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.

Subtopic 1 - Relations between consumer demographic and risk of defaulting

Background/introduction of problem:

We want to analyze how the consumer attributes like age, gender, income and family status relate to the likelihood of loan default. This will allow us to know which characteristics have high or low risk of defaulting.

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.

Data analysis steps:

In order to understand the following problem, we must analyze the relationship between the TARGET column within the dataset and various attributes. To start, we must filter the dataset such that within the CODE_GENDER column, we want to exclude any values that are not ‘M’ or ‘F’. When grouping the data in order to find the value count of the values in CODE_GENDER, we realized that there were 4 rows in the dataset that were represented by the value ‘XAN’, which needed to be filtered out. Once the dataset was cleaned, we created another dataframe which stored the values in the CODE_GENDER column (which were ‘M’ and ‘F’) and the percentage of clients that had a target value of 1 and a target value of 0 for each of the respective gender groups. Once this data was obtained, we were able to plot it onto a bar graph (see Appendix A-1). The bar graph was the best way to show how the composition of the clients varied between both genders. With the percentages labeling the bars, we are able to draw clear conclusions and properly visualize how the percentage of clients of each gender varied between the two target values.

When analyzing the clients' age, specifically the DAYS_BIRTH column, all values are in the negatives. In order to make this information comprehensible, we needed to convert all negative values to positive values using absolute function. Afterwards, we converted and rounded the values representing the clients' ages from days to years and saved the respective values in a new column called AGE. We also created two arrays, one representing the bins that will store the age values and another representing the labels for the age groups. From there, we used the cut function to store the values from the AGE column into their respective bins and categorize them accordingly. Then we grouped the values in the bin according to their values in the TARGET column, and found the percentage of each value for the respective category. Once this was stored in a dataframe, we then plotted it to get a bar graph displaying the categories, and the percentage each target value occupied for each category (see Appendix A-2).

When analyzing the relationship between the 'TARGET' variable and family status ('NAME_FAMILY_STATUS'). We realized there came a fifth group being the 'Unknown' values in the 'NAME_FAMILY_STATUS' column. These unclear values were then removed by filtering them because they were unnecessary for a proper investigation in the 'NAME_FAMILY_STATUS' column. We had to separate the dataset in order to remove rows in which 'NAME_FAMILY_STATUS ' had the label 'Unknown' in. This made sure that our analysis after that was all filtered would be more precise now. After the filtered version of the dataset, we got five bar graphs, each categorized by the relevant statuses and the percentage of clients with target values of one and zero (refer to Appendix A-3).

The analysis classifying the clients according to their income helps us understand the following problem. First we started by binning the 'AMT_INCOME_TOTAL' data to create a column called

'INCOME_GROUP'.A bar plot was then created to show the visualization of the distribution of loan defaults and non-defaults for different incomes. The analysis shown in A-4 showed relationships between changes in loan default rates and differences with incomes. It provided useful information about how income can affect the likelihood of a loan default, this can also give us info on the decision-making processes.

Main results/conclusions:

This data analysis broke down some important insights into the variables affecting loan defaults. Appendix A-1's gender analysis provided a visual representation of the percentage of loan defaults for each gender by showing the clients differences between male and female categories. With a focus on age, Appendix A-2 showed how different age groups relate to the risk of loan defaults. The family status analysis, in Appendix A-3, shows the impact of family status on loan default rates by categorizing the clients with different statuses. Lastly, the distribution of loan defaults within income groups was shown by the analysis in Appendix A-4, which gave us a visualization on the relationship between income levels and default rates. The results of this analysis showed us some factors impacting loan defaults, which helps with the decision-making process. From the analysis, we can see that there are no outliers, and the few rows that needed to be removed did not affect the rest of the analysis.

Subtopic 2 - Client financial behaviors

Background/introduction of problem:

This subtopic focuses on concluding previous loan applications to identify patterns and indicators for loan default in current applications. It involves examining historical data such as previous loan outcomes, interest rates, loan amounts, down payment rates, and other relevant factors to predict and prevent future default. This will allow lenders to perceive a more accurate assessment of the risk associated with new loan applications by making clear patterns and factors that are indicative of potential default.

