Credit EDA Case Study Analysis

Data Science 1130: Final Project Report

By: Aum Patel, Myra Rasaiah, Yuri Matienzo, Xidong Liu, Krupali Patel

Overall Research Topic

The investigative focus is what consumer attributes and loan attributes influence the tendency of default.

Subtopics

Relations between consumer demographic and risk of defaulting	Historical Loan Application Outcomes and Default Patterns
Analyze how the consumer attributes like age, gender, income and family status relate to the likelihood of loan default. Helps in knowing which characteristics have high or low risk of defaulting.	Discussing whether there is an existing relationship between financial defaulting and client financial behaviours. Helps in identifying which clients are more likely to default.

Subtopic 1:

Relations between consumer demographic and risk of defaulting

Background and Analysis Steps

Background

Analyze how the consumer attributes like age, gender, income and family status relate to the likelihood of loan default. Helps in knowing which characteristics have high or low risk of defaulting.

Hypothesis

A clients attributes such as their age, gender, income, and family status, has a significant influence on the likelihood of loan default, showing that certain characteristics such as younger age, lower income, and specific family statuses will be associated with a higher risk of default, while higher income and stable family statuses will be linked to a lower likelihood of default.

Cleaning Process

Identified if there were any unfamiliar or missing values within the relevant columns

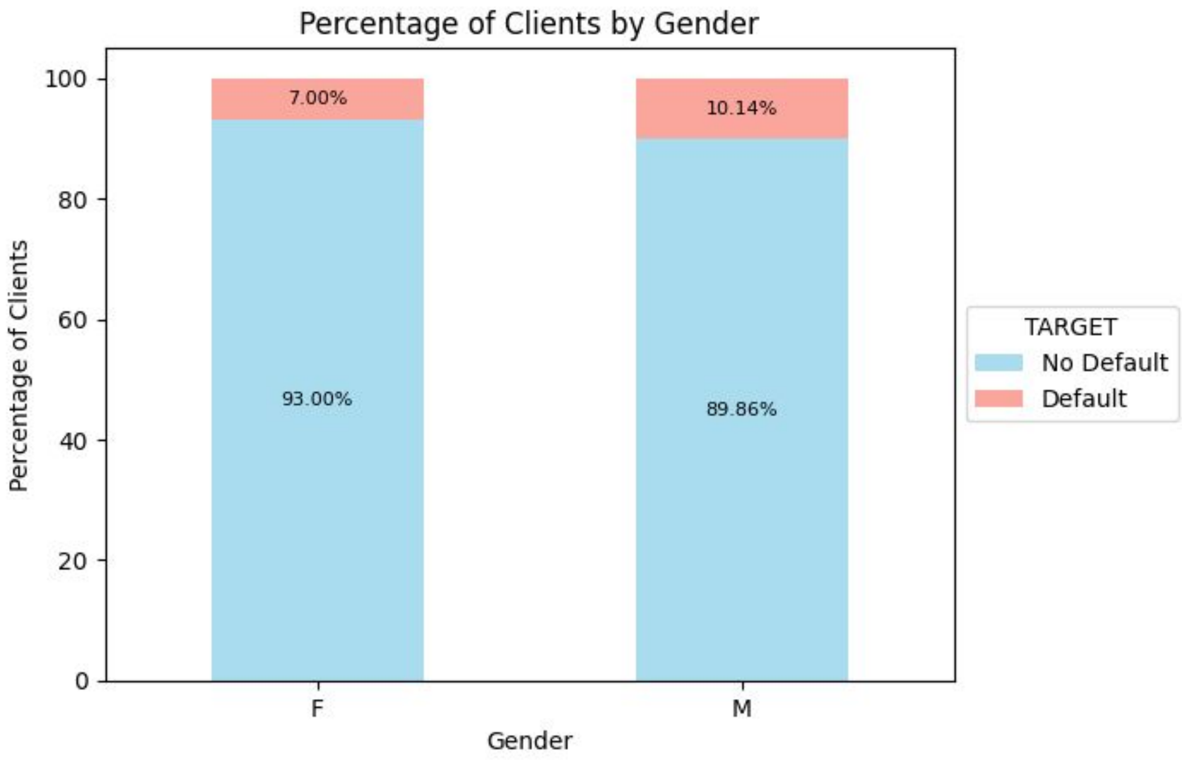
Gender versus Target Value

Cleaning/organizing data:

