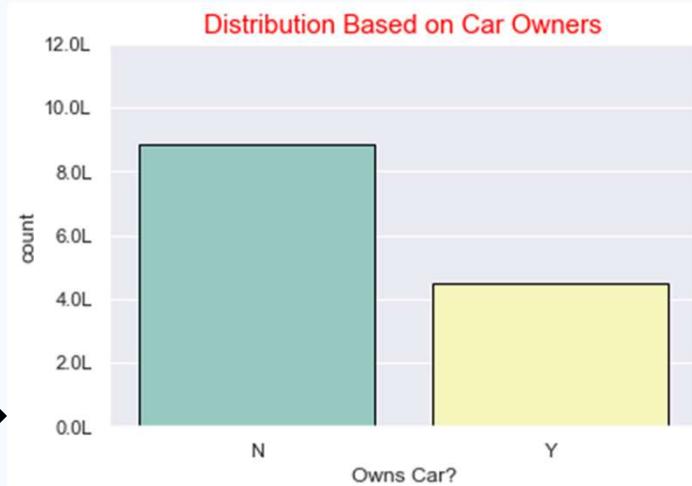


Mid income range & secondary educated people are potential customers

Preferred Customers - IV

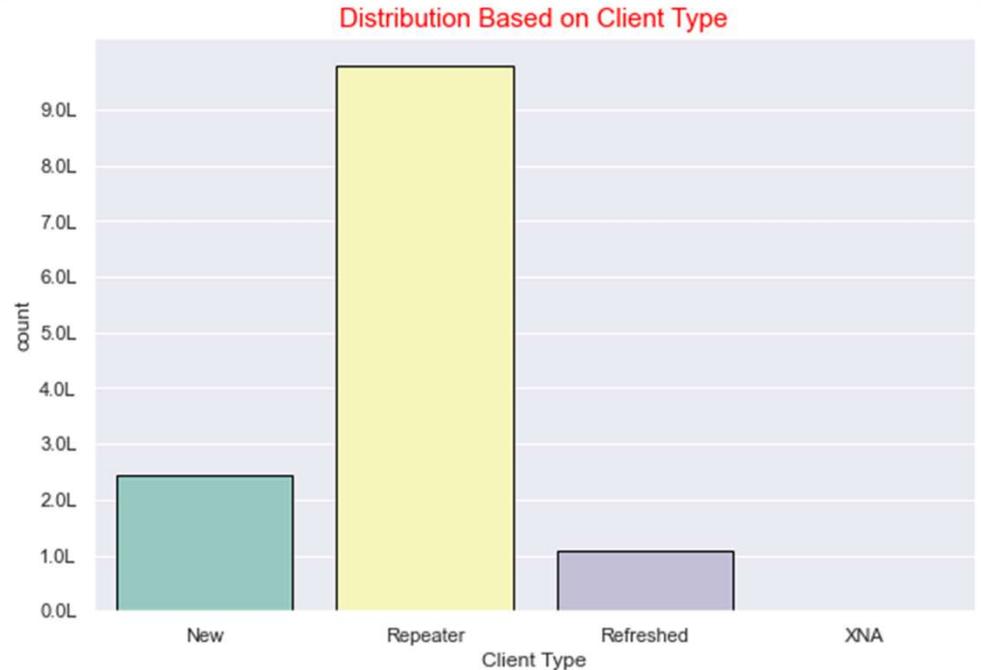
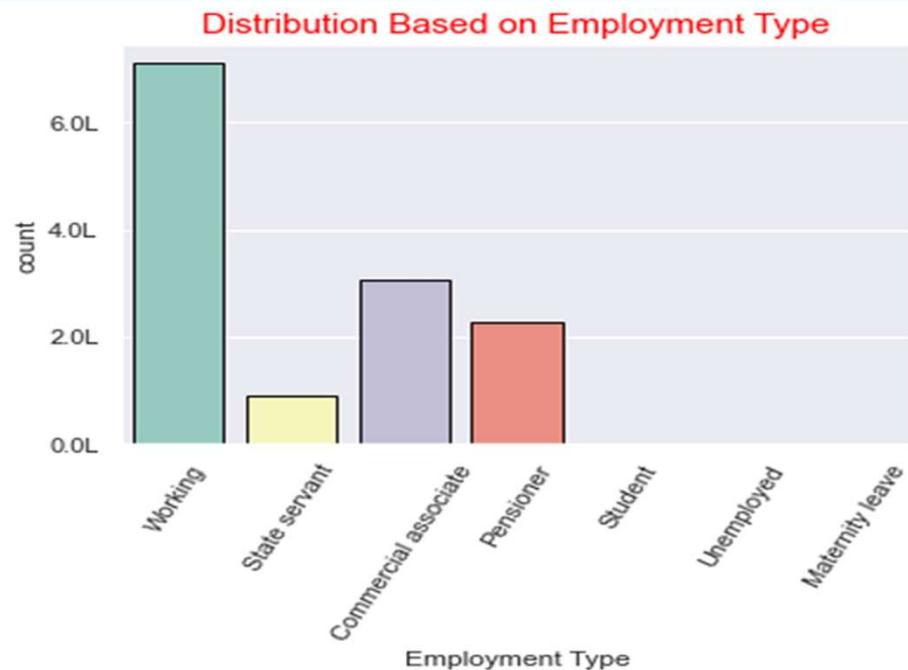


