Loan Approval Data Analytics Report

By: Hassan Raza

Dataset

- Loan Approval Classification Dataset
- Data source: Kaggle
- 45000 rows
- Categorical and continuous variables
- Demographic, socio-economic background, and loan details
- Target Variable: loan_status
  - 0 for denied, 1 for approved

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

Choice of Variables

- Gender
  - Association can mean gender bias
- Education
  - Indicator of current and future financial stability
- Loan Interest Rate
  - Benefit for loan giver
- Loan as Percentage of Income
  - Larger vs. smaller loans
- Credit Score
  - Trustworthiness of applicant

Research Questions

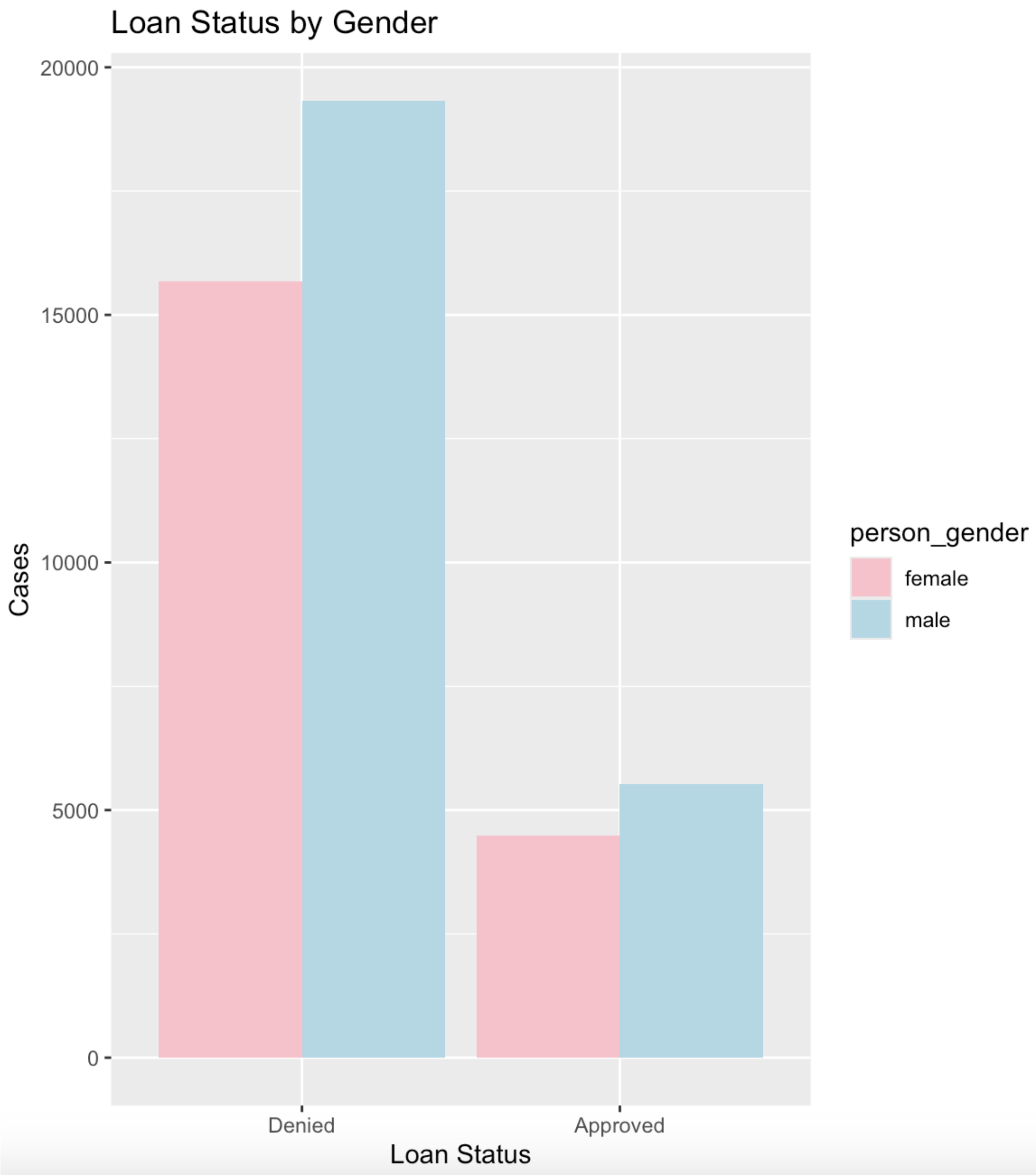
- Is there any association between gender and loan approval?
    - Do males tend to get more loans approved or females?
- How is an individual’s education level associated with loan approval?
    - Does higher education mean a higher likelihood of loan approval?
- Does a higher loan interest rate influence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