- Remove rows that contained value ‘XAN’, only look at rows containing values ‘M’ and ‘F’

Analyzing the graph:

- The percentage of female clients is higher than the percentage of male clients when calculating the likelihood to not default
- The percentage of male clients is greater than the percentage of female clients when calculating the likelihood to default



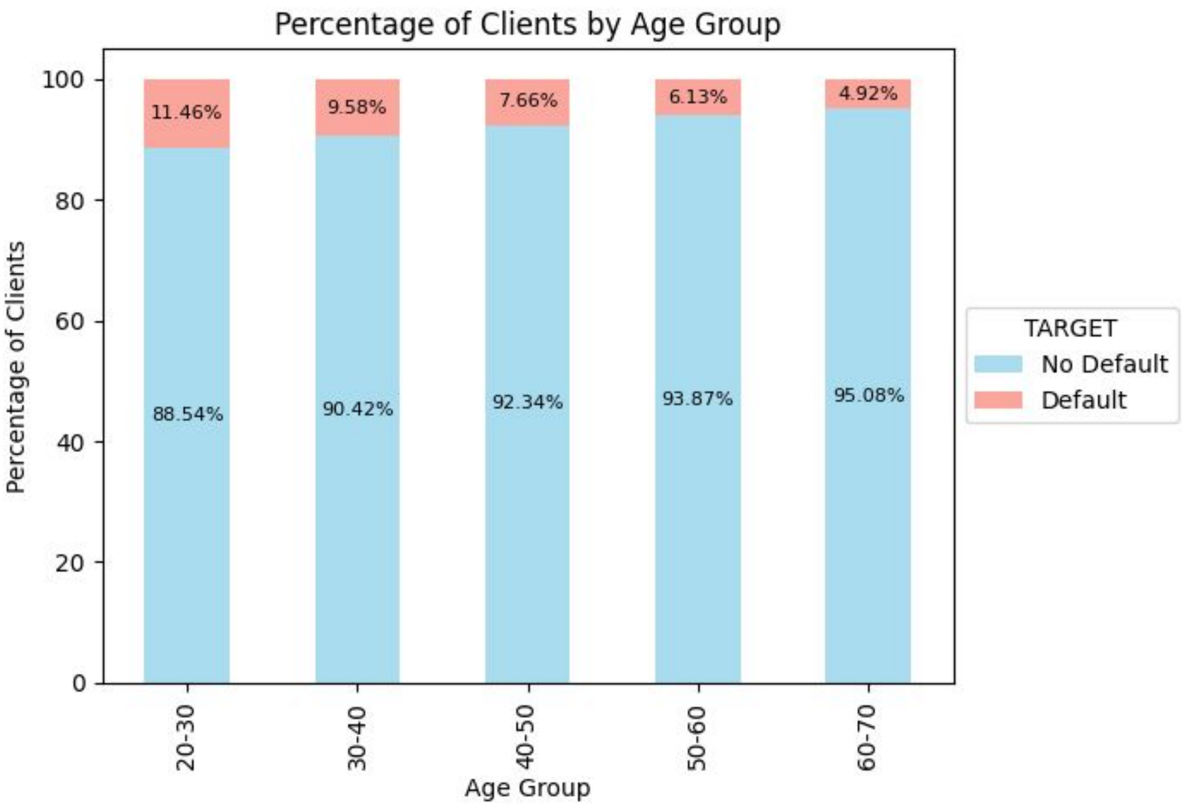
Age Group versus Target Value

Cleaning/organizing data:

- No missing data within the data frame
- Needed to convert the column DAYS_BIRTH from days to years and store age in year of clients in another column

Analyzing the graph:

- As the age group increases, then the percentage of the likelihood of clients not defaulting increases.
- The percentage of clients likely to default is decreasing over the years



