CS 571 - Data Visualization and Exploration

Spring 2025 - UMass Amherst

Project Proposal

Project Metadata

Project Title: *CreditCanvas* – Interactive Credit Data Visualization

Group Members:

- **Ruchi Gupta** | Email: ruchigupta@umass.edu | SPIRE ID: 32824484
- **Shoubhit Ravi** | Email: sravi@umass.edu | SPIRE ID: 33644772
- **Suryam Gupta** | Email: suryamgupta@umass.edu | SPIRE ID: 31142337

Project Repository: GitHub - CreditCanvas

GitHub Usernames:

- **Ruchi Gupta:** `RuchiGupta20`
- **Suryam Gupta:** `Suryamgupta25`
- **Shoubhit Ravi:** `shoubhitravi`

Background and Motivation

In today's financial landscape, credit plays a crucial role in determining access to loans, mortgages, and various financial opportunities. However, understanding credit data can be complex due to the vast number of factors influencing credit scores, lending decisions, and financial risk.

Our motivation for this project stems from a desire to make credit data more transparent and accessible through interactive visualizations. By leveraging modern web-based visualization techniques, we aim to uncover key trends, relationships, and disparities within credit data, helping users explore patterns such as:

- **Credit score distribution:** Understanding how credit scores are distributed across different demographics and income levels.
- **Loan approval trends:** Analyzing factors that influence loan approval rates and potential biases in lending.
- **Debt and repayment patterns:** Visualizing the relationship between debt levels, interest rates, and repayment behaviors.
- **Impact of economic factors:** Exploring how external factors like inflation, unemployment rates, and market trends affect creditworthiness.

All of us share a strong interest in the intersection of technology and finance, with some of us even having an academic background in finance. This naturally led us to choose credit data as our focus. Credit is fundamental to long-term financial goals, from buying a home to securing business loans. Personal finance planning is something we all care about, and we recognize how understanding creditworthiness can empower individuals to make better financial decisions.

Additionally, we noticed that there aren't many free and accessible tools available for credit data visualization. Most existing platforms that provide detailed credit analysis are either paid, expensive, or restricted, limiting the ability of everyday users to explore their own financial standing. We wanted to change that by developing an open, user-friendly, and interactive tool that makes credit insights available to everyone, regardless of financial background.

With this in mind, we emphasized the importance of making credit trends and insights intuitive and interactive. Our goal is to create a tool that helps users enhance their understanding of credit, whether it's for improving their own credit standing, making informed financial choices, or simply exploring how credit works at a broader level. By bringing data to life with engaging visualizations, we hope to bridge the gap between complex financial data and everyday decision-making.

Beyond this course, we see this project as something we could potentially pursue for research, exploring areas like algorithmic fairness in lending, biases in credit scoring models, and financial literacy through visualization. By making credit data easier to understand, we hope to empower users with the knowledge they need to take control of their financial future, fostering better financial literacy and accessibility in an increasingly data-driven world.

Project Objectives

Our project focuses on visualizing credit data to uncover meaningful insights and make complex financial concepts more accessible. Through interactive visualizations, we aim to explore key questions that help users better understand credit trends, creditworthiness, and financial decision-making.

Primary Research Questions:

- How are credit scores distributed across different demographics, income levels, and regions?
- What factors most strongly influence credit scores?
- What factors most strongly influence loan approvals and denials?
- Are there any noticeable biases or disparities in lending decisions?
- How do debt levels and repayment behaviors impact credit scores over time?
- What external and economic factors, such as region, inflation and unemployment rates, affect credit trends?

Goals and Expected Outcomes:

- **Identify Patterns in Credit Data:** Our visualizations will help users explore how various financial behaviors influence credit scores and lending decisions.
- **Enhance Financial Literacy:** By making credit data more intuitive, we aim to provide users with a clearer understanding of what impacts their creditworthiness.
- **Support Data-Driven Decision-Making:** Users will be able to analyze trends and patterns to make informed financial choices, whether for personal finance planning or business decisions.
- **Increase Accessibility to Credit Insights:** Unlike existing paid tools, our project provides a free and interactive way for individuals to visualize and engage with credit data.

By achieving these objectives, we hope to create a tool that doesn't just answer research questions but genuinely helps people make sense of their credit data in a way that feels practical and personal. Whether someone is trying to improve their credit score, understand why a loan application was denied, or simply get a clearer picture of how financial decisions impact their future, our visualizations aim to provide clarity.

Credit plays a role in so many major life decisions, such as buying a home, securing a car loan, and starting a business, yet most people don't have easy access to intuitive tools that explain how it works. By making these insights accessible and interactive, we want to empower users with the knowledge they need to take control of their financial future, make informed choices, and feel more confident navigating the world of credit.

Data

- Datasets:
  - Standard/Good/Poor Credit Score Rating:
    - https://www.kaggle.com/datasets/parisrohan/credit-score-classification?select=train.csv
  - Credit Score Rating, categories for separate domains (clothing spend, education, rent, etc.)
    - https://www.kaggle.com/datasets/conorsully1/credit-score
  - Approved/not approved for credit card, other fields like debt to income ratio
    - https://www.kaggle.com/c/GiveMeSomeCredit/overview
  - Loan-Approval-Prediction-Dataset
    - https://www.kaggle.com/datasets/architsharma01/loan-approval-prediction-dataset/data
  - Loan-Approval-Prediction.csv
    - https://github.com/prasertcbs/basic-dataset/blob/master/Loan-Approval-Prediction.csv
  - Lending interest rate (%)
    - https://data.worldbank.org/indicator/FR.INR.LEND
  - Credit Lending Interest Rates
    - https://www.kaggle.com/datasets/tarique7/credit-lending-interest-rates
- Other helpful sources:
  - What Is the Average Credit Score in the US?
    - https://www.experian.com/blogs/ask-experian/what-is-the-average-credit-score-in-the-u-s/
  - The 20 Most Relevant Credit Score Statistics in 2023
    - https://www.creditstrong.com/credit-score-statistics/
  - What's in my FICO® Scores?
    - https://www.myfico.com/credit-education/whats-in-your-credit-score
  - Borrower risk profiles
    - https://www.consumerfinance.gov/data-research/consumer-credit-trends/student-loans/borrower-risk-profiles/
  - Average Credit Score in US: FICO and VantageScore Breakdowns
    - https://www.lendingtree.com/credit-repair/credit-score-stats-page/
  - Data Warehouse and Visualizations for Credit Risk Analysis
    - https://blogs.oracle.com/database/post/data-warehouse-and-visualizations-for-credit-risk-analysis
- Other sources/websites for obtaining data:
  - Experian
  - FICO
  - VantageScore
  - USA Today
  - Urban Institute

Data Processing

Before creating meaningful visualizations, we need to preprocess the credit data to ensure it is clean, structured, and optimized for analysis. This involves several key steps, including data cleaning, transformation, integration, and feature extraction.

Is the Data Ready to Use?

We have found relevant datasets on **Kaggle**, including datasets with credit score history and Loan approval datasets, which provide information on factors influencing credit decisions. While these datasets offer a strong starting point, we will also explore additional data sources from **Experian, FICO, and VantageScore** to correlate parameters with credit trends, loan approvals and denials, and the factors influencing credit history. Since data from these sources may not always be readily available in structured formats, additional cleaning and processing may be required.

What Level of Cleanup is Needed?

- **Handling Missing Values:** We aim to handle missing values by either imputing them with relevant statistical measures (mean, median) where it makes sense or removing incomplete entries where important fields are missing to maintain consistency. For example, if an entry is missing an annual income value but has other relevant credit history data, we may impute the missing value with the median income from similar profiles. However, if a record lacks critical fields like both credit score and loan repayment status, it may be removed to prevent misleading insights.
- **Removing Incorrect, Corrupted, or Incomplete Data:** In addition to handling missing values, we will identify and remove data that appears incorrect or corrupted, such as negative income values, implausibly high loan amounts, or inconsistent credit score records.
- **Standardizing Categorical Variables:** Ensuring consistency in categorical data, such as loan types, credit history categories, and borrower demographics, to facilitate meaningful comparisons. For instance, different datasets may label loan types inconsistently, such as "Home Loan" vs. "Mortgage" or "Auto Loan" vs. "Car Loan." We will standardize these labels to maintain uniformity.
- **Ensuring Numerical Consistency and Data Transformation:** Normalizing income levels, credit scores, and loan amounts to a uniform scale for accurate visualization. Additionally, we will check for unit consistency, such as verifying whether income is recorded in annual or monthly figures, loan amounts are in thousands or full values, and interest rates are expressed in decimals or percentages. We will also apply log transformations or feature scaling where necessary to make data more interpretable in visualizations.
- **Data Integration:** Since we are using multiple data sources, we will integrate them into a single, cohesive dataset. This involves merging data from Kaggle, Experian, FICO, and VantageScore to create a more comprehensive view of credit trends. We will align data fields, resolve discrepancies, and ensure that variables from different sources are comparable.

- **Handling Duplicated Values:** Identifying and removing duplicate records to avoid skewed results, especially in datasets where multiple entries for the same individual may exist due to repeated credit checks or loan applications.
- **Sampling and Balancing Data:** If datasets contain a disproportionate number of certain credit score groups or loan outcomes, we may need to apply sampling techniques to ensure fair representation. For instance, if a dataset has significantly more approved loans than rejected ones, we may downsample the majority class or upsample the minority class to maintain balanced insights.
- **Removing Features that Could Introduce Bias**: While considering as many factors as possible is important, we don't want to consider features that may derive implicit biases in code. Objectively, credit should be measurable for everyone equally regardless of their gender, level of education, etc.

What Quantities Do We Expect to Derive?

To enhance visualization and analysis, we plan to compute and derive various metrics that provide deeper insights into credit trends, loan approvals, and financial behavior. These derived attributes will help uncover meaningful patterns and relationships within the data.

Derived Attributes

We can derive new attributes from existing ones using various transformation techniques:

- **Changing Attribute Type:**
  - Converting credit score from a numerical value to categories like poor, fair, good, and excellent for easier interpretation.
  - Transforming loan amounts into bins such as small (<$10,000), medium ($10,000–$50,000), and large (>$50,000) to analyze borrowing trends.
- **Acquiring Additional Information:**
  - Enriching datasets with economic indicators like inflation rates or unemployment statistics for better contextual analysis of credit behavior.
- **Using Arithmetic, Logical, or Statistical Operations:**
  - **Computing ratios:**
    - Credit utilization ratio = total credit used / total credit limit to assess financial health.
    - Debt-to-income ratio = total debt / annual income to evaluate borrowing capacity.
  - **Difference calculations:**
    - Difference between requested loan amount and approved loan amount to assess lending patterns.
    - Change in credit score over time to visualize credit improvement or deterioration trends.
  - **Averaging attributes:**
    - Mean interest rates across different credit score categories to analyze borrowing costs.
    - Average repayment time for different loan types to understand payment behaviors.

We will categorize the dataset attributes into the following types to ensure proper handling for visualization:

- **Categorical (No Implicit Ordering):**
  - Loan type (e.g., home loan, auto loan, personal loan)
  - Lender type (e.g., bank, credit union, online lender)
  - Employment status (e.g., employed, self-employed, unemployed)
  - Homeownership status (e.g., own, rent, mortgage)
- **Ordinal (Implicit Ordering but No Arithmetic Operations):**
  - Credit score category (e.g., poor, fair, good, excellent)
  - Loan approval status (e.g., denied, conditionally approved, fully approved)
  - Education level (e.g., high school, bachelor's, master's, PhD)
- **Quantitative (Ordered and Supports Arithmetic Comparison):**
  - Credit score (numerical value)
  - Annual income ($ value)
  - Loan amount requested/approved ($ value)
  - Interest rate (% value)
  - Debt-to-income ratio (calculated metric)
  - Repayment period (in months/years)

How Will We Implement Data Processing?

