MIT 5032 Group Project

Group B

Caroline R. Marley Nicholas, Ibrahim Kondeh, Shohruz Junaidov,

Sam Harkins, Colby Heiden

Analysis of Factors Influencing Loan Approval

I. Problem Statement, Business Questions, and Related Hypothesis

Loan approvals are critical for both individuals and the economy as a whole since they directly affect borrowers' capacity to attain financial goals like purchasing a home, supporting school, or establishing a business. Understanding the factors influencing loan approval decisions is critical for lenders, borrowers, and policymakers alike. Lenders must minimize risk while making loans to eligible candidates. Borrowers need to know how to increase their chances of approval, and policymakers can benefit from insights that assist build a more equitable and inclusive lending climate.

Initially, the goal of this analysis was to investigate factors influencing loan defaults. However, upon reviewing the available data, it became clear that there was insufficient information to properly model and analyze loan defaults, as key variables such as loan payment history or delinquency data were not included. Given these limitations, we decided to shift the focus to a more feasible and relevant business objective: analyzing loan approval outcomes. The purpose of this project is to investigate and analyze the important elements that influence loan approval decisions, with the goal of providing data-driven answers to various critical business concerns. These questions assist clarify the relationships between borrower characteristics, loan attributes, and approval outcomes, resulting in a more complete understanding of the dynamics involved in loan approval procedures.

Preliminary Hypothesis and Intuition

Business Questions

The analysis aims to answer several business questions:

1. How does annual income affect loan approval likelihood, particularly for borrowers with varying levels of employment experience?
2. To what extent does a borrower's credit score influence the probability of loan approval, especially for those with shorter credit history lengths?
3. What is the relationship between homeownership status and the likelihood of loan approval, considering the impact of credit score?

4. How does the loan interest rate affect approval, particularly for borrowers with high loan-to-income ratios?
5. How does the purpose of the loan influence approval rates, especially for educational loans?

Business Hypotheses:

1. **Hypothesis 1:** Borrowers with lower annual income are less likely to have their loans approved, with this effect being more pronounced for individuals with less employment experience.
   Categories analyzed: Income (person_income), Employment Experience (person_emp_exp). Types: Continuous numerical, discrete numerical.

2. **Hypothesis 2:** Borrowers with lower credit scores are less likely to have their loans approved, especially among those with shorter credit history lengths.
   Categories Analyzed: Credit Score (credit_score), Credit History Length (cb_person_cred_hist_length). Types: Continuous numerical, discrete numerical.
3. **Hypothesis 3:** Borrowers who own homes are more likely to have their loans approved compared to those who rent, particularly among borrowers with higher credit scores.
   Categories Analyzed: Homeownership Status (person_home_ownership), Credit Score (credit_score). Types: Categorical nominal, continuous numerical.
4. **Hypothesis 4:** Higher loan interest rates decrease the likelihood of loan approval, with this effect being stronger for borrowers with higher loan-to-income ratios.
   Categories Analyzed: Loan Interest Rate (loan_int_rate), Loan-to-Income Ratio (loan_percent_income). Types: Continuous numerical, continuous numerical.
5. **Hypothesis 5:** Loan intent affects the likelihood of loan approval, with education loans being more likely to be approved for borrowers with higher annual incomes, especially for those with a longer credit history.
   Categories Analyzed: Loan Intent (loan_intent), Annual Income (person_income), Credit History Length (cb_person_cred_hist_length). Types: Categorical nominal, continuous numerical, discrete numerical.

II. Data Source and Description:

The data source was a Kaggle dataset for loan information involving 45,000 borrowers. This dataset includes various borrowers' characteristics and loan attributes to help analyze factors influencing loan approval. It is a public data set available for use.

The key borrower characteristics analyzed include age, credit score, employment status, annual income, and homeownership status. These variables provide insights into the financial background of each applicant. Loan attributes such as loan amount, loan intent, interest rates, and loan-to-income ratios are also included to understand the features of each loan.

Using functions like dim and colnames, the data dimensions and column names were defined. The dataset originally had 45,000 rows and columns with no missing values. The columns included variables like person age, income, employment experience, credit score, and loan details such as amount, interest rate, and loan status. These steps are further described in the preprocessing section below.

The dataset provides a solid foundation for understanding the factors that determine loan approval outcomes, with no data completeness issues due to the absence of missing values. However, it is important to note that the dataset is synthetic, meaning it was generated to simulate real-world data. This may introduce limitations related to its representativeness and applicability to actual loan approval scenarios.

III. Data Preprocessing Steps

Each hypothesis required different preprocessing steps that are outlined in each hypothesis. For variables like employment experience, income, credit history, credit score group, and interest rate, it was necessary to group them into categories to better understand the dataset.

Additionally, since we discovered the dataset was synthetic, there were some outliers in the values of certain variables. For example, there were rows with 125 years of work experience and income amounts over one million dollars, which seems unrealistic. Therefore, for analysis, the dataset was filtered to exclude rows where person_income > 500000, person_emp_exp > 50, or person_home_ownership = "OTHER". After filtering, a random sample of 10,000 rows was selected for further analysis. In using a smaller sample of the data, we hoped that the visualizations would better prove or disprove the hypotheses.

IV. Analysis, Approach, Outcomes, and Insights

Hypothesis 1: Impact on Loan Approvals Based on Annual Household Income and Previous Employment Experience

Hypothesis one looks at the relationship between loan approval rates, income, and employment experience. The logic surrounding this hypothesis is that those with greater amounts of income and more work experience would seem more reliable to pay back loans.

Data Analysis and Preparation

To better understand the data, we thought it would be helpful to create bins for income and employment experience. Income was split into 5 categories: "Very Low" ($8000-$47,483), "Low" ($47,483-$67,023), "Moderate" ($67,023-$94,615), "High" ($94,615-$132,684), "Very High" ($132,684+). Employment Experience was split into 4 categories: 0-10 years, 10-20 years, 20-30 years, and 30 years or more.