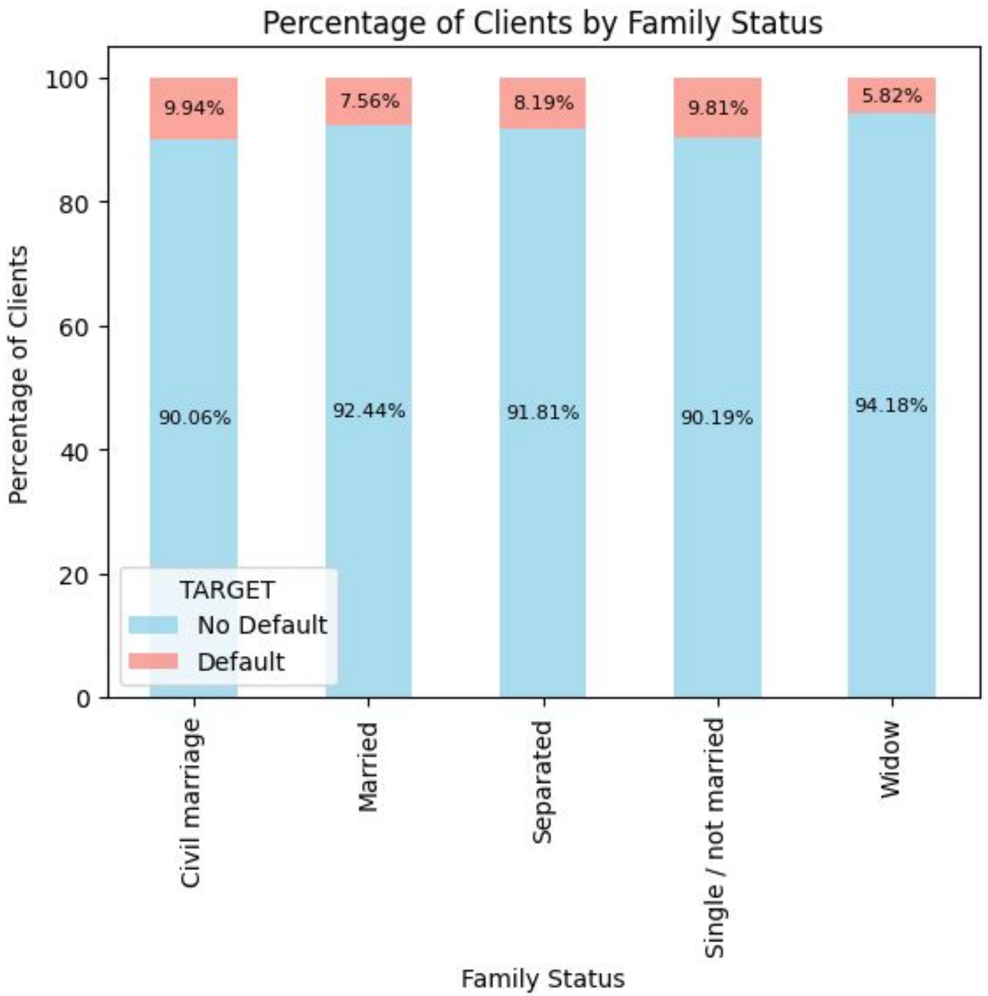
Family Status versus Target Value

Cleaning/organizing data:

- Remove rows that contained value ‘Unknown’

Analyzing the graph:

- The largest category of clients who are at risk of not defaulting is in the category of ‘Widow’, while the smallest category are those with the status of ‘Civil marriage’
- difference in the percentage of clients whose statuses are ‘Married’ and ‘Single/not married’