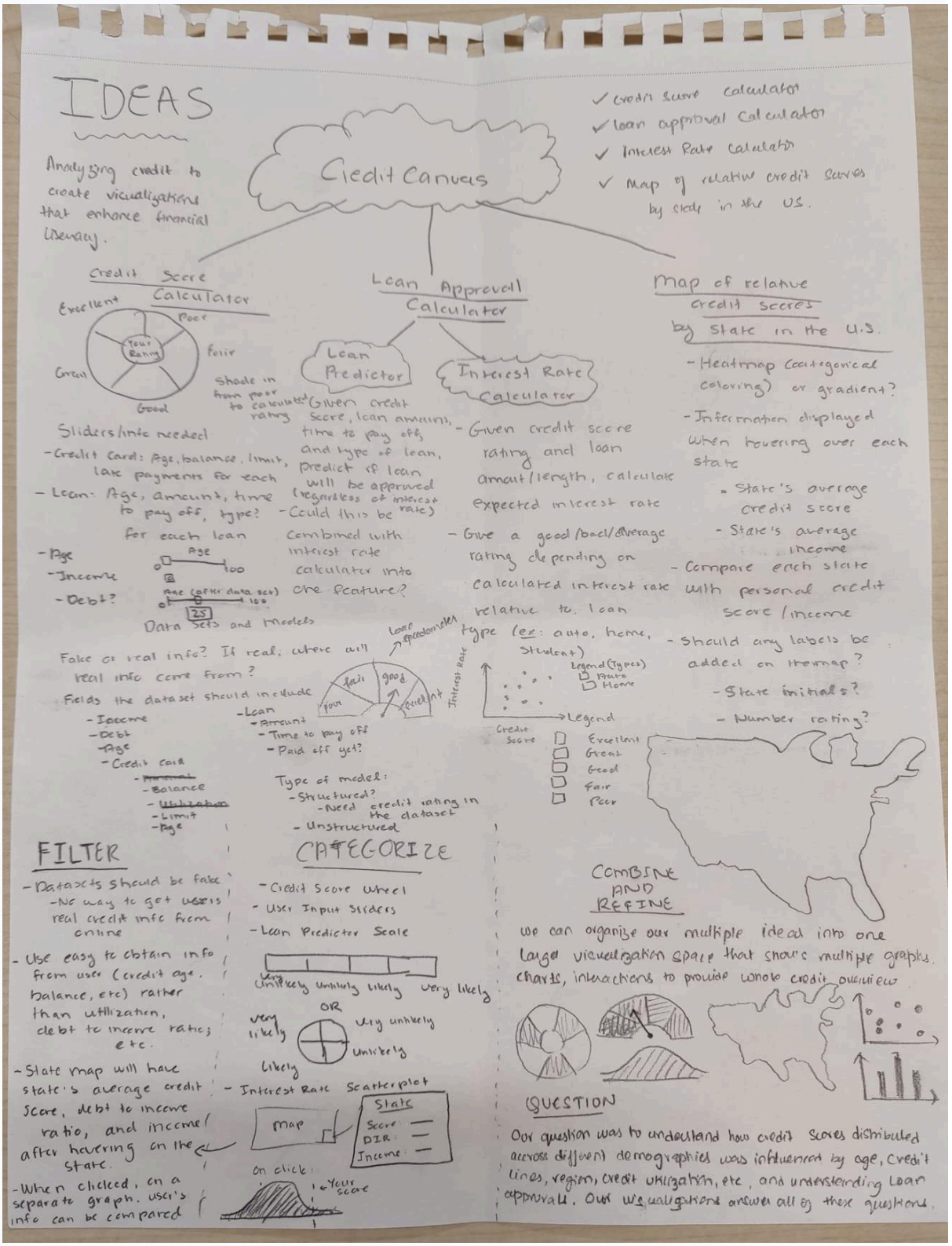
We will use **Python** for data preprocessing, leveraging libraries such as:

- **Pandas** for data manipulation and cleaning
- **NumPy** for numerical transformations
- **Scikit-learn** for scaling, encoding, and imputation
- **Matplotlib and Seaborn** for preliminary data exploration

