Loan Approval Data Analytics Report

Hassan Raza

College of Information, University of Maryland

INST 447: *Data Sources and Manipulation*

Matthew Patrick

December 18, 2024

Section 1. Problem Definition

The overarching problem explored in this study is to identify various characteristics of individuals and the loan itself that increase the likelihood of a loan getting approved. The answer to this problem can be utilized in understanding how loans are approved, what one can do to increase their chances of loan approval, and whether the system of approval is subject to bias.

Data and Variables

Kaggle Link: https://www.kaggle.com/datasets/taweilo/loan-approval-classification-data

The dataset upon which the analysis is done is called Loan Approval Classification dataset and is publicly available on Kaggle. Each row of the data is a recorded instance of a case of an approved or denied loan, alongside various categorical and numerical variables pertaining to the characteristics of the loan seeker and the loan itself. The variables initially within the dataset included age, gender, education, income, employment experience, home ownership, loan amount, loan intent, loan interest rate, loan as percentage of income, credit history, credit score, previous loan defaults, and finally, the target variable of loan status, which could be approved or denied.

However, among these variables, the gender, education, loan interest rate, loan as percentage of income, and credit score were considered variables of interest to be compared with the target variable of loan status. Gender was chosen among the demographic variables to find if there is a relationship with loan status. If so, this would constitute unfair bias on the part of loan approvers and would mean the approval system needs to be reviewed. Education was used as it is a good predictor of both financial stability and maturity, which can influence whether a loan is approved or not. The interest rate was chosen to see whether financial institutions are only approving loans that are of high interest rate. Both credit score and loan as percentage of income were used to see whether they are significant metrics used by loan approvals to make decisions.

Research Questions

While developing the framework for the study, the following research questions were developed to find association between the variables of interest and the outcome variable of loan status:

- Is there any association between gender and loan approval? Do males tend to get more loans approved or females?
- How is an individual's highest education level associated with loan approval? Does higher education mean a higher likelihood of loan approval?
- Does a higher loan interest rate influence loan approval? Specifically, does a higher interest on loan mean that the loan is more likely to be approved?
- What is the relationship between loan to income percentage and loan approval? Meaning does a loan that is lower compared to income get approved more often?
- Does a high credit score individual generally have their loan approved?

This study aims to answer these questions through exploring the relationship between each of the variables and loan status using descriptive statistics, visualizations, and hypothesis testing.

Section 2. Data Description

Variable Descriptions

The following are the variables of interest for the study and their description as provided by the data source:

person_gender (string, categorical): Gender of the applicant person_education (string, categorical): Highest education level of applicant

loan_int_rate (float): Loan interest rate

loan_percent_income (float): Loan amount as a proportion of annual income, loan/income ratio

credit_score (integer): Credit score of the applicant

loan_status (integer): Target variable, loan approval status: 1 = approved; 0 = rejected

Descriptive Statistics

Loan Status by Applicant Gender. The first relationship viewed through descriptive statistics is loan status by gender of the applicant. Although both loan status and person_gender are categorical variables, at this point within the data, the loan status is an integer with 1 for approved and 0 for denied. Calculating the mean of loan status by another categorical variable will provide the proportion of approved loans for each category.

Person_gender	variable	n	mean	sd
female	loan_status	20159	0.222	0.416
male	loan_status	24841	0.222	0.416

We can see that the mean loan status, which is the loan approval proportion, for both males and females is 0.222. This means that within the dataset, there does not seem to be any significant difference observed between proportions of approved loans approved between males or females.

