How to Identify Bad Loans?

Tshin Qi Zhou

Scope

- Problem Statement
- Background
- Objectives
- Limitation
- Methodology
- Dataset Examination
- Target Variable Definition
- Summary, Analysis, Insight
- Recommendations

Problem Statement

How do banks distinguish “good” from “risky” borrowers?

Compounded by multi-dimensional nature of credit risk, data constraints, and methodological heterogeneity

feature-selection challenges as datasets contain multiple dimensions and **a lack of standardised feature sets** produce inconsistent performance.

no single established method to predict loan defaults.

defaults often represent **< 20%** of the samples in the dataset.

Background

- Smaller banks systematically avoid riskier lending and contract their loan portfolios more in response to adverse shocks *e.g. G. K. Bhaumik & J. Piesse (2005) and European Central Bank (2021)*

- A part of the risk comes from loan defaults.

Objective

To analyse conditions in which loans could default.

To predict which loan is at high risk of becoming non-performing

To propose business solutions using insights from analysis.

Metric: Recall and F_1 score

Limitation

1. Insufficient data for Joint loan applications; scope of analysis covers only personal loans
2. No data beyond 2020; dataset covers loans only from Sep 2007 – Sep 2020
3. Lack of membership data; each loan is taken as a ‘fresh’ loan from a new borrower

Methodology

To extract the dataset

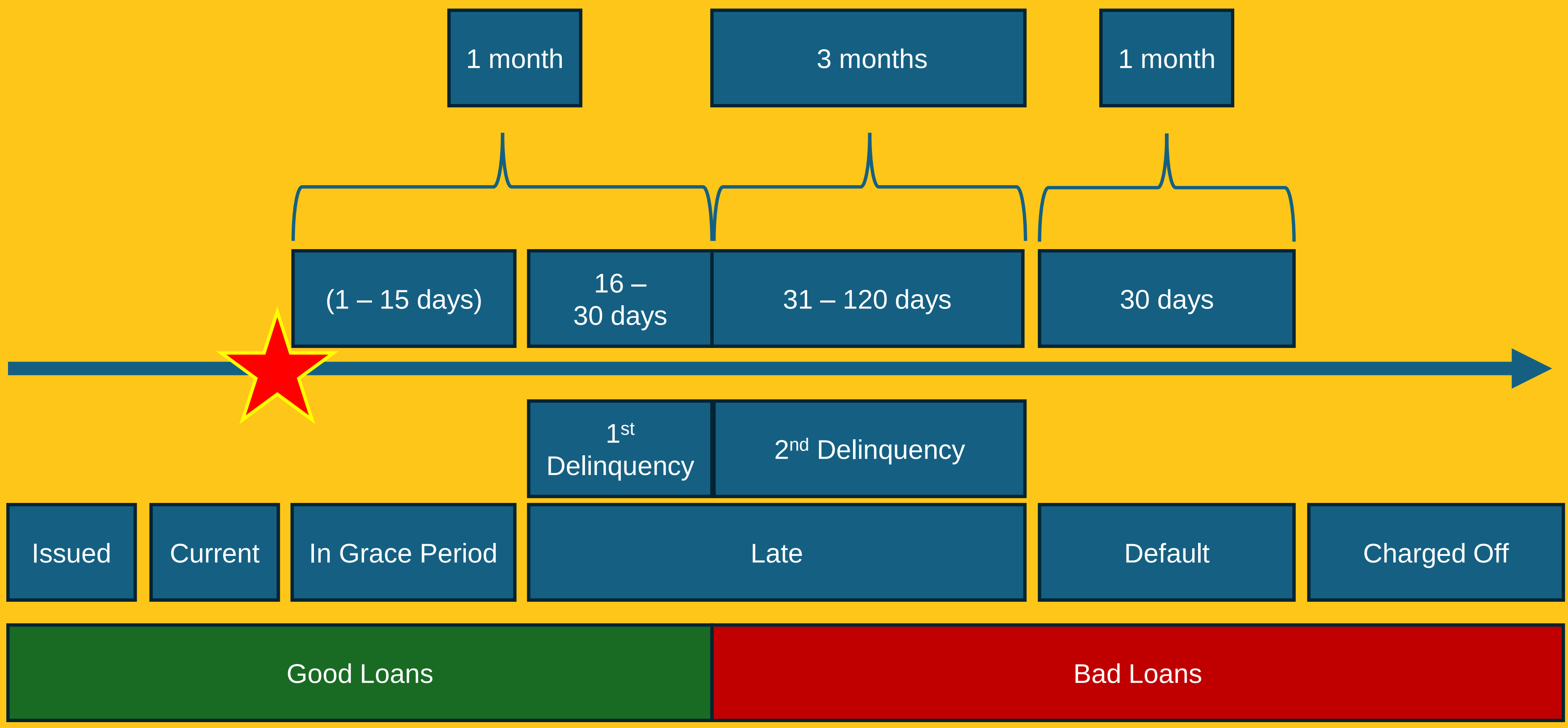
- Data preparation
- Exploratory Data Analysis
- Feature Engineering
- Modelling