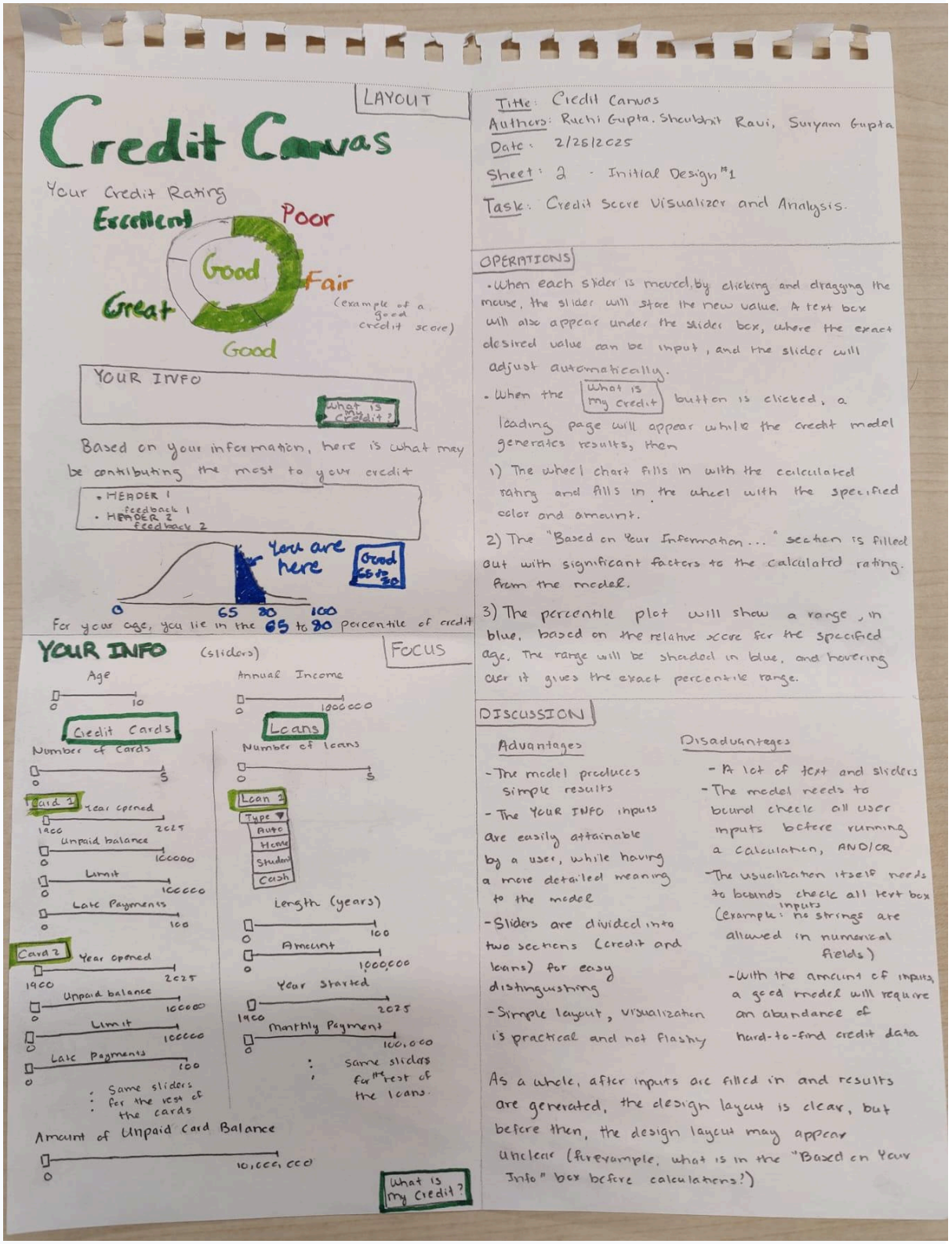
---

Visualization Design

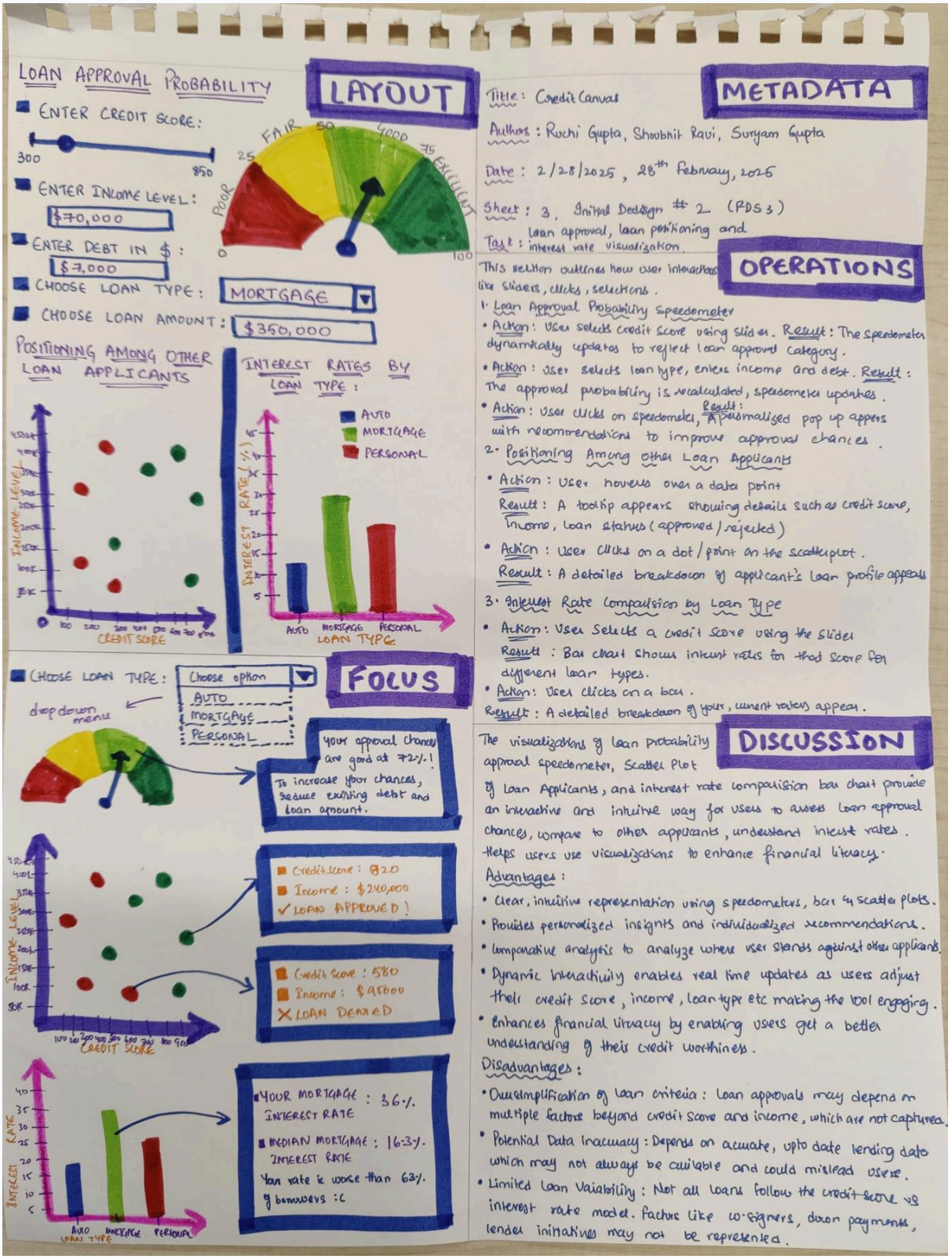
Sheet 1: Brainstorm



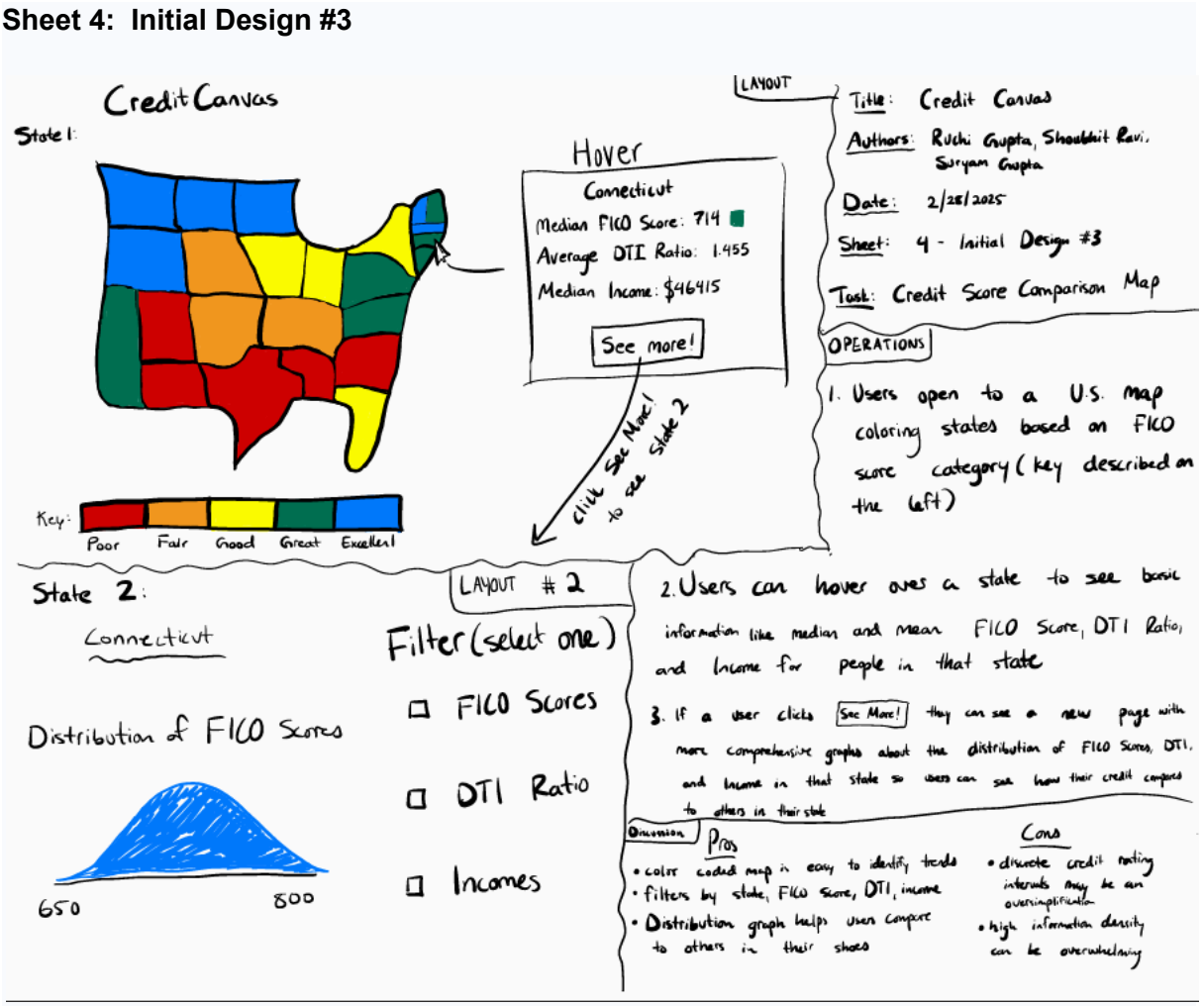
Sheet 2: Initial Design #1



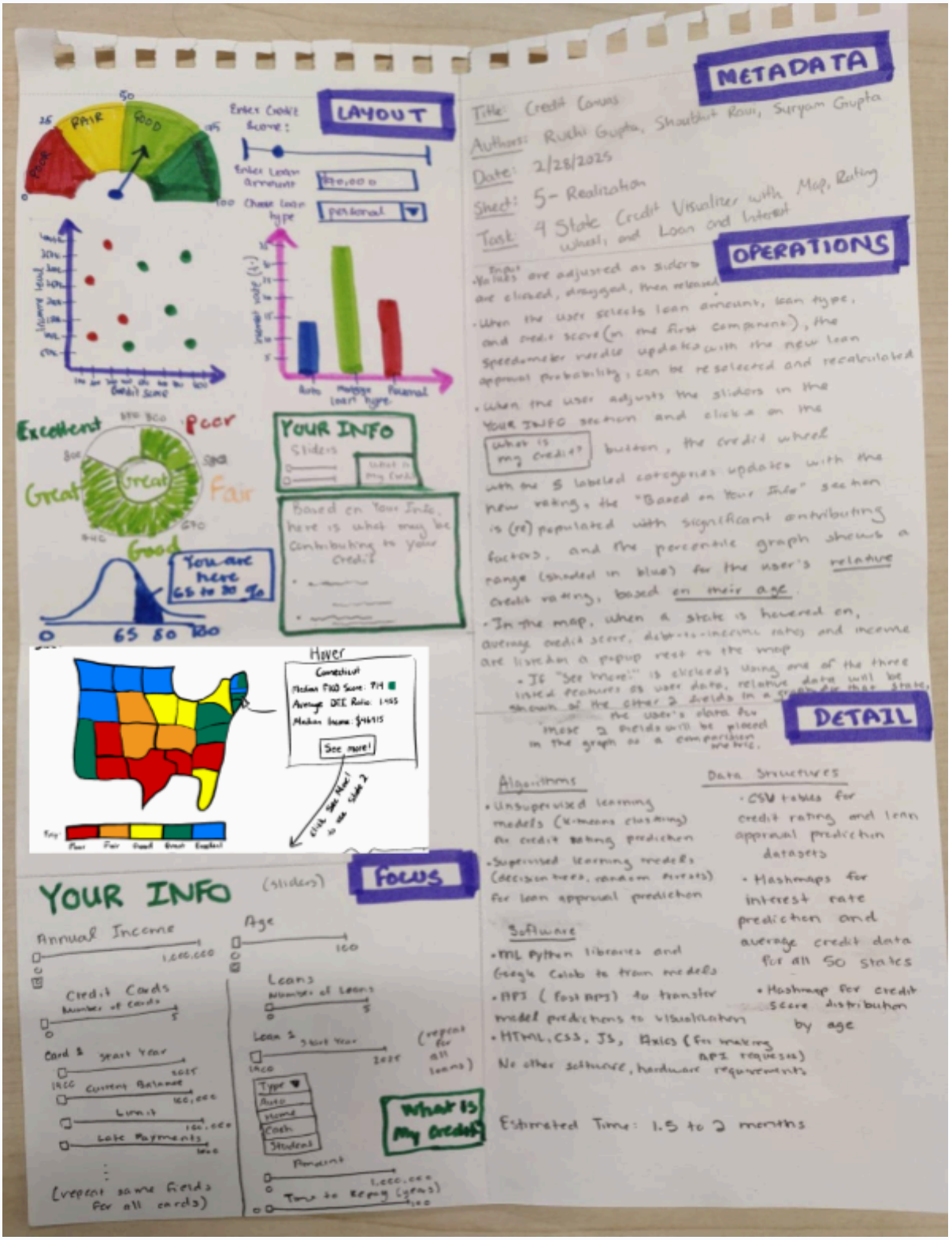
Sheet 3: Initial Design #2



Sheet 4: Initial Design #3



Sheet 5: Final Realization



Justifying Our Design and Visual Encoding Choices

The visual encodings in this credit visualizer, color, shape, size, and interactivity, make it easy to explore financial data in a way that feels natural and engaging. Each choice is

intentional to help users quickly grasp key insights. Color makes information intuitive. The red-to-green gradient instantly signals credit risk levels. Scatter plots and bar charts use distinct colors to differentiate loan types and FICO scores, making comparisons effortless. Shape and size guide users in understanding relationships. Donut charts break down credit distributions, bar charts emphasize approval probabilities, and scatter plots reveal patterns in loan outcomes.

Interactivity takes this a step further by making the experience dynamic. Sliders let users adjust inputs like income and credit card usage, with immediate updates to their insights. Hovering over states on the map brings up regional credit trends, adding context without overwhelming the screen. The “See More” button keeps things clean while still allowing users to dive deeper when they want to. These thoughtful design choices create a smooth, informative experience that makes it easy to understand personal credit standing.

Color Encoding

Color helps users quickly interpret their standing. The red-to-green gradient on the credit rating scale makes risk levels instantly recognizable. Scatter plots and bar charts use distinct colors to separate loan types and FICO scores for easy comparison.

Shape and Size Encoding

Shapes and sizes make comparisons straightforward. Donut charts break down credit distributions. Bar charts visually emphasize loan approval probabilities. Scatter plots highlight trends in credit scores and outcomes.

Interactivity

Interactive elements keep the experience engaging. Sliders let users adjust factors like income and credit cards, showing real-time updates. Hovering over states on the map reveals regional credit trends. The “See More” button keeps the interface clean while allowing deeper exploration.

Must-Have Features

1. Credit Rating

Understanding one's credit rating is essential for financial planning and decision-making. The credit wheel graph provides a visual representation of a user's credit standing, making it easy to interpret. By incorporating a backend model that evaluates factors like age, income, credit history, and loan details, the system offers a reliable credit score classification. Additionally, the percentile graph compares the user's credit rating to others in the same age group, helping them see where they stand and identify areas for improvement.

Credit Wheel Graph

5 sections: poor, fair, good, great, and excellent, arranged clockwise.
Color fill: based on a calculated rating, sections from poor up to the calculated rating are filled with a specific color.
Ratings and colors:

- Poor: red-orange
- Fair: orange-yellow
- Good: yellow-light green
- Great: green
- Excellent: dark green

Rating Calculation

A backend trained model predicts credit ratings based on age, annual income, credit history, and loan history. The visualization fetches model results via an API and categorizes the user’s rating as poor, fair, good, great, or excellent.

User Inputs for Calculation (Each Controlled by a Slider)

- Age
- Annual income
- Number of credit cards
    - For each card:
        - Year started
        - Current balance
        - Limit
        - Late payments
- Number of loans
    - For each loan:
        - Year started
        - Amount
        - Time to pay off
        - Type (dropdown: home, auto, student, or cash)

Additional Insights

- Significant factors affecting credit rating are highlighted below the credit wheel.
- A credit percentile graph displays the user's percentile range based on their age, with poor at the 0th percentile and excellent at the 90th or 99th percentile. Since credit ratings are categorical and discrete, the percentile must be a range rather than a single value.
- The percentile range is shaded on the graph, and users can hover over it to see the exact percentile range.

2. Loan Approval and Interest Rate

Loan approval chances and interest rates impact major financial decisions, such as buying a home or car. This feature allows users to gauge their approval probability in real time, compare themselves with past applicants, and understand how credit scores influence interest rates. The loan approval speedometer provides a simple way to interpret approval chances, while the scatter plot helps users see where they stand among other applicants. The interest rate comparison bar chart further informs users about potential savings based on their credit score.

Loan Approval Probability Speedometer

- User inputs: credit score, income level, current debt, and loan type.
- The speedometer updates dynamically to categorize loan approval chances as:
    - Poor (red)
    - Fair (yellow)
    - Good (light green)
    - Excellent (dark green)
- Clicking on the gauge provides personalized recommendations to improve approval odds, such as reducing debt or increasing income.

Positioning Among Other Loan Applicants (Scatter Plot)

- Axes:
    - X-axis: credit score
    - Y-axis: income level
- Data points:
    - Green dots = approved applicants
    - Red dots = rejected applicants
- Clicking on a dot reveals details about that applicant's profile and loan outcome, allowing users to compare their chances with real applicant trends.

Interest Rate Comparison by Loan Type (Bar Chart)

- Axes:
    - X-axis: loan type (auto loan, mortgage, personal loan)
    - Y-axis: interest rate (%)
- Bars represent interest rates for each loan category based on the user's selected credit score.
- Clicking on a bar reveals:

    - The median interest rate for that loan type
    - A comparison of the user's rate vs. other borrowers (e.g., "Your rate is better than 65% of borrowers.")

3. Credit Comparison Map

Financial conditions vary across states, making it useful to compare credit trends regionally. The credit comparison map provides an overview of credit scores, debt-to-income ratios, and income levels by state. Users can hover over a state to see basic financial statistics and, if needed, explore deeper insights through a detailed state-specific distribution graph. This feature helps users compare themselves with residents of different states and better understand their financial position.

State 1: Overview Map

- Color-coded by state, based on median or mean FICO scores, using either discrete categories or a color gradient.
- Hovering over a state displays preliminary financial statistics, including:
    - FICO score (mean or median)
    - Debt-to-income (DTI) ratio
    - Income (mean or median)
- A "See More" button allows users to access State 2, which provides a more detailed breakdown of financial trends within the selected state.

State 2: In-Depth Financial Distribution Graphs

- Users can choose to view a distribution graph for one of the following financial metrics in the selected state:
    - FICO scores
    - DTI ratio
    - Income
- The graph shades a percentile range corresponding to the user's data. Hovering over the shaded region reveals where the user falls within the state's distribution.
- Users may also set a fixed value for one financial feature (such as income) to compare themselves with others in similar financial situations.

---

4. Credit Prediction

There are not many tools online that allow someone to not only check their credit score for free, but also visualizing how changing certain financial factors can affect it and by how much. The credit prediction is visualized as a speedometer, where different areas represent different “ratings” of credit. There are four areas on the speedometer corresponding to the following credit ratings:

- Red: “Poor” (< 580 credit score)
- Yellow orange: “Fair” (580-669 score)
- Green: “Good” (670-739 score)
- Dark green: “Excellent” (> 740 score)

While credit scores in real life have five categories, the dataset did not provide enough data for “Very Good” and “Excellent” to be two separate categories, even with balancing. Therefore, for the visualization, we decided to merge the two into one category.

Users input a series of parameters such as their age, marital status, income, and debt. When the user clicks on a button, the speedometer needle adjusts to the relative position of the exact predicted score. However, this isn’t displayed due to privacy reasons. Instead, in the middle, a text box will display the rating.

- User inputs: age, marital status, annual income, monthly expenses, debt, number of loans in the past five years, amount of loans to pay off, loan history (have they ever had a loan), employment status, residential status, debt, and bank history (length of credit history)
- The speedometer updates dynamically to categorize loan approval chances as:
    - Poor (red)
    - Fair (yellow)
    - Good (green)
    - Excellent (dark green)
- Clicking on the gauge provides personalized recommendations to improve credit score, such as getting a beginner card for those with “Poor” rating, or consistently paying off card balances and loans in the “Fair” and “Good” rating.

Optional Features

This section outlines enhancements that would improve the project but are not critical for its success. These could include:

- Interactive elements
- Additional filtering options
- More advanced analytical tools

Initial Design #3: Feature Implementation in State 2

In State 2 (the state users reach after hovering over a state and clicking *See More!*), we aim to implement a feature that allows users to set one financial attribute as a constant and compare themselves to others with similar financial profiles.

For example, if a user has a FICO Score of 700, they can set the FICO Score filter = 700 and see:

- The Distribution of DTI Ratios for individuals with a FICO Score of 700
- The Distribution of Incomes for individuals with a FICO Score of 700

Project Schedule

This section presents a structured timeline for project completion, breaking down tasks into weekly deadlines. It also defines individual responsibilities among team members to ensure a balanced workload and prevent last-minute work rush before the final deadline.

1. Week of 3/3 to 3/7: Find credit datasets, map data, and decide on models to be used

- Suryam: Find credit dataset for the credit score calculator, decide on credit score calculator model to be used
- Ruchi: Find credit/loan dataset for the loan approval calculator, decide on loan approval calculator model to be used
- Shoubhit: Obtain average credit score, debt to income ratio, and income for each of the 50 states

2. Week of 3/10 to 3/14: Create skeleton of visualization, with sections for the main components (credit score calculator, loan approval calculator, and map of relative credit scores in the U.S.) and buttons where, once clicked, will get usable results

- Suryam: Add section for the credit score visualizer, including the wheel graph and the percentile graph
- Ruchi: Add sections for loan approval and interest rate calculator
- Shoubhit: Add sections for map and additional information that would be obtained from the optional feature

3. Week of 3/17 to 3/21: Spring break

4. Week of 3/24 to 3/28: Finish up skeleton of visualization, add input fields for the credit score calculator, loan approval calculator, and interest rate calculator, and compile data for the map into a single database for easy access

- Suryam: Add user input fields (sliders) for credit score visualization
- Ruchi: Add user input fields for the loan approval and interest rate calculator
- Shoubhit: Compile data for the map into a single datasource, either internal to the visualization or external and accessible by an API

5. Week of 3/31 to 4/4: Train models for credit score calculator and loan approval calculator, including cleaning the datasets, and generate a scale of the U.S. map on the skeleton, with a feature to highlight each individual state.