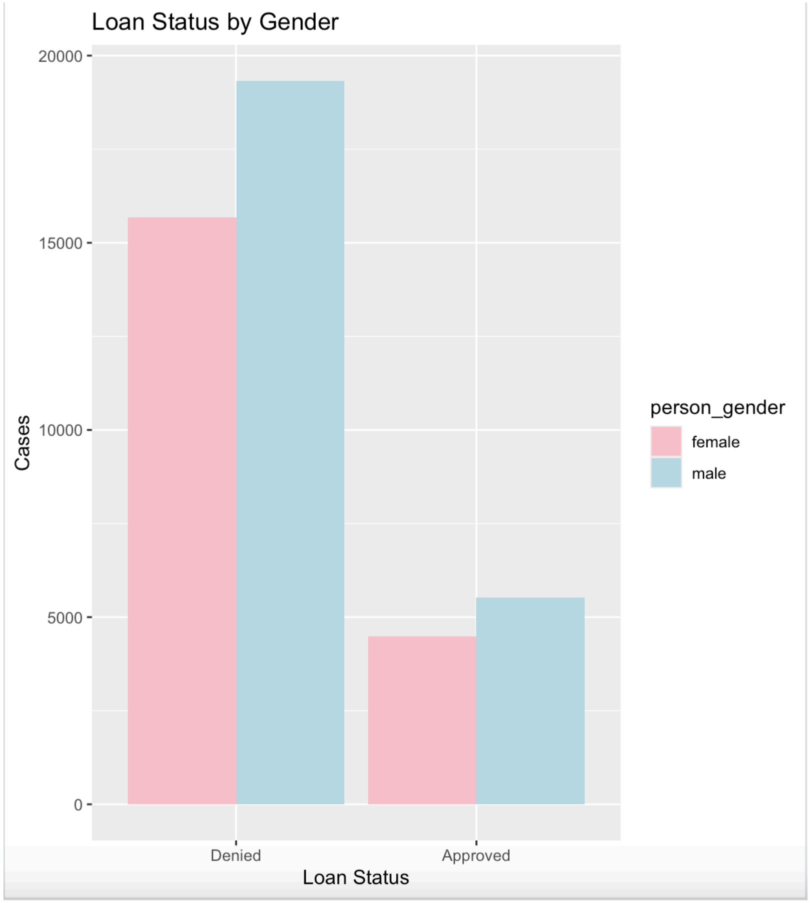

Figure 1

Here we can see that males took the lead in both denied loan and approved loan cases. This justifies the fact that both males and females had the same proportion of approved loans that even though males were approved more loans than females, they were also denied more loans, thereby preserving the proportions.

Loan Status by Applicant Education. The second relationship viewed through descriptive statistics is loan status by education attainment level of the applicant. Similarly, even when loan status and person_education are categorical variables, since loan status is an integer that can only be 1 or 0, calculating the mean provides the proportion of approved loans by education category.

Person_education	variable	n	mean	sd

Associate	loan_status	12028	0.22	0.414
Bachelor	loan_status	13399	0.225	0.418
Doctorate	loan_status	621	0.229	0.42
High School	loan_status	11972	0.223	0.416
Master	loan_status	6980	0.218	0.413

This shows that the highest loan approval rate, which is the mean, was for doctorate degrees (0.229), followed by bachelor (0.225), high school (0.223), associate (0.22), and then finally, master's degree (0.218).