Descriptive Statistics

Loan Status by Applicant Gender

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

- Mean is proportion of loan approval
- Same approval rate for both genders
- Males get more loans approved and more denied than females

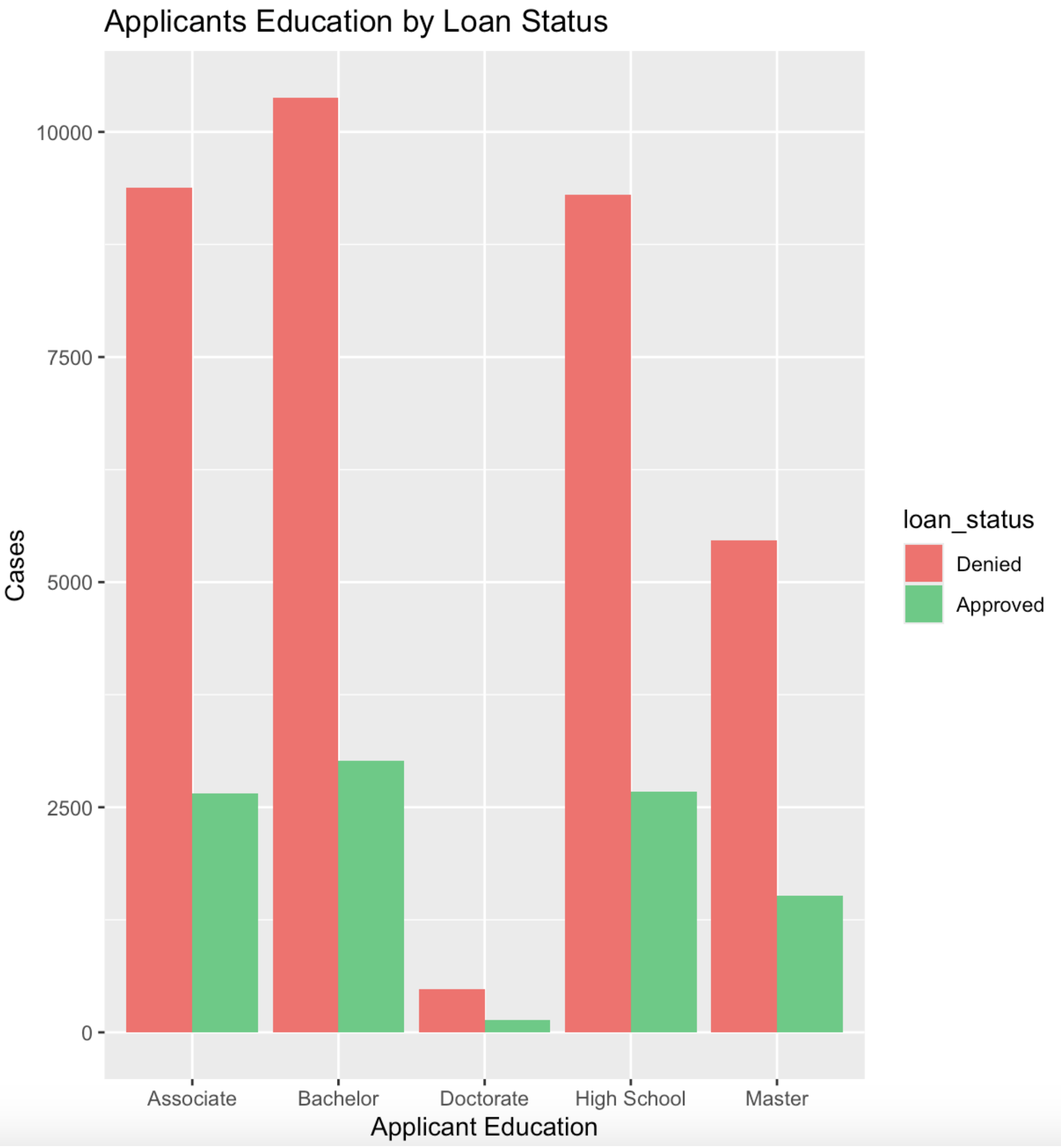


Descriptive Statistics

Loan Status by Applicant Education

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

- Mean is proportion of loan approval
- Doctorate has highest mean, associate lowest
- Bachelor has most loans