Insights Communication

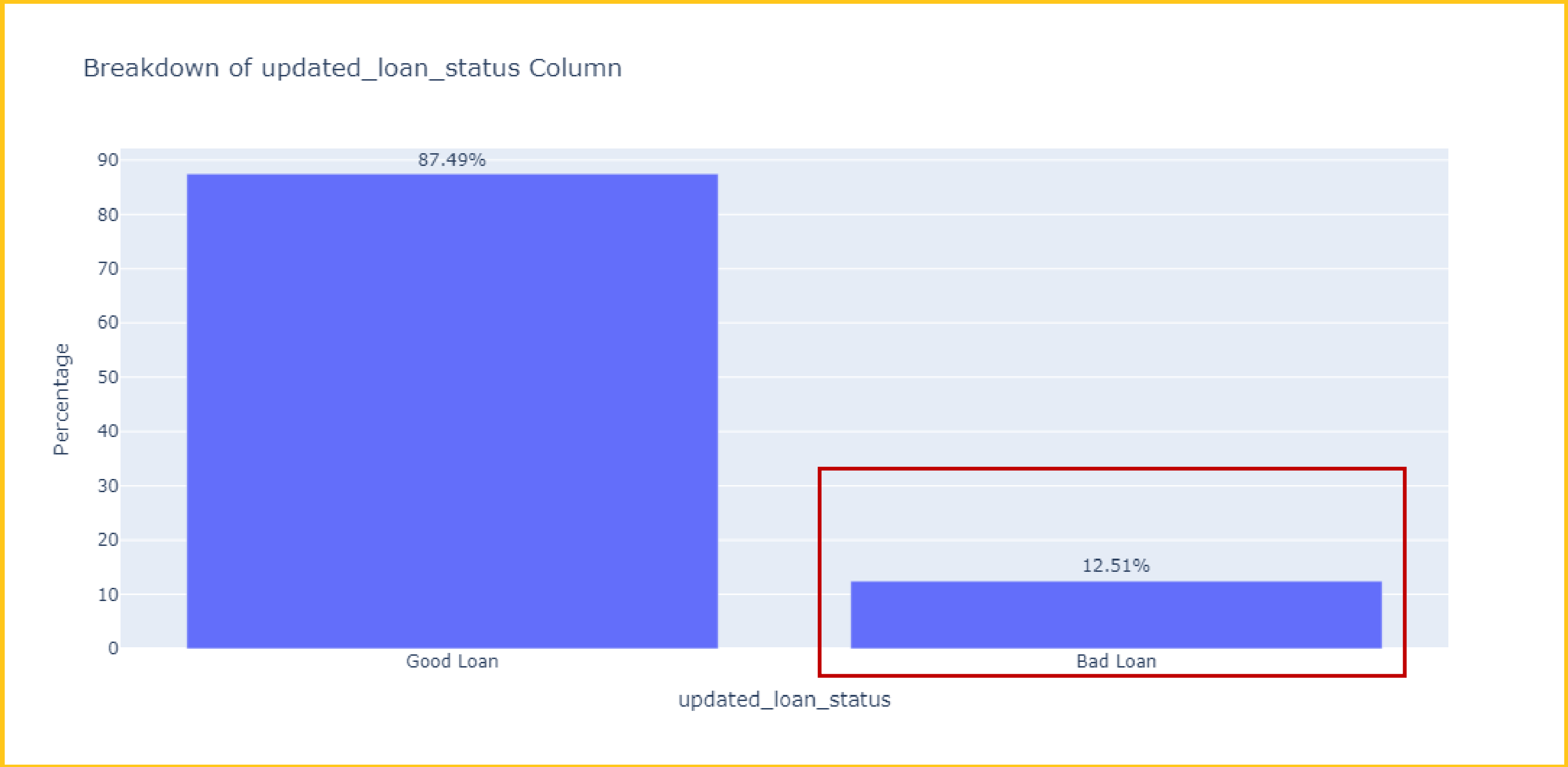
Dataset Examination

- LendingClub: peer-to-peer lending and Fintech services
- Taken from: https://www.kaggle.com/datasets/ethon0426/lending-club-20072020q1/data.
- 142 columns and 2,925,493 rows for loans from 2007 – 2020
- 4 main categories:
  1. Borrower's immediate financial indicators (e.g. income, debt-to-income ratio, FICO scores, etc)
  2. Indicators about the loan (e.g. interest rate, grade, purpose, etc).
  3. Borrower's financial history (e.g. revolving balance, revolving utilisation rate, and history of delinquency)
  4. Miscellaneous (e.g. residential state, employment length, home ownership, etc)

Defining the Target Variable



Distribution of Feature Variable



‘bad loans’ account for approx. 12.51% of all loans.

Quick Facts on Loans

2,264,576	$34.84 billion	$15,383.52	2019	2015
Total No. of Loans	Total Amount Loaned	Avg Loan Amount	Most No. of Loans in a Year	Most No. of Bad Loans in a Year

Average FICO and Annual Income

With higher FICO scores or higher annual income, likelihood of loan default decreases

Cat 1: Borrower's immediate financial indicators (e.g. income, debt-to-income ratio, FICO scores, etc)

Cat 2: Indicators about the loan (e.g. interest rate, grade, purpose, etc).

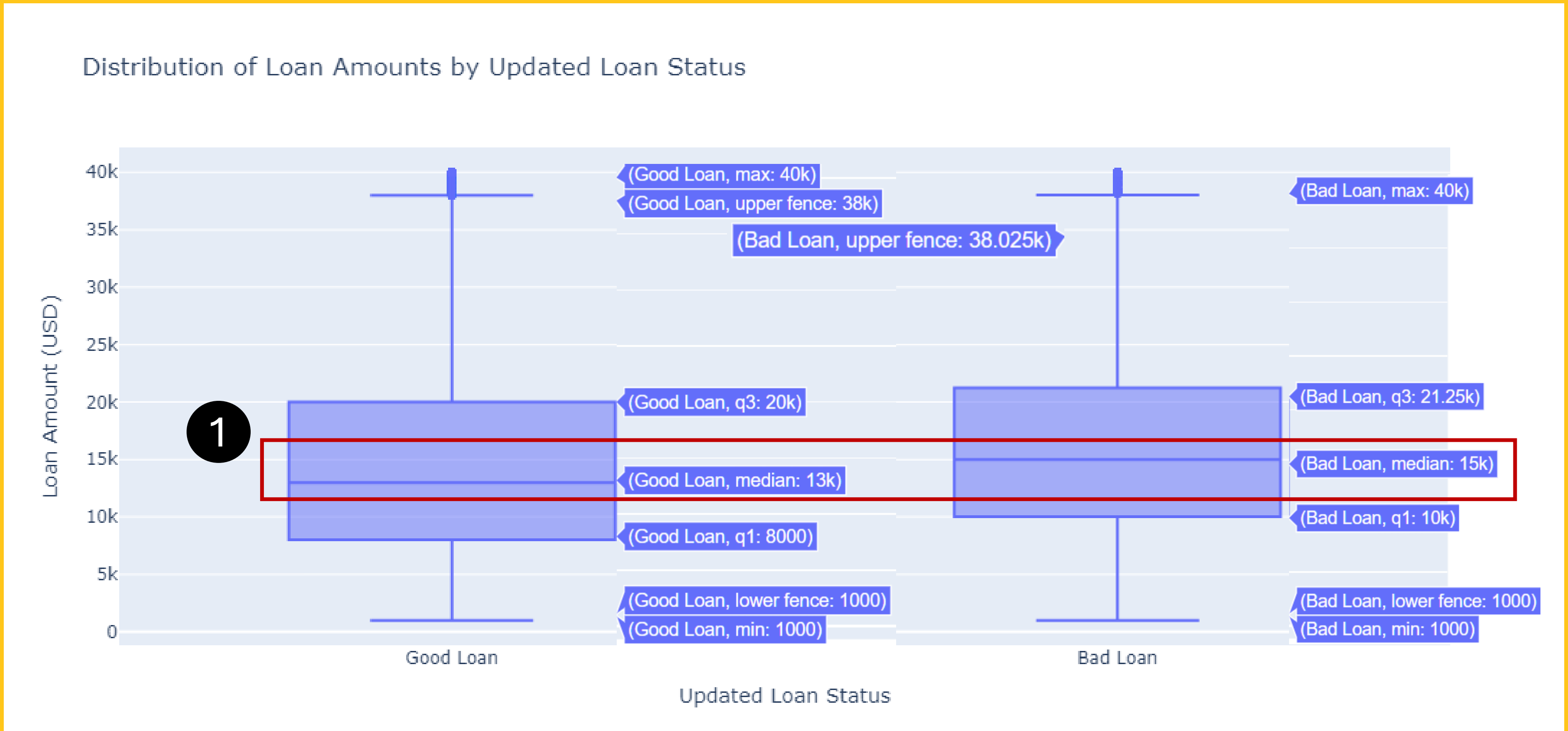
Cat 3: Borrower's financial history (e.g. revolving balance, revolving utilisation rate, and history of delinquency)