Findings

Figure 1 shows the distribution of income within the dataset. It is left skewed showing that most people in the dataset make between roughly $50,000 and $80,000. There are very few people in the dataset with an income over $300,000.

Figure 2 shows the percentage of loan approval based on income and employment experience. People with very low income tend to be more likely to be approved for a loan as compared to people with high or very high income. Additionally, the data shows that there are more people who were rejected for a loan than approved. The exception to this was people with low income and 30 or more years of work experience, where their loan approval percentage was higher than the rejection percentage. Overall, however, people with 30+ years of work experience have the highest percentage of loan approvals.

Putting together these graphs and learning this information was very surprising. What the graphs show goes against the first part of hypothesis one because we did not predict people with lower income to have higher loan approval percentages. There are a few different reasons as to why this could be the case. The first is that the dataset was synthetic instead of drawing from real world data. With randomly drawn up data, it is more difficult to predict the relationships between different variables. The second is that lenders could look at young people just entering the job field with low income and see high potential earnings. This could lead lenders to believe these people will likely pay back their loans if approved. The third being that people who need a co-signer on loans could ask someone with a good credit history to make the person asking for a loan look more reliable to pay back the loan. The fourth is people who ask for smaller loan amounts are more likely to be approved for loans as compared to people asking for larger amounts. It would make more sense for people with lower incomes to ask for smaller amounts of money. Lastly, as shown by figure 1, there are a lot more people who fall into the categories of very low income and low income. Since that is the case, it makes sense that there are more loan approvals for people in those income categories.

Conclusion

People with low income are eligible to apply for personal loans. Some lenders require borrowers to have a minimum amount of income that they disclose, while others do not. Lenders want to ensure that applicants have stable income to show that they will repay the loan amount that they have asked for in addition to interest. However, those with lower income tend to have higher interest rates on their loans as opposed to people with higher income. There are multiple other factors besides income like credit scores, DTI ratio, and documentation that play a factor in whether someone will be approved for a loan (US News). This would explain why people with low income still qualify for loans. Additionally, lenders look at borrowers' employment history to approve someone for a loan. Generally, two years of stable employment is all that is necessary, but lenders also look for time gaps in employment (MRC). Since all that is generally necessary for employment experience is 2 years, those with greater than 30 years have significantly more to show for their reliability in loan repayment.

Figure 1: Distribution of Annual Income

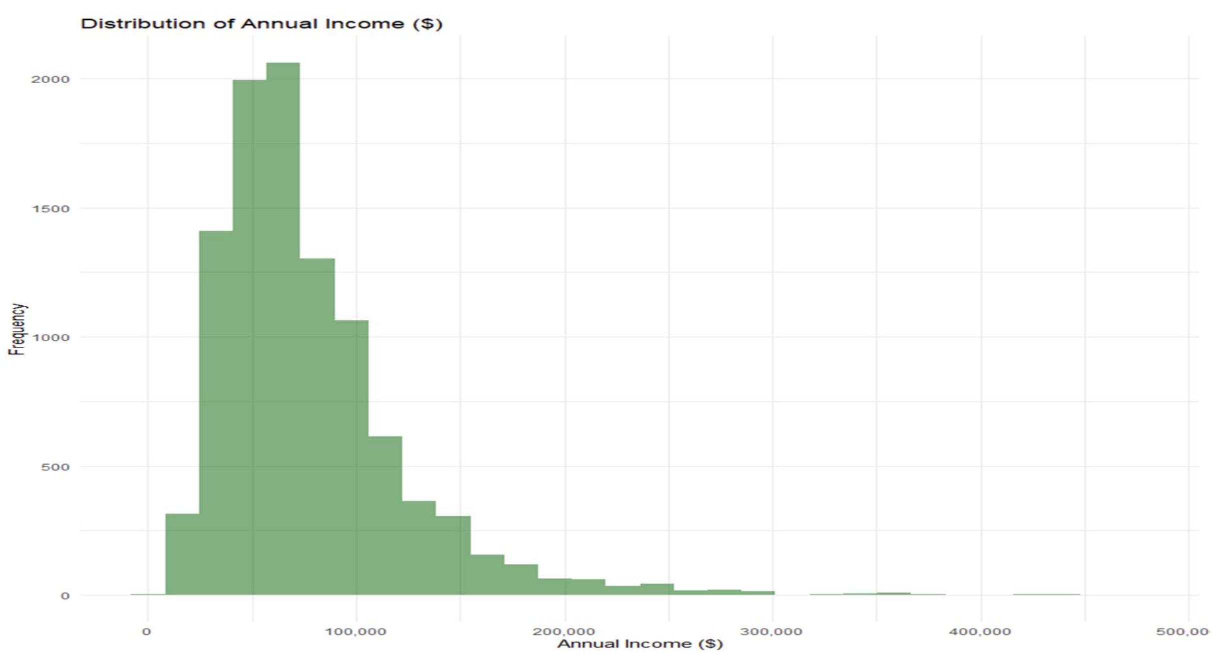
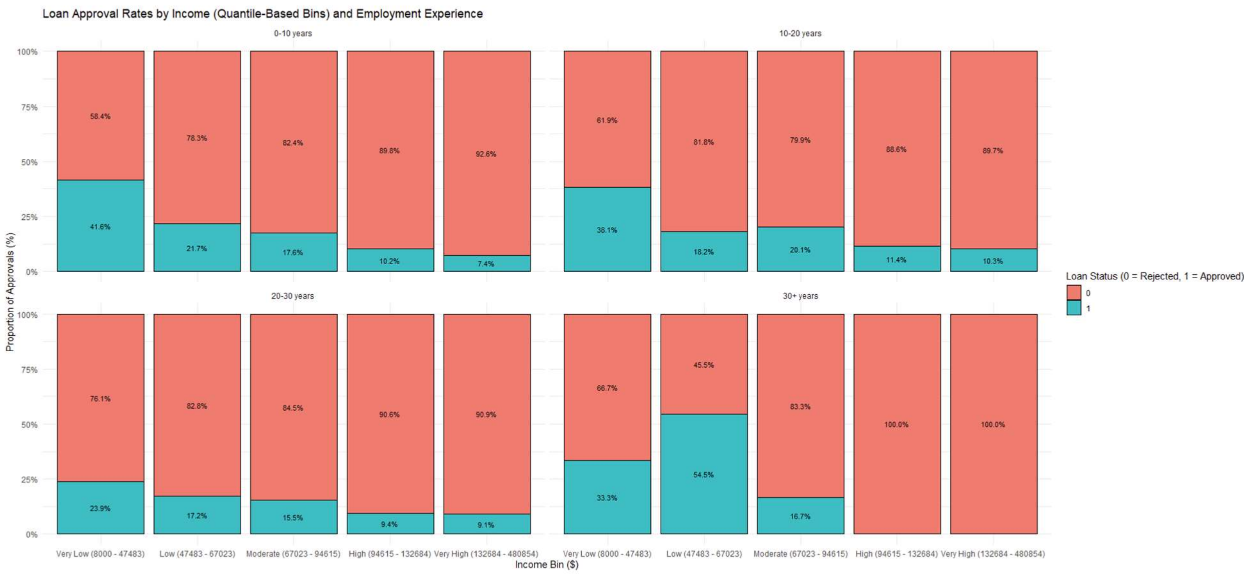