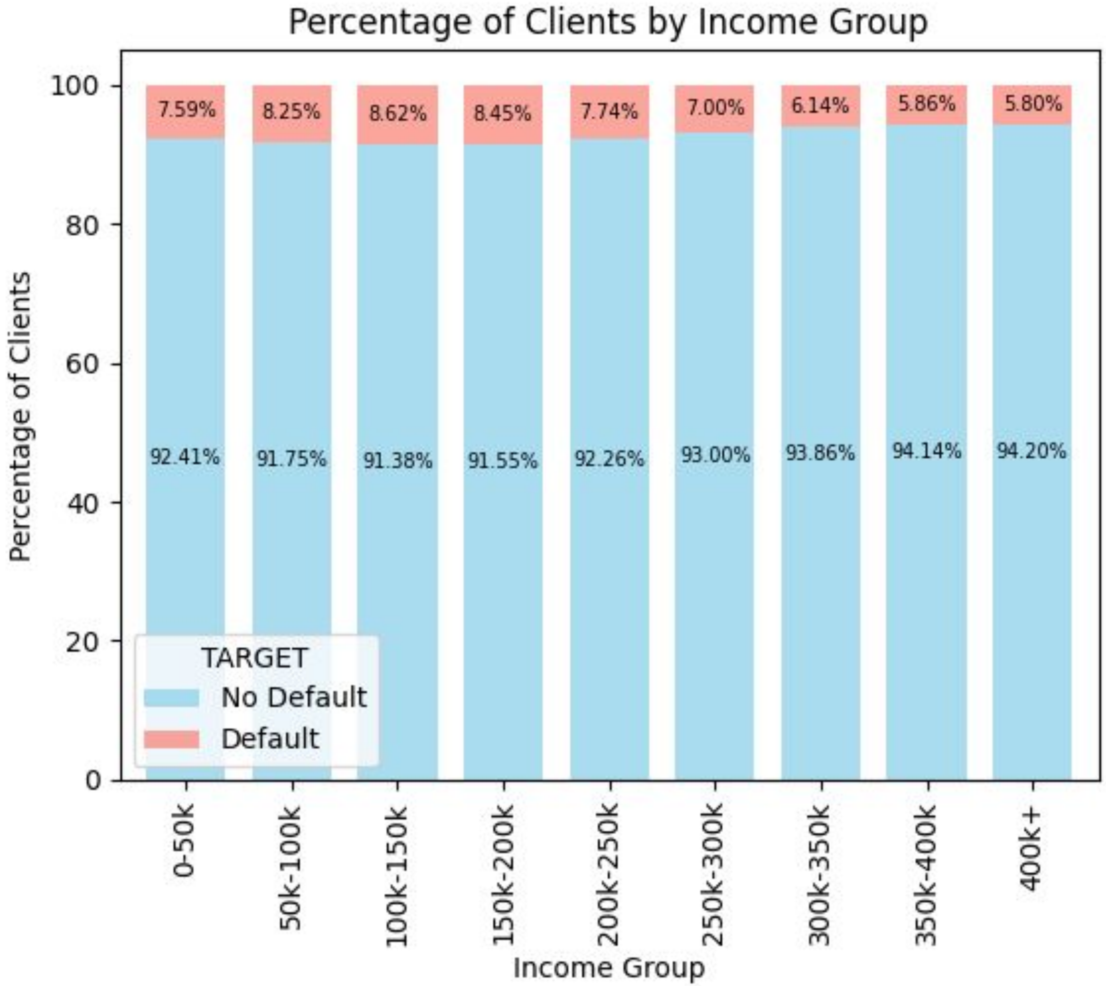
Income Group versus Target Value

Organizing data:

- Binned the 'AMT_INCOME_TOTAL' column to create a new column called 'INCOME_GROUP'.

Analyzing the graph:

- As the range of income of the client increases, they are less likely to default.
- There is a small decrease in the percentage of clients likely to default between the ranges of 0-50k and 100k-500k.



Main Results/Conclusion

- Gender analysis and graph shows the clients differences between male and female categories.
- Age group analysis and graph shows how different age groups relate to the risk of loan defaults.
- Family status analysis and graph shows the impact of family status on loan default rates by categorizing the clients with different statuses.
- Income group analysis and graph shows the relationship between income levels and default rates

Subtopic 2:

Historical Loan Application Outcomes and Default Patterns

Background and Analysis Steps

Background

This subtopic focuses on leveraging information from previous loan applications to identify patterns and indicators for loan default in current applications.

Hypothesis

It involves examining historical data such as previous loan outcomes, interest rates, approvals, refusals, and other relevant factors like Income, Debt, Down Payments, and Loan Amounts to predict and prevent future default.