Cat 4: Miscellaneous (e.g. residential state, employment length, home ownership, etc

Debt-to-Income Ratio

If a larger percentage of their income is used to finance debt, likelihood of loan default increases

Loan Amount

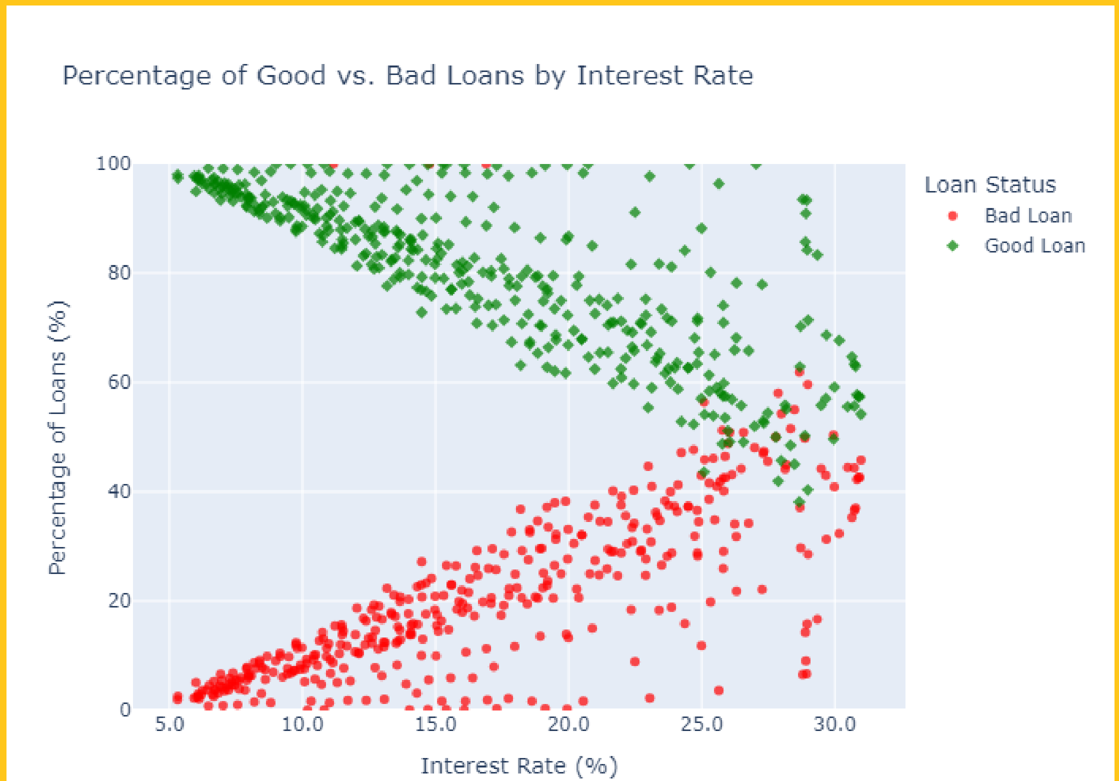


Notwithstanding the min and max amounts, across the board, bad loans tend to be higher. The median amount for a bad loan is 15.4% higher than a good loan.

Loan Amount

While loan amount for bad loans tend to be higher, loan amount is not a factor of whether borrowers is likely to default. In addition, there are no bad loans for many amounts between 35 – 40k.