Figure 2: Loan Approval Rates by Income and Employment Experience



Hypothesis 2: Impact on Approval Rates Based on Credit History Length and Credit Score

Hypothesis 2 looks into the correlation between credit history length, credit score, and approval status. The intent of this hypothesis is to find relations between the variables to understand the importance of having a good credit score and keeping it consistent over multiple years.

Data Analysis and Preparation

For hypothesis 2, we started by splitting the credit history length in groups by 10 years, leaving years 0-5 in a group by themselves to be as accurate as possible. This gives us bins of 0-5 years, 6-15, 16-25, and 26 plus. Along with the credit history lengths, we also have credit scores split into groups: Low (301-500), Medium (501-650), High (651-700), and Very High (701-850).

Findings

In Figure 3, we see the approval percentage is very high, which is a little interesting. My eyes were first drawn to the 16-25 credit history graph, looking at the low (301-500) credit score section. That section of the graph being 100% approved is interesting. As well as the

26+ credit history graphs the very high (701-850) credit score section has the highest rate of decline, which again we would think the opposite as borrowers. From these findings, we decided to take a step back and simplify our searches. We created Figure 4, which is just the approval rate based on the credit score bins. In this figure, we see a small difference based on the credit score bins. Overall, almost all the bins have the same percentage of approvals compared to denials, which makes our figure 4 even more confusing. Finding this odd and very far from my hypothesis, we took a deeper look to find the data set we used was synthetic data. Meaning, the data was generated, so it does not give an accurate representation of real-world happenings. By doing a correlation heat map, we are able to see which variables would cause noticeable changes in a potential hypothesis. The variables I looked at in my hypothesis were not shown to affect the overall loan status, so seeing very equal graphs over the board makes sense. In good faith of the hypothesis asked, we looked into other research with real life data to see if the findings could support my hypothesis. According to Bankrate, I was able to find that my hypothesis was correct. Although credit history length does not drastically change the credit score like payment history, it is still responsible for 15% of your credit history. Although 15% is not a lot, what you really learn is the habits over the years. For example, lenders are more likely to approve borrowers who can show their history of making payments, especially when they have a good credit history to show for it. A good credit score does a lot for borrowers, but 15% of what that is composed of is the length of credit history, which can easily be enough to make someone get approved or declined. All things consistent between 2 different individuals seeking a loan, 15% is enough to push someone to get accepted and push another to get declined.