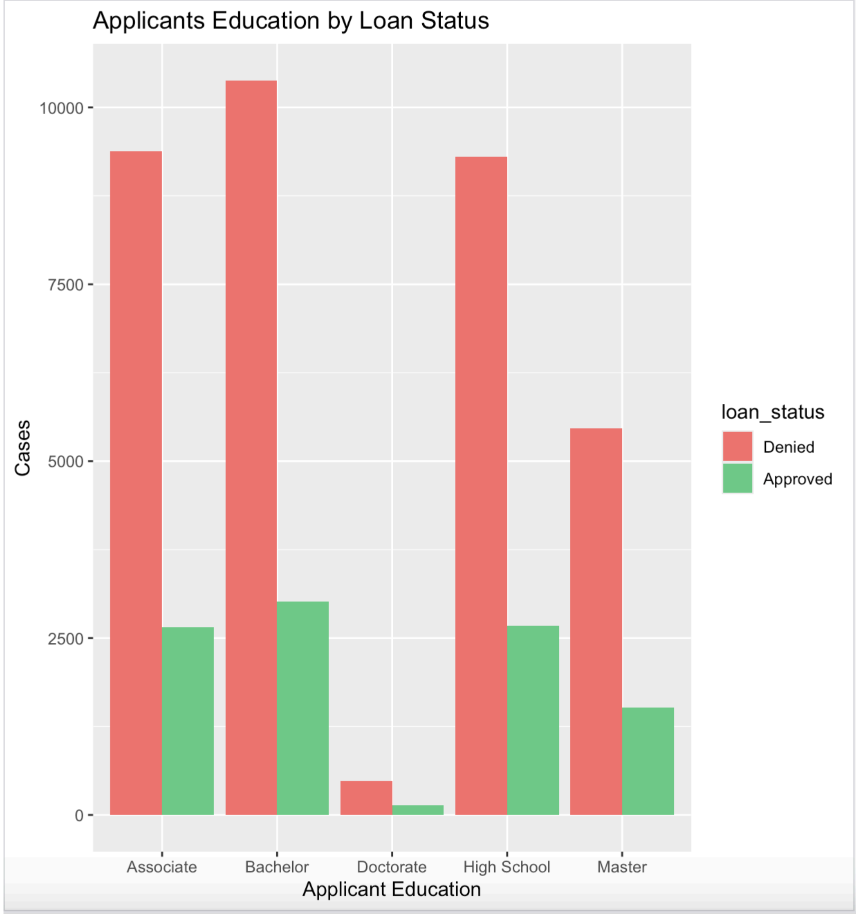

Figure 2

Firstly, it can be observed that for each education degree, the majority of the loans were denied. Secondly, it can be seen that if an education level had a higher amount of denied loans than others, it also had a higher amount of approved ones. This pattern is followed among all degrees. Holders of bachelor degrees had the most loans, approved and denied, followed by an approximate tie between associate and high school degrees, followed by master's degrees, and finally, doctorate degrees. It is interesting to note, that despite having the least amount of loans applied, doctorate degrees had the most approval proportionally. As for bachelor, associate, and high school degrees, their high application numbers can be justified by students taking loans for education and living arrangements during that time.

Loan Interest Rate by Loan Status. The third relationship explored using descriptive statistics is loan interest rate, which is a numerical variable, by loan status, which is a categorical variable.

Loan_status	variable	n	mean	sd
0	loan_int_rate	35000	10.5	2.73
1	loan_int_rate	10000	12.9	3.07

For denied cases, the mean interest rate on the loan was 10.5%, while for approved cases the mean interest rate was 12.9%. The approved cases also had a higher standard deviation, meaning the data was more spread out as compared to denied cases. A higher interest rate for financial institutions is a more profitable case, and therefore, may have influenced the application to be approved.

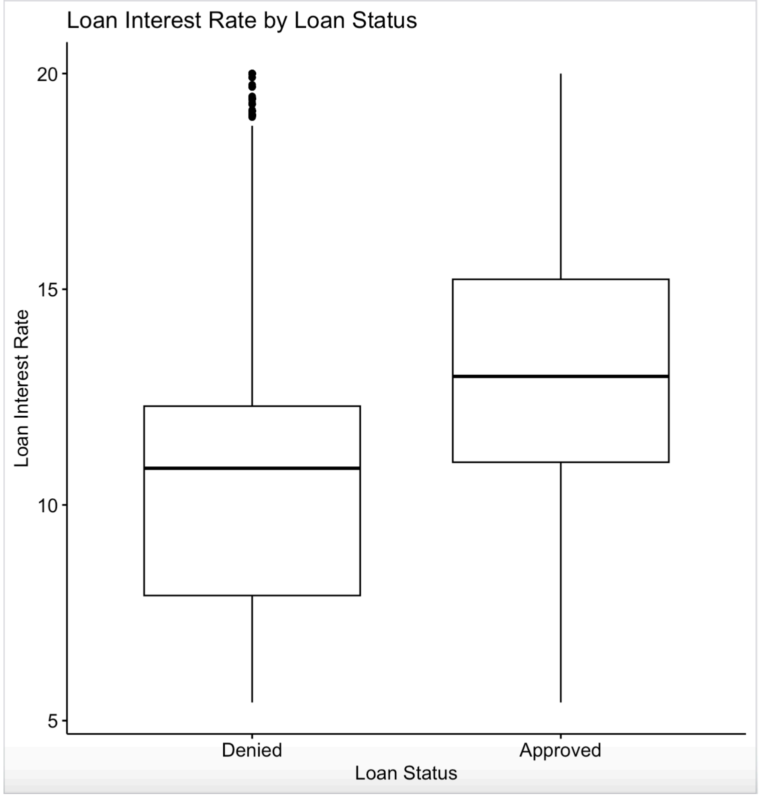

Figure 3

The upper whisker, upper and lower quartiles, and the median of loan interest rate was higher for approved cases as compared to denied ones. The few cases denied with high interest rates were considered outliers. Using the visualization, one can assume that loans with higher interest rates generally have a higher chance of getting approved.

Loan as Percentage of Income by Loan Status. Although the variable name suggests the value to be a percentage, the value of the variable is loan as proportion of income. It needs to be multiplied by a hundred to become a percentage.

Loan_status	variable	n	mean	sd
0	loan_percent_income	35000	0.122	0.071
1	loan_percent_income	10000	0.203	0.107

The mean of loan as proportion of income for approved cases (0.203) was greater than that of denied cases (0.122). Similarly, the standard deviation was also higher for approved cases, meaning the data point for approved cases was more spread out. We can make an assumption that within this dataset, as compared to income, larger loans have a higher likelihood of getting approved.