Apartment/ House Owners are more likely to apply for the loan



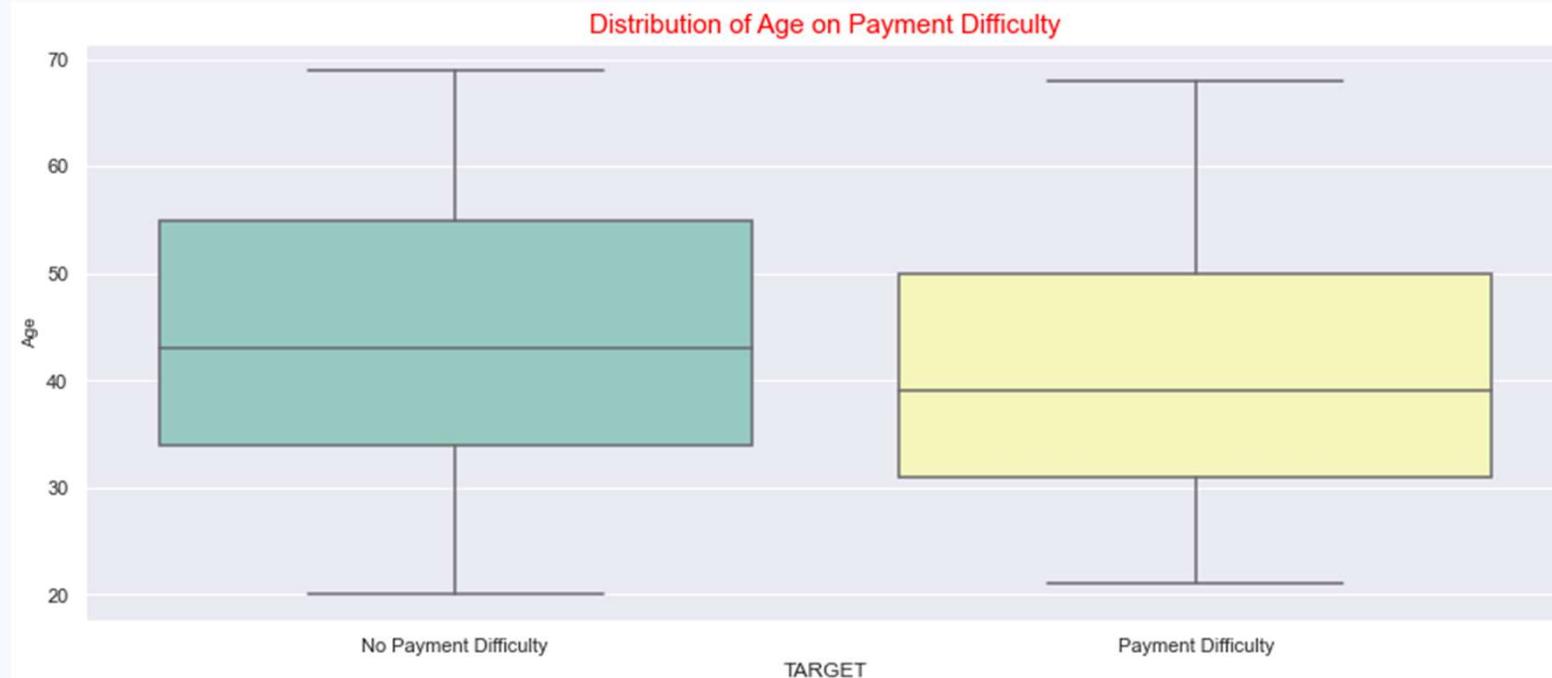
But not the car owners

Preferred Customers - V



Working people aswell as Existing customers are more likely to apply for the loan

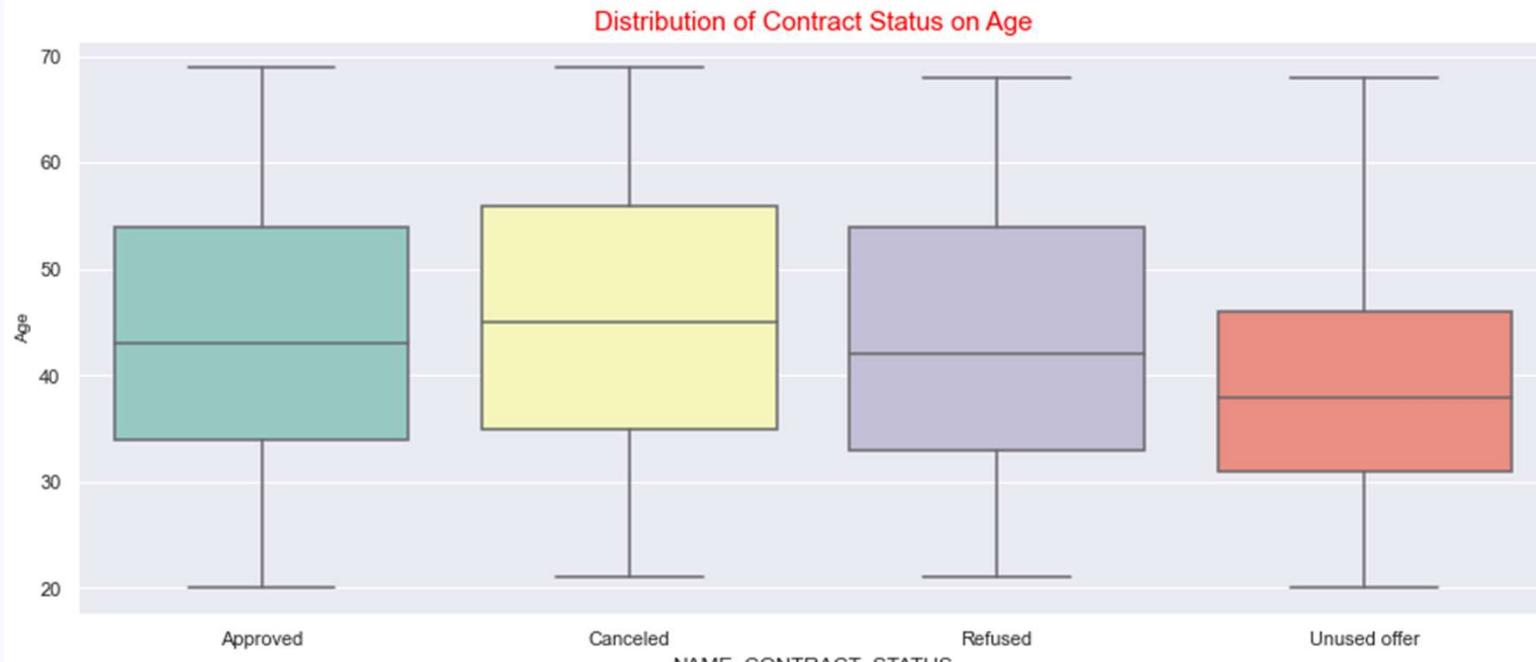
Alert! I



TARGET

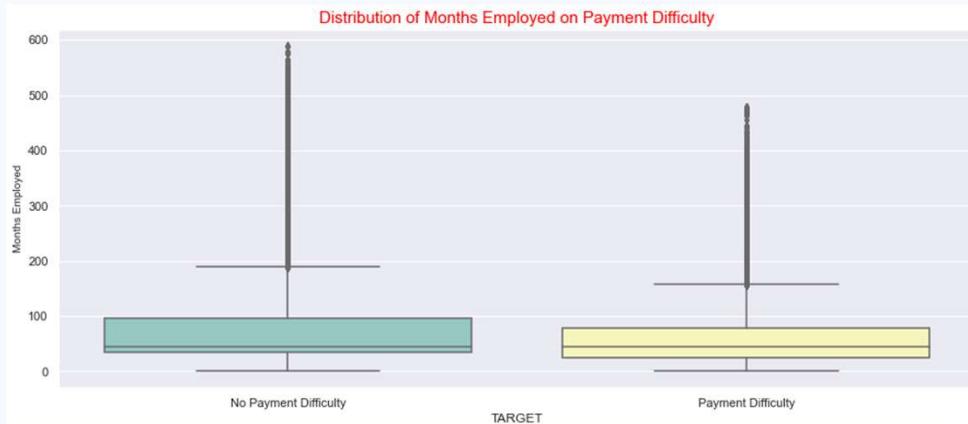
Young People tends to have more payment difficulty