Cleaning Process

Identified if there were any unfamiliar or missing values within the relevant columns

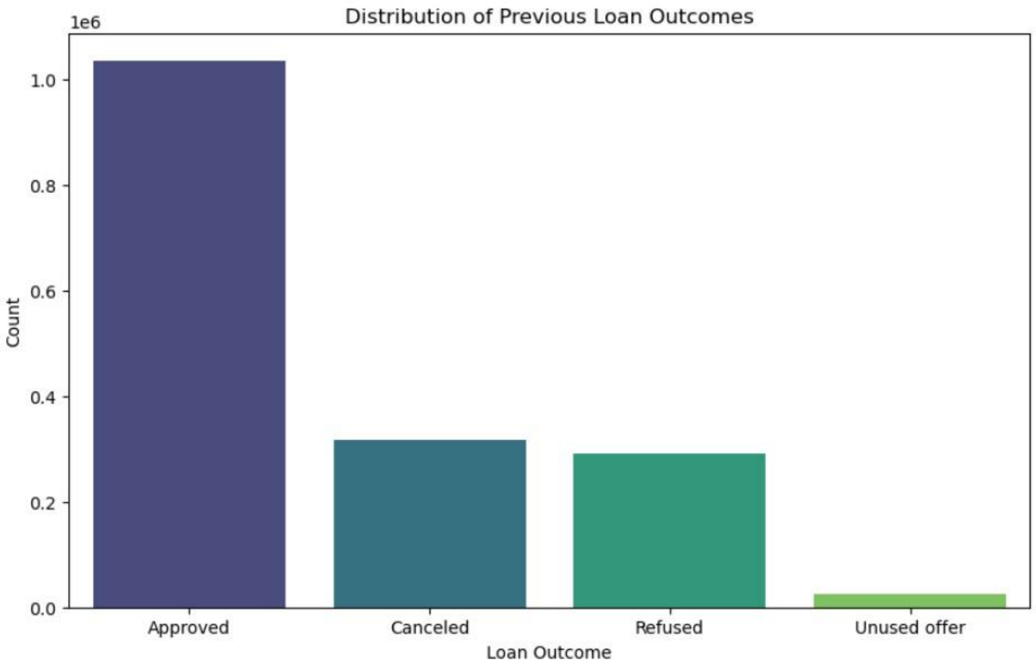
Analysis of Previous Loan Outcomes

Cleaning/organizing data:

- No missing data within the data frame.
- Used Bar plot for visualization.
- We use NAME_CONTRACT_STATUS from previous application dataset.

Analyzing the graph:

- The majority of previous loan outcomes were marked as "Approved," with a count of 1,036,781.
- "Canceled" and "Refused" are the next most common outcomes, with counts of 316,319 and 290,678, respectively.
- A good proportion of previous loan applications were successfully approved.



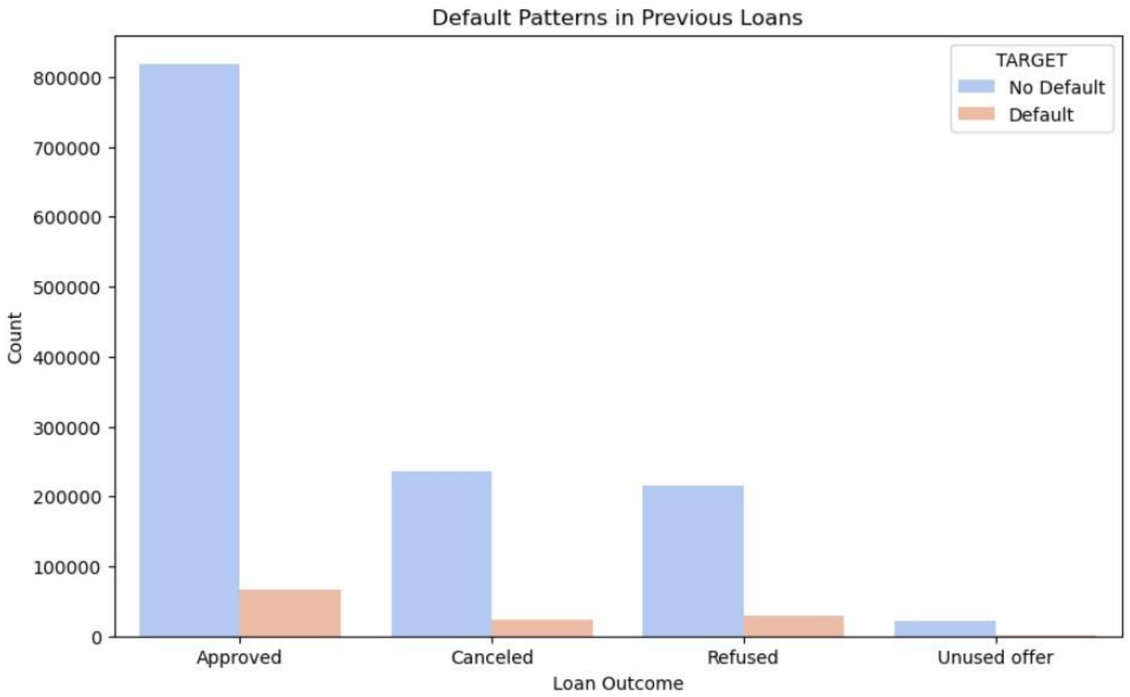
Analyze Default Patterns in Previous Loans.