- Suryam: Train credit score calculator model and clean credit dataset
- Ruchi: Train loan approval calculator and clean loan approval dataset
- Shoubhit: Generate the U.S. map on the visualization, with all states outlined and a feature to select each individual state on hover and click

6. Week of 4/7 to 4/11: Train, test, and fine tune the models for credit score calculator and loan approval calculator, and add hover feature to the map with a popup containing credit data for each state

- Suryam: Train, test, and fine tune the credit score calculator model

- Ruchi: Train, test, and fine tune the loan approval calculator model
- Shoubhit: Add feature that when a state on the map is hovered over, a popup containing average credit score, debt-to-income ratio, and income for the state is shown

7. Week of 4/14 to 4/18: Add a legend and coloring to the map based on categorical analysis of how "good" each state's average credit is, and finish up training, testing and fine tuning models for the credit score and loan approval calculators.

- Suryam: Finish up training, testing, and fine tuning the credit score calculator model
- Ruchi: Finish up training, testing, and fine tuning the loan approval model, and create the calculation for the loan interest rate section (should not be a trained model)
- Shoubhit: Add a legend and categorical coloring to the map based on how "good" each state's average credit is

8. Week of 4/21 to 4/25: Add the optional feature of comparing user's data to a state's average data based user's input data on one of three fields (credit score, income, or debt-to-income ratio), and write the API to get model results to the visualization

- Suryam: Write part of the API to get credit score model results to the visualization
- Ruchi: Write part of the API to get loan approval model results to the visualization, and add the interest rate calculation to the visualization
- Shoubhit: Work on the optional feature of comparing user's data to a state's data based on setting one of the three fields, and analyzing comparisons of the other two fields between the user and the state average

9. Week of 4/28 to 5/2: Continue and finish work on the optional map feature, add API calls to the visualization that get models' results upon user input and clicking the skeleton's buttons, and add information to the visualization about how the models generate results based on user input, also do the screencast on 5/2

- Suryam: Add API calls to the visualization to get the credit score model's results, and add info about how the credit score model generates its results and significant features that may impact the user's credit
- Ruchi: Add API calls to the visualization to get the loan approval model's results, and add info about how the loan approval model generates its results
- Shoubhit: Continue working on the optional map feature, as written in the week of 4/21 to 4/25

10. Week of 5/5 to 5/9: Error handling, compile and submit the final visualization

- Suryam: Error handling for models and user input
- Ruchi: Error handling for models and user input
- Shoubhit: Ensure visualization as a whole works as intended, and submit the final visualization

---

Project Milestone

Overview and Motivation

Credit plays a pivotal role in shaping an individual's financial future, from securing loans and mortgages to determining access to lower interest rates and favorable lending terms. Despite its importance, credit-related information is often hidden behind confusing jargon, fragmented platforms, and opaque evaluation systems.

Our motivation for **CreditCanvas** stems from the belief that understanding one's credit standing shouldn't be a privilege reserved for experts or those with access to premium tools. Instead, we want to make credit data more transparent, accessible, and interactive by using data visualization to demystify key credit concepts and trends.

By bringing together web-based visualizations and predictive modeling, our goal is to help users:

- Visualize their credit health
- Explore regional credit trends
- Understand what factors impact loan approval

We are driven by a shared interest in the intersection of finance and technology, and by the potential to use visualization as a bridge between complex data and everyday decision-making. Whether someone is trying to improve their credit score or simply understand the factors that influence it, **CreditCanvas** aims to provide meaningful, intuitive insights.

Related Work

Our inspiration came from multiple places:

- **Existing financial dashboards and credit tracking tools** like *Experian*, *Credit Karma*, and *banking app scorecards* offer snapshots of financial standing. However, these tools often limit transparency into why a score is what it is, or what can be done to change it.

- **Financial planning apps** often focus on transactions or budgeting rather than credit health. We found a gap in the market for a one-stop, visual-first credit insight platform.

- We were also influenced by the **MyFICO score breakdowns**, but aimed to move beyond static pie charts and offer dynamic, user-personalized experiences instead.

While individual tools exist to track financial behavior, few combine predictive modeling, credit education, and regional comparison into a single, visual-first experience. That's the niche **CreditCanvas** aims to fill.

Questions

From the start, our goal was to explore:

- How do credit scores vary across regions, demographics, and income levels?
- What factors most strongly influence credit scores and loan approval?
- Are there disparities or biases in loan outcomes?
- How do debt levels, income, and financial behavior interact?

As we progressed, we refined some questions and added new considerations, based on feedback from the Project Proposal:

- Instead of having users input their exact credit score, we plan to shift toward **categorical ranges** (for example, 300–579 = Poor), both for privacy and usability.

- We also simplified our state-level comparison by retaining only **State 1** (overview map) as a core feature. The more detailed breakdown (State 2) has been moved to the optional features list based on feasibility.

These refinements didn't change our core vision, but they made the product more focused, scalable, and accessible.

Project Progress

We've been steadily progressing on **CreditCanvas**, holding weekly meetings and tracking updates collaboratively via GitHub and shared planning documents.

Key Milestones Achieved So Far:

- **Finalized Datasets**: After receiving feedback to reduce scope, we narrowed down our dataset sources and cleaned them for focused, high-impact insights.

- **Preprocessing**: We implemented data cleaning pipelines using Python (Pandas, NumPy), including:

    - Handling missing or invalid values
    - Standardizing units and categories
    - Feature scaling and encoding

    - Creating derived attributes like debt-to-income ratio

- **Visualization Setup**: The HTML and JS frontend skeleton is built and running through VS Code Live Server. We've created distinct sections for:
    - The credit score visualizer and wheel
    - Loan approval prediction and interest rate comparison
    - State-level credit comparison

- **ML Model Training**: We built two machine learning models using Scikit-learn:
    - A **credit score classifier** based on user attributes
    - A **loan approval predictor** using applicant financials Models are trained, tested, and integrated with mock API endpoints for frontend testing.

Data

Revised Datasets and Sources

After revising our scope, we selected the following focused sources:

1. State-Level Financial Data

Compiled from multiple web sources, cleaned and joined manually into a single CSV.

- Average Credit Score and Credit Card Debt by State – Investopedia
- Per Capita Income by State (2021) – FRED

We computed derived metrics like:

- **Debt-to-Income Ratio**
- **Categorical credit score buckets (for example, Good, Fair, Poor)**

2. Loan Approval Data – Kaggle

- Loan Approval Dataset

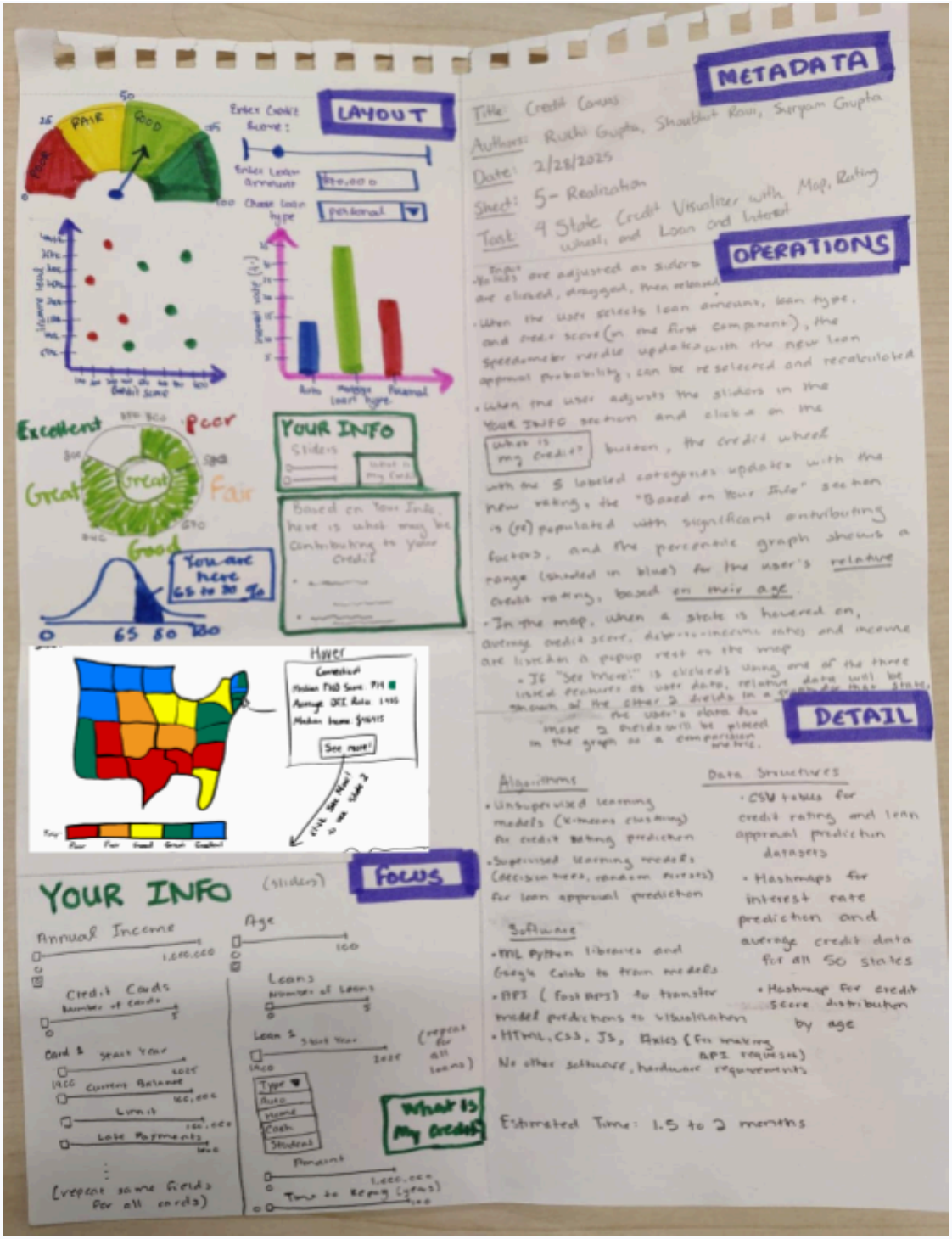
We split this dataset into two clean, labeled subsets:

- One for **loan approval prediction**
- One for **credit score prediction**

Preprocessing Steps

Based on our proposal plan, we implemented:

- Imputation of missing values where appropriate (for example, median income)
- Removal of corrupted entries (for example, negative loan amounts)
- Standardization of categorical variables (for example, employment status, loan type)
- Scaling of numerical features (income, score, balance)
- Feature creation such as computing credit utilization ratio, binning scores, tagging outliers
- Balancing for classification to avoid overfitting to approved loans



This is the final design document we have chosen to implement for the website.

Code Explanation

`CreditCanvas/data/`

This folder contains all finalized datasets used in the project. These datasets have been cleaned, preprocessed, and are ready for use in both visualization and model integration.

Files:

- `Combined_State_Financial_Profile.csv`
  Final merged file with state-level data, including average credit score, credit card debt, per capita income, and debt-to-income ratio. This dataset is the backbone of the US map visualization.
- `Loan Dataset.csv`
  The original raw dataset downloaded from Kaggle, containing user-level features relevant to both loan approval and credit scoring tasks.
- `Loan_Cleaned_Data.csv`
  A cleaned and filtered version of the loan dataset, used directly in machine learning model training and as a source for sampled scatterplot data.

These files represent the final outcome of our preprocessing pipeline, optimized for performance and structure in downstream analytics and visual interfaces.

`CreditCanvas/data_processing/`

This folder contains all scripts used to clean, transform, and structure our raw data sources. It is organized into subfolders for modularity and clarity:

- `state_data/`: Cleans and combines Investopedia and FRED data for state-level comparisons
- `loan_data/`: Preprocesses the Kaggle loan dataset into usable formats
- `credit_data/`: Constructs and filters features needed for credit score regression tasks

Example: `state_data/merge_and_clean_state_financial_data.py`

This script:

- Reads two separate CSVs containing average FICO scores and credit card debt, and per capita income.
- Renames columns and merges them on the common "State" column.
- Computes a derived metric, debt-to-income ratio, and reorders columns for consistency.
- Outputs the final `Combined_State_Financial_Profile.csv` file, which supports our US map visualization.

Example: `loan_data/preprocess_loan_data.py`

This script:

- Removes duplicates and handles any missing values.
- Filters for only the most relevant attributes (such as age, income, credit score).
- Converts categorical variables to lowercase for consistency.
- Saves the final cleaned dataset as `Loan_Cleaned_Data.csv`, used in both training and live prediction.

Similarly, the credit data preprocessing script processes numeric and categorical features, ensuring the model receives a clean, structured input format.

`CreditCanvas/assets/`

This folder contains the static data needed for rendering geographical visualizations.

- **`us-states.json`**: A TopoJSON file that contains the geometry definitions for all 50 U.S. states. This is essential for drawing the state map using D3. It enables interactive data binding by linking each shape with its corresponding financial metrics via the "State" property.

`CreditCanvas/docs/`

This folder is used to organize all documents related to the project's lifecycle:

- Initial project proposal
- Milestone writeups
- Final process book
- Design iteration notes and diagrams

Having all key documentation in one place ensures clarity during submission, presentation, and peer reviews.

`CreditCanvas/js/`

This directory contains the interactive JavaScript logic that drives the dynamic features of the website. The core of this code lives in `main.js`.