Alert! II

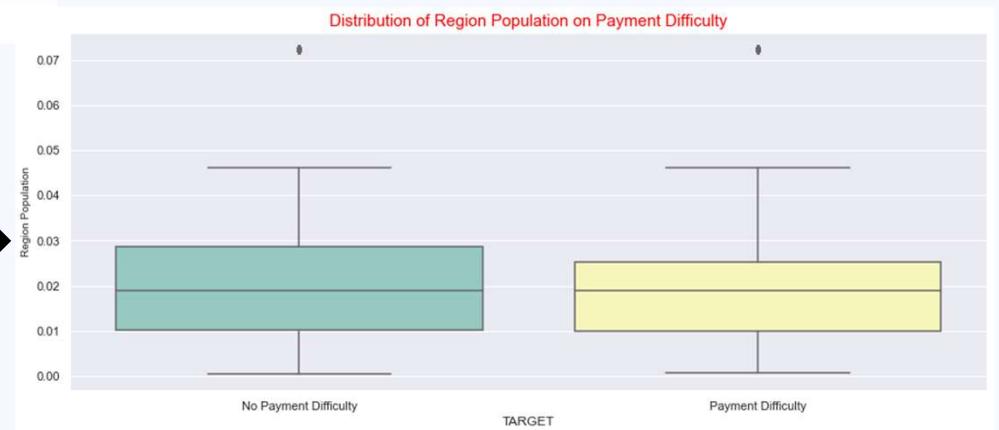


Young People often don't use the offered loan. Seniors either approve, cancel or reject the loan offer. They hardly leave that unused.