Cleaning/organizing data:

- No missing data within the data frame.
- Used Countplot Visualization.
- Merged the two dataframes on SK_ID_CURR.
- Renamed AMT_CREDIT as there are two of it in new dataframe.

Analyzing the graph:

- The people who were approved for the loan previously don't have difficulty in paying back.
- Clients with a history of approved loans are less likely to face payment difficulties.



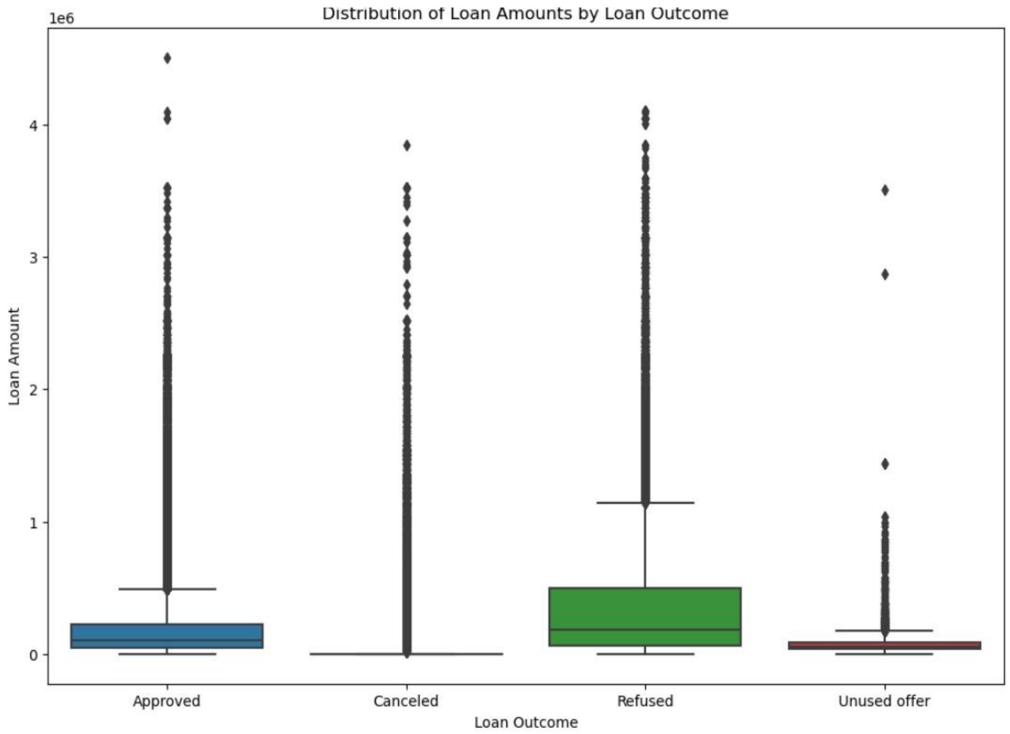
Analyze Loan Amounts by Outcome

Cleaning/organizing data:

- No missing data within the data frame.
- Used AMT_CREDIT_previous.
- We use a box plot to display the distribution of loan amount for different loan status.

Analyzing the graph:

- The loan amounts that are unusually high are Refused and Approved loans we get idea of appropriate loan amount. i.e the amount that will be easily approved by the institution.
- There are outliers in the plot.
- This plot helps institution to approve safe amount of loans.