Data analysis steps:

To analyze if a relationship exists between the outcomes of previous loan applications (approvals, refusals, cancellations) and missed payments, we must look at how the columns “TARGET” and “NAME_CONTRACT_STATUS” interact with each other. Before analyzing if such a relationship exists, a test must be conducted to determine if any applications are missing a contract status or target value, and therefore, should be removed prior to examining this relationship. Fortunately, no nan values could be found in the columns “NAME_CONTRACT_STATUS” or “TARGET”, signaling that no further cleaning needed to be done to proceed. Afterwards, a bar plot was made to count the number of occurrences of each contact type in the dataset created from “appication_data”. Reasons being to get a sense of how the contract types are distributed within the dataset, and more (see Appendix 2A-1). As the analysis is between the client’s contract status and the likelihood of default, the key columns to study are “NAME_CONTRACT_STATUS” and “TARGET”. However, as they come from different datasets, the 2 datasets had to be merged on their common variable, 'SK_ID_CURR'. Following these preparations, a count plot was created in order to compare and visualize the rate of default across each contract type (see Appendix 2A-2), allowing for conclusions to be drawn on how clients from each contract status fare when it comes to repaying their loans.

When analyzing the distribution of down payments by contract status, we had to merge the previous_application and application_data to be able to use ‘NAME_CONTRACT_STATUS’ and

‘AMT_DOWN_PAYMENT’ columns for the graph. Then with these 2 columns, we printed out the descriptive statistics which are the count, mean, standard deviation, minimum, 25th percentile (Q1), median (50th percentile or Q2), 75th percentile (Q3), and maximum. We used facet grids with density plots to visualize the comparison between the distribution of the down payments and different contract statuses as seen in B-1. There are 4 graphs displayed with each of them representing different contract statuses. This allows for a clear and direct comparison between each status since each subplot is separated from each other. We can see that the approved down payments have the most amount of density, refused payments as second, unused offers as third, and lastly canceled since it has no density at all. Density in this graph refers to the probability of the distribution of payment.

When analyzing the default rates payment range, we declare ‘down_payment_bins’ which is a list that defines the edges for categorizing the amount of down payments into different ranges. This is later used to categorize the ‘AMT_DOWN_PAYMENT’ into groups. After that, we create a new column in the ‘merge_data’ called ‘Down_Payment_Range’. This will contain values representing the down payment range for each ‘AMT_DOWN_PAYMENT’. Then, we calculate the mean of the ‘TARGET’ column for each value in the ‘Down_Payment_Range’ column through the ‘default_rates’ variable. This will contain the mean default rate for each downpayment range. We used a bar plot since it helps us visualize the default rates across different down payment ranges more easily. We know that the payment range is definite and the rates are numerical values which means that the bar plot is more effective. We see that each bar represents a down payment range and the y-axis gives us a pay range. The graph displays that the more down payment there is, the lower the default rate they receive as shown in B-2. After displaying the graph, we calculate the p-value by doing the Chi-Square Test to assess the significance of the differences in the default rates. If the given p-value is less than 0.05, it shows a significant connection between default status and the down payment range. Since the number is less than 0.05, there seems to be a level of significance involved with the difference in the rates as shown in B-2.

Main results/conclusions:

The factors influencing loan defaults were broken down into some significant discoveries by this data research:

From Appendix 2A-1, we can conclude that the majority of previous loan outcomes were marked as "Approved". "Canceled" was the next most common followed by "Refused" and lastly, "Unused offer". The high count of "Approved" outcomes suggests that a significant proportion of previous loan applications were successfully approved.

From Appendix 2A-2 , We can observe that most of the people who were approved for the loan previously don't have difficulty in paying back. Although this is also the case for the other contract status types, out of the 4 contract statuses, clients with a history of approved loans are the least likely to face payment difficulties in the current application.

In Appendix B-1, we can see that depending on the status of the contract, the resulting density will vary. The contracts that have an “Approved” status have the most probability of payment distribution hence the highest density. Then, the contracts that are “refused payments” come in second for density. After that, the “unused offers” are third then the “canceled” status is last in terms of density meaning they don’t have a chance with the payment distribution.

Appendix B-2 shows the different default rates depending on the down payment of the client. By observing the graph, we can see that the more a client puts down for the down payment, the lower the likelihood of them defaulting hence the lower default rate. On the very left of the bar plot, we can see the bar of clients with the least amount of down payment which makes their default rate high. Then, on the far right side of the bar plot, we can see that the bar of clients which have the highest amount of down payment have the lowest amount of default rate.