Alert! III

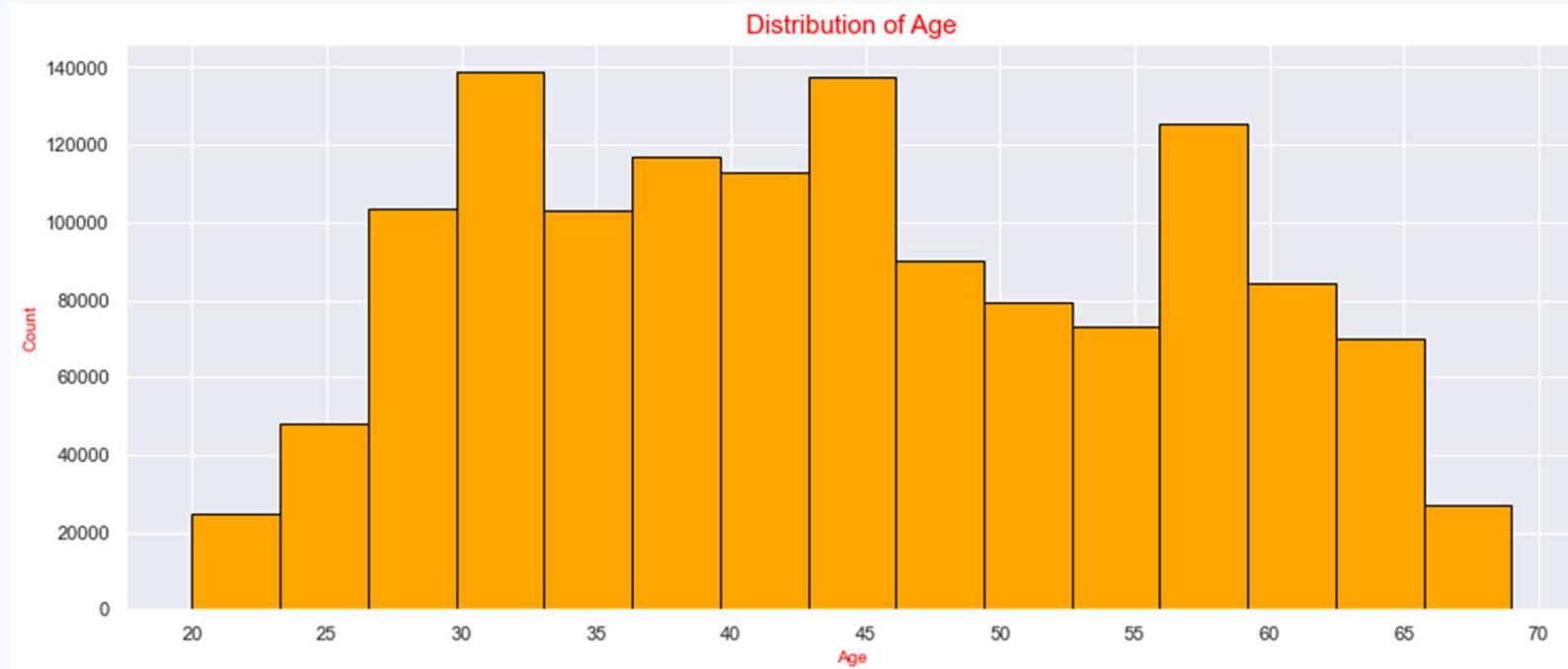


Less months of employment; more chances of payment difficulty



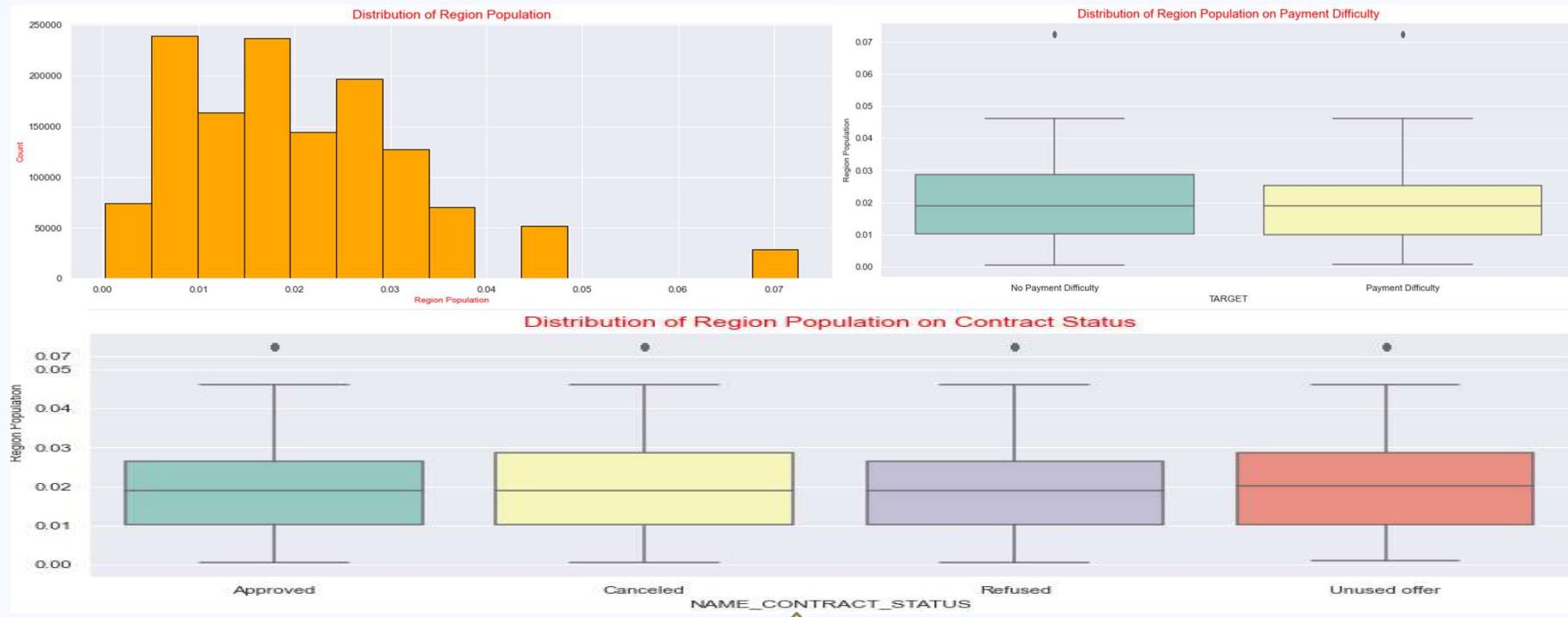
Living in less dense area; more chances of payment difficulty

Other Stats - The age factor



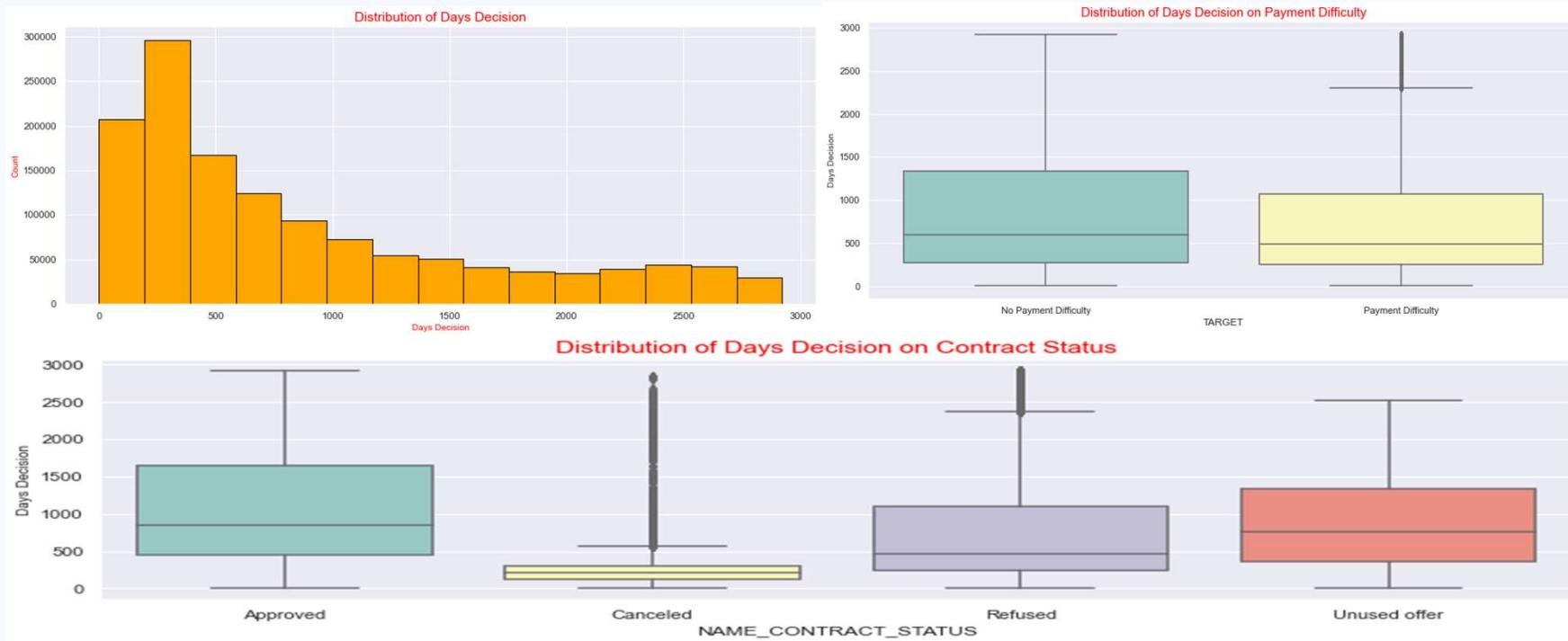
People in early 30's, Mid 40's and late 50's are more likely to apply for loan - this can be linked to important stages of a person's life - initial settlement, buying a house and education expenses for kids.