Interest Rate



Interest rate is a useful indicator for whether a borrower is likely to default

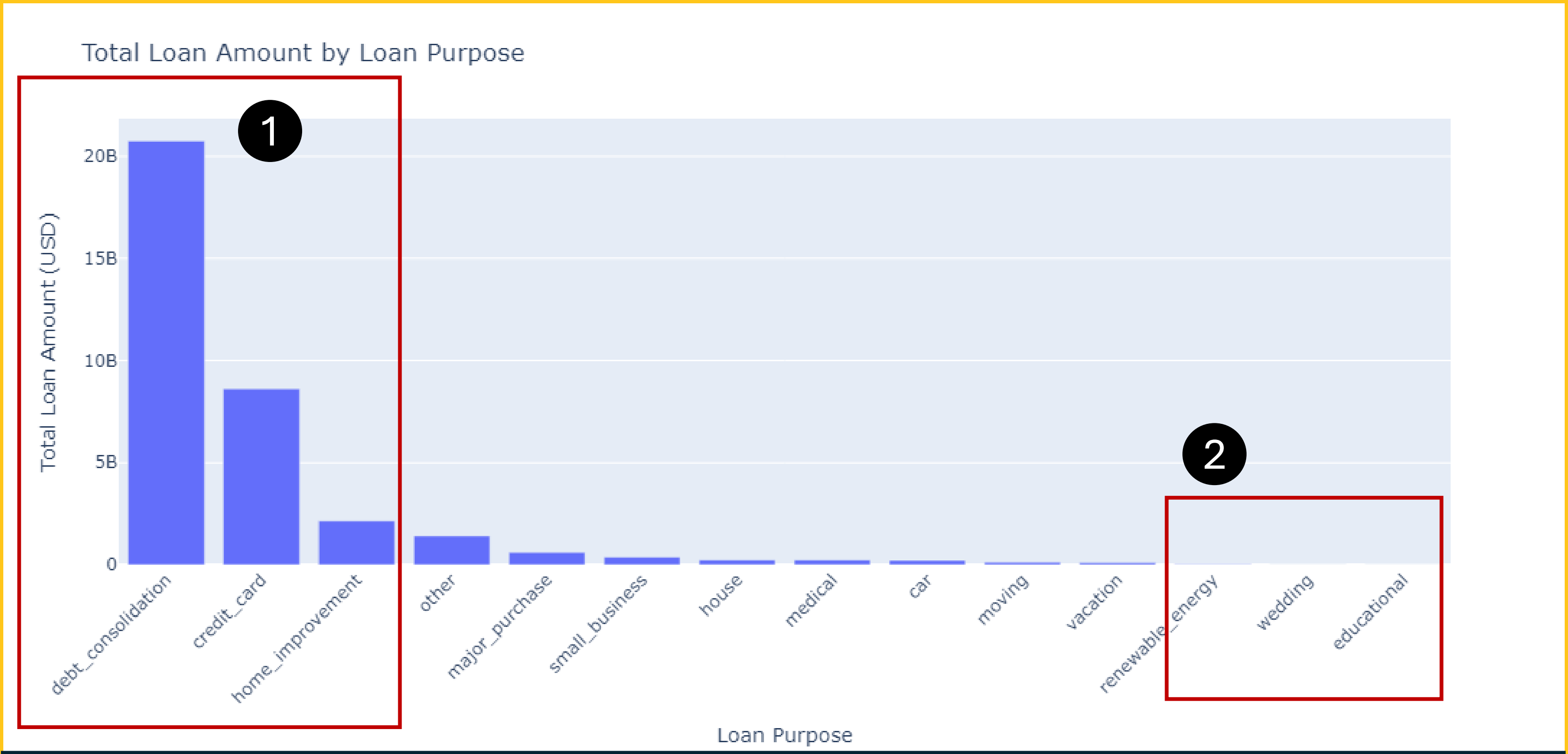
Cat 1: Borrower's immediate financial indicators (e.g. income, debt-to-income ratio, FICO scores, etc)

Cat 2: Indicators about the loan (e.g. interest rate, grade, purpose, etc).

Cat 3: Borrower's financial history (e.g. revolving balance, revolving utilisation rate, and history of delinquency)

Cat 4: Miscellaneous (e.g. residential state, employment length, home ownership, etc

Purpose



Majority of the loans were funding debt consolidation, credit card payment, and home improvement projects. Virtually no loans to renewable energy, wedding, and education.

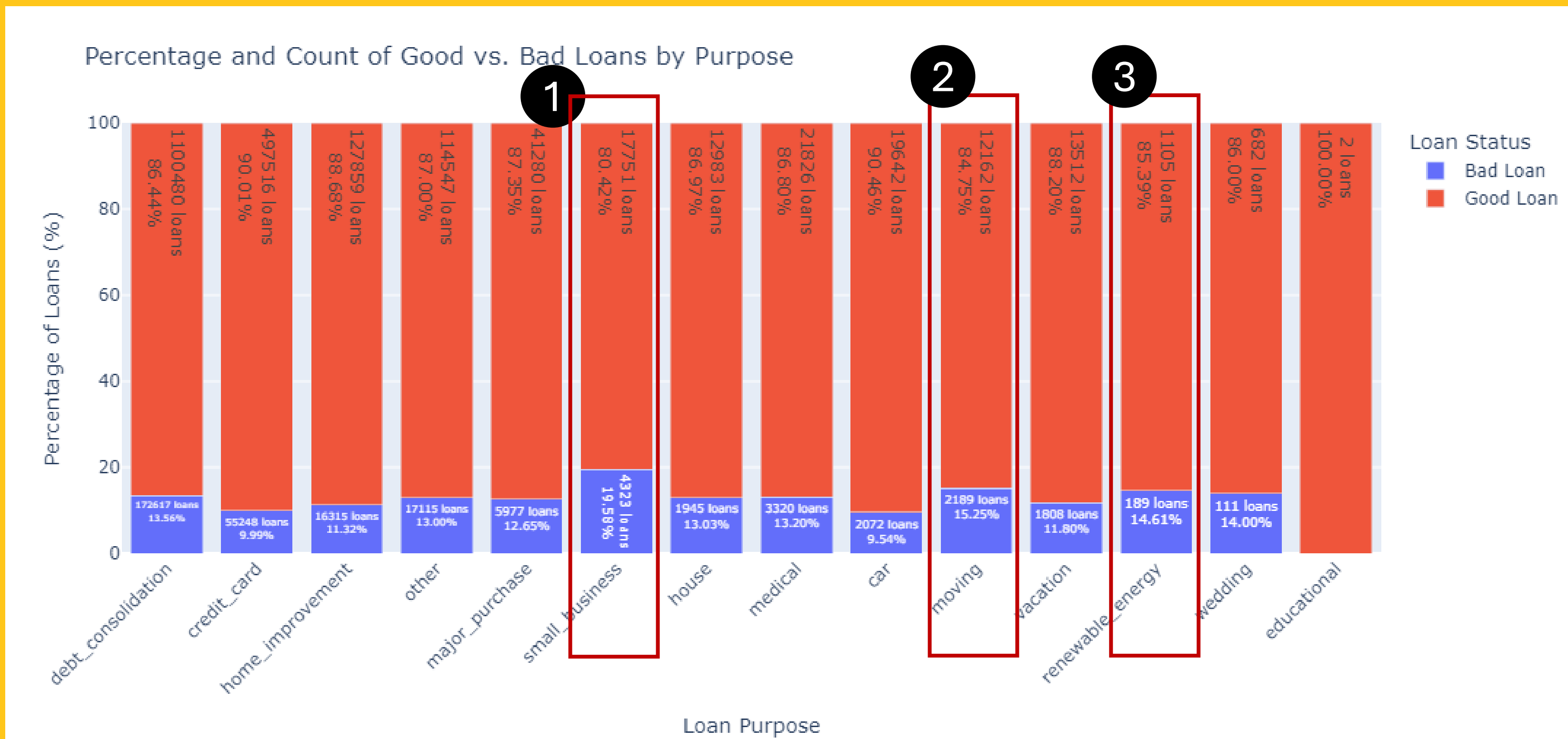
Cat 1: Borrower's immediate financial indicators (e.g. income, debt-to-income ratio, FICO scores, etc)

Cat 2: Indicators about the loan (e.g. interest rate, grade, purpose, etc).