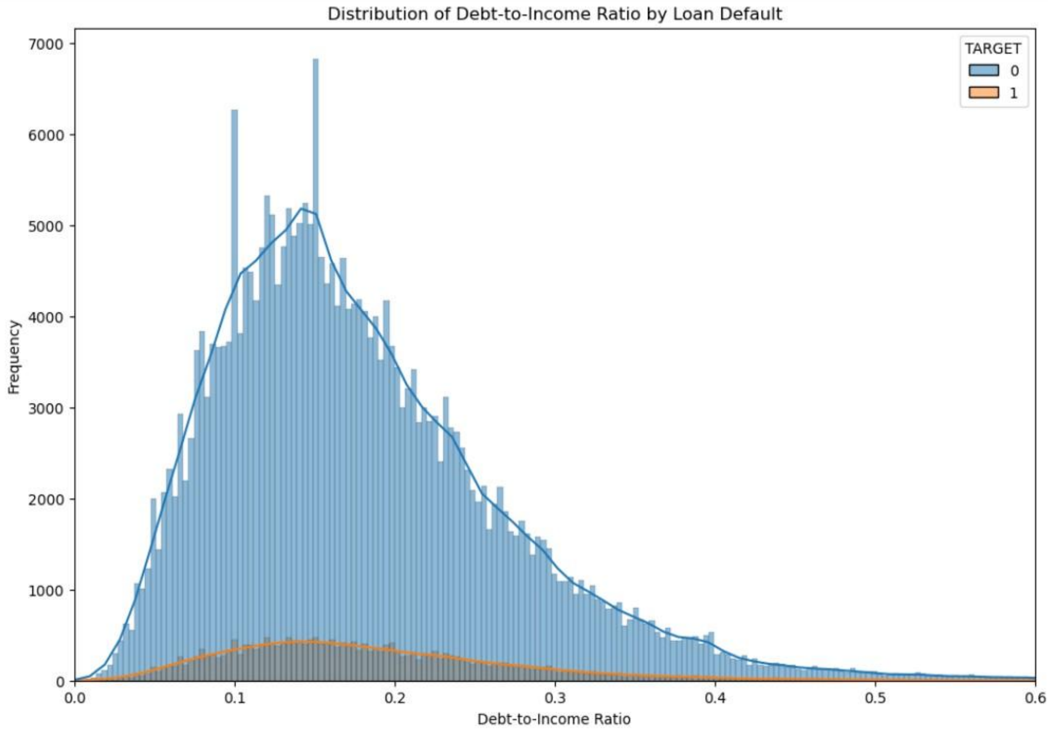
Debt-to-Income Ratio (DTI)

Cleaning/organizing data:

- We create a new column 'DTI' by dividing AMT_ANNUITY by AMT_INCOME_TOTAL.
- There were 12 missing values in relevant columns which were dropped.

Analyzing the graph:

- The frequency of different DTI ranges for clients with and without payment difficulties.
- Higher DTI values, as seen in the right-skewed portion of Target 0, might be indicative of a lower risk of default.
- The highest frequency occurs in the range of 0.1 to 0.2 on the x-axis.



Analyzing down payment amounts by contract status

Cleaning/organizing data:

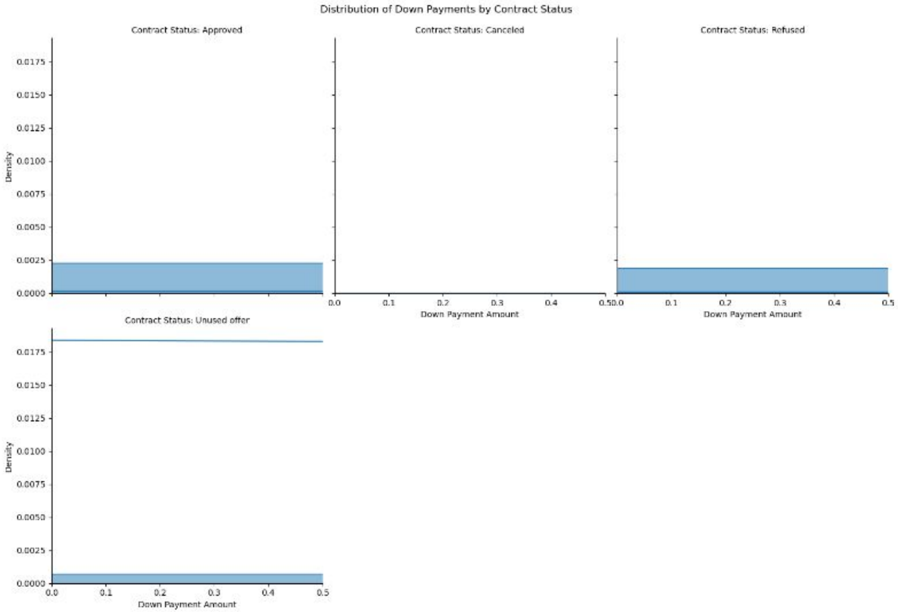
- Merged application_data and previous_application based on the 'SK_ID_CURR' column using inner join.
- We calculate descriptive statistics for 'AMT_DOWN_PAYMENT' column, grouped by the 'NAME_CONTRACT_STATUS.