Figure 3: Loan Approval Rate by Credit Score and Credit History Length

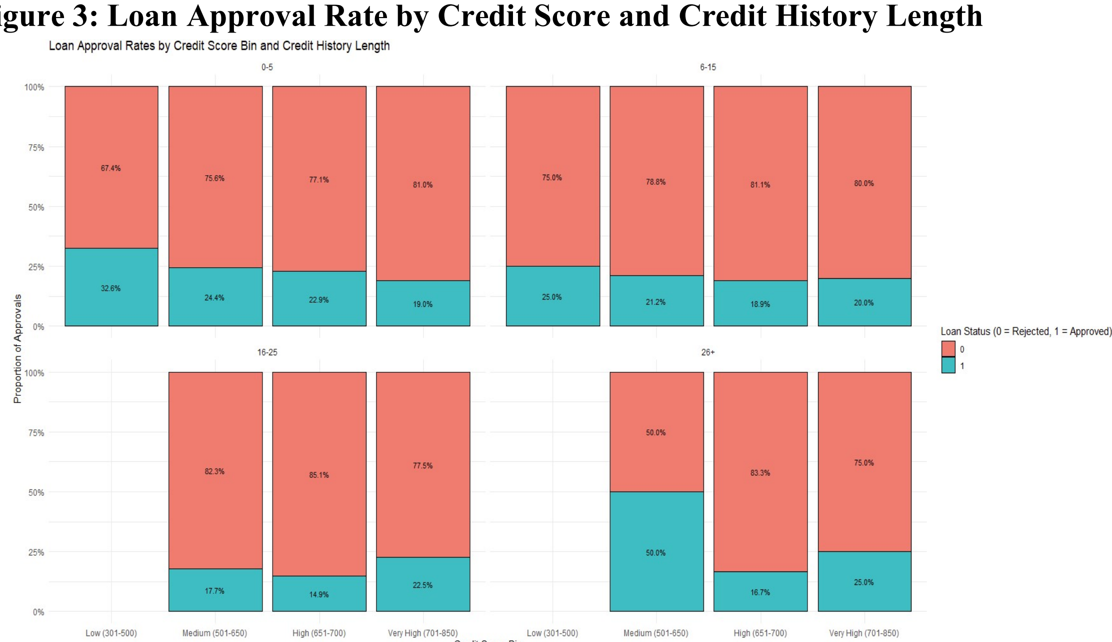
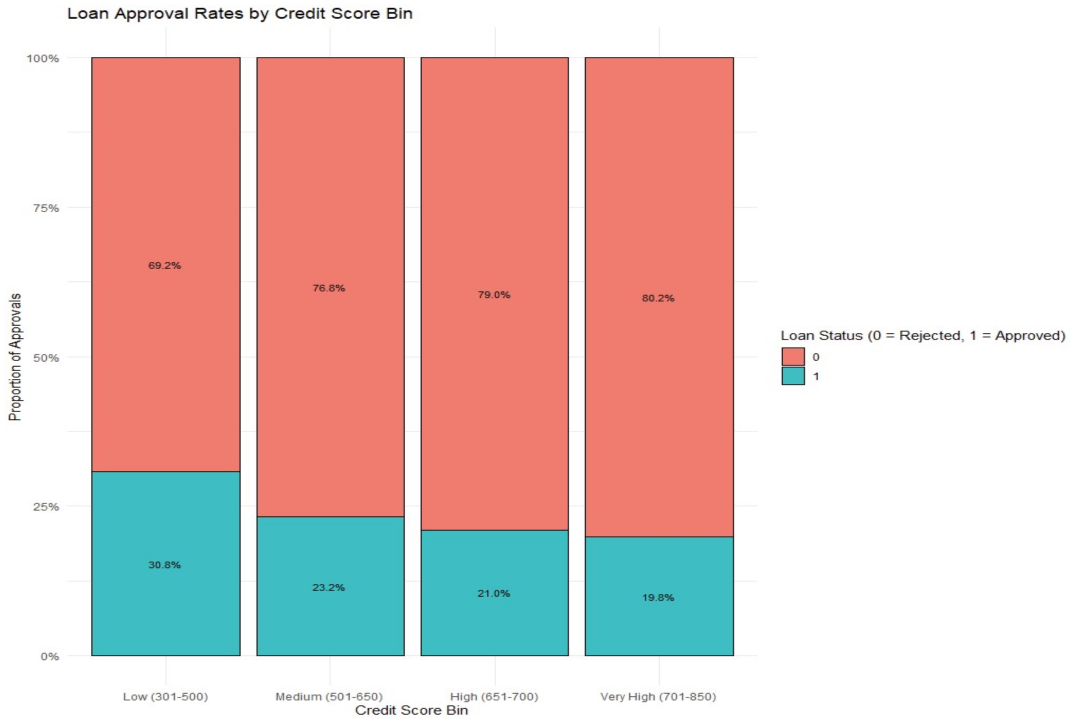


Figure 4: Loan Approval Rates by Credit Score Bin



Hypothesis 3: Impact of Loan Interest Rates and Loan-to-Income Ratios on Approval Likelihood

Loan approvals are influenced by several factors, including interest rates and the loan-to-income (LTI) ratio. Hypothesis 3 explores how higher loan interest rates impact the likelihood of loan approval, with a stronger effect on borrowers with higher LTI ratios. The goal is to analyze these variables in-depth and assess their combined influence on loan approval outcomes.

Data Analysis and Preparation

To analyze the impact of interest rates and LTI ratios, two key variables were selected: loan_int_rate (loan interest rate) and loan_percent_income (loan-to-income ratio). These variables were binned into categories to allow for better comparisons. Loan interest rates were categorized as Low (5–8%), Moderate (8–12%), High (12–15%), and Very High (15%+). Similarly, LTI ratios were divided into Low (0–10%), Moderate (10–20%), High (20–30%), and Very High (30%+). This categorization facilitated the aggregation of approval rates within each group for a more nuanced understanding of their relationships.

The summary statistics reveal important characteristics of the dataset. Loan interest rates ranged from 5.42% to 20%, with a median and mean of 11.01%. Most interest rates were clustered between 8.59% (1st quartile) and 12.99% (3rd quartile). Loan-to-income ratios ranged from 0 to 63%, with a mean of 13.89% and a median of 12%. The majority of borrowers had LTI ratios between 7% and 19% (1st and 3rd quartiles), indicating that most loans were applied for by individuals with moderate borrowing-to-income levels.

Findings

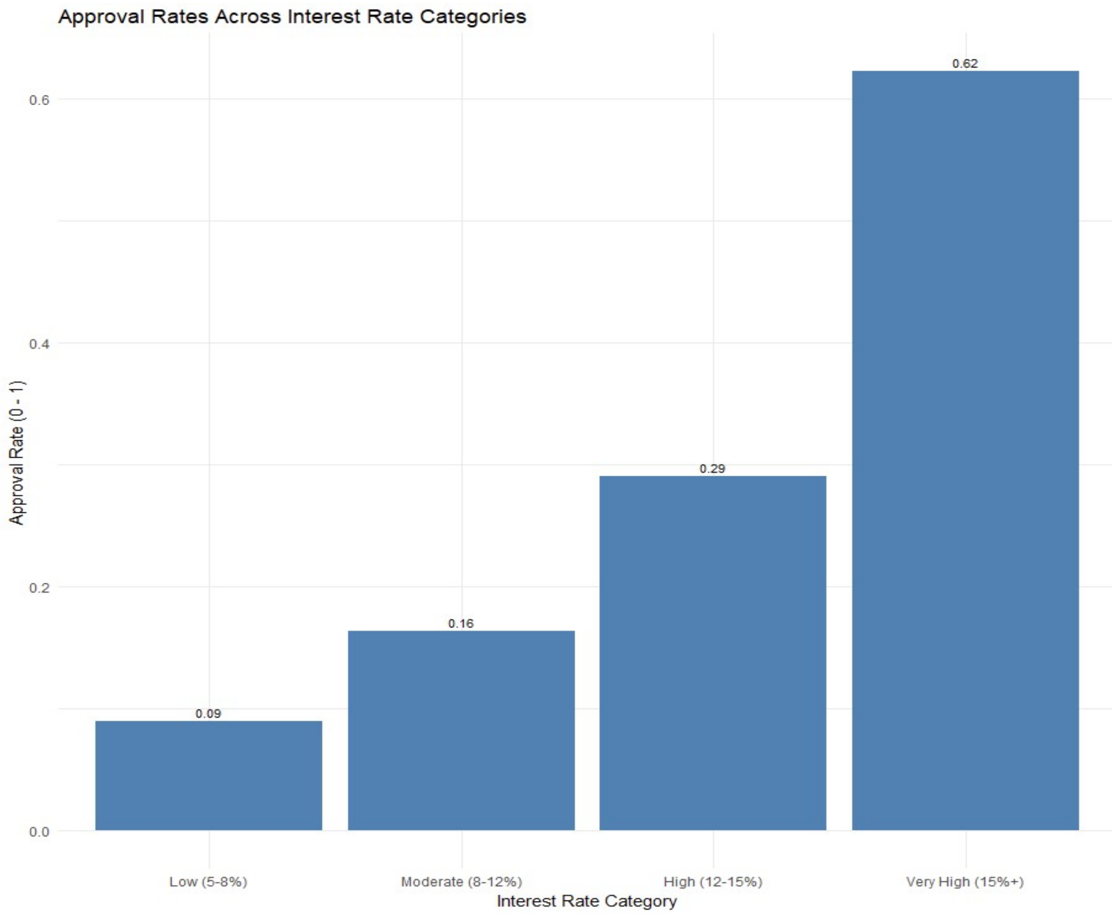
1. Distribution of Interest Rates and LTI Ratios