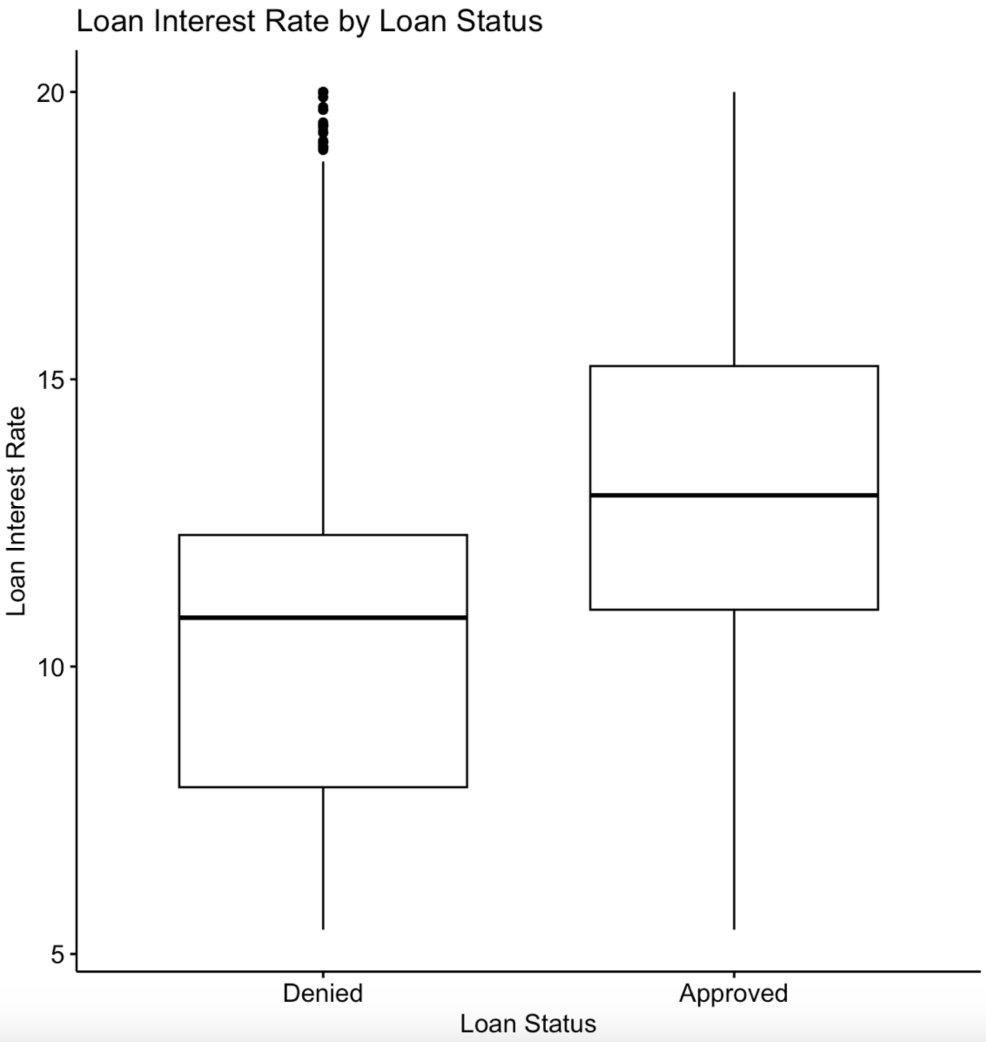


Descriptive Statistics

Loan Interest Rate by Loan Status

Loan_status	variable	n	mean	sd
Denied	loan_int_rate	35000	10.5	2.73
Approved	loan_int_rate	10000	12.9	3.07