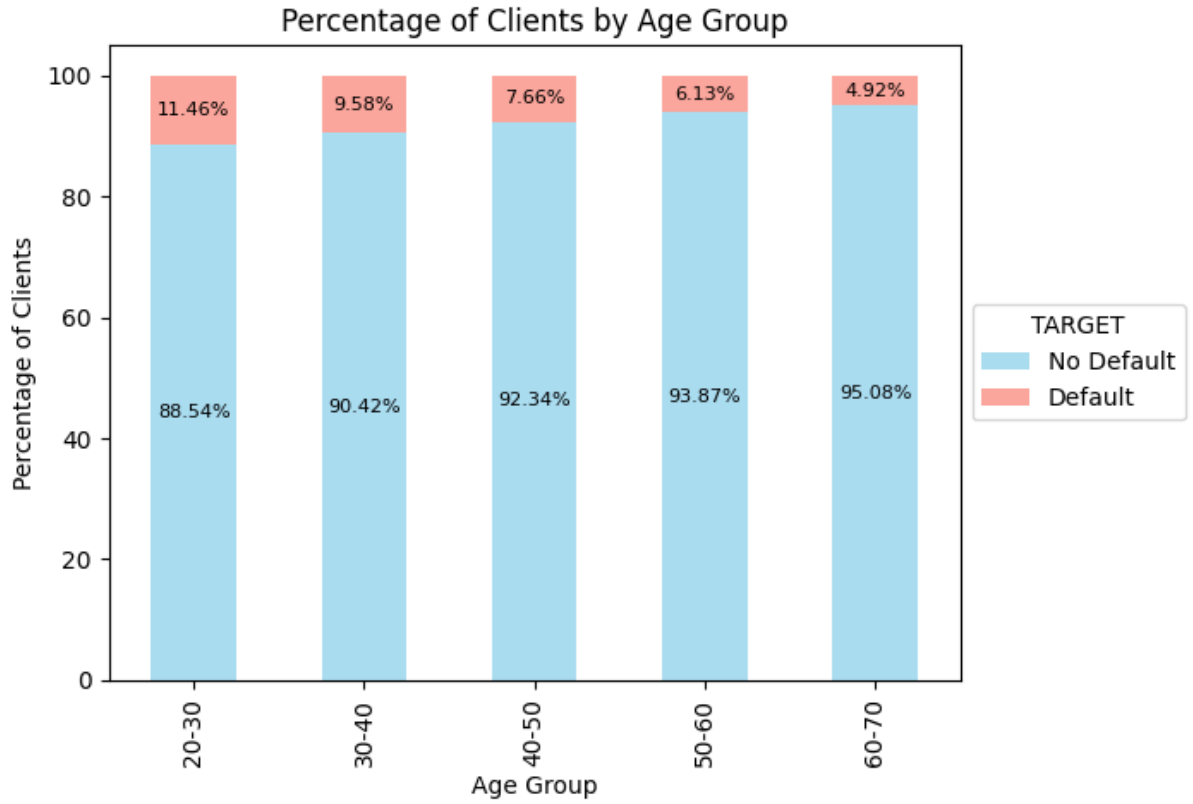
Drawbacks of analysis performed and any concerns:

The data cleaning process does not remove all the outliers for us and we cannot exclude these Loan statuses. From Appendix 2A-3 we use a box plot to display the distribution of loan amount for different loan statuses**,** we can observe that the absence of a box for Canceled loans and the scattering of points may suggest that Canceled loans do not follow a clear pattern in terms of loan amounts. Exceptionally high loan amounts that were approved previously, can be depicted by the outliers on the box plot of Approved loan amounts . This implies that the institution has a flexible range that may be due to specific or unique circumstances. Outliers can be found for Refused loans indicating that there are cases when the institution refused loans for falling outside a certain range. This can be done to avoid risk or may be due to specific financial criteria. These outliers can have positive and negative impacts on the statistical interpretation.