Other Stats - Place of living



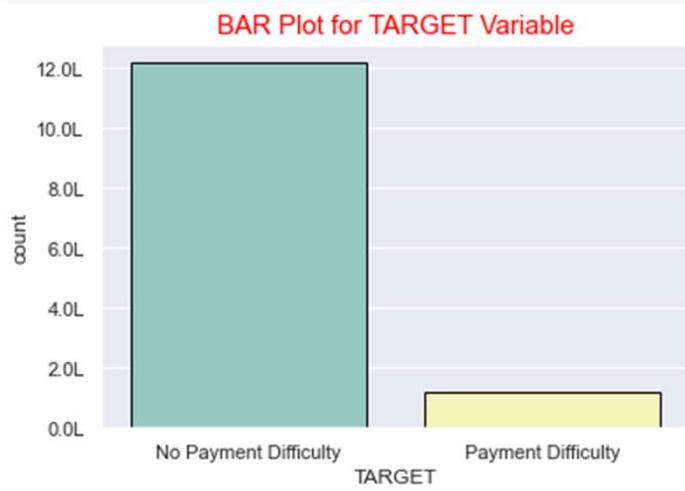
Majority customers for the client lives in area where population is less however people living in higher population regions may face less payment difficulties

Other Stats - Decision Time



Decisions should never be made in haste. It is noticed that applications with higher processing time (with the previous application) have less chances of payment difficulty. However, in general this time is already too high.

Other Stats - Payment Difficulty

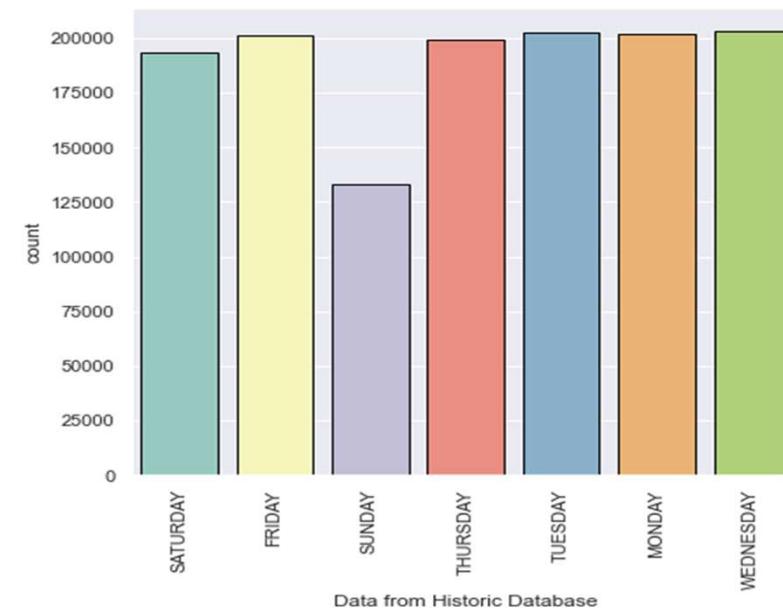
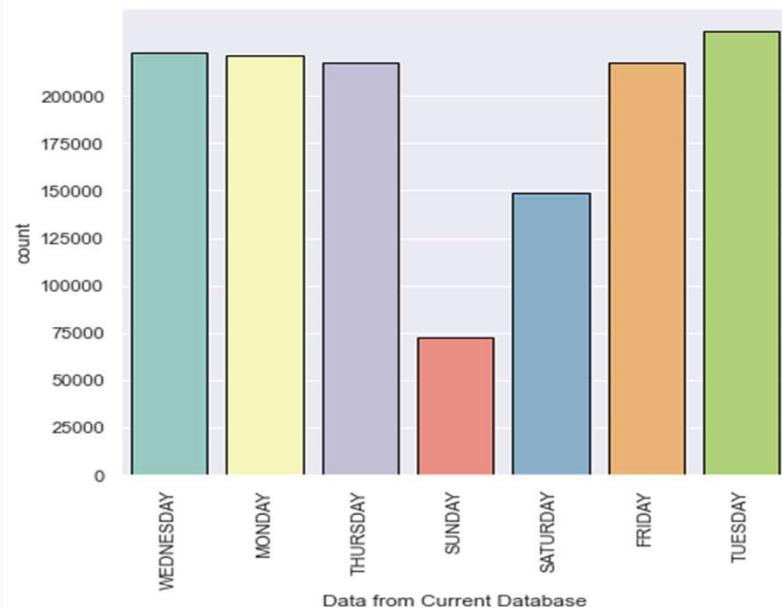


Less than 9% of existing applicants face
Payment difficulty.
Imbalance is Good!