Cat 3: Borrower's financial history (e.g. revolving balance, revolving utilisation rate, and history of delinquency)

Cat 4: Miscellaneous (e.g. residential state, employment length, home ownership, etc

Purpose



While most borrowers loaned money for debt consolidation, small business loans were most likely to result in bad loans, followed by moving and renewable energy. It is misleading to think that educational loans are safest with 100% repayment because there were just 2 such loans out of 2.6 million loans.

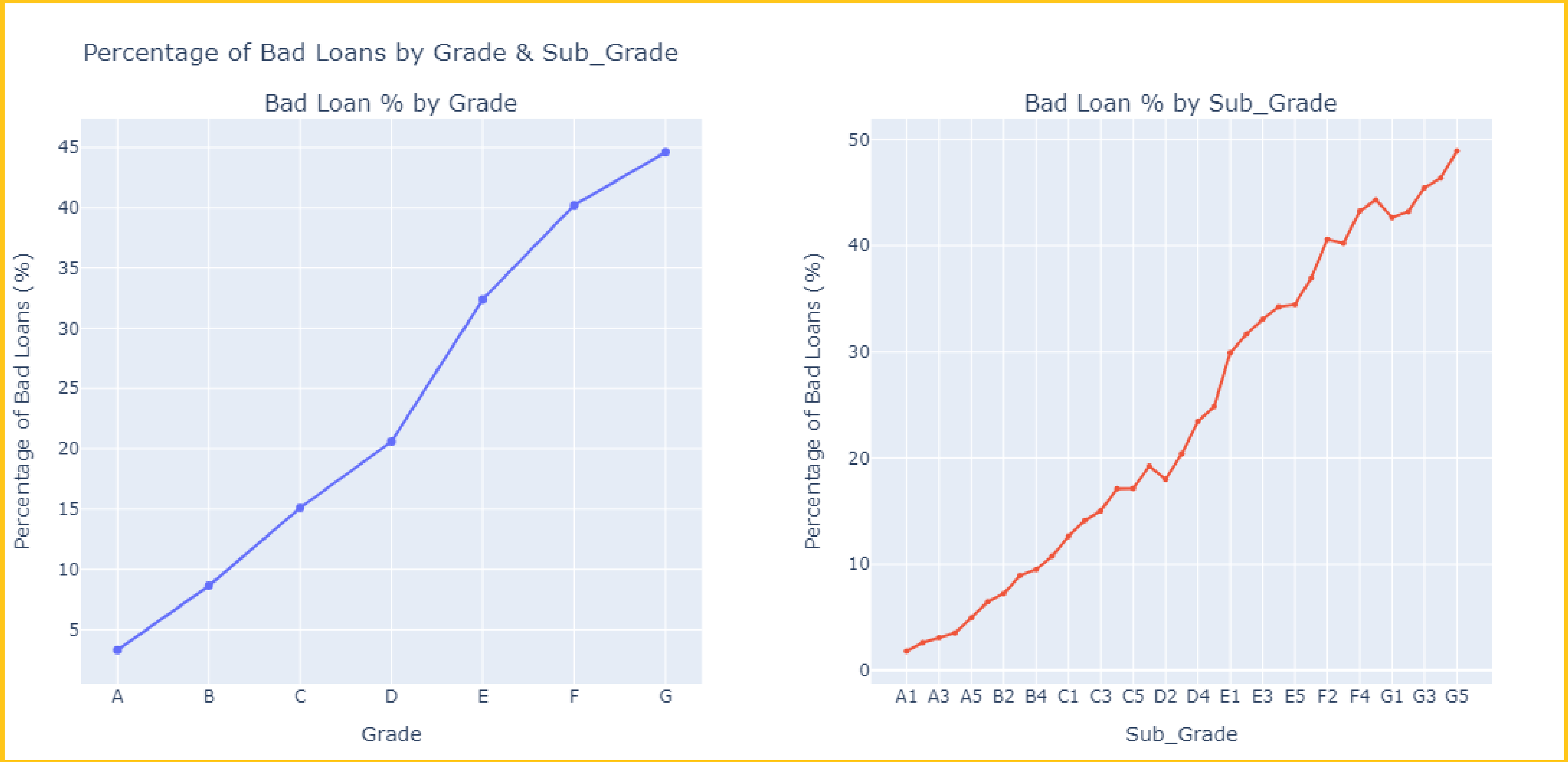
Cat 1: Borrower's immediate financial indicators (e.g. income, debt-to-income ratio, FICO scores, etc)

Cat 2: Indicators about the loan (e.g. interest rate, grade, purpose, etc).

Cat 3: Borrower's financial history (e.g. revolving balance, revolving utilisation rate, and history of delinquency)

Cat 4: Miscellaneous (e.g. residential state, employment length, home ownership, etc

Loan Grade and Sub-Grade



Similarly, loan grade and sub-grade are useful indicators for whether a borrower is likely to default