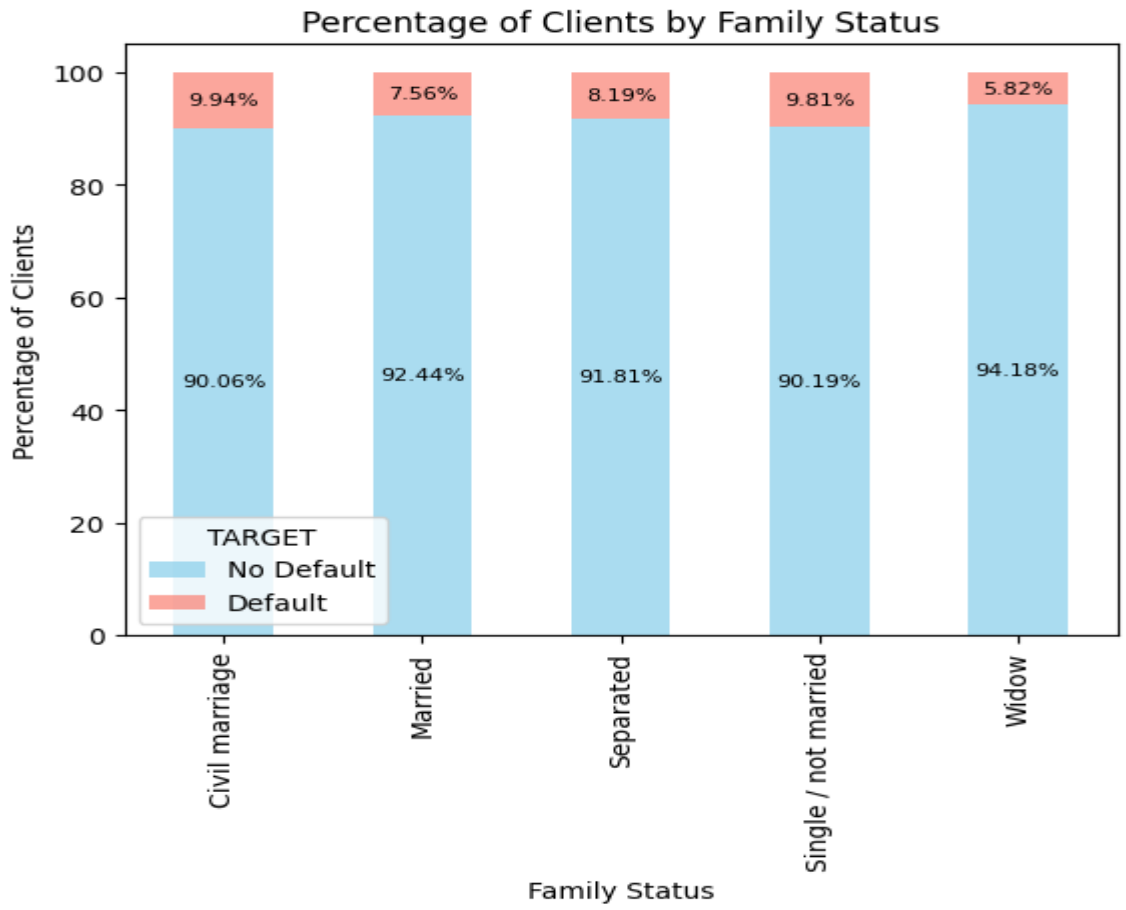
Appendix

A-1

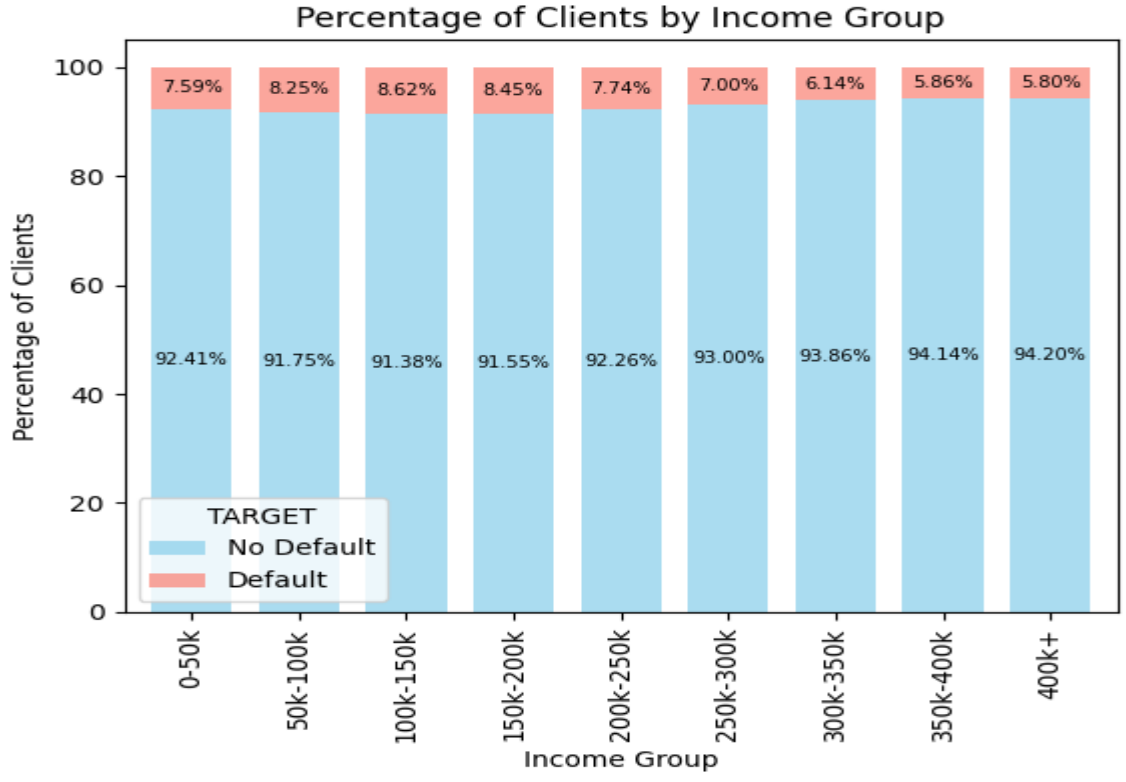
A-2



A-3


Percentage of Clients by Family Status
Percentage of Clients
100
80
60
40
20
0
9.94%
7.56%
8.19%
9.81%
5.82%
90.06%
92.44%
91.81%
90.19%
94.18%
TARGET
No Default
Default
Civil marriage
Married
Separated
Single / not married
Widow
Family Status

