Data Description:

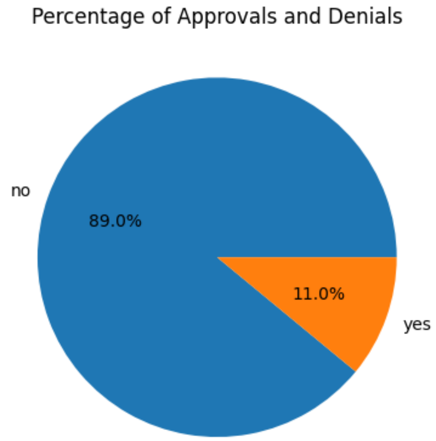
The data contains 100,000 data values that are representative of different customers. Building from that, the data is comprised of 14 variables that consist of two identifier variables (User ID and applications), also; the data has 7 categorical variables (Including the Approved variable) and 5 numerical variables (Including the bounty variable) that encompass a profile to determine if a lender should approve or deny a customer for a loan.

Analysis on All Data:

Initially I removed the User ID and applications variables from the data frame for ease of manipulation and cleaning. Also, I dealt with the missing values within the Employment_Sector column. I input the mode of the column for the NA values because it is efficient at preserving the distribution and presence of the highest Employment_Sector value. My last step for manipulating the data was engineering two new variables:

- **Monthly_income_after_housing** → represents the portion of income left over after paying monthly housing bill.
- **Percentage_income_housing_payment** → represents the percentage of total income is spent of housing payment each month.

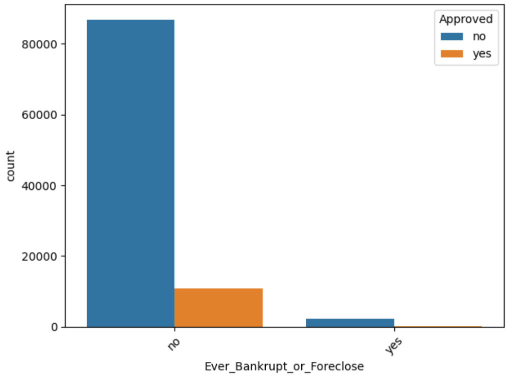
These new variables bring more emphasis to the customer spending habits in relation to income. One main observation lender look at is a customer's spending habits and if they are responsible with their money. After cleaning the data, I first looked at the approval rate (in pie graph) and total revenue.

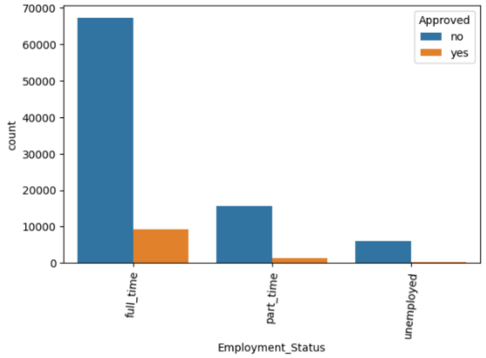


Total Revenue: $2.64 Million

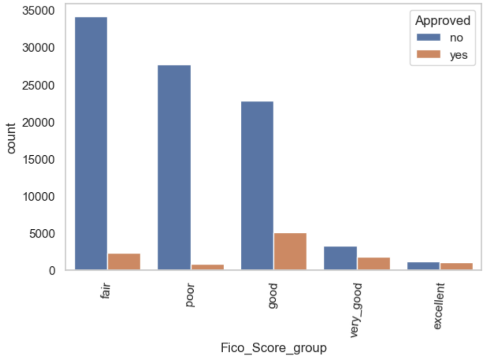
This presents a major point of interest because we can see from the pie chart that most of the applicants being sent out to lenders are getting denied and hindering the revenue opportunities from approved applicants.