The distributions of loan interest rates and LTI ratios provide critical context for the findings. The histogram of interest rates indicates a concentration of loans in the Moderate (8–12%) range, with fewer loans in the Very High category. Similarly, most LTI ratios are below 30%, suggesting that borrowers tend to request loans proportional to their income. This cautious borrowing behavior likely contributes to the overall approval trends observed in the dataset.

2. Impact of Loan Interest Rates on Approval Rates

The data shows an interesting trend: loan approval rates increase as interest rates get higher. For loans with low interest rates (5–8%), the approval rate is just 8.9%. This goes up to 16.3% for moderate interest rates (8–12%) and 29% for high interest rates (12–15%). Surprisingly, loans with very high interest rates (15%+) have the highest approval rate of 62.3%.

Figure 5: Approval Rates Across Interest Rate Categories

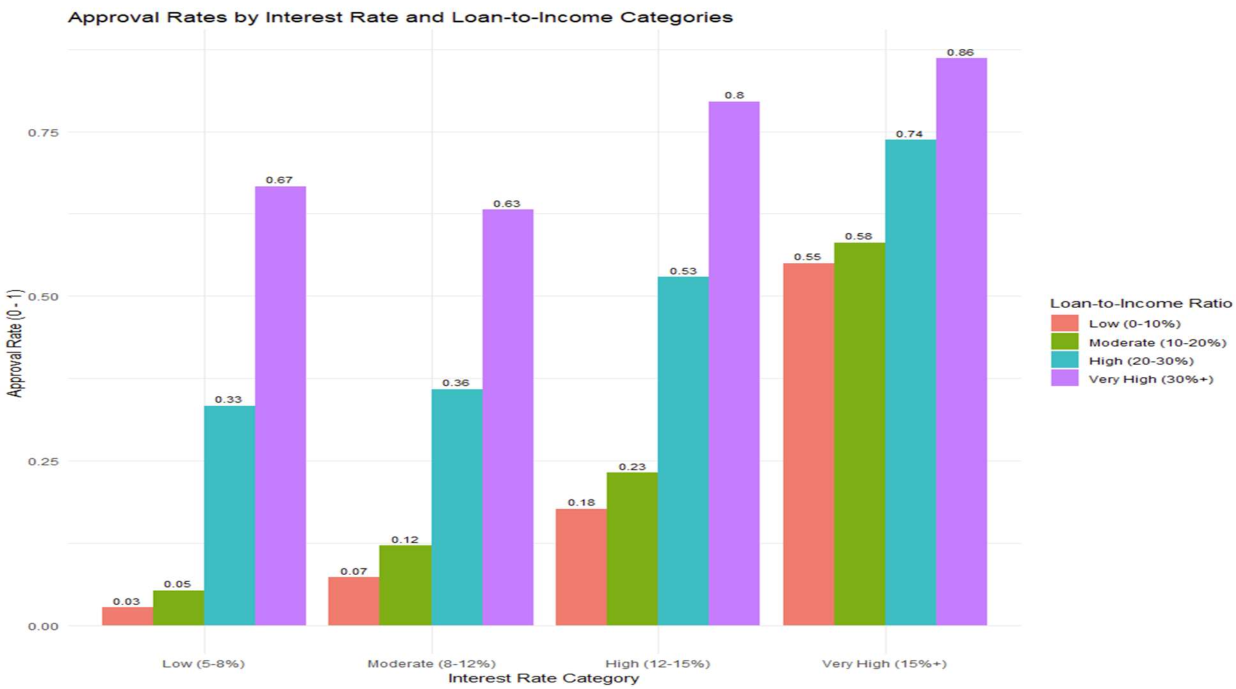


At first, this might seem unusual because higher interest rates are usually seen as risky. But this trend suggests that loans with very high interest rates are approved carefully and only for borrowers who are likely to have strong financial stability. These borrowers may have higher incomes or valuable assets that make them less risky for lenders, even at higher interest rates. Essentially, lenders are willing to take the chance because these borrowers seem like safe bets despite the high cost of the loan.