A-4

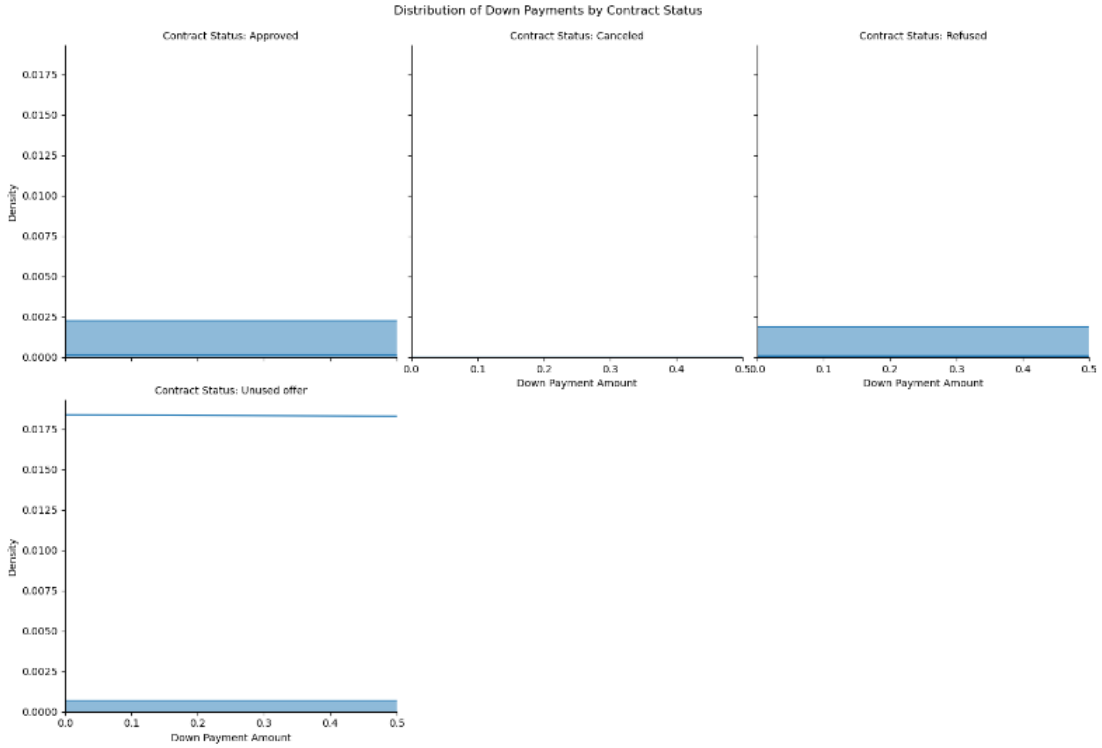

Percentage of Clients by Income Group
Percentage of Clients
100
80
60
40
20
0
7.59%
8.25%
8.62%
8.45%
7.74%
7.00%
6.14%
5.86%
5.80%
92.41%
91.75%
91.38%
91.55%
92.26%
93.00%
93.86%
94.14%
94.20%
TARGET
No Default
Default
0-50k
50k-100k
100k-150k
150k-200k
200k-250k
250k-300k
300k-350k
350k-400k
400k+
Income Group

B-1

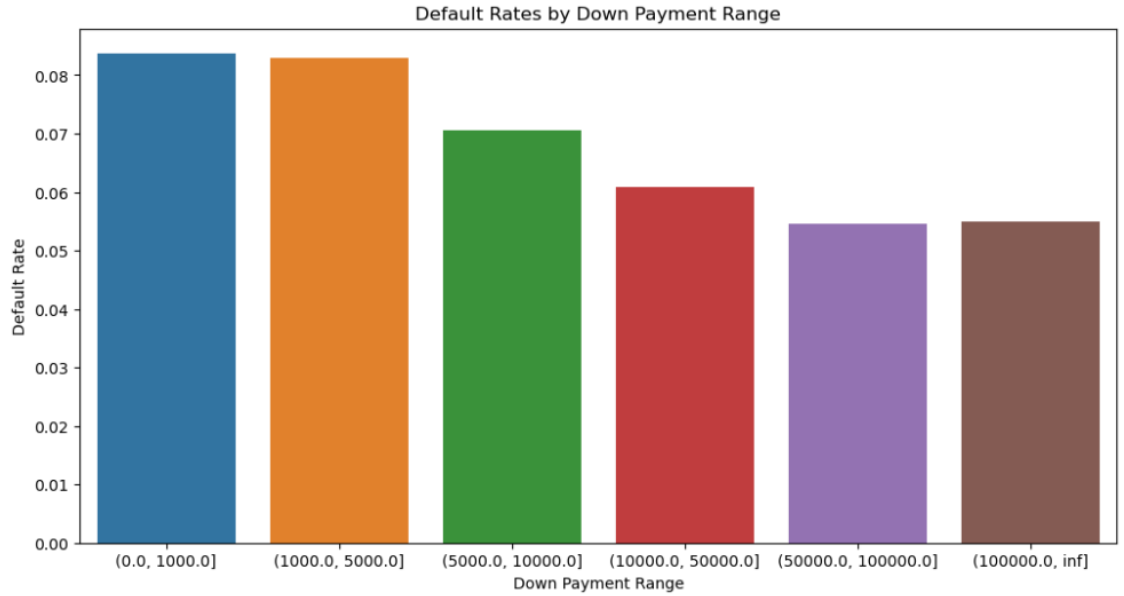
```
                      count          mean           std  min  25%     50%  \
NAME_CONTRACT_STATUS
Approved           568197.0   6832.369469   19304.373164 -0.9  0.0  2322.0
Canceled              536.0  21642.580410  101140.746442  0.0  0.0     0.0
Refused             74778.0   7040.091841   29282.486057  0.0  0.0     0.0
Unused offer        20650.0      1.252809     158.320930  0.0  0.0     0.0

                       75%        max
NAME_CONTRACT_STATUS
Approved            8302.5  3060045.0
Canceled               0.0   918000.0
Refused             6583.5  2475000.0
Unused offer           0.0    22500.0

<Figure size 1200x800 with 0 Axes>
```