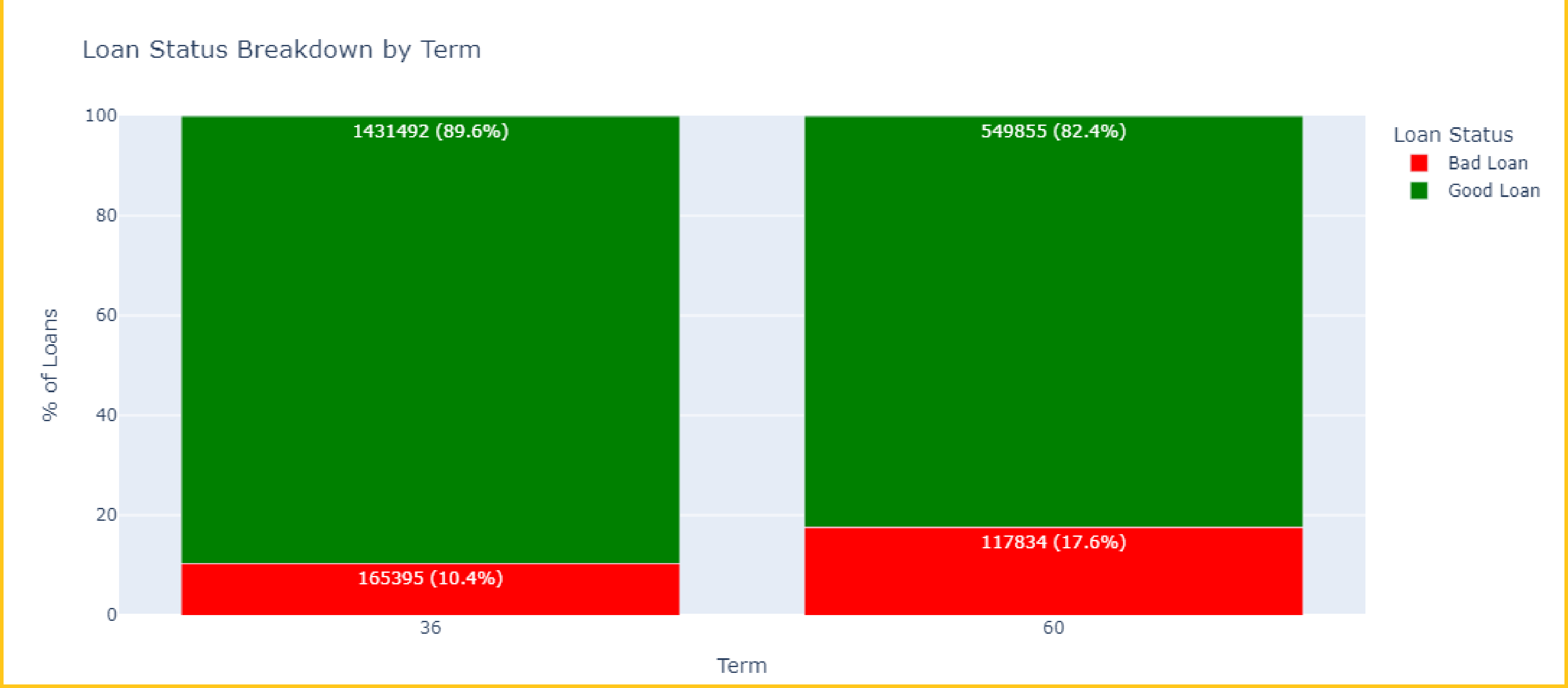
Cat 1: Borrower's immediate financial indicators (e.g. income, debt-to-income ratio, FICO scores, etc)

Cat 2: Indicators about the loan (e.g. interest rate, grade, purpose, etc).

Cat 3: Borrower's financial history (e.g. revolving balance, revolving utilisation rate, and history of delinquency)

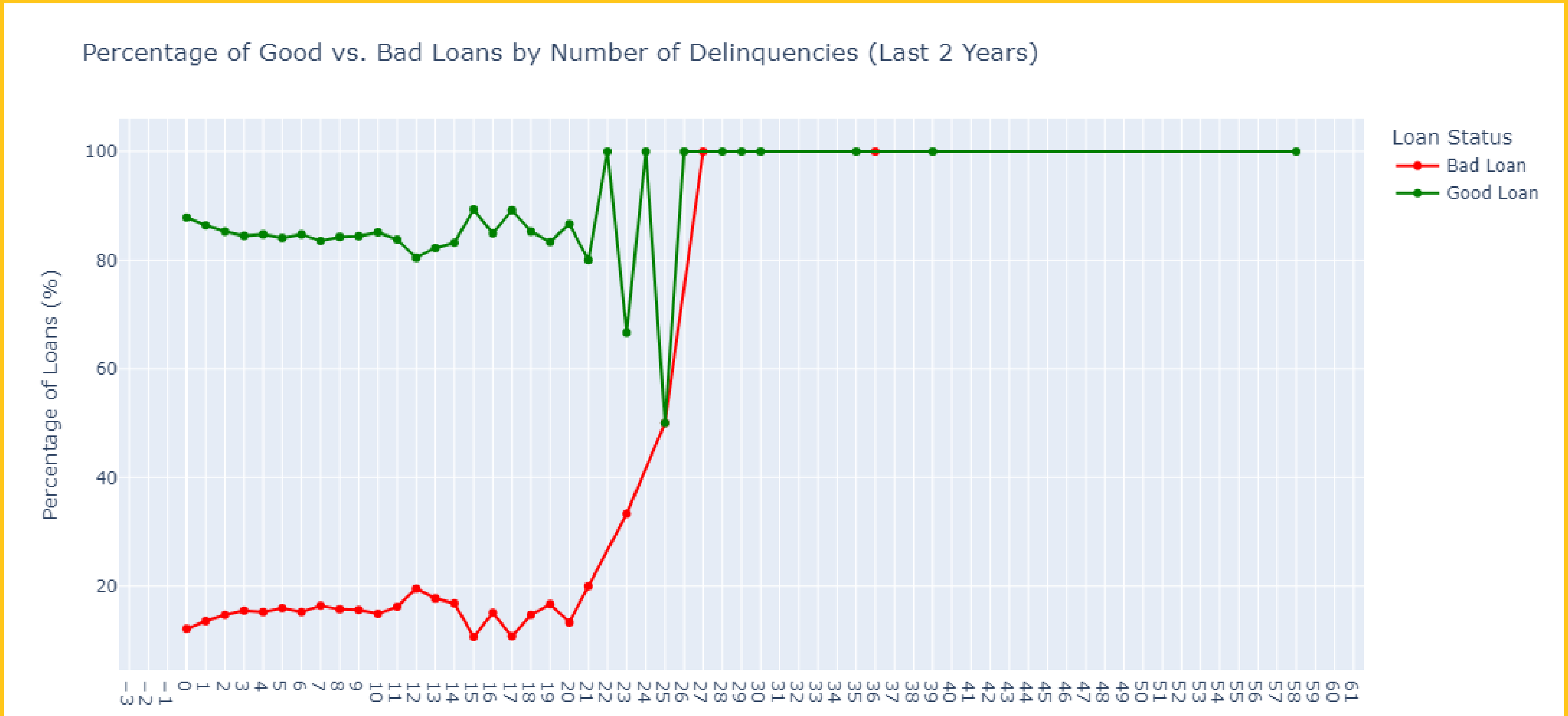
Cat 4: Miscellaneous (e.g. residential state, employment length, home ownership, etc

Borrowing Term



Borrowers who chose a 60-month loan is more likely to default in their loan.