3. Interaction Between Interest Rates and LTI Ratios

The combined effect of interest rates and LTI ratios on approval rates reveals significant interactions. For borrowers with higher LTI ratios, approval rates are consistently lower across all interest rate categories. For example, at Low (5–8%) interest rates, approval rates drop from 3% for borrowers with LTI ratios below 10% to just 0.03% for those with LTI ratios above 30%. A similar trend is observed at Moderate (8–12%) and High (12–15%) interest rates, where higher LTI ratios are associated with lower approval rates. However, for Very High (15%+) interest rates, approval rates improve significantly, even for borrowers with LTI ratios above 30%, reaching 86.1%.

Figure 6: Approval Rates by Interest Rate and Loan-to-Income Categories



This trend suggests that loans with higher interest rates are more likely to be approved for borrowers demonstrating financial stability, such as strong income or collateral. Lenders may use such selective approvals to mitigate the risks associated with high-interest loans.

Conclusion

The analysis challenges the idea that higher interest rates always reduce the chances of loan approval. Instead, it shows that loans with very high interest rates (15%+) are actually more likely to be approved, especially for borrowers with high loan-to-income (LTI) ratios. This suggests that lenders may approve these loans for borrowers who appear financially stable or have strong profiles, even if the interest rates are high.

For borrowers, it’s important to keep LTI ratios low and try for loans with lower interest rates to improve approval chances. Lenders, on the other hand, should focus on making their approval criteria fair and consistent, ensuring that they’re evaluating risk properly across all interest rate levels. This analysis reminds us that understanding the details of a borrower’s financial situation is key to making better lending decisions.

Hypothesis 4: Impact of Homeownership and Credit Score on Approval Likelihood

Hypothesis 4 states that borrowers who own homes are more likely to have their loans approved compared to those who have a mortgage, followed by those who rent, particularly among borrowers with higher credit scores. We created this hypothesis as research suggests that homeownership can positively impact loan approval by showcasing a history of responsible credit management. Consistently making mortgage payments on time could boost credit score, which in turn makes an applicant appear as a reliable borrower to lenders when applying for new loans. We agreed that owning a home can signal financial stability, enhancing the chances of loan approval. Research also suggests that “credit score and credit history are one of the primary factors that lenders consider” when determining the result of the loan application. To test this hypothesis, our team conducted a detailed analysis on our chosen dataset involving several steps to prepare and analyze the data.

Data Analysis and Preparation

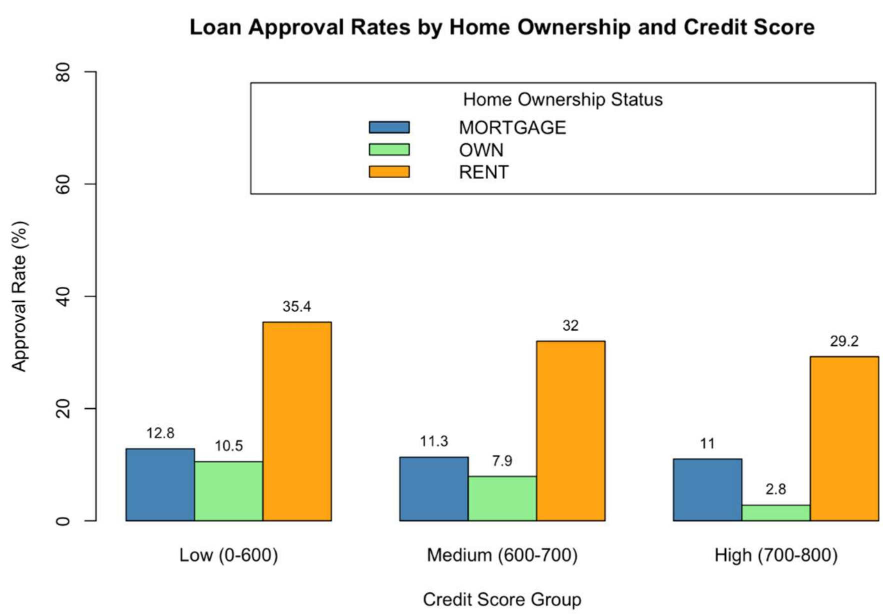
First, we decided to filter the dataset to remove rows where homeownership was classified as “other”. The quantity of this category was much smaller in comparison to the others and gave a misleading percentage. This step was essential to focus our analysis on the primary categories of Own, Rent, and Mortgage. Additionally, we created a binary variable to indicate the loan approval status, with 1 representing approved loans and 0 representing rejected loans. This binary variable allowed us to compute approval rates more effectively. Next, we categorized the credit scores into three distinct groups: Low (0-600), Medium (600-700), and High (700-800). This binning was important for us to analyze the impact of credit scores on loan approval rates across different home ownership statuses. We then calculated the loan approval rates by home ownership status and credit score group and converted these rates into percentages for easier comparison. The resulting data was visualized in a bar plot, which showed the loan approval rates by home ownership and grouped by credit score. The plot used different colors to distinguish between the categories of homeownership, providing a clear visual representation of the data.

Findings

The analysis of this synthetic dataset revealed that borrowers who rent homes generally exhibit a higher loan approval rate compared to those who have a mortgage or own. This finding does not support the hypothesis, as both the home ownership and mortgage categories have lower approval rates compared to the rent category. Additionally, within each home ownership category, borrowers with lower credit scores tended to have higher approval rates, compared to those with higher credit scores. (See figure 7)

These results highlight some unexpected trends that contradict common assumptions, particularly that homeownership is generally seen as an indicator of financial stability, which often leads to higher approval rates. It is also important to note that this is a synthetic dataset and does not fully encapsulate the complexities and nuances of real-world approval processes. Many other factors must be taken into consideration when determining the reasoning behind approvals/rejections. Therefore, further validation with real-world data would be more suitable to determine these trends.

Figure 7: Loan Approval Rates by Home Ownership and Credit Score



Conclusion

Although our analysis and findings do not follow our hypothesis or logical reasoning, it is important to understand how we can use R to find inaccuracies and unexpected outcomes from datasets. For example, our examination of the relationship between home ownership status and loan approval rates went against our hypothesis. Several factors can contribute to such discrepancies. One significant factor is the use of synthetic datasets, which may not perfectly represent real-world scenarios and can introduce biases and inaccuracies. Using R for data analysis involves rigorous data manipulation, cleaning, and statistical testing. Each step must be carefully executed to ensure accuracy and reliability.