- Mean and median interest rate higher for approved cases
- Concentrated around higher value for approved cases
- Higher interest rate in denied considered extreme outliers

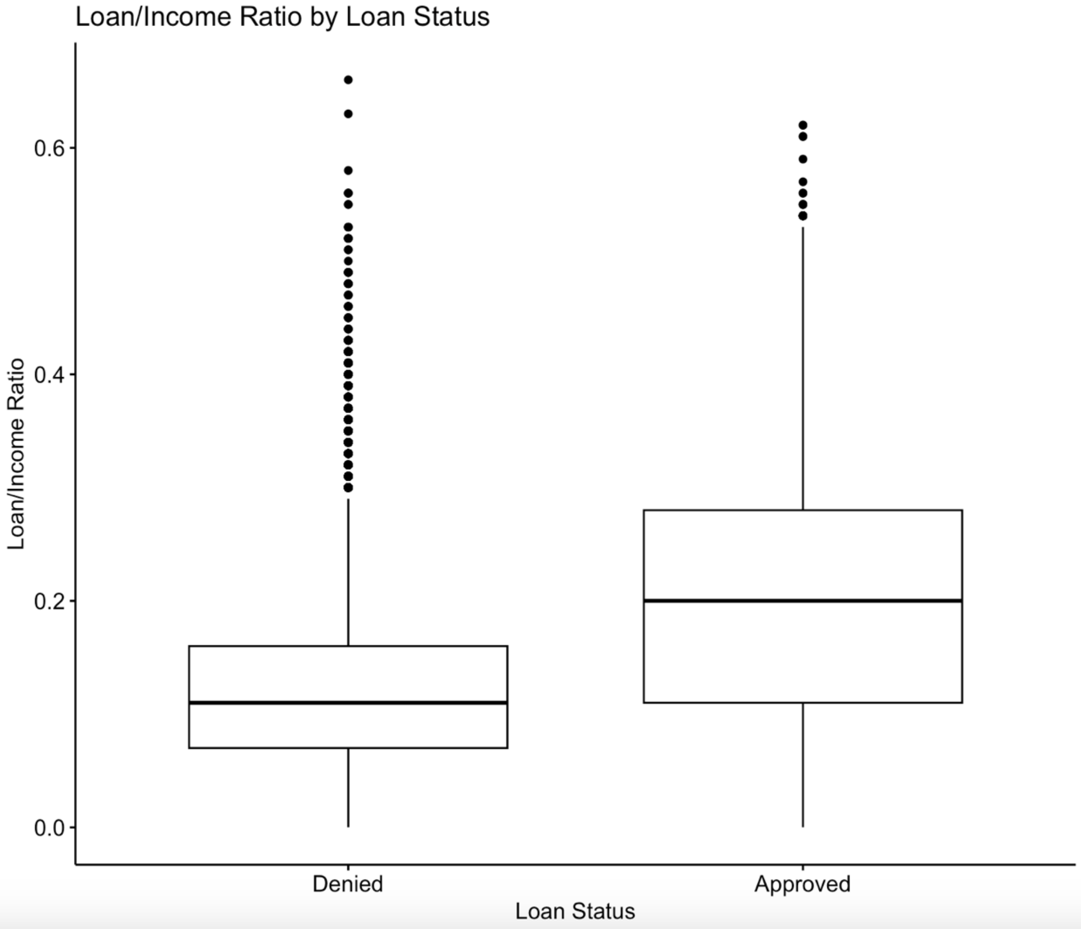


Descriptive Statistics

Loan as Percentage of Income by Loan Status

Loan_status	**variable**	**n**	**mean**	**sd**
Denied	loan_percent_income	35000	0.122	0.071
Approved	loan_percent_income	10000	0.203	0.107

- Mean and median proportion greater for approved cases
- Denied proportions are narrowly concentrated
- Denied proportions of > 0.3 considered extreme outliers