`main.js`: Functionality and Interactive Data Structures

The `main.js` file contains all front-end logic tied to interactivity, D3 rendering, and form behavior. It is modular and sectioned as follows:

1. US State Financial Map

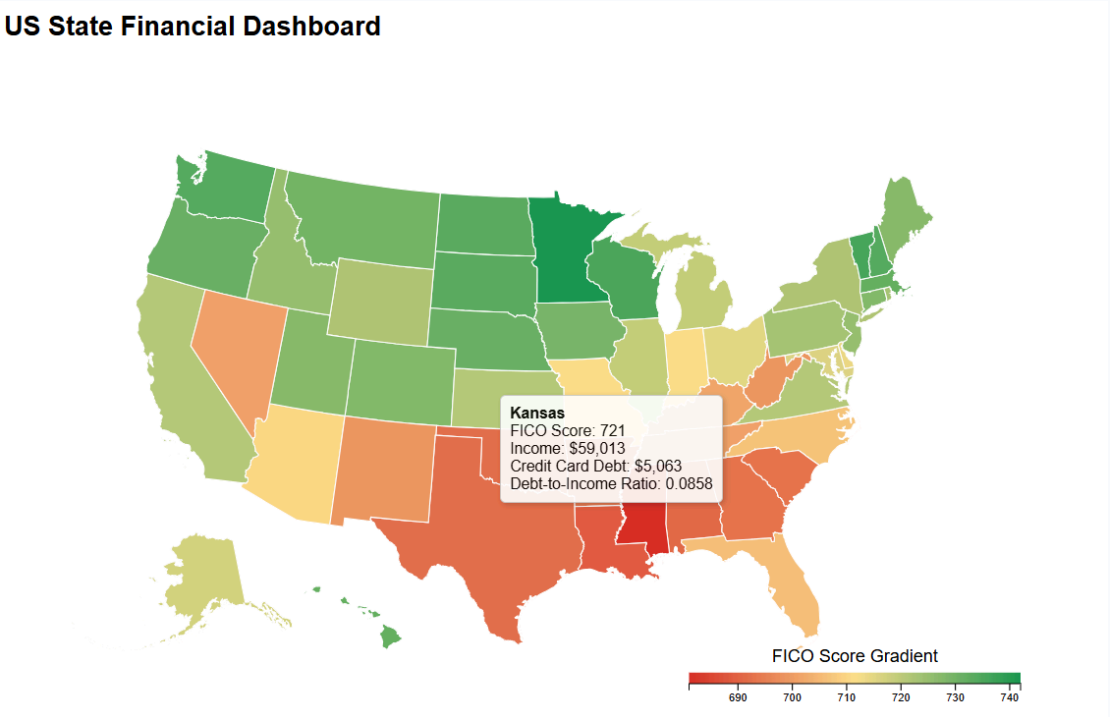
This section of the code creates a choropleth map of the United States using D3 and TopoJSON. Each state is filled with a color that corresponds to its **average FICO score**, scaled using a linear gradient from red (low scores) to green (high scores).

Key implementation details:

- Uses a `geoAlbersUsa()` projection and path generator to render state boundaries.
- Loads `us-states.json` for geometry and `Combined_State_Financial_Profile.csv` for financial values.
- Maps each state to its financial data using a `stateFinancialDataMap` object, enabling efficient data lookup.
- Includes a gradient legend to help users understand score ranges across the country.

- **Interactivity**:
  Hovering over any state triggers a tooltip that displays:
  - State name
  - Average FICO score
  - Average income
  - Credit card debt
  - Debt-to-income ratio

This provides instant, location-specific insight, making the map highly informative and user-friendly.



2. Loan Approval Predictor with Speedometer Gauge

This section captures form inputs filled by the user related to age, income, debt, and marital status. The values are structured into a JSON object and sent via a `POST` request to the Flask endpoint `/predict`.

Backend response:

- The backend returns a **predicted loan approval probability**, calculated using a logistic regression model trained on the cleaned Kaggle dataset.

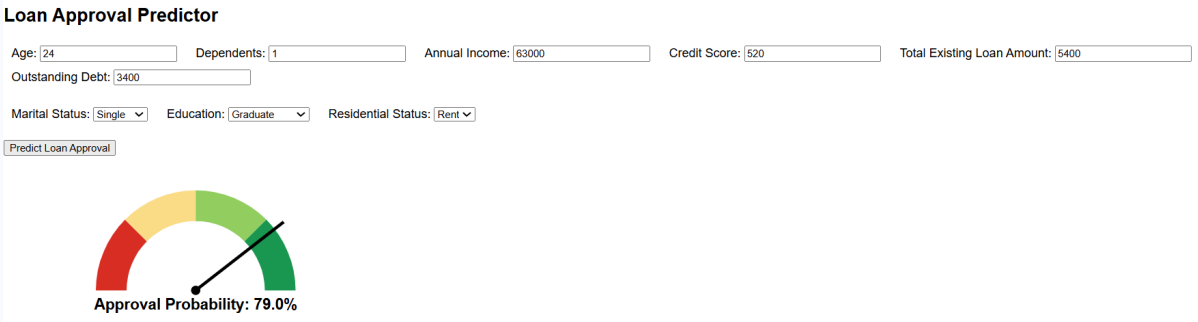
Frontend visualization:

- This prediction score is displayed using a **speedometer-style D3 gauge**, which is both visually engaging and informative.

Gauge features:

- Divided into 4 labeled regions: **Poor, Fair, Good, Excellent**
- The needle rotates based on the score returned by the model
- On-hover tooltips offer **personalized tips** for each range, such as suggestions to reduce debt or maintain financial consistency.

This interaction transforms complex model outputs into easy-to-understand visual feedback, creating a seamless and educational experience for the user.



3. Scatterplot: Credit Score vs Income Visualization

This section is designed to let users explore how credit score correlates with income and loan approval status.

Functionality:

- On clicking the **"Generate Random Dataset"** button, the frontend sends a `GET` request to the `/scatter-sample` backend endpoint.
- The backend returns 30 randomly selected entries from the loan dataset, each containing a credit score, income, and loan approval outcome.

Rendering:

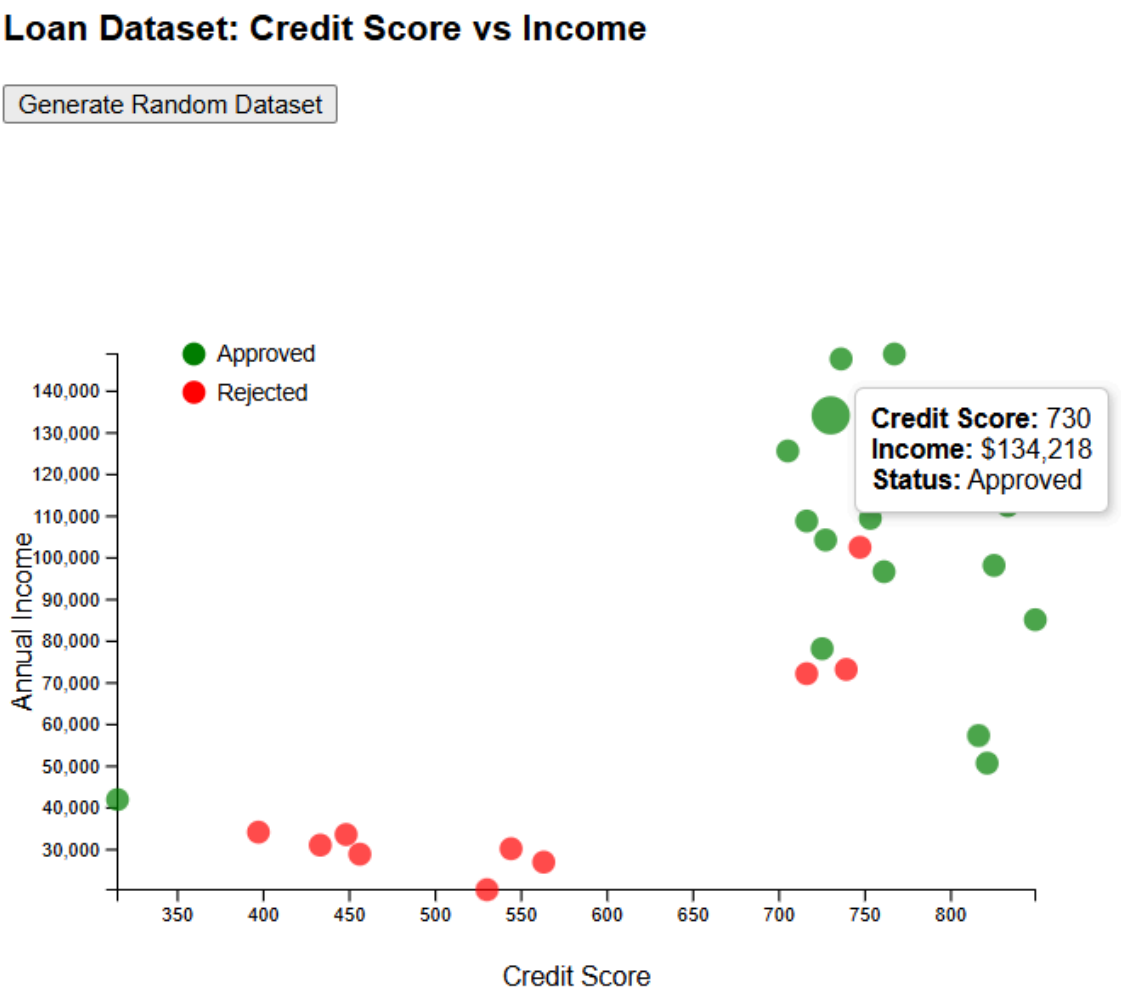
- Each point is plotted using D3 with:
    - X-axis: Credit Score
    - Y-axis: Annual Income
    - Color: Green for approved, Red for rejected

Axes are scaled dynamically based on data ranges returned by the backend.

Interactivity:

- **Hovering over any data point shows a tooltip** with detailed information about that user:
    - Exact credit score
    - Income
    - Loan approval decision

This interactive, user-driven scatterplot allows users to visually analyze patterns and correlations between variables, making data exploration intuitive and engaging.

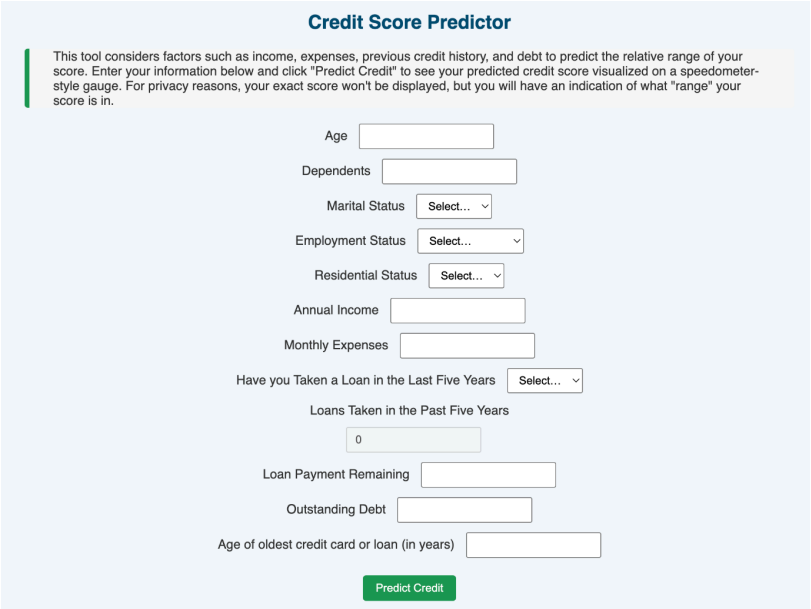


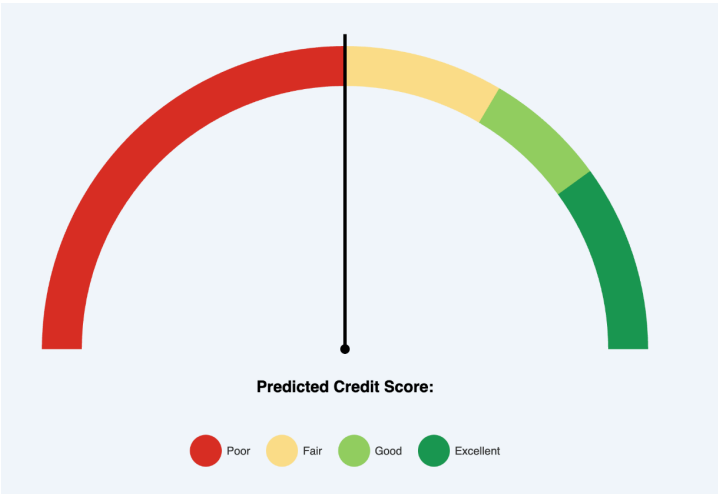
4. Credit Prediction with Speedometer Gauge

This section is meant to help users understand factors that contribute to credit score and a perspective on their own credit profile, without breaching privacy.

User values are first input in the frontend's main page. Then, when "Predict Credit" is pressed, these are condensed into a JSON object as the payload and sent as a POST request to the **/credit** endpoint. This will then access the pretrained credit score model and run the new payload as a prediction test point. The predicted score is returned from this endpoint and used to move the speedometer needle in the credit prediction visualization.

The needle rotates based on the input parameters, and the "Predict Credit" button can be pressed again to get the updated results. The user can also hover on top of any of the four labeled regions ("Poor", "Fair", "Good", "Excellent") and see feedback on how to improve their credit.





`CreditCanvas/index.html`

The `index.html` file provides the structural layout and form elements for the application. It contains:

- The main **U.S. state map container** (`#map`)
- A full **loan approval form**, complete with dropdowns and numeric fields
- A **gauge container** for displaying the predicted loan approval probability
- A **scatterplot area** and generate button
- A second **form for credit score prediction**, capturing financial behavior details like income, expenses, loan amount, and more

Styling and layout:

- Tooltip styling is defined inline for responsiveness
- Elements are spaced using simple margin and padding rules
- The entire layout is optimized for modularity, allowing future visualizations or form components to be plugged in with ease

Scripts are loaded using CDN links to D3 and TopoJSON, with `main.js` loaded at the bottom to bind functionality to HTML elements.