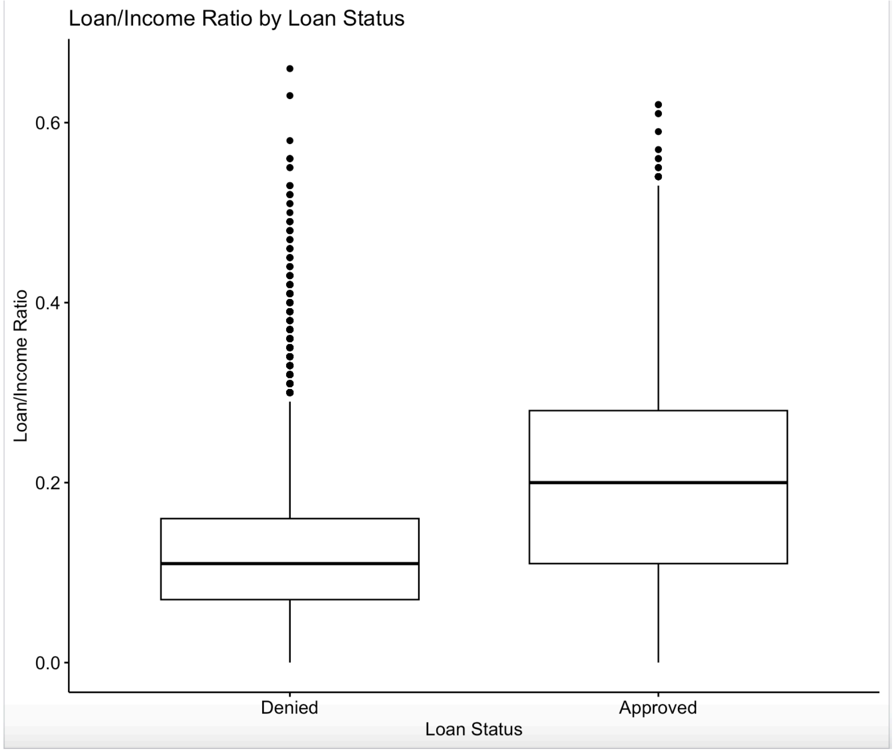

Figure 4

This visualization corroborates the fact that approved cases tend to have a higher loan to income proportion, as the upper and lower quartiles, median, and upper whisker are higher for approved cases than the denied ones. Furthermore, any cases that have a high proportion and are denied are considered outliers.

Credit Score by Loan Status. The last relationship explored through descriptive statistics is credit score, a numerical variable, by loan status, the categorical, outcome variable.

Loan_status	variable	n	mean	sd

0	credit_score	35000	633.	50.5
1	credit_score	10000	632.	50.3

For denied cases, the mean credit score was 633, which was 1 point higher than those of approved cases, which was 622. The standard deviation was also very slightly higher for denied cases. This is opposite of what one might assume of credit score being an indicator of a trustworthy applicant for loaners.