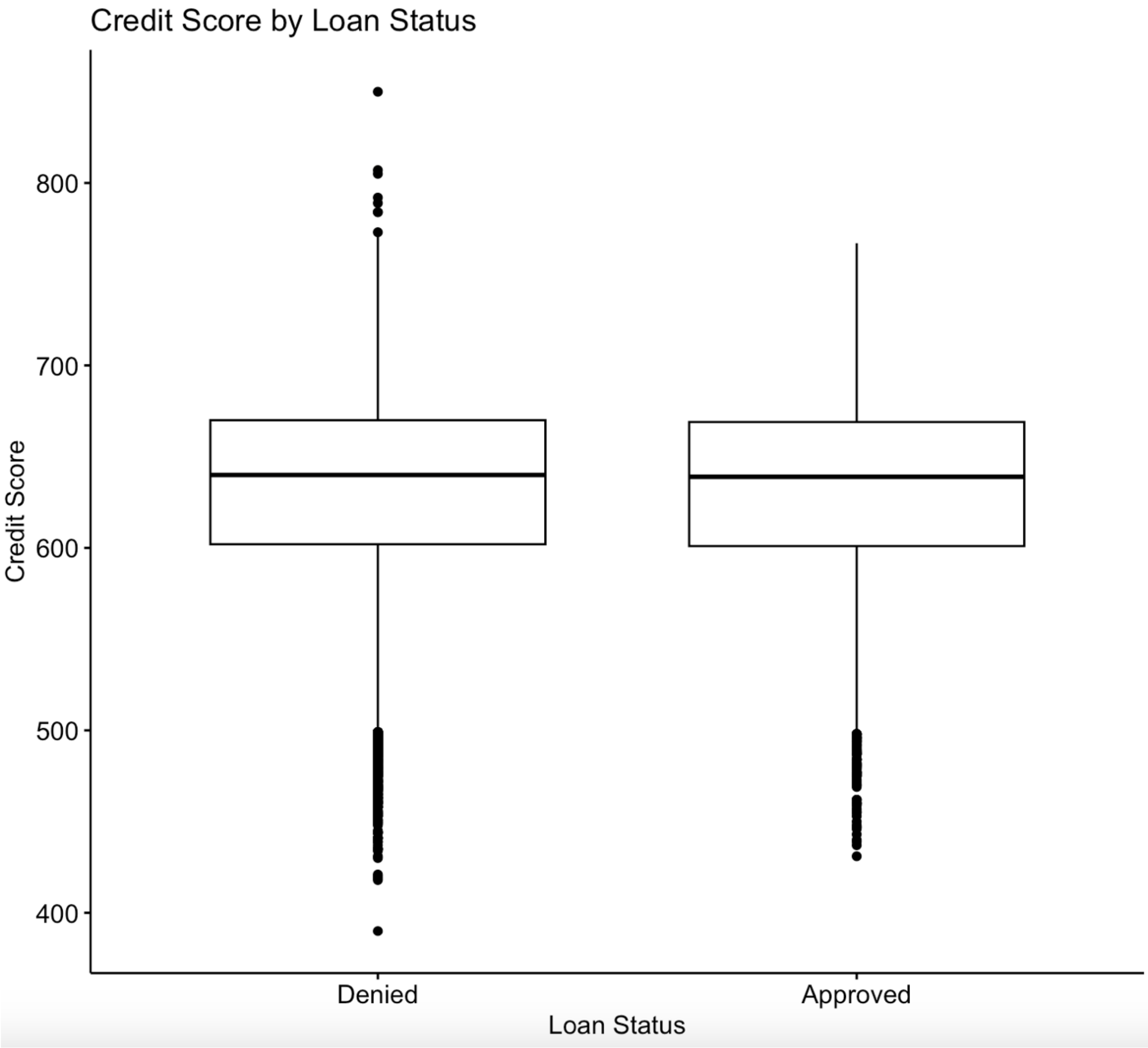


Descriptive Statistics

Credit Score by Loan Status

Loan_status	variable	n	mean	sd
Denied	credit_score	35000	633	50.5
Approved	credit_score	10000	632	50.3

- Mean credit score higher for denied cases
- Credit score near 800 in denied considered extreme outliers
- Boxplots very similar in shape and size



Hypotheses

Gender and Loan Status

H_0: No association between applicant gender and loan status

H_1: Significant association between applicant gender and loan status

Hypotheses

Education and Loan Status

H_0: No association between applicant education and loan status

H_1: Significant association between applicant education and loan status

Hypotheses

Loan Interest Rate and Loan Status

H_0: No association between loan interest rate and loan status

H_1: Significant association between loan interest rate and loan status

Hypotheses

Loan as Percentage of Income and Loan Status

H_0: No association between loan as percentage of income and loan status

H_1: Significant association between loan as percentage of income and loan status