`CreditCanvas/models/`

This folder houses all ML-related artifacts and backend logic used for model training, prediction, and inference.

Backend Server: Flask (`app.py`)

The Flask application is the backend engine that powers all ML predictions and data delivery.

- **Flask App Initialization**
  The server is initialized with `Flask(__name__)`, and `CORS()` is enabled to allow JavaScript requests from other ports.
- **Route `/credit`**

  Accepts user form data, rescales it to **look like the random forest regression model**, and sends it to the random forest regression model, which returns the credit_score back as a decimal number.

- **Route `/predict`**
  Accepts user form data, applies trained label encoders, and sends the features into the **logistic regression model**. It returns the **approval probability** in JSON format, which is rendered as a speedometer gauge on the frontend.

- **Route `/scatter-sample`**
  Returns a JSON array of 30 randomly sampled data points from `Loan_Cleaned_Data.csv`. These are used in the scatterplot to help users understand approval patterns across income and credit score levels.
- **Route `/predict-credit-score`**
  Accepts another set of inputs for a separate model that predicts the user's **credit score** using a random forest regressor. The predicted score is displayed in the credit score prediction section of the webpage.

All models are loaded from `.pkl` files using `joblib`, and the Flask server runs locally in debug mode.

Training Scripts

- **`loan_prediction_model.py`**
  - Trains a logistic regression model on the loan dataset
  - Encodes string fields using `LabelEncoder` and saves them for future prediction
  - Reports accuracy, ROC AUC, and classification metrics

- **`credit_score.py`**
  - Trains a random forest model to predict credit score as a regression task
  - Evaluates performance with MAE
  - Saves the trained model for use in the Flask app

Together, this stack powers an **interactive, data-driven credit analysis website** that allows users to:

- Visually explore state-level financial health
- Submit personal data to see model-driven loan predictions
- Understand financial trends via interactive plots and tooltips

---

How to Run the Project

Required Dependencies

To set up the backend environment, make sure the following Python libraries are installed:

- flask
- flask-cors
- pandas
- scikit-learn
- Joblib
- Xgboost
- pickle

You can install them using pip:

pip install flask
pip install flask-cors
pip install pandas
pip install scikit-learn
pip install joblib

Pip install xgboost

Pip install pickle

OR you can run `pip install -r requirements.txt` to install them all at once, since they are all in the requirements.txt file

---

Step-by-Step Instructions

1. **Clone the repository and navigate into the project folder:**

   ```
   git clone https://github.com/RuchiGupta20/CreditCanvas.git
   cd creditcanvas
   ```

2. **Start the Flask backend:**

   ```
   cd models
   python app.py
   ```

   This will launch the Flask development server on http://localhost:5000 and enable three backend endpoints:

   - /predict
   - /scatter-sample
   - /predict-credit-score

3. **Open the frontend:**

   Go back to the root directory and locate the index.html file.
   Use the **Live Server** extension in VS Code to launch it:
   Right-click on index.html and choose "Open with Live Server."

Interactive Features Walkthrough

- **State Map Visualization**
  Hover over any state to view average FICO score, income, credit card debt, and debt-to-income ratio. The tooltip updates in real time, making the map highly

informative and interactive.

- **Loan Approval Prediction (Speedometer Gauge)**
  Fill out the loan prediction form with user attributes like age, income, credit score, loan amount, etc. Click the "Predict Loan Approval" button to receive a prediction. The result is shown on a color-coded speedometer with a pointer and tooltip-based advice.

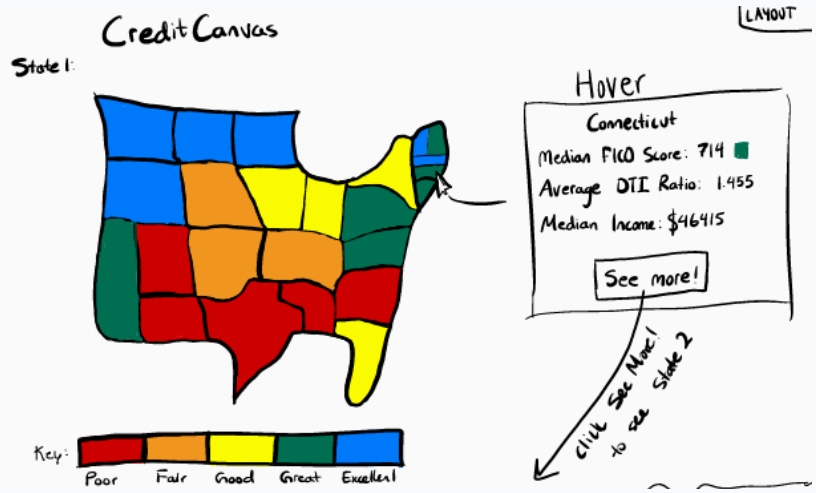
- **Scatterplot Generation**
  Click the "Generate Random Dataset" button to load 30 random entries from the loan dataset.Dots represent users (green = approved, red = rejected). Hovering over a dot shows details like credit score, income, and approval status.

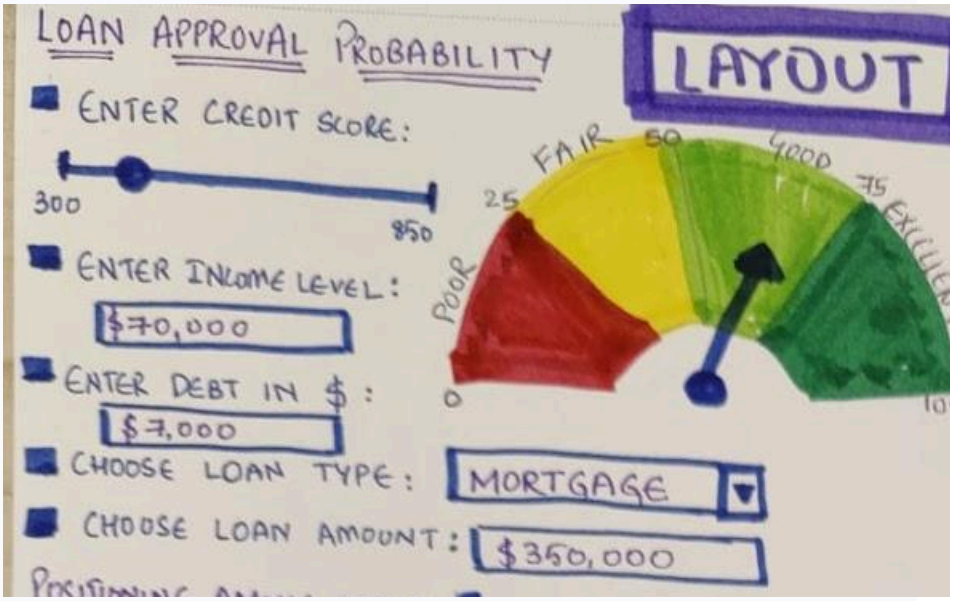
Meeting Schedule and Notes

- 3/7 meeting:
    - We discussed datasets and decided on a maximum of two datasets to use for credit approval and loan approval model training, and the sources for statewide credit data. The initial expectation was that credit approval and loan approval would have separate datasets, as from first glance, we thought the datasets were aimed towards two separate fields (all credit versus loans).
        - Credit approval dataset: https://www.kaggle.com/datasets/parisrohan/credit-score-classification
        - Loan approval dataset: https://www.kaggle.com/datasets/architsharma01/loan-approval-prediction-dataset
    - For credit data by state, we combined two separate datasets:
        - FICO and credit card debt data by state: https://www.investopedia.com/average-credit-scores-by-state-5105100
        - State income data (2021):https://fred.stlouisfed.org/release/tables?rid=110&eid=257197&od=2021-01-01#

We did recognize that the state income data may be inaccurate, since it is four years outdated. However, we didn't address it immediately, as our main priorities were to start data cleaning, model training, and website skeleton building.

When time permits, we may be able to project the state income data from 2021 to 2025 by adding four years worth of compounded interest to all of the 2021 incomes. This won't be the most accurate solution, but if we aren't available to find more recent state income data (which at the moment seems unrealistic, after having already searched for quality free data), this is a workable temporary solution.

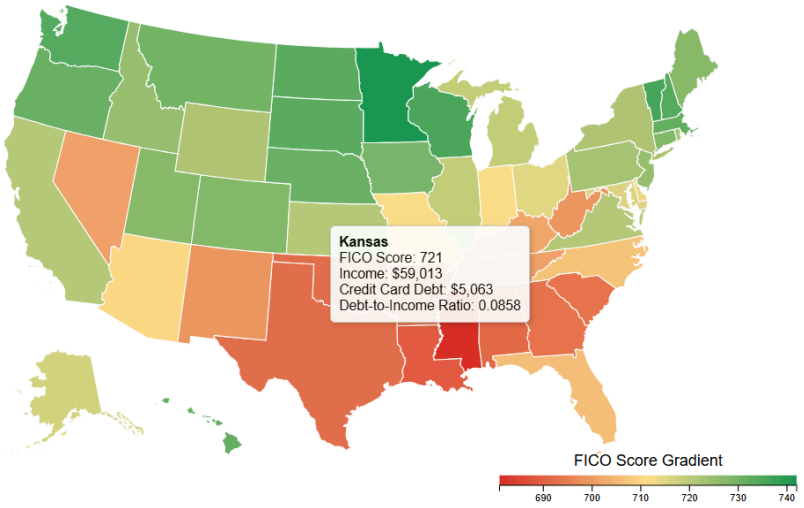
- 3/14 meeting: We worked almost exclusively on the front end skeleton. Right now, we focused on getting a minimalistic design, with all of the user fields accessible and the base outlines of the initial designs.
    - Initial designs (from the project proposal):
        - State data
        - 

        - Loan approval wheel
        - 

        - Credit approval wheel
        - 

    - Initial implementations (from the website)

US State Financial Dashboard

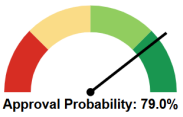


Loan Approval Predictor

Age: 24 Dependents: 1 Annual Income: 63000 Credit Score: 520 Total Existing Loan Amount: 5400
Outstanding Debt: 3400

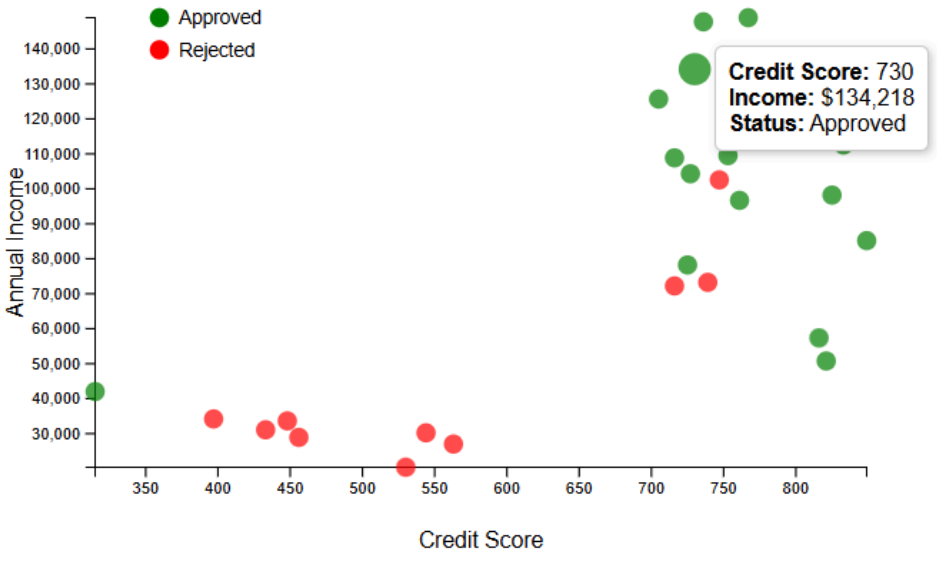
Marital Status: Single Education: Graduate Residential Status: Rent

Predict Loan Approval



Loan Dataset: Credit Score vs Income

Generate Random Dataset



- 4/4 meeting:
    - We narrowed down the number of datasets being used for all training to one. The loan approval dataset would also be used for credit approval.
        - As training occurred, it became clear that the credit approval dataset chosen had too many fields that were not only hard to obtain user information for (as in, a user would not be able to quickly identify a value for the field), there were so many fields during model training that results were slightly inaccurate.
        - We discussed that high dimensional data was necessary and useful for making an accurate prediction, but having too many fields can mess up predictions, especially if the fields don't provide any additional insight to making predictions about credit approval.
    - We also debriefed results for the loan approval model and created the API for the website frontend to access the model's results and make predictions on user-input data points. The loan approval component of the web page updates whenever new user data is input, and a form is submitted.
- 4/10 meeting:
    - This week was spent improving the frontend interface and getting an idea for credit approval models.
    - Frontend improvements
        - Added interactivity to the map: Whenever a state on the map is hovered on, the state changes color and there is a note next to the state with exact average FICO, income, and debt. This is more useful than the original draft of the website, which just had the outline of the map.
            - States are also shaded in a color hue scale based on average FICO score (red is poor, fair is orange, good is green, etc.)
        - Animated loan approval wheel: This is best seen when there have been previous data and predictions made. Whenever new data is input, and the web form is submitted, once the API returns the new loan approval prediction, there is a needle on the wheel that will move to the new prediction.
    - Credit approval model
        - The driving factor behind what model is chosen depends on the desired output. Do we want an exact credit score prediction, or do we just want a relative range (i.e. Poor, Fair, Good, Very Good, or Excellent). This is given that our dataset has a credit score value for every data point.
            - If we need the exact credit score prediction, we discussed that we can use logistic regression with the credit score field as labels, and ensure the dataset is quantified and scaled appropriately beforehand
            - If we need a relative credit rating and not an exact number, we discussed that we could convert the quantitative credit scores into categorical format and run a logistic regression or random forest algorithm, or we could ignore the labels altogether and run a k-means clustering where k = 5. The k-means model would cluster every user into one of 5 categories, which could

then be mapped to a categorical credit rating based on the average credit rating of the cluster.
- Originally, a draft was made using k-means clustering. However, since clusters are randomly labeled when the algorithm reruns, it is a bit more time consuming to keep track of which clusters represent poor/fair/good credit, etc.
- Since there is a lot of data available (~52,000 data points), using logistic regression to predict a more exact credit may be more useful to a user than just a credit range, so we will most likely choose logistic regression going forward.
- Going forward, we will also work on updating the UI of the website to seam the different features together into a single comprehensive package

Final Project Process Book Updates

Overview and Motivation

CreditCanvas began with a simple but powerful motivation: make credit information easier to understand and access for everyone. In today's financial world, credit scores impact nearly every major life decision, from buying a home to applying for a student loan. Yet most people don’t fully understand how credit scores are calculated, what influences loan approvals, or how their financial standing compares to others.