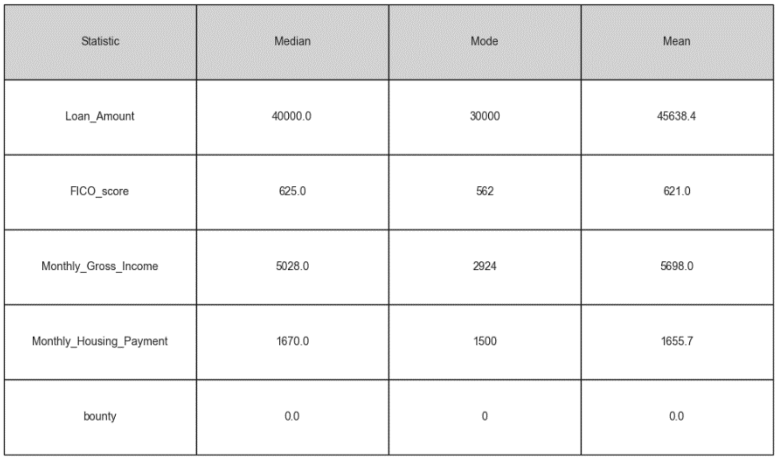
Next, I dove into the representation of the categorical variables when applicants are approved and denied. First, the variables that do not see much variation in determining the approval of a loan: Reason, Employment_Sector, and Loan_Amount. These variables saw a relatively even distribution in the approval quantities throughout their values creating a notion that lenders do put much weight on a customer’s reason for taking out a loan and their occupation alone. Next, the variables that appear to have the most leverage in approvals when connotated as a certain value are: Ever_Bankrupt_or_Forclose, Fico_score_group, and Employment_Status. As seen in the graphs these variables saw a high approval rate for certain characteristics. This aligns with loan approval standards because these variables revolve around a major contribution to loan approval—credit and income history.





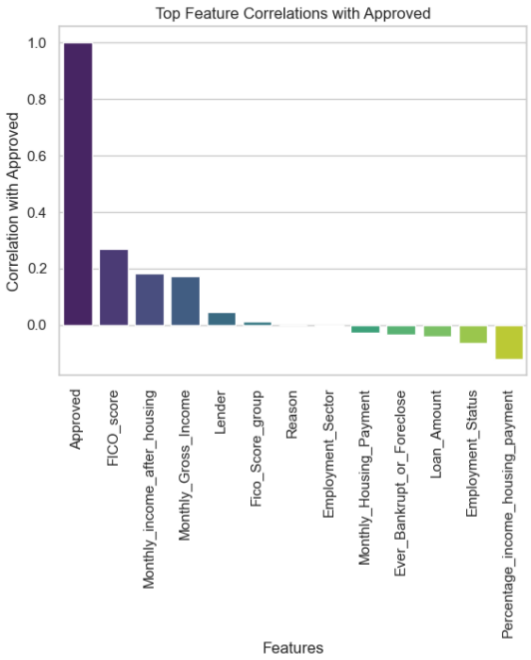
After that, I looked at the importance of numerical columns in the dataset. I calculated key summary statistics to get a grasp on the distribution, spread, and central tendency. Looking at the statistics I can see that FICO_score and Monthly_House_Payment has a relatively normal distribution and establish central tendency with a similar median and mean. On the other hand, Loan_Amount and Monthly_Gross_Income skew to the right presenting the idea of these values have a lower mean than median and castrates more of a peak.





Statistic	Median	Mode	Mean
Loan_Amount	40000.0	30000	45638.4
FICO_score	625.0	562	621.0
Monthly_Gross_Income	5028.0	2924	5698.0
Monthly_Housing_Payment	1670.0	1500	1655.7
bounty	0.0	0	0.0

The last analysis I conducted was looking at the correlation of all variables with the Approved variable. The correlation between each variable with the Approved column presents a linear relationship between each variable with the approval decision. The graph to the right shows the correlation coefficients. From this graph it is evident that variables pertaining to credit history and income (original and newly created) have the strongest relationship negatively or positively with the Approval process.



From this analysis, the most important variable to help with decision making process revolve around the financial stature of the customer (FICO score, income, and expenses). At the same time, all variables seem to show some strength in creating a decision so no variable will be excluded from the analysis.

<u>Lender Analysis:</u>

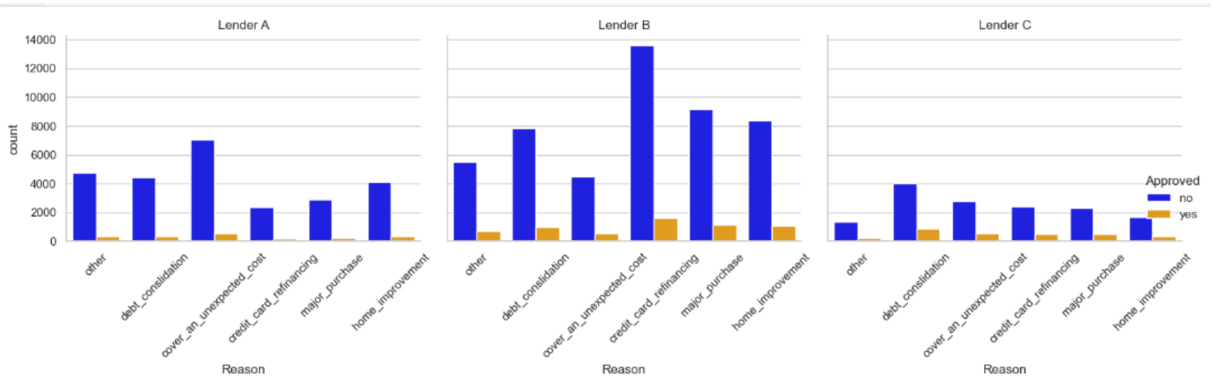
There are three lenders present within the data set: A, B, C. I will conduct a similar analysis as seen on the overall data set.