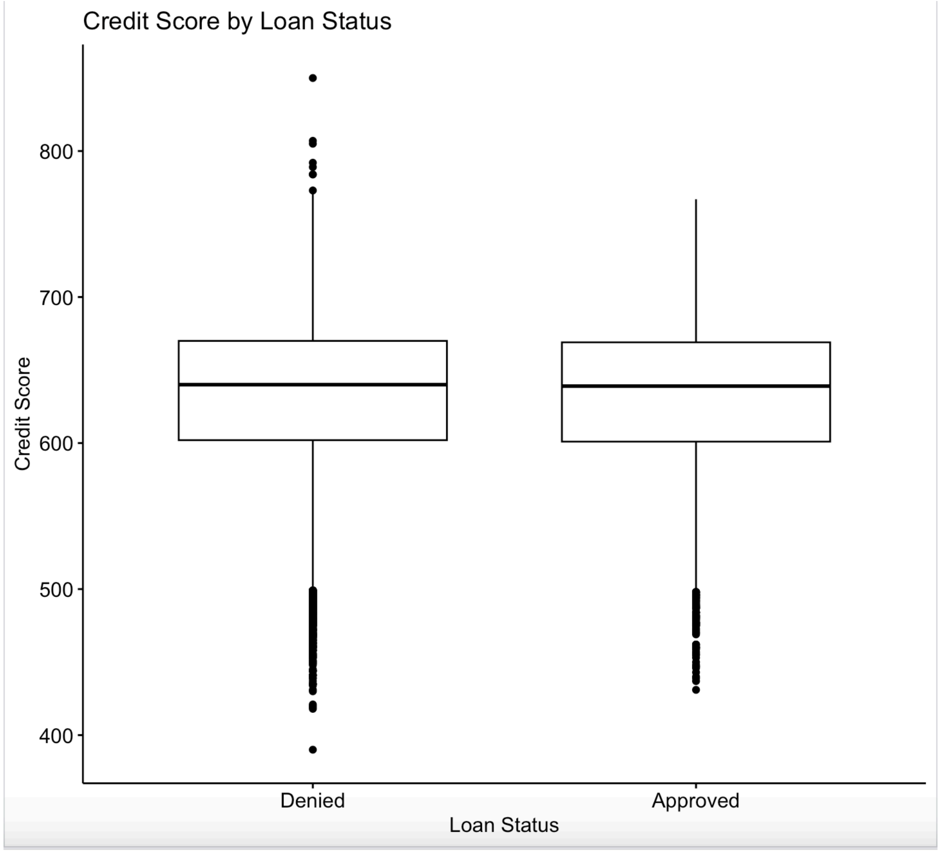

Figure 5

It can be observed that the median and upper and lower quartiles for approved cases is slightly lower than the denied cases. Furthermore, it is interesting to note that credit scores near 800 are considered outliers if they fall under the denied category. However, this graph does not point to a significant association between loan status and credit score.

Section 3. Hypotheses

For each of the five research questions, there were null and alternative hypotheses developed for the relationship between the mentioned variables and loan status.

Gender and Loan Status

The null hypothesis developed for the first research question pertaining to gender and loan status was H_0: there is no association between applicant gender and loan status. The alternative hypothesis was H_1: there is a significant association between gender and loan status.

Education and Loan Status

The null hypothesis developed for the second research question pertaining to education level and loan status was H_0: there is no association between applicant education level and loan status. The alternative hypothesis was H_1: there is a significant association between applicant education level and loan status.

Loan Interest Rate and Loan Status

The null hypothesis developed for the third research question pertaining to loan interest rate and loan status was H_0: there is no association between the loan's interest rate and loan status. The alternative hypothesis was H_1: there is a significant association between the loan's interest rate and loan status.

Loan as Percentage of Income and Loan Status

The null hypothesis developed for the fourth research question pertaining to the loan as a proportion of applicant's income and loan status was H_0: there is no association between the loan as a proportion of applicant's income and loan status. The alternative hypothesis was H_1: there is a significant association between the loan as a proportion of applicant's income and loan status.

Credit Score and Loan Status

The null hypothesis developed for the last research question pertaining to the applicant credit score and loan status was H_0: there is no association between the applicant credit score and loan status. The alternative hypothesis was H_1: there is a significant association between the applicant credit score and loan status.

Section 4. Data Analysis Procedure and Methods

The data analysis procedure mainly entailed performing hypothesis testing to test the five hypotheses created for each of the five research questions. The methods included performing specific statistical tests suitable to the variables being analyzed to find any significant association between them. If the association is found, the null hypothesis is rejected, otherwise, the null hypothesis fails to be rejected.

Gender and Loan Status

Since both person_gender and loan_status are categorical variables, the Pearson's Chi-Square test is implemented to find any association between the variables. The assumptions of the test are that the data points are independent and the sample size is not too small. Due to utilizing between-group design, the independence assumption is satisfied, as each case appears only once in the dataset and does not influence other data points. Next, a 2x2 contingency table was created using both variables, as both variables only had two conditions. Since each cell in the table had more than five observations, the sample size assumption was also satisfied. The Chi-Square test was then run using the chisq.test() function on the whole dataset to retrieve results for association.

Education and Loan Status

Similarly, since both person_education and loan_status are categorical variables, the Pearson's Chi-Square test is once again utilized to find any association. The assumption of independence was once again satisfied through between-group design, as each case appears only once in the dataset and does not influence other data points. Next, a 5x2 contingency table was created using both variables as person_education had five conditions, while loan status had two. Again, since each cell in the contingency table had at least five observations, the sample size assumption was also satisfied. The Chi-Square test was then run using the chisq.test() function on the whole dataset.

Loan Interest Rate and Loan Status

Since loan status is a categorical variable and loan interest rate is a continuous variable, the initial approach was to use Independent Samples t test. The independence assumption was satisfied, as each case appears only once in the dataset and does not influence other data points. After that, the assumption for no extreme outliers was tested. There were 70 outliers identified, all in the loan_status category of denied. However, none of them were extreme so the assumption was met. Next, the normality assumption was tested by first getting a random sample of size 5000 for each condition of loan_status. Then, the Shapiro-Wilk test was run, which found a p-value of 1.09e-25 for denied and 4.04e-21 for approved, which is smaller than 0.05. After confirming that data points were not close to the trend line of the qq plot (see Figure 6), the assumption was violated.