Nonetheless, encountering results that defy the hypothesis is a valuable part of the scientific process. Although we did not find success in our findings for hypotheses 4, valuable methods were used that could be utilized in other datasets that require data validation.

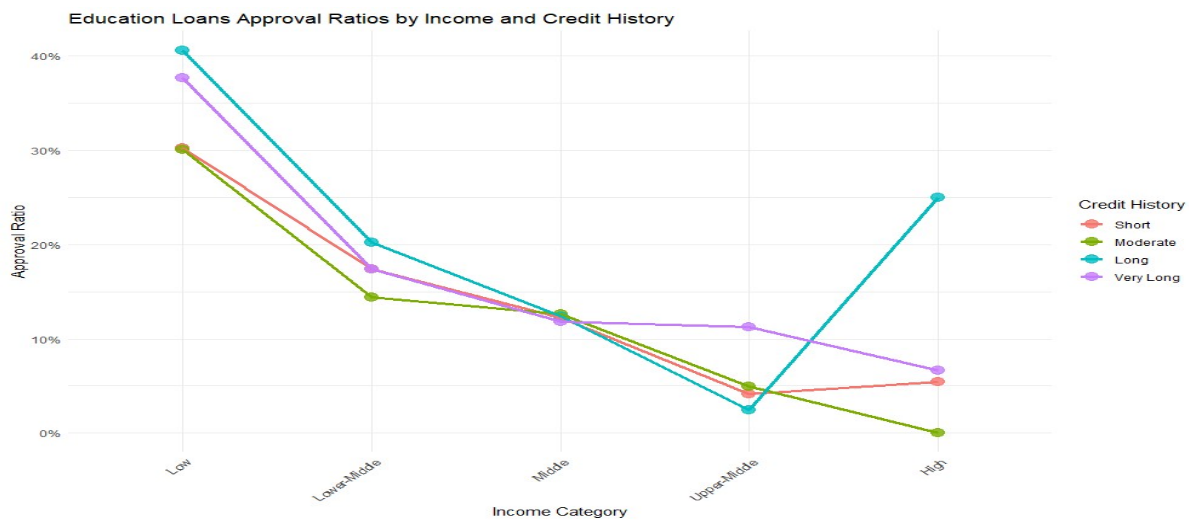
Hypothesis 5: Impact on Education Loans Based on Credit History Length and Annual Household Income

The hypothesis holds that loan intent influences the likelihood of loan approval, with education loans more likely to be approved for borrowers with higher annual salaries, particularly for those with a longer credit history. However, the results from Graph X contradict this expectation. The data reveals that borrowers in the lower-income category with longer credit histories have the highest approval ratios, exceeding 40%, while borrowers in the higher-income category show a significant increase in approval only when their credit history is very long (above 8 years). For income groups in the middle range, approval ratios drop below 10% across all credit history lengths, indicating that neither higher income nor credit history consistently guarantees approval. (*See figure 8)*

These insights suggest that approval decisions for education loans may prioritize borrowers from lower-income groups, particularly those with well-established credit histories, rather than strictly favoring higher-income applicants. For borrowers with short or moderate credit histories, approval ratios remain low across all income groups, further emphasizing the importance of credit history length in influencing outcomes.

In summary, this divergence from the hypothesis can be attributed to targeted lending policies aimed at supporting borrowers from underserved communities and lower-income groups, who may otherwise lack access to higher education financing. As the **National Center for Education Statistics (n.d.)** highlights, borrowers from underserved backgrounds, such as Black and Native American populations, often face systemic financial barriers but are more likely to be approved for loans due to initiatives designed to increase educational equity. These programs focus on addressing disparities in income and access, explaining why borrowers with lower incomes and longer credit histories experience higher approval rates, contrary to the hypothesis' original premise.

Figure 8: Education Loans Approval Ratios by Income and Credit History



Key Learnings

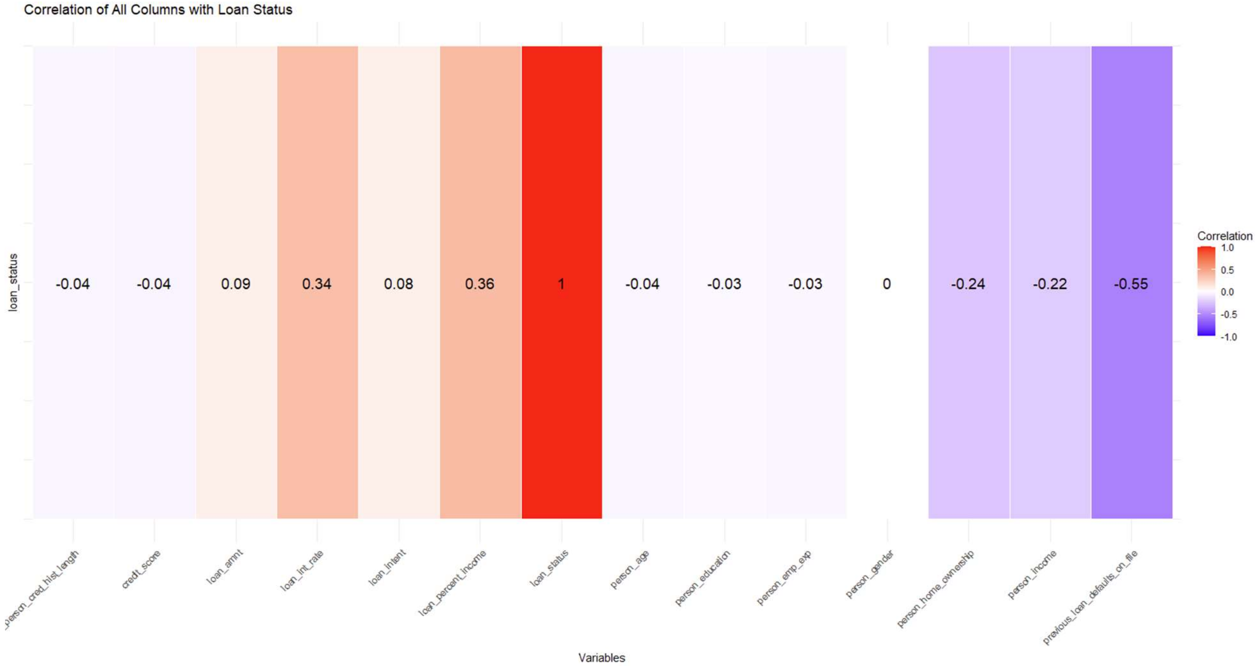
Our comprehensive analysis of loan approval factors revealed several unexpected insights that challenge conventional assumptions about lending practices:

1. **Income and Employment Experience**: Contrary to initial expectations, borrowers with lower incomes, particularly those with extensive work experience (30+ years), demonstrated higher loan approval rates. This suggests that lenders value long-term employment stability and potential future earning capacity over current income levels. The analysis highlights that factors beyond simple income metrics play crucial roles in loan approval decisions.
2. **Credit History and Scores**: While traditional wisdom suggests that longer credit histories and higher credit scores guarantee loan approval, our synthetic dataset revealed nuanced and counterintuitive patterns. The findings underscore the complexity of credit evaluation, indicating that lenders consider multiple interconnected factors beyond straightforward credit score classifications.
3. **Interest Rates and Loan-to-Income Ratios**: Surprisingly, loans with very high interest rates (15%+) showed the highest approval rates, especially for borrowers with higher loan-to-income ratios. This challenges the assumption that high-interest loans are universally riskier. Instead, it suggests that lenders may use high-interest rates as a risk management strategy for borrowers they perceive as financially stable.
4. **Homeownership and Loan Approval**: Our analysis contradicted the hypothesis that homeownership significantly improves loan approval chances. Renters unexpectedly

showed higher approval rates across different credit score categories, highlighting the limitations of using homeownership as a sole indicator of financial reliability.

5. **Educational Loan Dynamics**: Education loan approvals demonstrated a unique pattern, with lower-income borrowers having higher approval rates, particularly those with longer credit histories. This insight potentially reflects targeted lending policies aimed at increasing educational access for underserved communities.

Figure 9. Correlation of All Columns with Loan Status



Summary

The data analysis revealed some interesting insights about the factors that influence loan approval decisions. The most significant findings are:

Previous loan defaults have the strongest negative correlation with loan approval. This suggests that a borrower's credit history and past repayment behavior are critical considerations for lenders.

Loan-to-income ratio and loan interest rate also show moderate positive correlations with approval. Lenders seem to weigh these factors when assessing a borrower's financial stability and ability to repay.

Factors like homeownership status and income level have weaker or even negative correlations with loan approval. This indicates these may not be as important as commonly assumed.

The loan amount requested does not appear to be the single most important factor, contrary to initial expectations.

Conclusion

The analysis highlights the complex and multifaceted nature of loan approval decisions. While traditional financial metrics like credit history and debt-to-income ratios matter, lenders seem to place the greatest emphasis on a borrower's track record of repaying past loans.

This underscores the importance for borrowers to maintain a clean credit history and avoid any previous loan defaults. Factors like interest rates and loan-to-income ratios should also be carefully considered, as they appear to influence approval likelihood.

At the same time, the data suggests that homeownership and income level may not carry as much weight as commonly believed. This challenges some common assumptions and highlights the need for a more holistic and nuanced approach to loan evaluation.

Overall, the insights from this analysis can help both borrowers and lenders better understand the key drivers of loan approval decisions. By focusing on the most impactful factors, they can work together to create a more transparent and equitable lending process.

Recommendations

Based on the insights from the data analysis, here are some key recommendations for both borrowers and lenders:

For Borrowers:

- Focus on maintaining a clean credit history with no previous loan defaults, as this seems to be the single biggest factor impacting loan approval.
- Be mindful of your loan-to-income ratio, as this has a moderate positive correlation with loan approval.
- Consider the loan interest rate, as higher rates also correlate positively with getting approved, though this may come at a higher cost.
- Homeownership and income level appear to be less critical factors, so don't obsess over these as much.

For Lenders:

- Heavily weigh a borrower's history of previous loan defaults, as this has the strongest negative correlation with approval.

- Carefully evaluate the loan-to-income ratio and interest rate, as these moderately positive factors can provide insights into the borrower's financial stability.
- Don't overly prioritize homeownership or income level, as the data suggests these may not be as predictive of creditworthiness.
- Strive for a balanced and transparent approval process that considers the full financial profile of the borrower.

By focusing on these key factors, both borrowers and lenders can work together to create a more equitable and effective lending environment.

References

Anonymous . *How Credit Impacts Your Loan Approval*. Retrieved 2024, July 11. Default. https://mortgage.sirva.com/mortgage-resources/sirva-mortgage-blog/details/2024/07/11/how-credit-impacts-your-loan-approval

Black, M. (2024, April 26). *How length of Credit History Affects Your Score*. Bankrate.

Geffner, Marcie. "Here's What You Should Know about Getting a Personal Loan with Low Income." *U.S. News* , 5 Apr. 2023, money.usnews.com/loans/personal-loans/articles/can-you-get-a-personal-loan-with-low-income.

Lucas, Tim. "Getting a Conventional Loan If You Have an Employment Gap." *Mortgage Research Center*, 8 Nov. 2023, www.mortgageresearch.com/articles/conventional-loan-employment-gap/.

National Center for Education Statistics. (n.d.). *Student loan debt*. U.S. Department of Education. Retrieved December 1, 2024, from https://nces.ed.gov/fastfacts/display.asp?id=900

Retrieved December 1, 2024, from https://www.bankrate.com/personal-finance/credit/length-of-credit-history-credit-score/