History of Delinquency



A borrower's track record is useful in determining whether a loan is likely to default. Past 20 delinquencies, there is a sharp increase in number of bad loans with each increasing delinquency.

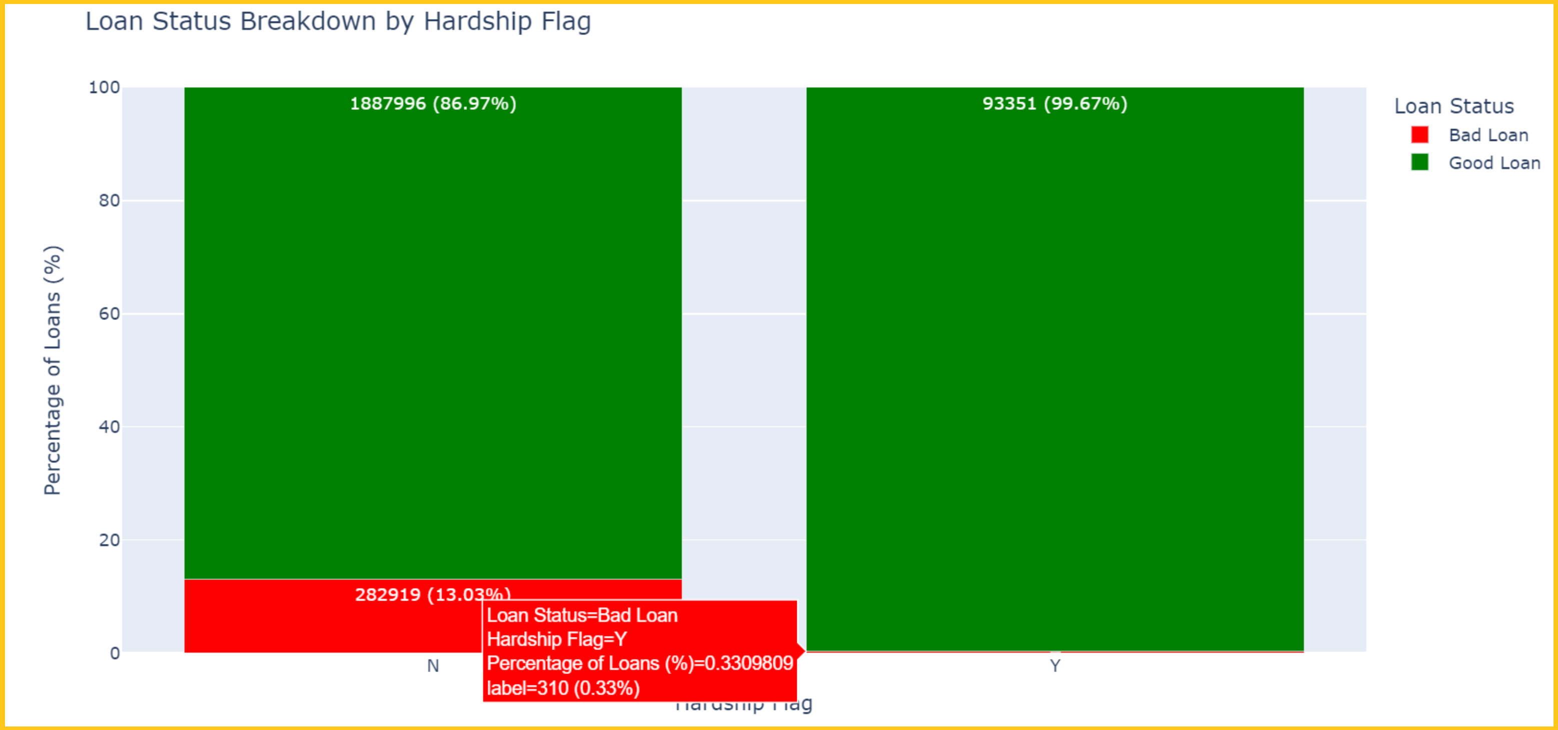
Cat 1: Borrower's immediate financial indicators (e.g. income, debt-to-income ratio, FICO scores, etc)

Cat 2: Indicators about the loan (e.g. interest rate, grade, purpose, etc).

Cat 3: Borrower's financial history (e.g. revolving balance, revolving utilisation rate, and history of delinquency)

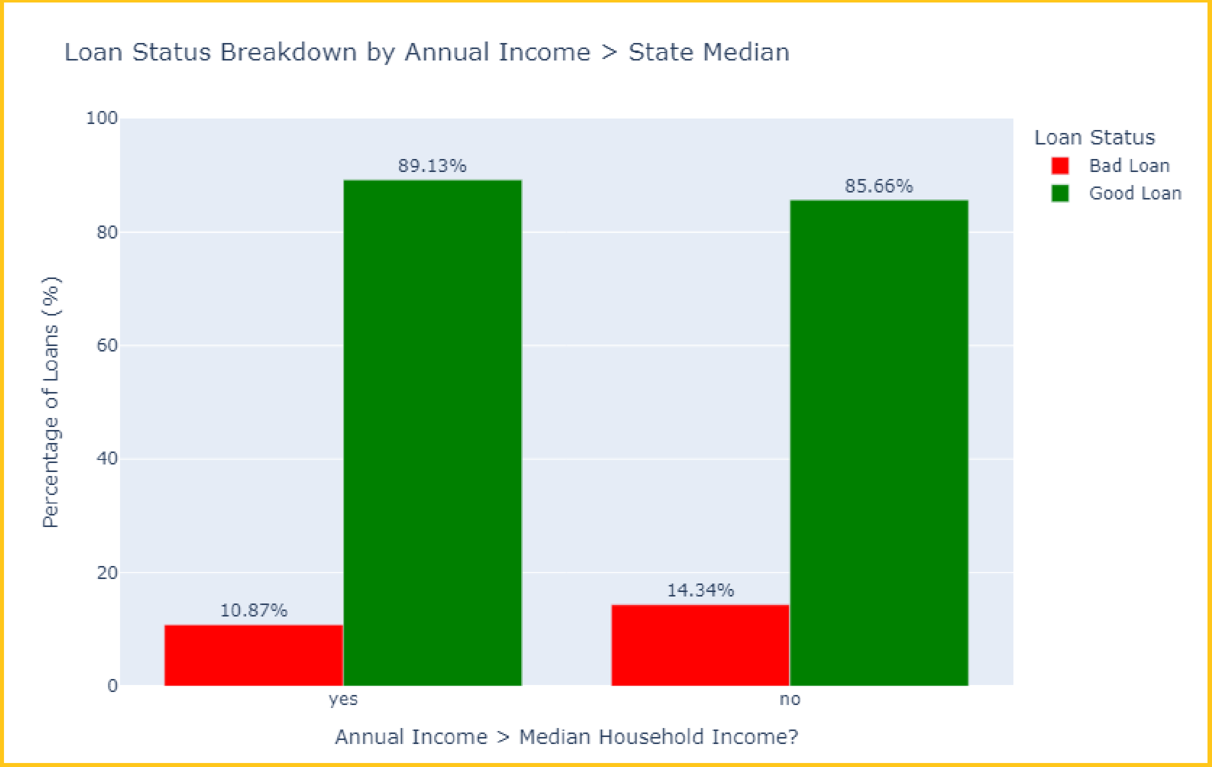
Cat 4: Miscellaneous (e.g. residential state, employment length, home ownership, etc