Hypotheses

Credit Score and Loan Status

H_0: No association between credit score and loan status

H_1: Significant association between credit score and loan status

Data Analysis Methods

Pearson’s Chi-Square Test

- Used for:
  - Gender and Loan Status
  - Education and Loan Status
- Used because:
  - Both variables are categorical
- Assumptions:
  - Independent data points (between-group design)
  - > 5 observations per contingency table cell

Data Analysis Methods

Mann-Whitney U Test

- Used for:
  - Loan Interest Rate and Loan Status
  - Loan as Percentage of Income and Loan Status
  - Credit Score and Loan Status
- Used because:
  - Independent Samples t Test can’t be used
  - Shapiro-Wilk Test of normal distribution fails
- Assumptions:
  - Independent data points (between-group design)
  - Dependent variable is continuous
  - Group distributions have similar shapes (boxplots)

Results

Gender and Loan Status

Pearson’s Chi-Square Test

- X-squared = 0.014909
- Degree of freedom = 1
- p-value = 0.9028 > alpha-level = 0.05
- Fail to reject null hypothesis
- No significant association between person_gender and loan_status

	Denied	**Approved**
female	15651	4485
male	19304	5515

Contingency table

Results

Education and Loan Status

Pearson’s Chi-Square Test

- X-squared = 2.0143
- Degree of freedom = 4
- p-value = 0.7331 > alpha-level = 0.05
- Fail to reject null hypothesis
- No significant association between person_education and loan_status