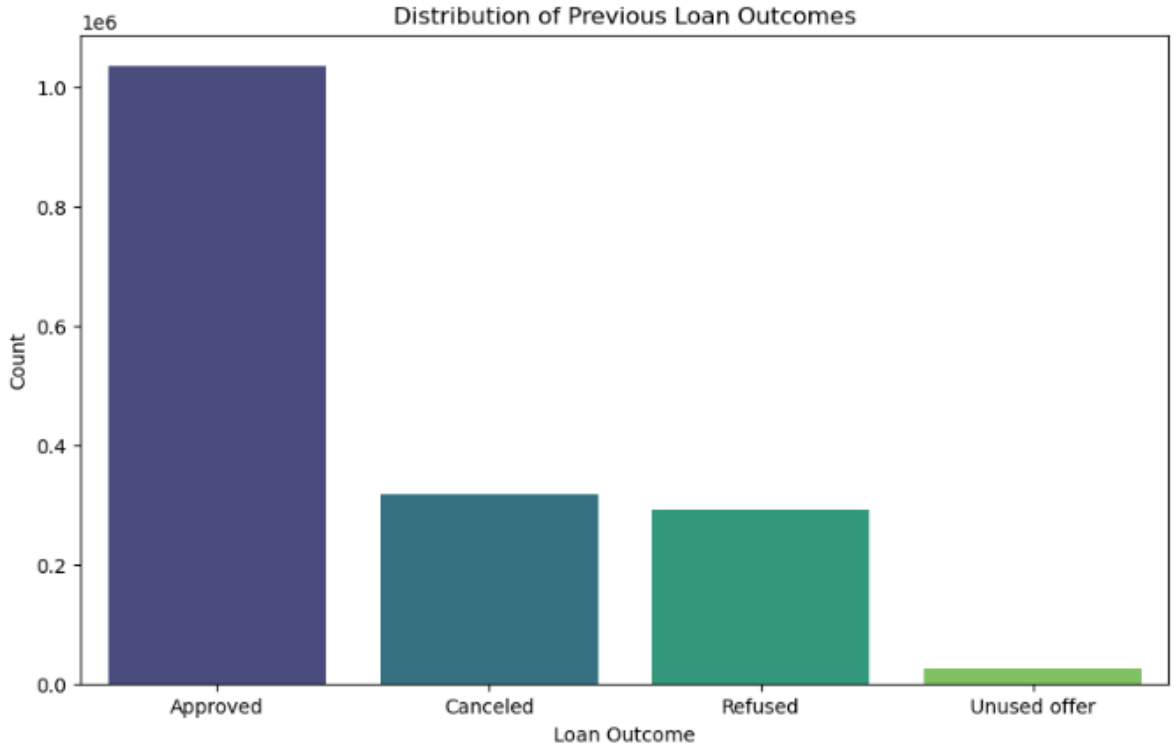


B-2

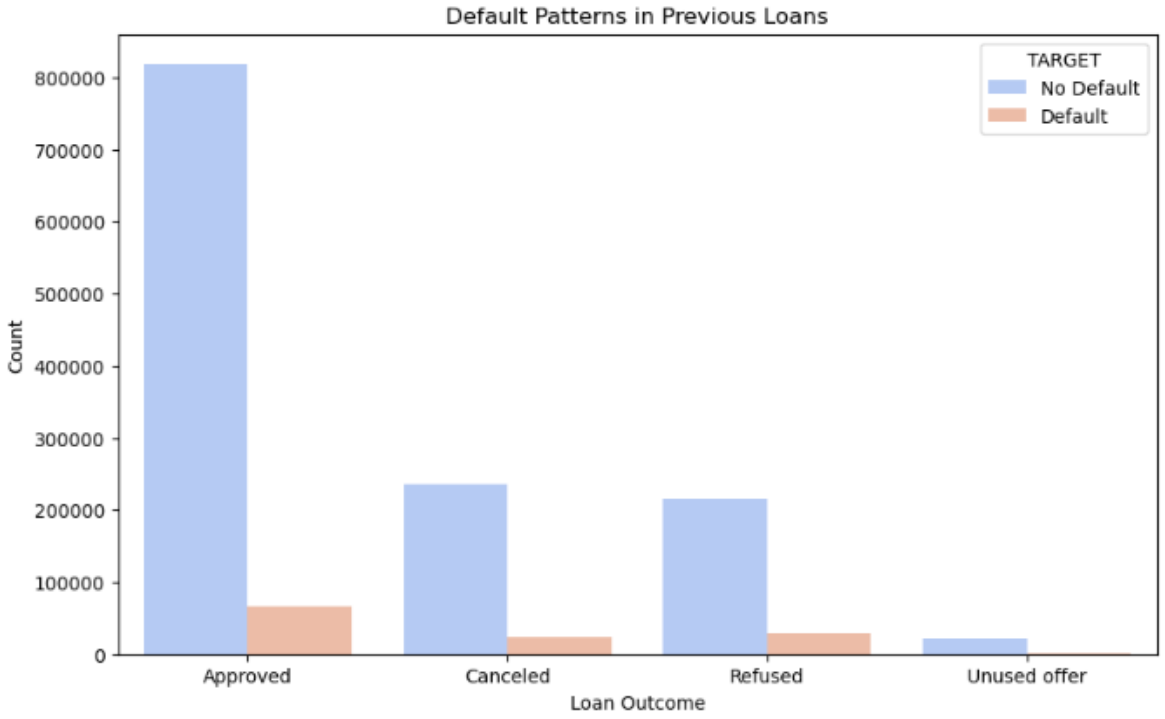



```
Chi-square value: 500.19917417872585
P-value: 7.232105926297502e-106
The difference in default rates between down payment groups is statistically significant.
```