Therefore, a non-parametric test was run in the form of Mann-Whitney U test. The test assumed independence which was already satisfied. The dependent variable (loan_int_rate) was also continuous. Boxplot of loan_int_rate and loan_status was observed (see Figure 3) for group distribution having similar shape. The size, shape, and positioning of the boxes did differ, however, Mann-Whitney U test would, in this case, compare medians rather than overall distributions.

Loan as Percentage of Income by Loan Status

Similar to loan_int_rate, loan_percent_income is a continuous variable, and therefore, to test with categorical variables of loan_status, the initial approach of Independent Samples t test was applied. The independence assumption was satisfied, as each case appears only once in the dataset and does not influence other data points. After that, the assumption for no extreme outliers was tested. There were 44 extreme outliers identified, all in the denied loan status category. The resulting data frame was saved to extreme_outliers so that these rows could be removed from the loan_data dataframe. Once the rows were removed, the test was run again and no outliers were found. Then, the normality assumption was tested by getting a random sample of size 5000 for each condition of loan_status, as per maximum sample size for the Shapiro-Wilk test. The test found a p-value of 4.57e-41 for denied category and 1.64e-29 for approved. Since both were less than 0.05, confirmed by the data points not landing on the trend line of the qq plot (see Figure 7), the assumption was violated.

The non-parametric Mann-Whitney U test was applied once more. The assumption of independence was already satisfied. The dependent variable (loan_percent_income) was continuous. Boxplot of loan_percent_income and loan_status was observed (see Figure 4) for group distribution having similar shape. The size, shape, and positioning of the boxes did differ, however, Mann-Whitney U test would, in this case, compare medians rather than overall distributions.

Credit Score by Loan Status

Finally, the initial approach of Independent Samples t test was taken once more for the categorical variable of loan_status and continuous variable of credit_score. The independence assumption was satisfied, as each case appears only once in the dataset and does not influence other data points. Next, the extreme outliers assumption test resulted in one extreme outlier identified in the denied loan_status category. The resulting data frame was saved to extreme_outliers so that this row could be removed from the loan_data dataframe. Once the row was removed, the test was run again and no outliers were found. The normality assumption was once again violated when the Shapiro-Wilk test, run with the random sample of size 5000 for each loan_status condition, gave p-values of 7.80e-27 for denied and 1.52e-27 for approved. Since both of these were less than 0.05, confirmed by the data points not landing on the trend line of the qq plot (see Figure 8), the assumption was violated.

Again, non-parametric Mann-Whitney U test was used, with independence and continuous dependent variable assumptions satisfied. The size, shape, and positioning of the boxplots of credit_score and loan_status (see Figure 5) are very similar, the assumption of group distributions being similar in shape was satisfied.

Section 5. Test Results

Gender and Loan Status

	Denied	Approved
female	15651	4485
male	19304	5515

data: contingency_table

X-squared = 0.014909, df = 1, p-value = 0.9028

The Chi-Square test results found a X-squared value of 0.014909, degree of freedom of 1, and p-value of 0.9028. The X-squared value is the difference between the observed frequency and the frequency if gender and loan status were considered independent. The X-squared value is quite low, meaning that the difference was quite small and is close to the difference expected if the null hypothesis was true. The df value is the degree of freedom and is calculated by subtracting 1 from each dimension of the contingency table and multiplying them together. Here, a 2x2 contingency table would become 1*1 which would equal 1. The p-value describes the probability of observed difference occurring if the null hypothesis was true. Since, the p-value of 0.9028 is greater than the standard alpha-level of 0.05, we fail to reject the null hypothesis, meaning that there is no significant association between gender and loan status.

Education and Loan Status

	Denied	Approved
Associate	9365	2650
Bachelor	10368	3018
Doctorate	477	142
High School	9292	2671
Master	5453	1519

```
data:  contingency_table
X-squared = 2.0143, df = 4, p-value = 0.7331
```

The Chi-Square test found the X-squared value of 2.0143, df of 4, and p-value of 0.7331. The X-squared value is the difference between the observed statistic and what could be expected under the null hypothesis. The X-squared value here is still small, albeit higher than the one for gender and loan status. The degree of freedom (df) is 4 now, as the 5x2 contingency table becomes 4*1, which is 4. The p-value shows that there was an about 73% probability of the observed difference occurring under the null hypothesis. Since 0.7331 is greater than the standard alpha-level of 0.05, we fail to reject the null hypothesis, meaning there was no significant association found between education and loan status.