We noticed that while financial literacy is becoming more important, most existing tools are either behind a paywall or too technical to be user-friendly. Platforms like Credit Karma or FICO dashboards show numbers, but they rarely explain *why* those numbers matter or how they are derived. We wanted to go beyond static scorecards and offer something that was interactive, visual, and grounded in real data.

With our shared interest in both finance and data visualization, we saw an opportunity to bring those worlds together. Our goal was to create a tool that helps people explore their credit health, understand regional trends, and make better-informed decisions, all in one place. Through engaging visualizations, predictive modeling, and clear design, we want users to walk away feeling more confident and in control of their financial future.

Related Work

We drew inspiration from several real-world tools and concepts when building CreditCanvas.

- **Credit tracking apps** like Credit Karma and Experian provided the foundation. These apps are helpful but often act like black boxes, offering little clarity on how scores are calculated or how users compare to others. We wanted to make that process more transparent.

- **Financial planning apps** like Mint or YNAB focus heavily on budgeting and transactions but often skip over credit health entirely. We saw a gap where credit-related visuals and recommendations could be integrated in a more meaningful way.

- **MyFICO's score breakdowns** inspired us to visualize the weight of different factors in an interactive format, rather than a static pie chart.

In class, we explored several principles around effective data encoding, color usage, interactivity, and narrative storytelling. These helped shape the way we structured our dashboard. For example, we applied perceptual guidelines when designing our red-to-green color scales for risk levels, and we used tooltips and sliders to enhance interaction without overwhelming the user.

Questions

At the beginning of the project, our guiding questions were:

- How do credit scores vary by geography, income, and demographics?
- What features most strongly influence credit scores and loan approvals?
- Are there biases or disparities in credit-related decisions?
- How do debt levels and financial behavior relate to credit risk?

As the project evolved, so did our questions. Based on peer and instructor feedback, we refined our focus:

- We shifted from asking for exact credit scores to using score ranges. This helped maintain user privacy while still allowing predictive modeling.

- Our original two-level state comparison map was streamlined into one overview map with hover-based tooltips. The more detailed view became optional, making the interface cleaner and more approachable.

- We started exploring more user-focused questions, like: "Where do I stand compared to others?" and "What can I change to improve my approval odds?"

These changes kept our project grounded in real user needs, while also making the product more scalable and intuitive.

Data

We sourced data from a mix of trusted financial and government platforms, primarily Kaggle, FRED, and Investopedia.

State-Level Data:

- Average FICO scores and credit card debt (Investopedia)
- Per capita income by state (FRED)

We cleaned and merged these into a single CSV file, from which we computed metrics like:

- Debt-to-income ratio
- Categorical score buckets (Poor, Fair, Good)

Loan Approval and Credit Score Data:

- Kaggle's Loan Approval Prediction dataset was our core source
- We split this into two cleaned subsets for training: one focused on loan approvals, the other on credit scoring

Cleaning Process:

- Used Python with Pandas, NumPy, and Scikit-learn
- Imputed missing values (median income, etc.)
- Removed corrupted records (like negative loan amounts)
- Standardized categorical fields (employment status, marital status)
- Scaled numerical features and tagged outliers

These curated datasets now support both our models and interactive visualizations.

Exploratory Data Analysis

We began by examining our data with traditional Python visualizations: Seaborn histograms, Matplotlib boxplots, and pairplots.

Some initial findings:

- A strong correlation between income and approval, but with clear exceptions
- Debt-to-income ratio had high predictive power for approvals
- Credit scores varied widely even among similar income brackets

This analysis helped shape both our model features and visualizations. We dropped low-informational fields and added derived metrics like credit utilization ratio.

From a design brainstorming standpoint, our early sketches were more ambitious than practical. We initially wanted:

- A full “credit wheel” prediction tool
- A state-level breakdown with comparisons by credit score, income, and debt
- Multiple views for FICO score distributions and user percentile ranges

Eventually, we narrowed this down to three core views that complemented one another: the state map, the approval predictor, and the scatterplot. These were chosen for their unique angles on the data and their combined power to support both personal reflection and macro-level insight.

Design Evolution

In early sketches, we imagined a dozen disconnected visualizations. But we quickly realized that overlapping information, cluttered layouts, and repeated user inputs made for a confusing experience. So we stepped back and re-centered the design on cohesion and clarity.

We ultimately committed to three core visualizations:

1. **The U.S. State Map**
   This visualization was built using D3 and TopoJSON. Each state is filled with a color from a red-to-green gradient representing average FICO scores. Hovering over a state reveals:
   - Average income
   - Average credit card debt
   - Debt-to-income ratio
     The gradient legend at the bottom helps users interpret the range, and the entire map is rendered using the geoAlbersUSA projection. This became our anchor for regional comparison.

2. **Loan Approval Speedometer**
   The second view centers on user inputs for age, income, credit score range, existing debt, and marital status. These values are sent to a Flask backend which runs a logistic regression model trained on Kaggle data. The output is visualized as a dynamic D3 speedometer gauge with color-coded zones:
   - Poor (red)
   - Fair (yellow)
   - Good (light green)
   - Excellent (dark green)
     Hovering over each segment reveals additional recommendations (e.g.,

"Reduce debt to increase chances"). This visualization emphasizes real-time feedback and ease of interpretation.

3. **Scatterplot of Credit Score vs Income**
   Our third major view helps users explore loan approval trends among real applicants. Users can click “Generate Random Dataset” to load 30 records from the Kaggle dataset. Each dot represents an applicant:

    - X-axis: Credit score
    - Y-axis: Income
    - Color: Green (approved), Red (rejected)
      Hovering over a point reveals the applicant’s exact income, credit score, and approval result. Once clicked, the point shrinks slightly to indicate it has been explored.

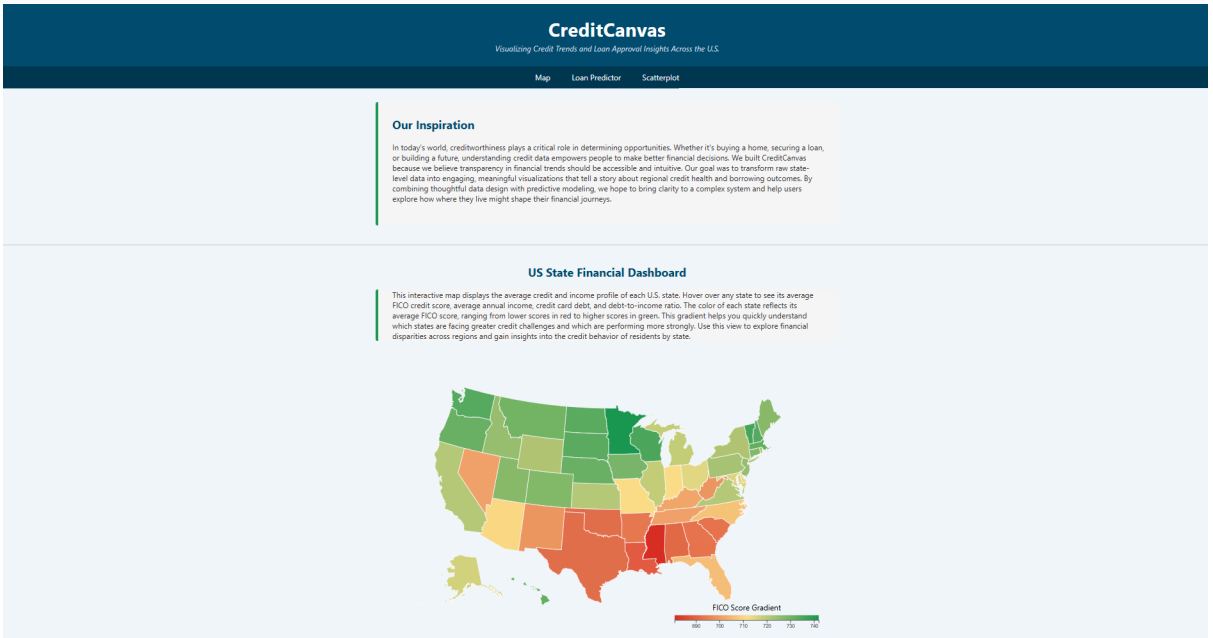
4. **Credit Prediction Speedometer**
   The final visualization focuses on predicting a user's credit score based on financial and demographic inputs. Unlike traditional tools that merely display a score, our visualization offers an interactive credit wheel (built with D3) that categorizes the user's credit health into four bands:
    - Poor (red)
    - Fair (yellow)
    - Good (light green)
    - Excellent (dark green)

   Users enter information such as age, income, debt, loan history, and account tenure. These inputs are sent to a Flask backend, where a machine learning model (Random Forest Regressor) estimates their credit score. The result is shown as a rotating needle over the wheel, accompanied by a textual label (e.g., “Good”).

   Hovering over each colored region reveals actionable credit improvement tips based on the predicted score bracket. While the exact number is not shown to protect sensitive estimations, the visual range and feedback give users meaningful insight into their standing and steps they can take to improve it.

By paring down our design to these three visualizations, we avoided duplication, made inputs more streamlined, and gave users three distinct yet complementary views into credit and approval trends. The final version feels cohesive and intuitive, with interaction and guidance built into each view.

Implementation



We designed CreditCanvas with three main interactive visualizations, each serving a unique purpose while working together to tell a cohesive story about credit trends and loan approval. At the core of our design is a clean, navigable interface with a top navigation bar

that anchors the experience. The nav bar includes links to each of the three primary views: **Map**, **Loan Predictor**, and **Scatterplot**, making it easy for users to move between insights without getting lost.

1. U.S. State Financial Dashboard

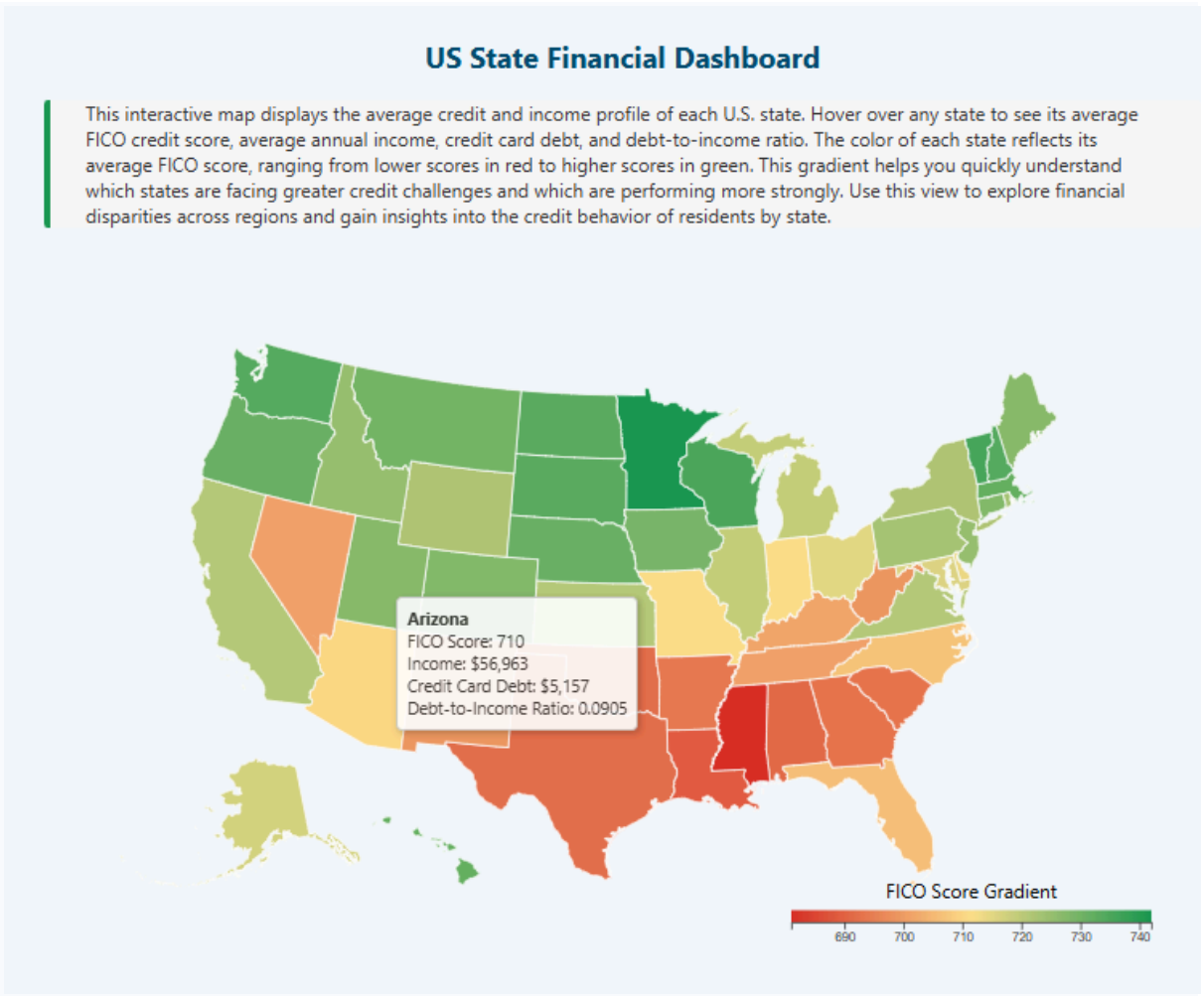
This view provides a big-picture look at the average financial profile of each U.S. state. Implemented using D3.js and a **TopoJSON file of U.S. states**, the map uses a **geoAlbersUSA projection** to render state boundaries accurately. Each state is color-coded using a **red-to-green gradient**, where red indicates lower average FICO scores and green represents higher scores. The gradient legend at the bottom of the map helps users interpret score ranges with clarity.

