TU Wien

Information Visualization

Part 1

Group №21

1. Topic

Visual exploration of credit risk and loan approval decisions in consumer lending.

When you apply for a loan or credit card, the bank makes a decision that can feel like a mystery approved or denied, often with little explanation. Behind the scenes, lenders evaluate dozens of factors: your income, credit history, existing debts, employment status, and more. But how exactly do these pieces fit together? Which factors matter most? And are there patterns in who gets approved versus rejected?

In this project, we set out to answer these questions by visually exploring a realistic loan approval dataset. Our goal is to open the black box of lending decisions and make credit risk assessment more transparent and understandable. By analyzing the data through visualizations, we can uncover the hidden patterns that drive approval outcomes and gain insight into how banks think about risk.

2. Dataset Description

Name: “Realistic Loan Approval Dataset – US & Canada” (also referred to as the Synthetic Loan Approval Dataset).
Source: Kaggle dataset with 50,000 loan applications and 20 attributes (customer_id + 18 predictors + 1 target).

Each row corresponds to a single loan application for a credit card, personal loan, or line of credit. The data is synthetic but generated based on real-world banking criteria and lending policies from the US and Canadian markets. The dataset implements realistic approval rules, for example thresholds on debt-to-income ratio, the impact of past defaults or delinquencies, and the influence of credit score bands. The target variable loan_status encodes whether the application was approved (1) or rejected (0).

Because the dataset is synthetic, no actual customer information is exposed, but the underlying logic is designed to approximate real credit risk decision-making. This makes it suitable for educational use, exploratory data analysis, and model development.

3. Data Attributes

Attribute	Description	Type / Level
customer_id	Unique applicant identifier	Nominal identifier
age	Age in years	Numeric, discrete (ratio)
occupation_status	Employment category (e.g., Employed, Self-Employed, Student)	Nominal
years_employed	Years in current job	Numeric, continuous (ratio)
annual_income	Yearly income in USD	Numeric, continuous (ratio)
credit_score	Credit score (approx. FICO range 300–850)	Numeric, discrete; treated as interval/ordinal
credit_history_years	Length of credit history	Numeric, continuous
savings_assets	Savings or liquid assets	Numeric, continuous
current_debt	Existing outstanding debt	Numeric, continuous
defaults_on_file	Any past default on file (0/1)	Binary
delinquencies_last_2yrs	Number of late payments in the last two years	Numeric, discrete count
derogatory_marks	Serious negative credit marks	Numeric, discrete count
product_type	Type of product (Credit Card, Personal Loan, Line of Credit)	Nominal
loan_intent	Stated loan purpose (Education, Business, Debt Consolidation, etc.)	Nominal
loan_amount	Amount of credit requested	Numeric, continuous
interest_rate	Annual interest rate (%)	Numeric, continuous

debt_to_income_ratio	Debt-to-income ratio	Numeric, continuous
loan_to_income_ratio	Loan amount relative to income	Numeric, continuous
payment_to_income_ratio	Estimated monthly payment relative to income	Numeric, continuous
loan_status	Loan outcome: Approved (1) or Rejected (0)	Binary target

There are no missing values; the table is complete. The values are generated to be realistic and include a number of edge cases, such as very high debt-to-income ratios or very low credit scores. These extreme values are not data errors but domain-meaningful observations, so they should not be removed as simple outliers in the analysis.

For our visual analysis we may apply several transformations to improve interpretability:
• Binning or log scaling of highly skewed monetary variables such as annual_income, loan_amount, savings_assets, and current_debt.
• Expressing ratio variables (debt_to_income_ratio, loan_to_income_ratio, payment_to_income_ratio) as percentages for clearer axis labels.
• Grouping age and credit_score into categorical bands when appropriate (e.g., "young / middle-aged / senior" or "poor / fair / good / excellent credit") to support aggregated comparisons.

4. Data Access

The dataset is publicly available on Kaggle under the title "Realistic Loan Approval Dataset – US & Canada". It can be accessed by searching for this title on Kaggle and downloading the CSV file provided by the author. The dataset is released under a Creative Commons licence suitable for educational and non-commercial use.

Link - https://www.kaggle.com/datasets/parthpatel2130/realistic-loan-approval-dataset-us-and-canada?resource=download