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

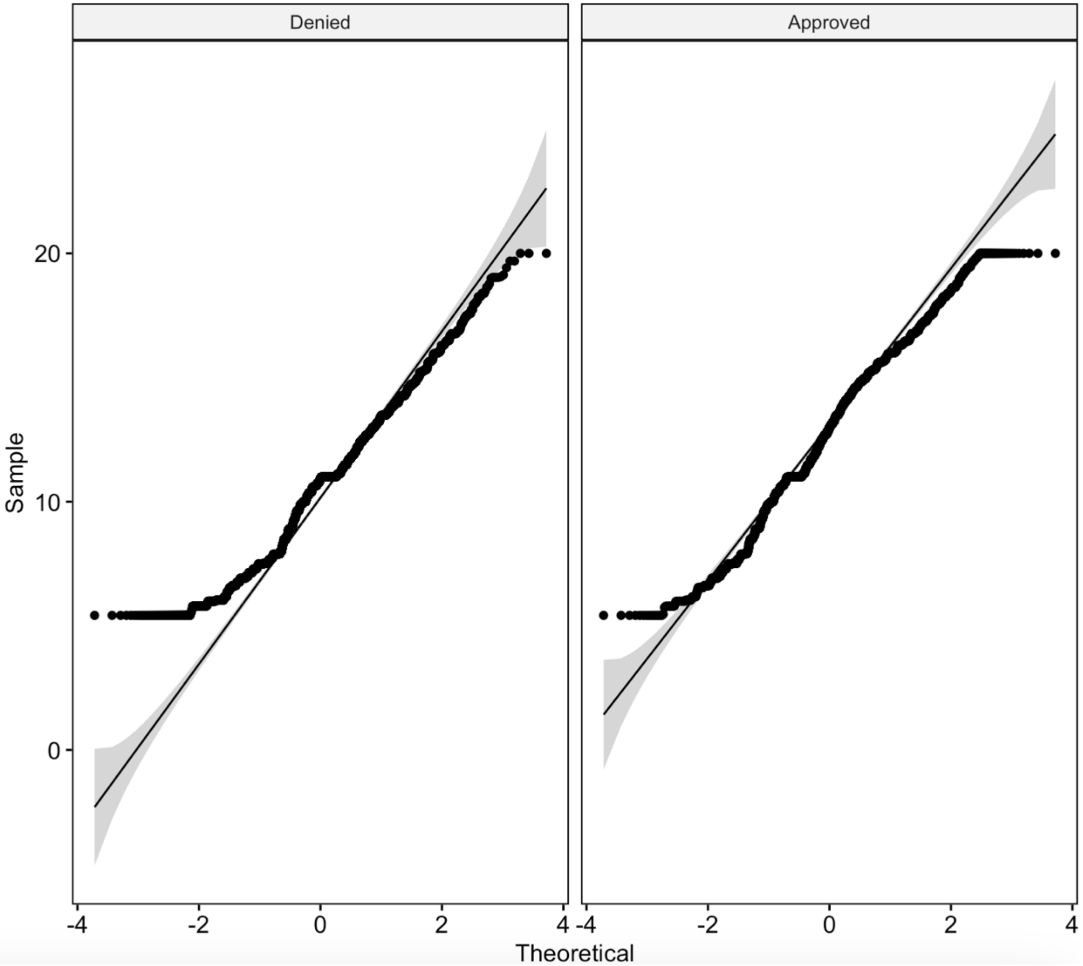
Contingency table

Results

Loan Interest Rate and Loan Status

Mann-Whitney U Test

- W = 99217532
- p-value < 2.2e-16
- p-value < alpha-level = 0.05
- Reject null hypothesis
- Significant association between loan_int_rate and loan_status



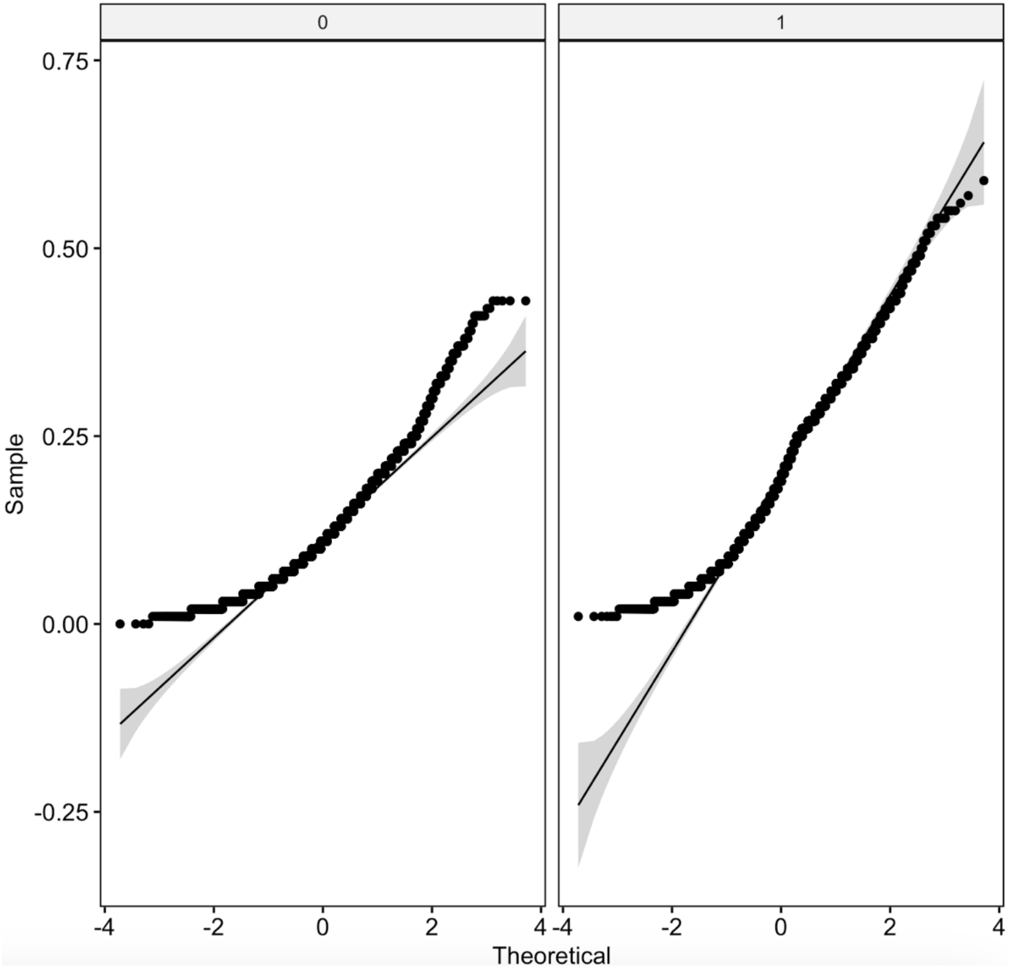
qq plot trend line for data distribution