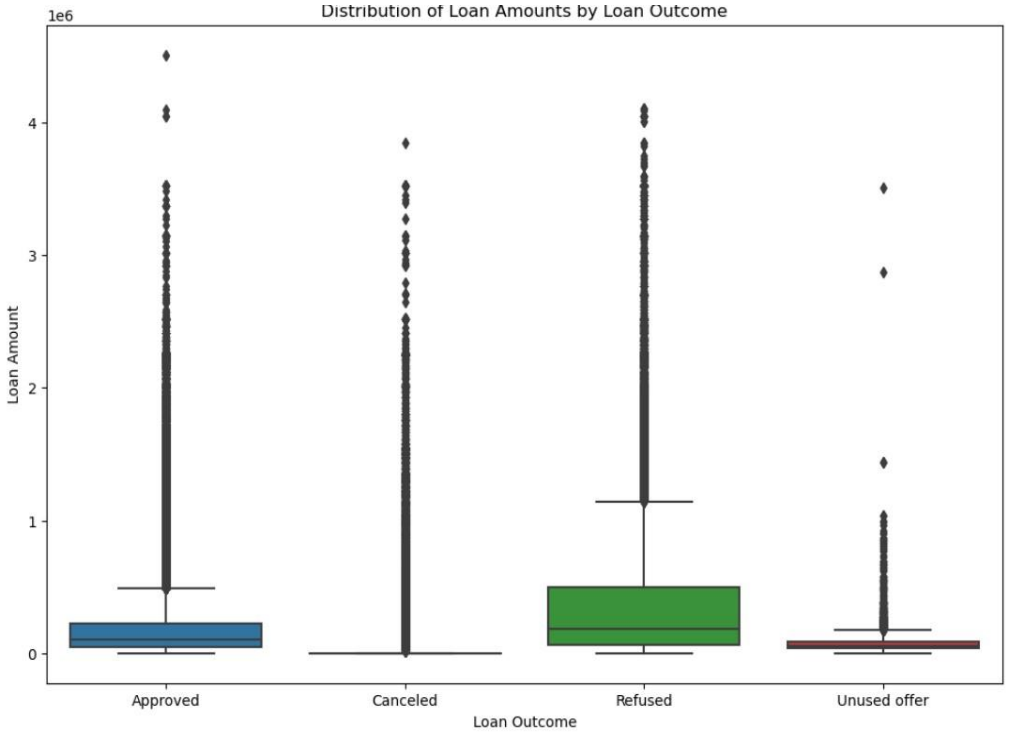
2A-1


Distribution of Previous Loan Outcomes
1e6
1.0
0.8
0.6
0.4
0.2
0.0
Count
Approved
Canceled
Refused
Unused offer
Loan Outcome

2A-2


Default Patterns in Previous Loans
TARGET
No Default
Default
800000
700000
600000
500000
400000
300000
200000
100000
0
Count
Approved
Canceled
Refused
Unused offer
Loan Outcome

2A-3


Distribution of Loan Amounts by Loan Outcome
1e6
4
3
2
1
0
Loan Amount
Approved
Canceled
Refused
Unused offer
Loan Outcome