Hardship Flag



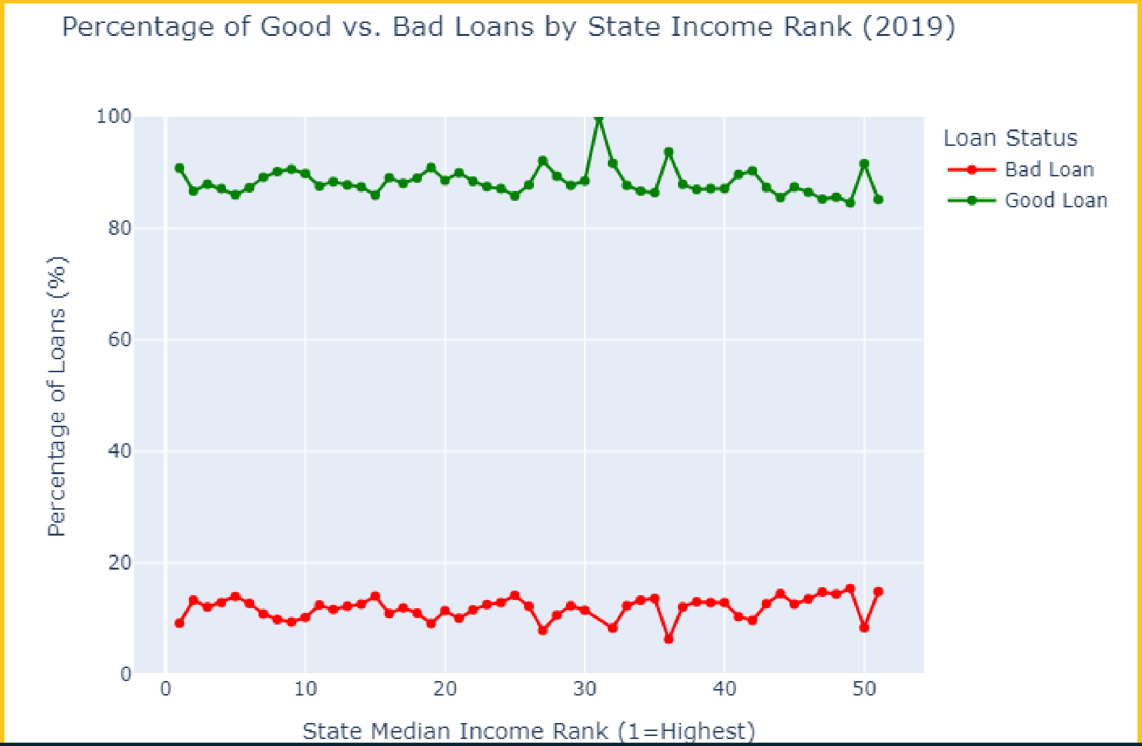
Borrowers who had a hardship flag were 913 times less likely to have a bad loan, compared to borrowers without a hardship flag.

Median household income based on residential State



Borrowers whose annual income were lower than their state's median household income had a higher chance of defaulting

Median household income based on residential State



The higher (i.e. wealthier) the state's median household income, the higher the likelihood of a good loan

Key Observations

1. Relationship between interest rate, loan grade, and loan status:
    a. Are bad loan more likely to be allocated a higher interest rate or poor loan grade or
    b. Are poor loan grades and high interest likely to cause loans to default?

2. While loan amount for bad loans tend to be higher, loan amount is not a factor of whether borrowers is likely to default.

3. Borrowers who had a hardship flag are more unlikely to default on their loans.

4. Borrowers who chose a 60-month loan is more likely to default in their loan.

Machine Learning Models

XGBoost

Logistic Regression Model

XGBoost Model

A good performing model for predicting bad loans would be able to

1. **Recall**: identify loans that were predicted to be good but eventually defaulted (i.e. false negatives) 70% of the time **and**
2. **Precision**: identify good loans correctly 70% of the time

Modelling – Logistic Regression

	Precision	Recall	F_1
Base Model	0.30	0.74	0.43
'best_log_reg_model' (i.e. Hyperparameter Tuned Model) n_iter = 6	0.30	0.74	0.43
Threshold Calibration for 'best_log_reg_model' (precision > 0.7)	0.25	0.00	0.00

While the recall for the base model was >70%, the precision was low and could not be increased after adjustments to the model was made. In fact, the performance dipped. Hence, it is **not a good model for predicting bad loans.**

Modelling – XGBoost

	Precision	Recall	F_1
Base Model	0.53	0.89	0.66
‘best_xgb_model’ (i.e. Hyperparameter Tuned Model)	0.53	0.89	0.66
Threshold Calibration for ‘best_xgb_model’ (precision > 0.7)	0.70	0.80	0.75

The recall for the base model was >70% and precision further improved after adjustments to the model was made. Hence, it is **a good model for predicting bad loans.**

Recommendation

1. Technical
   a. XGBoost machine learning model is preferred
   b. Further study on the causal relationship between economic indicators and loan credit default

2. Business
   a. Small banks should encourage borrowers to flag repayment issues as early as possible
   b. Further study on the assumption that small banks are more risk-averse as it might not be true for all small banks