Average Approval Rate Lender A: 11.0%

Average Approval Rate Lender B: 7.1%

Average Approval Rate Lender C: 17.1%

Building from this, Lender A and B see a much higher quantities of customers compared to Ledner C. But at the same time the variation in approved categorical values appears consistent throughout the three vendors. Except for the reason variable with Lender A showing unexpected costs, Lendor B having credit card refinancing, and Lender C favoring debt consolidation as the prominent reason (This can be seen in the graph below). On top of that, the distribution of numerical variables throughout the three venders remains intact from the overall analysis and does not sway too heavily in a difference for any of the three lenders.



Lastly, comparing the correlation between all variables and approved for each of the lenders showcases a similar standpoint as the correlation of the overall dataset. The order of importance through correlation is consistent within each of the three lenders with only minor alterations. This can be reasoned because approving loans comes with strict guidelines. However, with the subtle differences potential focal points could be given to each lender as their main priority when it comes to what a customer is desiring their loan for.

Predictive Classification Model:

The model used to predict the ideal customer for each lender is a Random Forest Classifier. This model generates various decision trees and creates predictions from an overall assumption from the trees.

For this model, I split all three of the lenders' data into three different data sets and conducted an 80/20 split into test and train sets. From there I fit the data to Random Forest Classifier with the default parameters.

Findings:

After running the model, I crafted an ideal customer for each of the lenders.

Lender A:

- **Reason**: Debt Consolidation
- **Loan Amount**: 20,000
- **FICO score**: 689
- **FICO score group**: 2
- **Employment Status**: full-time
- **Employment Sector**: Communication Services
- **Monthly Gross Income**: 18,536
- **Monthly Housing Payment**: 2,302
- **Ever Bankrupt/Foreclosed**: No
- **Monthly Income After Housing**: 16,234
- **Percentage Income Housing Payment**: 12.42

Lender B:

- **Reason**: Other
- **Loan Amount**: 80,000
- **FICO score**: 813
- **FICO score group**: 0
- **Employment Status**: full-time
- **Employment Sector**: Energy
- **Monthly Gross Income**: 13,475
- **Monthly Housing Payment**: 1,116
- **Ever Bankrupt/Foreclosed**: No
- **Monthly Income After Housing**: 12,309
- **Percentage Income Housing Payment**: 8.65

Lender C:

- **Reason**: Credit Card Refinancing
- **Loan Amount**: 10,000
- **FICO score**: 588
- **FICO score group**: 1
- **Employment Status**: full-time
- **Employment Sector**: Health Care
- **Monthly Gross Income**: 8,277
- **Monthly Housing Payment**: 500
- **Ever Bankrupt/Foreclosed**: No
- **Monthly Income After Housing**: 7,777
- **Percentage Income Housing Payment**: 6.04

These predictions provide a foundation for grouping customers into certain lenders. These predictions tested against a subset of data from their respective lenders saw very positive approval rating:

- **Lender A**: 64% approval for ideal customer
- **Lender B**: 51% approval for ideal customer
- **Lender C**: 70% approval for ideal customer

These numbers shed light on the importance of providing the correct customer to the correct lender.

Furthermore, looking at maximizing approval rate and revenue. If taking an average of all approval rates for the test subjects, there is an overall approval rate of 61.67% -- over a 50% increase in approval rating. Not only that, but a increase in approval rating also directly correlates with a vast increase in overall revenue.

- **Lender A (55000 customers):** (55000 * 0.64) * 250 = **$8,800,000**
- **Lender B (27500 customers):** (27500 * 0.51) * 350 = **$4,908,750**
- **Lender C (17500 customers):** (17500 * 0.70) * 150 = **$1,837,500**
- **Total Revenue** = 15.55 million
- **Previous Revenue** = 2.64 million

- **Increase in Revenue = 12.91 million (589%)**

These findings provide exciting insight into the opportunity to increase approval rating and revenue through strategically sorting customers to the right vendor. In the future, this model has potential to improve on the decision-making process within the pairing of customers with lenders. However, there is still room for improvement due to the nature of the data set size and limitation in variables present. On top of that, the next step would be generating a concrete range of variable values for each lender to swiftly pinpoint each customer to the perfect lender to increase approval ratings and total revenue.