Loan Interest Rate and Loan Status

QQ Plot Trend Line for Data Distribution

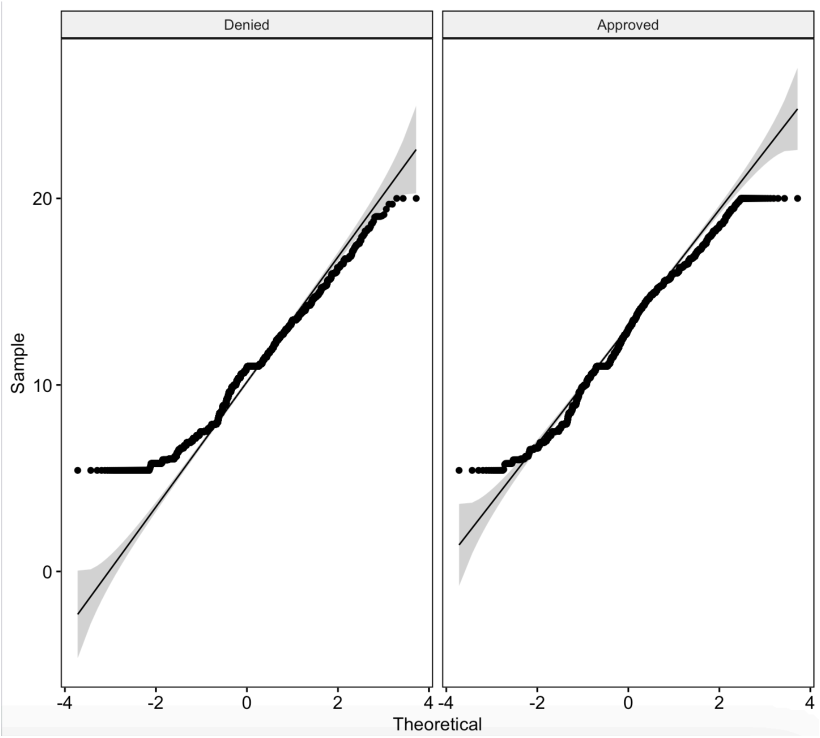

Figure 6

```
Wilcoxon rank sum test with continuity correction

data:  loan_int_rate by loan_status

W = 99217532, p-value < 2.2e-16

alternative hypothesis: true location shift is not equal to 0
```

The results of the Mann-Whitney U test had a W value of 99217532, which is the difference of ranks between both loan_status conditions. This W value is quite high, meaning the

loan_status group ranks are highly separated. Furthermore, the test resulted in a p-value of less than 2.2e-16, which is less than the standard alpha-level of 0.05. Therefore, we are able to reject the null hypothesis, meaning there is a significant association between loan_int_rate and loan_status variables.

Loan as Percentage of Income by Loan Status

QQ Plot Trend Line for Data Distribution

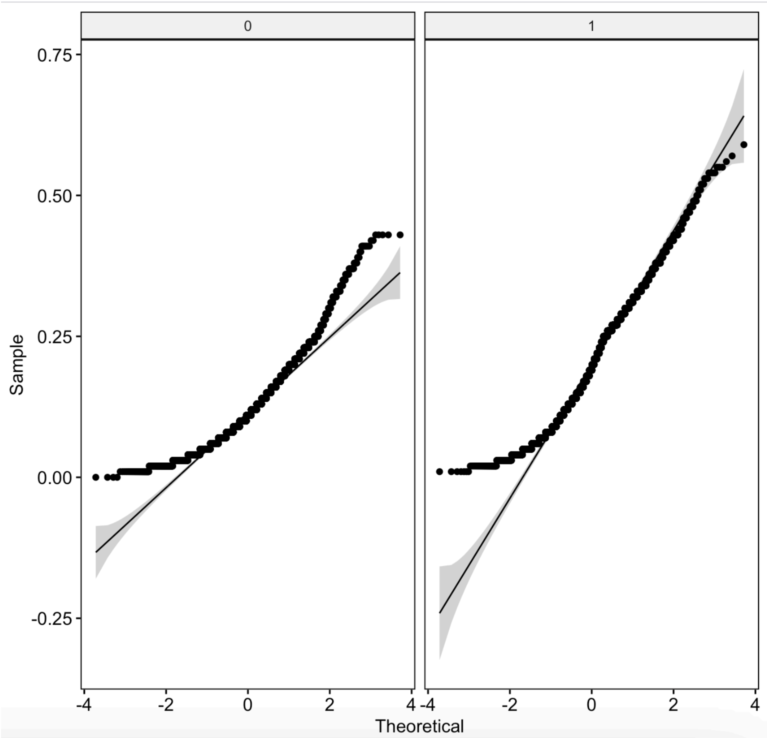

Figure 7

```
Wilcoxon rank sum test with continuity correction

data:  loan_percent_income by loan_status

W = 95811341, p-value < 2.2e-16

alternative hypothesis: true location shift is not equal to 0
```