Results

Loan as Percentage of Income and Loan Status

Mann-Whitney U Test

- W = 95811341
- p-value < 2.2e-16
- p-value < alpha-level = 0.05
- Reject null hypothesis
- Significant association between loan_percent_income and loan_status



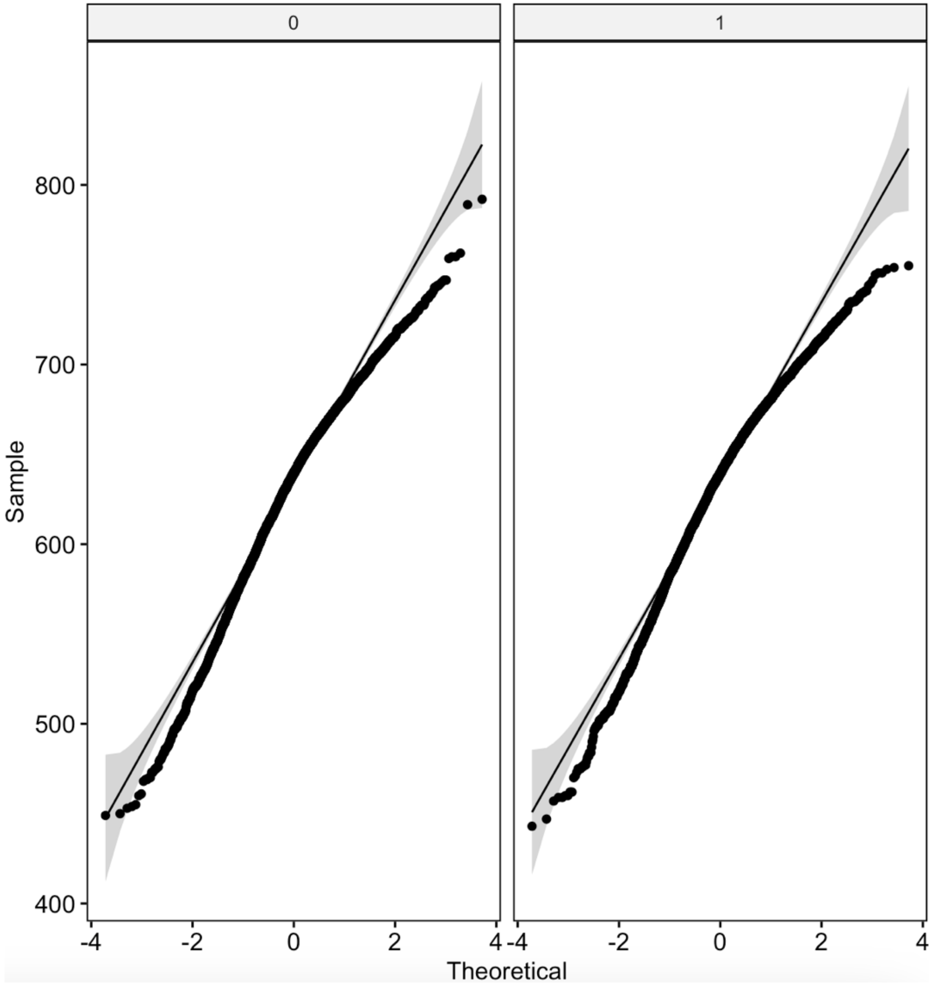
qq plot trend line for data distribution

Results

Credit Score and Loan Status

Mann-Whitney U Test

- W = 176825677
- p-value = 0.07313 > alpha-level = 0.05
- Fail to Reject null hypothesis
- No significant association between credit_score and loan_status



qq plot trend line for data distribution

Interpretation

✗

No association

- Gender -> no bias in approving loans
- Education -> indicator of financial stability, should be used
- Credit Score -> indicator of trustworthiness, should be used

✓

Significant association

- Loan interest rate -> loaners want higher return on investment
- Loan as Percentage of Income -> long time commitment to paying interest