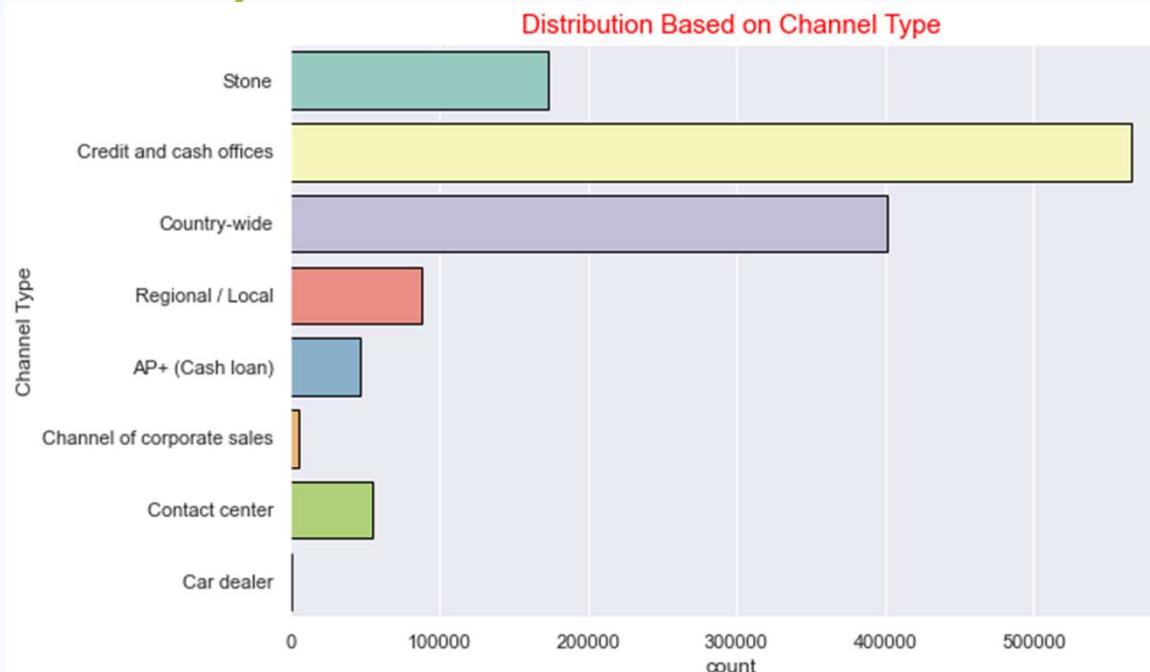
Other Stats - Processing on Weekends

Distribution Based on Days of Week



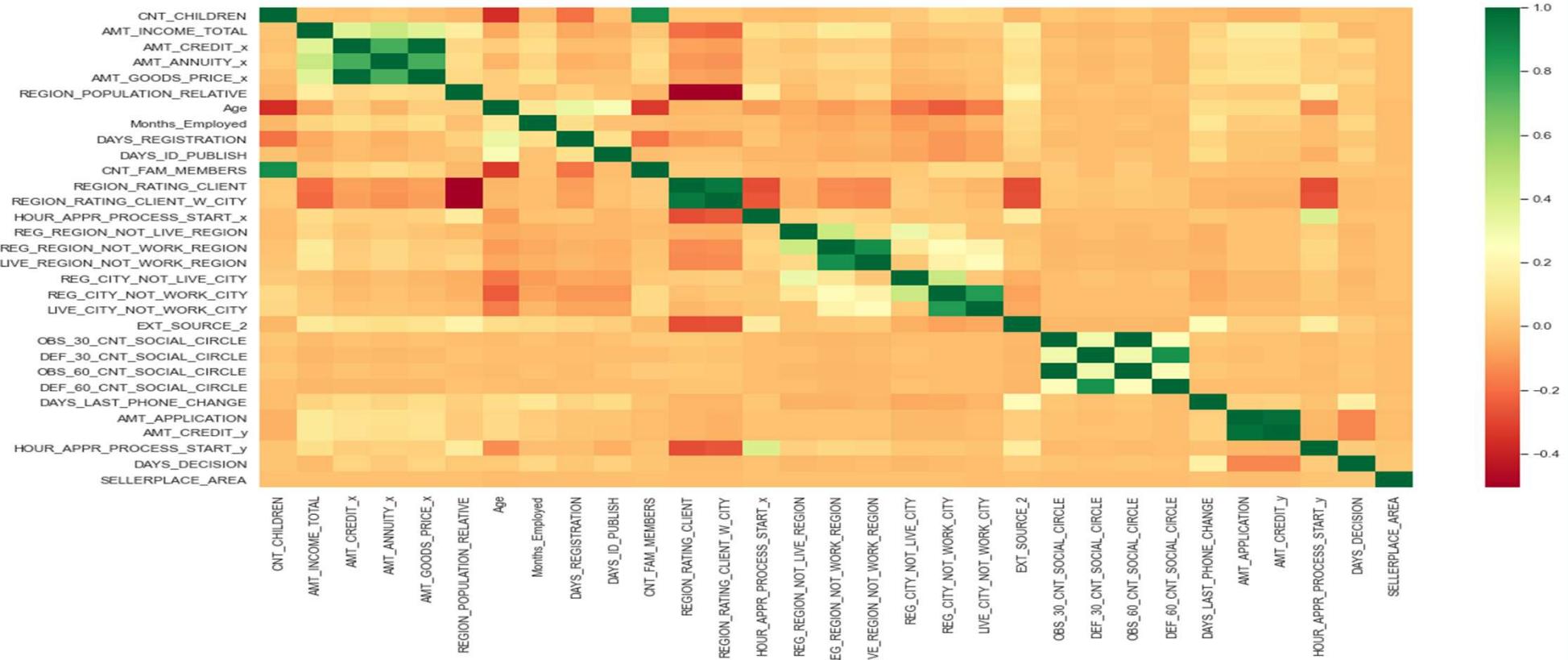
Very less application processing is done on Sundays & Saturdays. It looks like the bank used to be aggressive on Saturdays in the past.

Other Stats - Important Channels



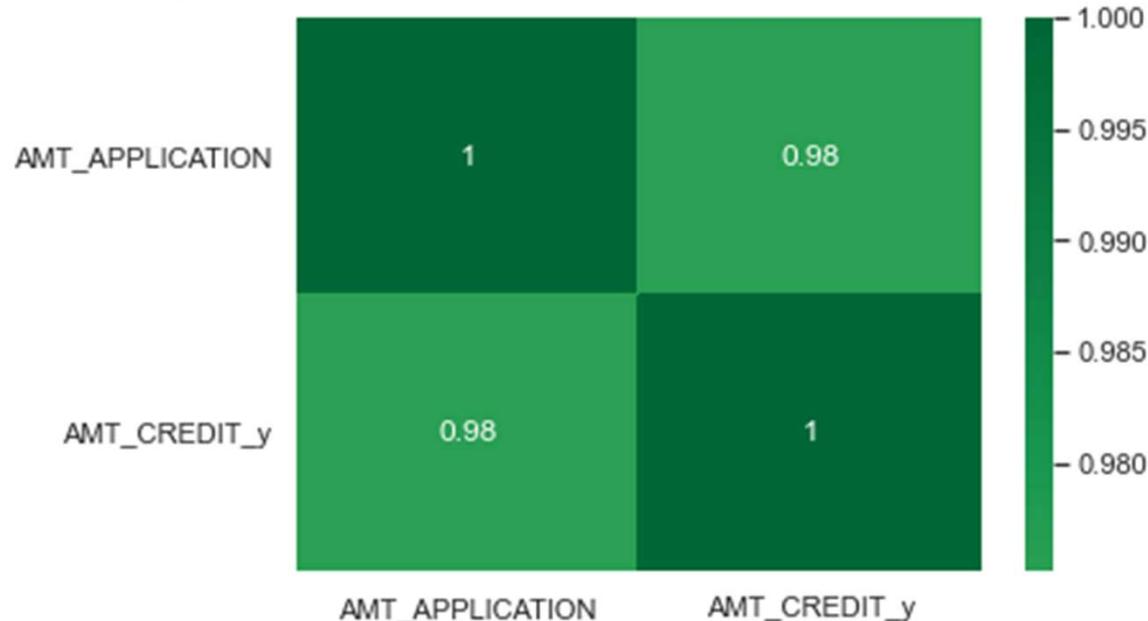
The major source for the application is Cash & Credit offices of the client. Other channels are not that effective.