The results of the Mann-Whitney U test had a W value of 95811341, which is the difference of ranks between both loan_status conditions. This W value is quite high, meaning the

loan_status group ranks are highly separated. Furthermore, the test resulted in a p-value of less than 2.2e-16, which is less than the standard alpha-level of 0.05. Therefore, we are able to reject the null hypothesis, meaning there is a significant association between loan_percent_income and loan_status variables.

Credit Score by Loan Status

QQ Plot Trend Line for Data Distribution

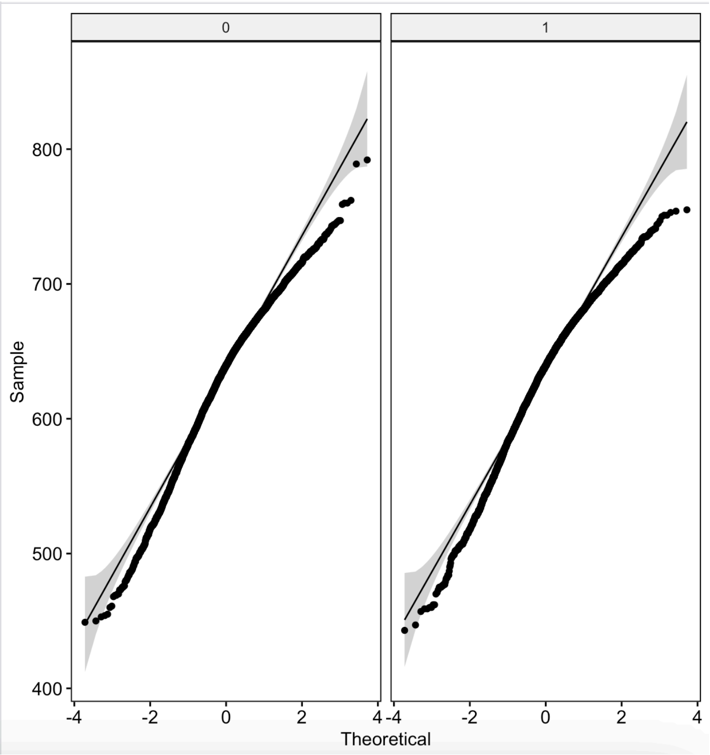

Figure 8

```
Wilcoxon rank sum test with continuity correction

data:  credit_score by loan_status

W = 176825677, p-value = 0.07313

alternative hypothesis: true location shift is not equal to 0
```

The results of the Mann-Whitney U test had a W value of 176825677, which is the difference of ranks between both loan_status conditions. This W value is quite high, meaning the

loan_status group ranks are highly separated. Furthermore, the test resulted in a p-value of 0.07313, which is greater than the standard alpha-level of 0.05. Therefore, we fail to reject the null hypothesis, meaning there is no significant association between credit_score and loan_status variables.

Section 6. Interpretation and Decision-making

Through the hypothesis testing and analysis above, it is found that an applicant's gender, educational attainment level, and credit score are not significantly associated with loan status. Therefore, an applicant's gender, education, or credit-score does not provide an edge or create a higher likelihood of getting their loan approved. However, through the testing, it was concluded that the interest rate of the loan and the loan amount as a proportion of applicant's income were significantly associated with loan approval status. Through descriptive statistics and boxplot visualizations, it was found that loans with a higher interest rate and proportionally large loans compared to applicant income generally had a higher likelihood of getting approved. Therefore, it can be seen that financial institutions like higher return on their investment through interest rate and higher amount of loan for more money earned.

Firstly, it is refreshing to note that the loan approval system is not subject to gender bias. However, the loan approvals system should take applicant education and credit score into account more. Education level of an applicant is a strong indicator of current and future financial stability of the applicant, which will lower the risk of the applicant defaulting on the loan. Similarly, credit score is another strong indicator of an applicant's trustworthiness to pay back the loan with appropriate interest. As for individuals seeking to get loans approved, applying for a higher loan amount and accepting a higher interest rate will likely reach that outcome.