Analyzing the graph:

- Density refers to the probability of the distribution of payment.
- The density of the graph varies depending on contract status
- Approved down payments have the most amount of density, refused payments as second, unused offers as third, and cancelled as no density.

NAME_CONTRACT_STATUS	count	mean	std	min	25%	50%	75%	max
Approved	568197.0	6832.369469	19304.373164	-0.9	0.0	2322.0	8302.5	3060045.0
Canceled	536.0	21642.580410	101140.746442	0.0	0.0	0.0	0.0	918000.0
Refused	74778.0	7040.091841	29282.486057	0.0	0.0	0.0	6583.5	2475000.0
Unused offer	20650.0	1.252809	158.320930	0.0	0.0	0.0	0.0	22500.0

<Figure size 1200x800 with 0 Axes>



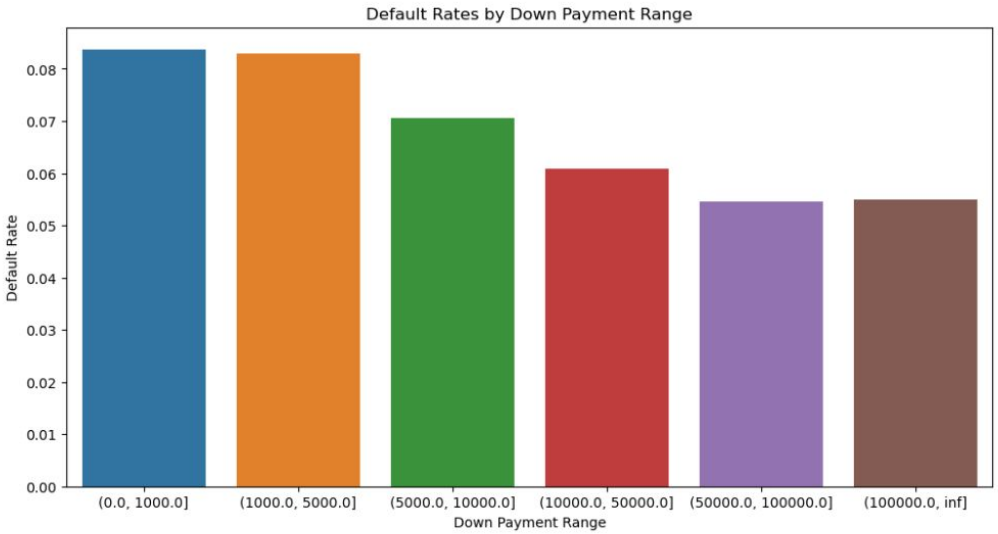
Default rates by down payment

Cleaning/organizing data:

- Create bins to categorize down payment amounts into different ranges.
- We create a new column, 'Down_Payment_Range', to the df.
- Calculate the mean of the 'TARGET' column for each down payment
- We calculate the p-value using chi-square test

Analyzing the graph:

- The default rate varies throughout the different down payment range
- The higher the down payment, the lower the default rate
- If the $p < 0.05$, it indicates that the difference in the default rates is significant




```
Chi-square value: 500.19917417872585
P-value: 7.232105926297502e-106
The difference in default rates between down payment groups is statistically significant.
```

Subtopic 2: Main Results/Conclusion

Debt to Income Ratio VS Defaulting: clients with higher income + lower annuities tended to be the least likely to default.

Contract Status VS Defaulting: contract status least likely to default is 'Approved'.

Default Rates VS Down Payment: positive relationship exists between higher down payments and lower the default rates.

Thank You!