Other Stats - Heatmap of Numeric Data



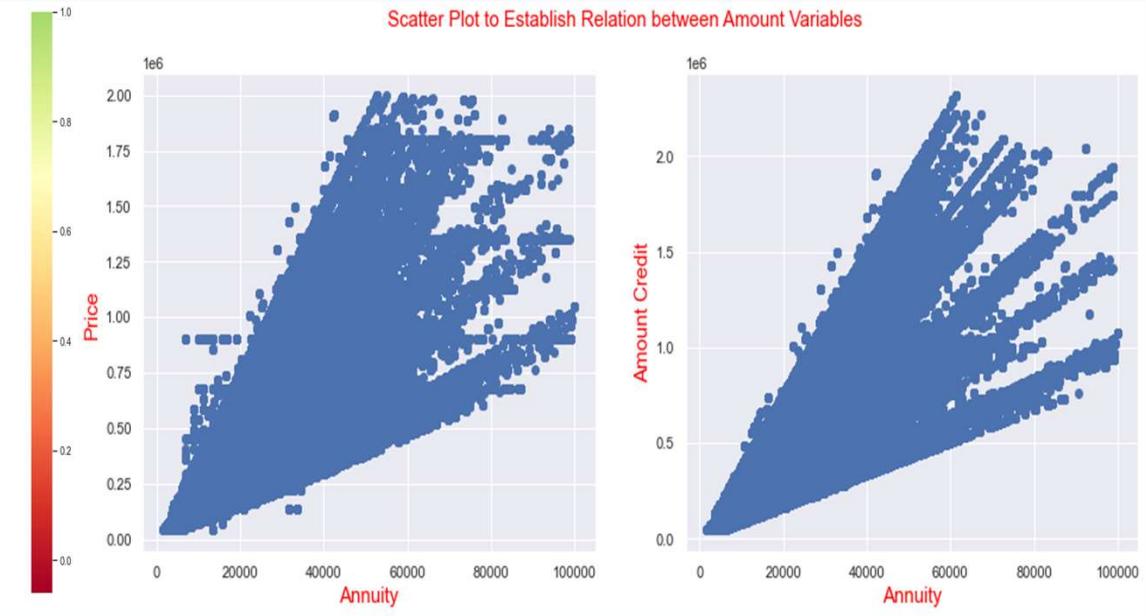
Heat Map for all Numeric Variables. Please refer next page for details.

Other Stats - Application Amount Vs Credit Amount



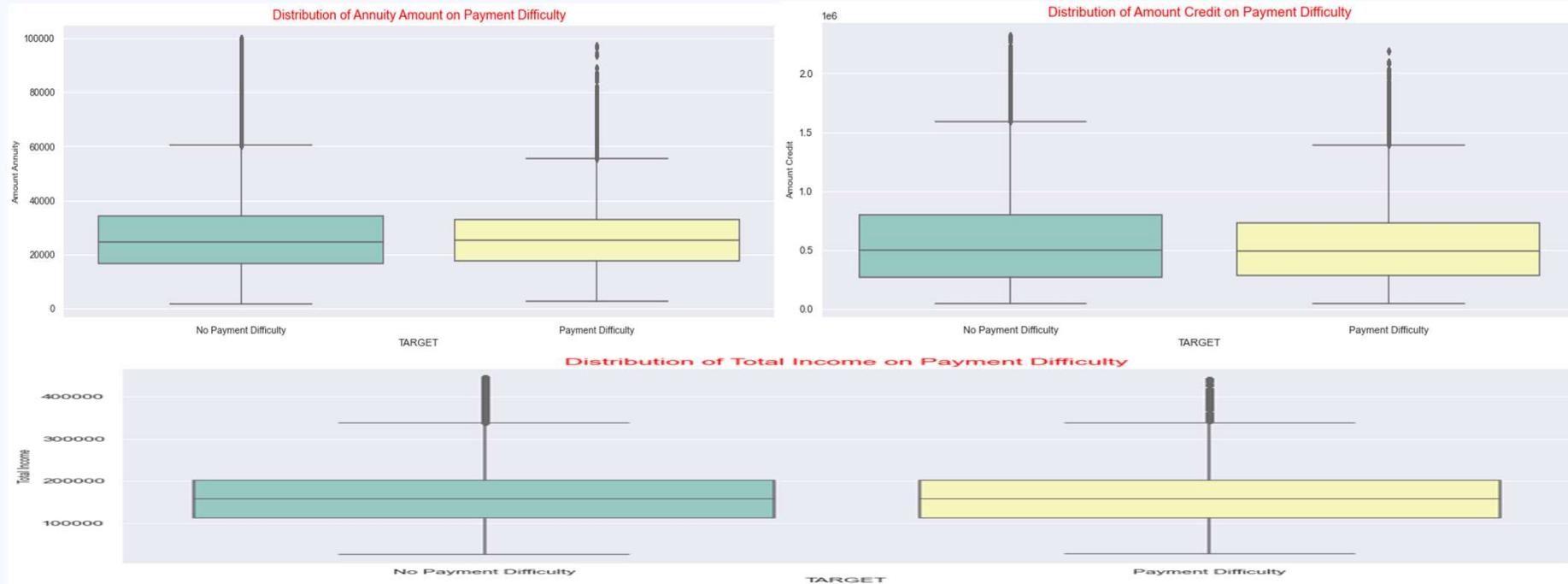
There is a very high correlation between the historic data for Amount application & Amount Credited. This means that bank normally approves the loan for the requested amount.

Other Stats - Understanding Amount Variables....1



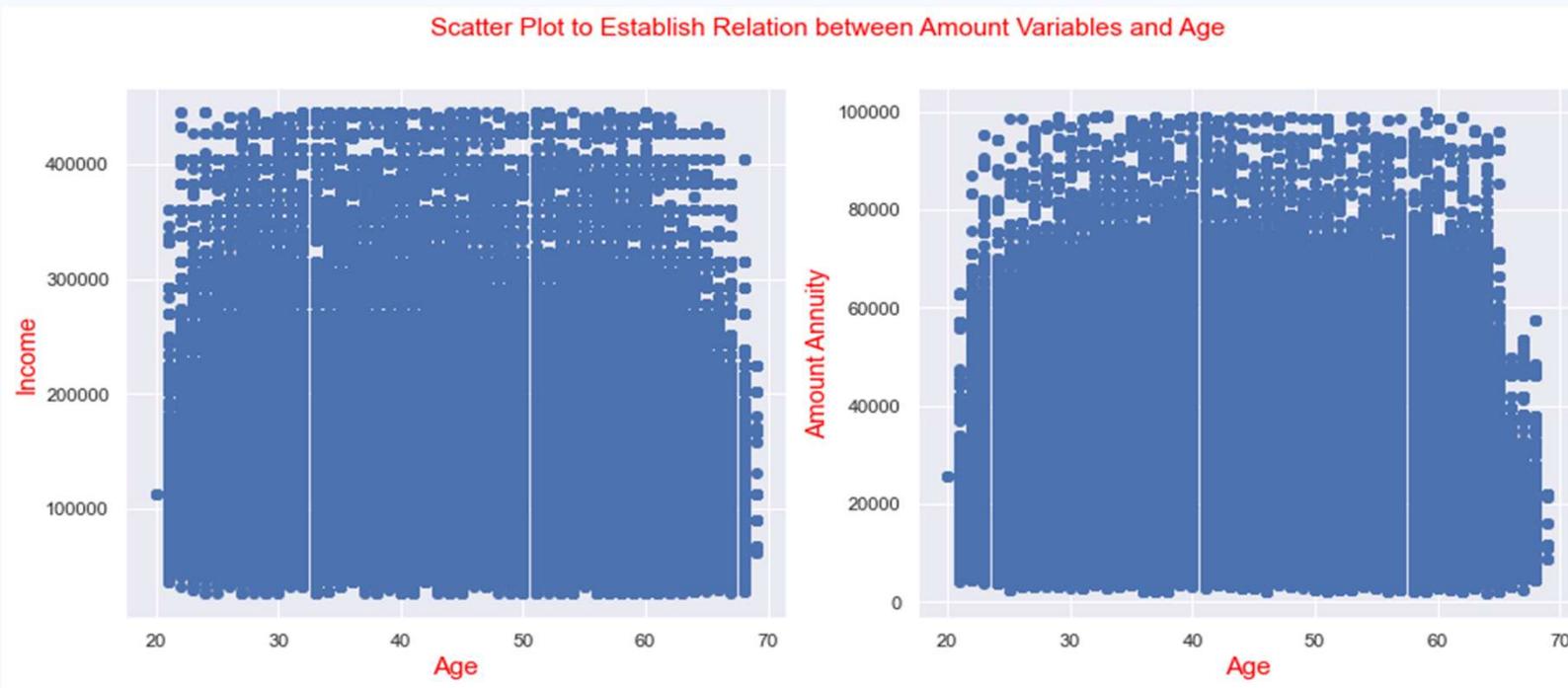
Credit Amount, Annuity Amount & Goods Price have high positive correlation. The same linear relation is clear from the scatter plot.

Other Stats - Understanding Amount Variables....2



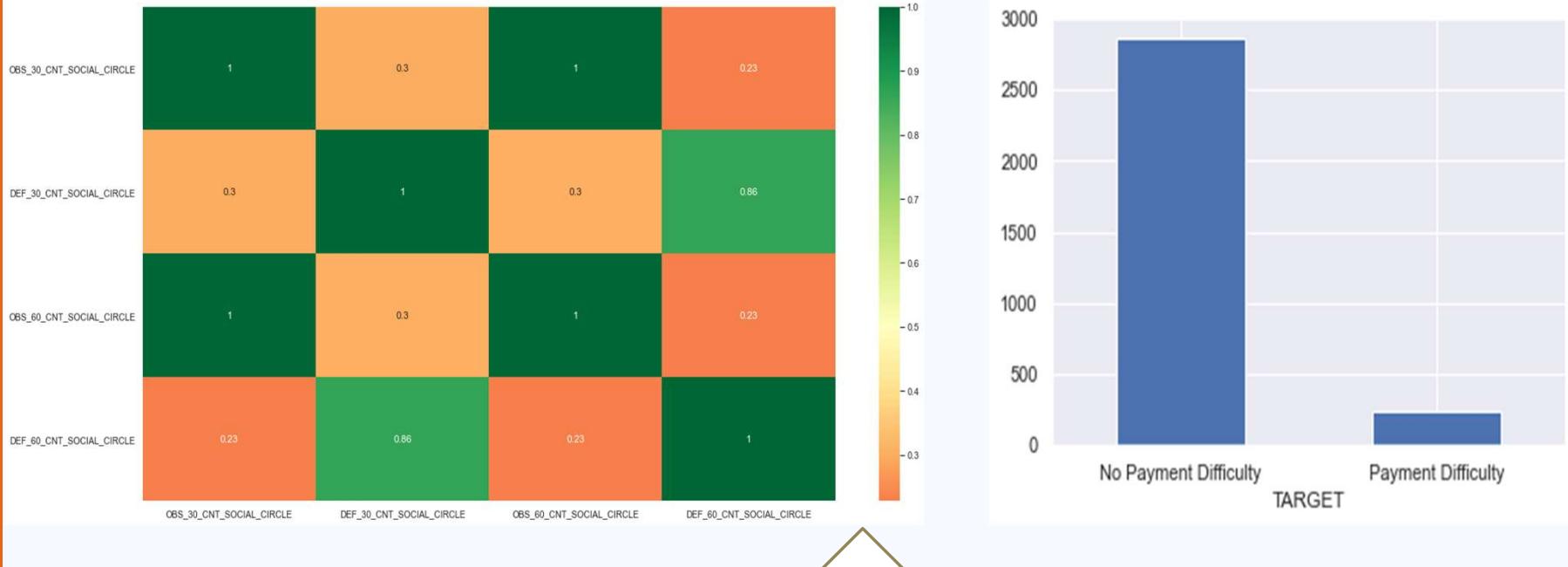
Inspite of high correlation between amount variables, Payment Difficulty is not dependent or very less dependent on Income, Amount Annuity or Amount Credit.

Other Stats - Understanding Amount Variables....3



It looked that Income & Annuity has negative correlation with Age, but the relation is actually scattered.

Other Stats - Impact of Social Surrounding



Social surrounding has very high correlation for 30 DPD and 60 DPD. However, there is no such observation of increase in payment difficulty for people with social surrounding where the 30/ 60 DPD is high. Infact beyond a certain value of surrounding; payment difficult becomes Zero.

IMPORTANT VARIABLES

1. Gender & Marital Status
2. Education & Income
3. Existing customer or New
4. Age
5. Payment Difficulty
6. Area of living
7. Channels
8. Application Amount & Amount Credited of Previous Application

Thank You