Hovering over any state reveals a tooltip with:

- Average FICO credit score
- Average income
- Average credit card debt
- Debt-to-income (DTI) ratio

This interactivity helps users explore regional disparities and raise questions like “Why is average debt higher in this state?” or “How does my state compare to neighboring ones?” All financial metrics are derived from merged datasets (Investopedia and FRED), preprocessed and unified using Python scripts. We also carefully chose our color scale to ensure perceptual clarity, using a colorblind-friendly palette where green hues subtly intensify with better scores.

This map isn’t just decorative, it’s where many users begin their journey with CreditCanvas, offering a grounded understanding of regional credit health before they dive into personal predictors.

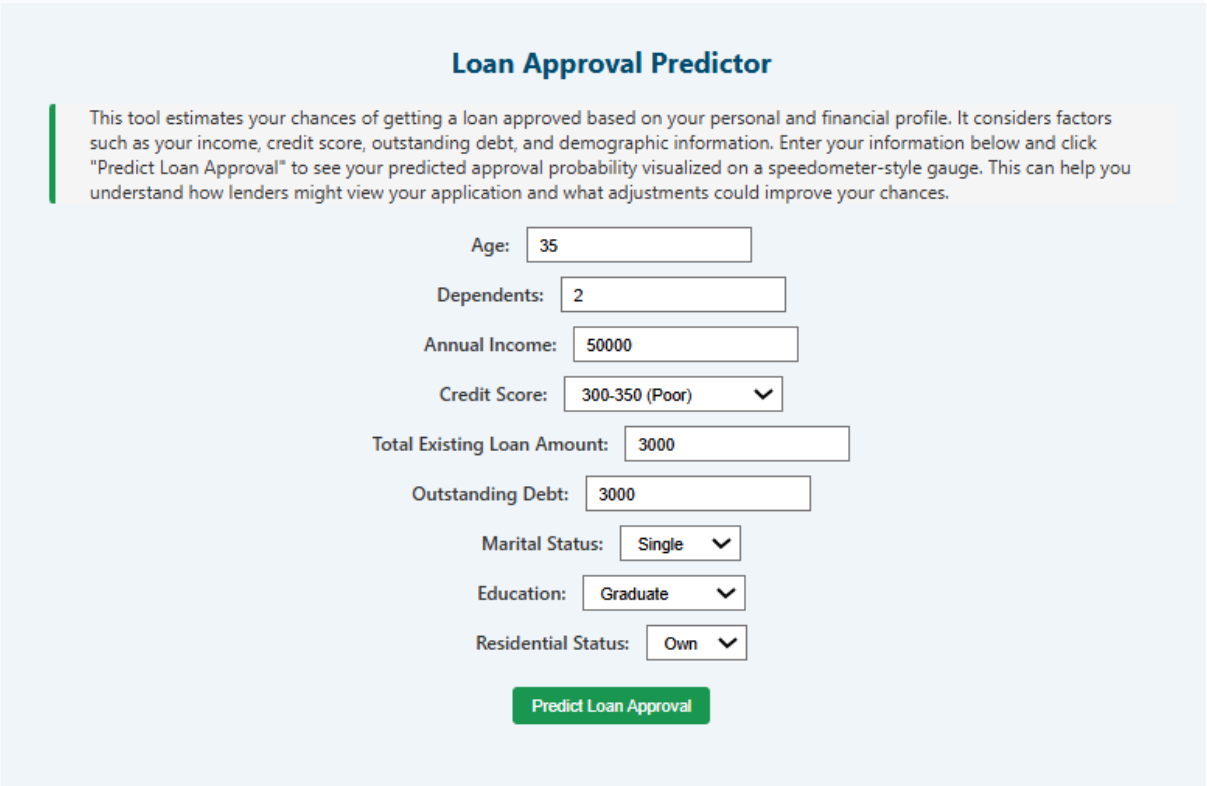


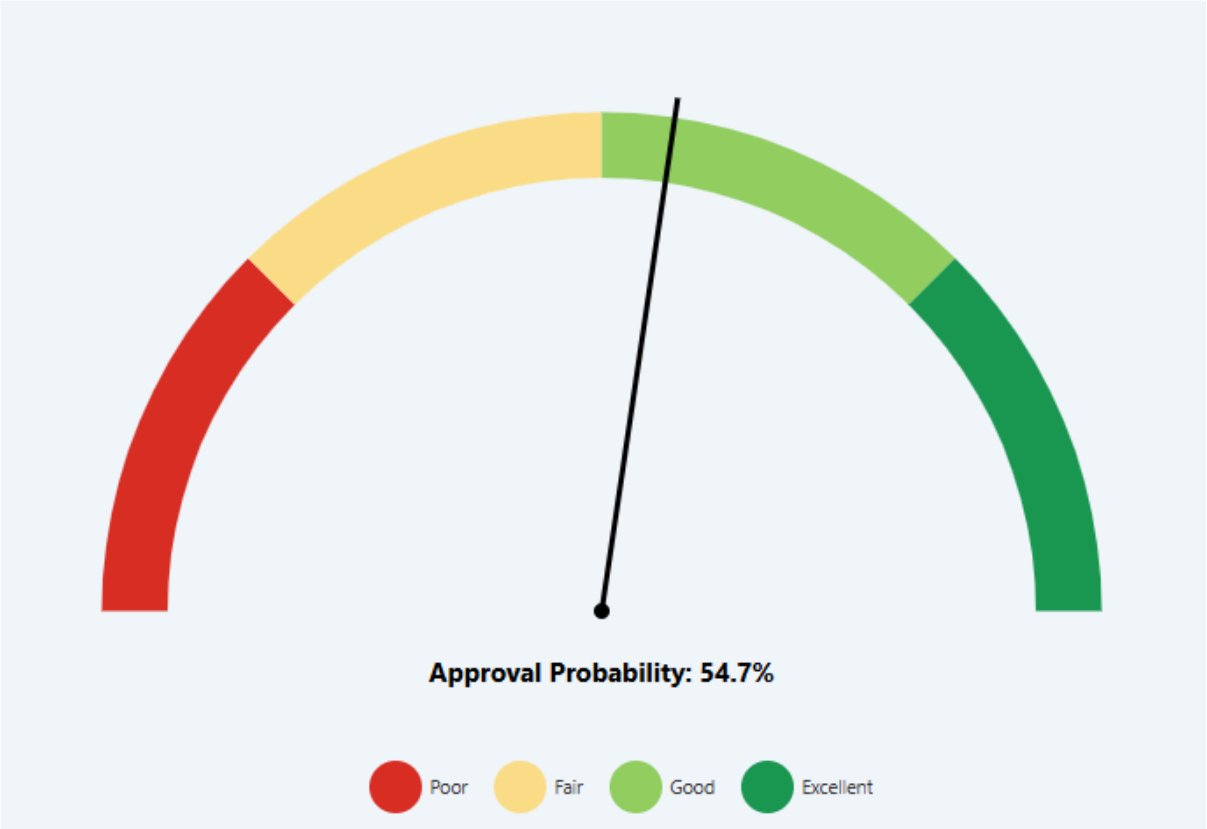
2. Loan Approval Predictor (Credit Score Speedometer)

This visualization allows users to estimate their loan approval probability based on inputted personal and financial data. The form fields include:

- Age
- Number of dependents
- Annual income
- Credit score (as a range, e.g., 300–350)
- Existing loan amount
- Outstanding debt
- Marital status
- Education level
- Residential status

For **privacy and usability reasons**, users select their credit score from categorical ranges rather than entering a specific value. On the backend, we use the **midpoint of the selected range** to pass into the model. The model itself is a **logistic regression classifier**, trained on a cleaned Kaggle loan approval dataset.



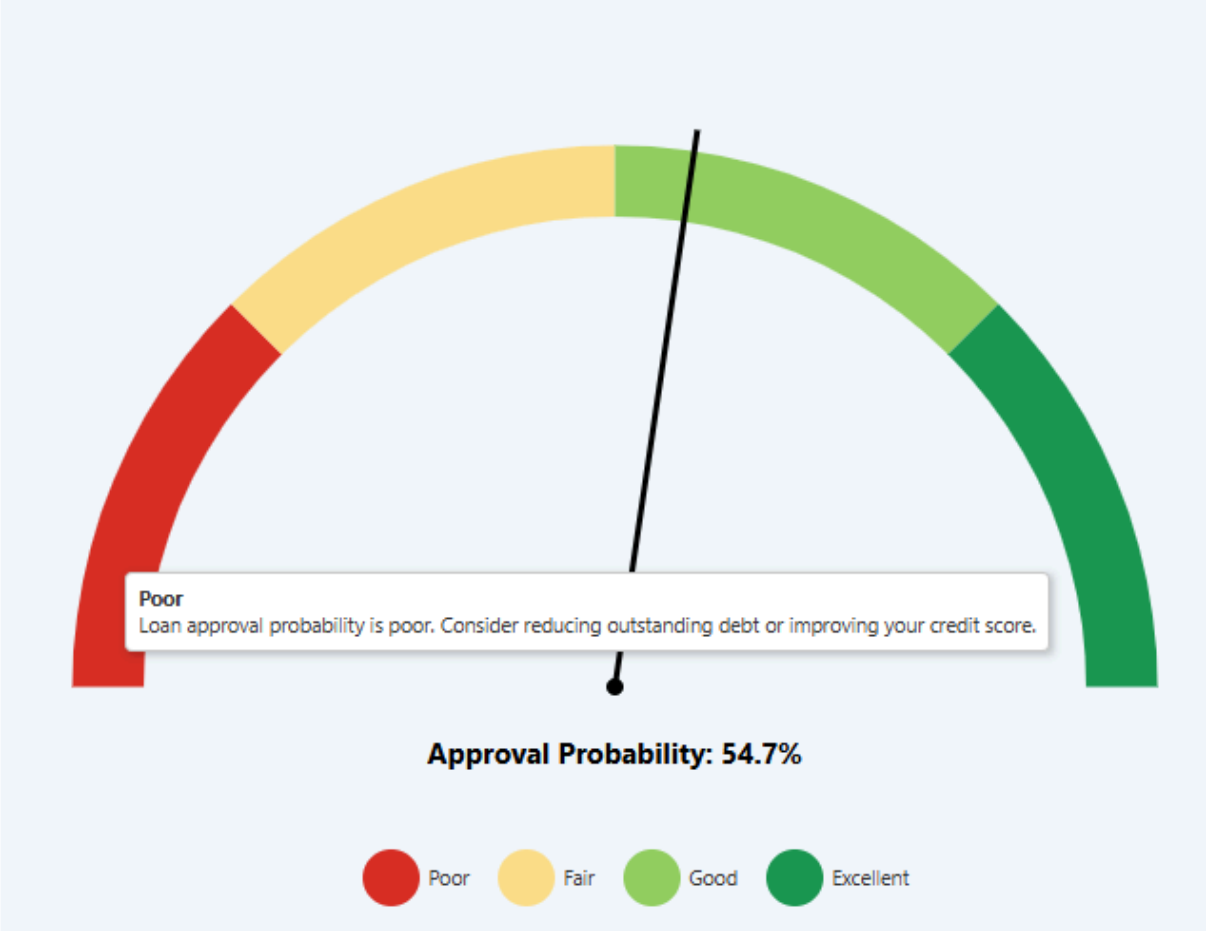


Upon clicking “Predict Loan Approval,” the data is sent to a **Flask endpoint** (`/predict`) via a POST request. The backend returns a probability score, which is visualized using a **D3-powered speedometer gauge**. The gauge is split into four color-coded regions: red

(poor), yellow (fair), light green (good), and dark green (excellent), providing an intuitive view of the result. The needle animates to its correct position, and tooltips offer recommendations for improving approval chances (e.g., reducing debt or increasing income).

To prevent user error, we implemented input clamping and validation: credit scores are restricted to standard ranges (300–850), and fields like income and debt cannot be negative.

We also added annotations and legends directly under the gauge to help users interpret the result, explaining what each color means and how the prediction was calculated.



3. Scatterplot: Credit Score vs. Income

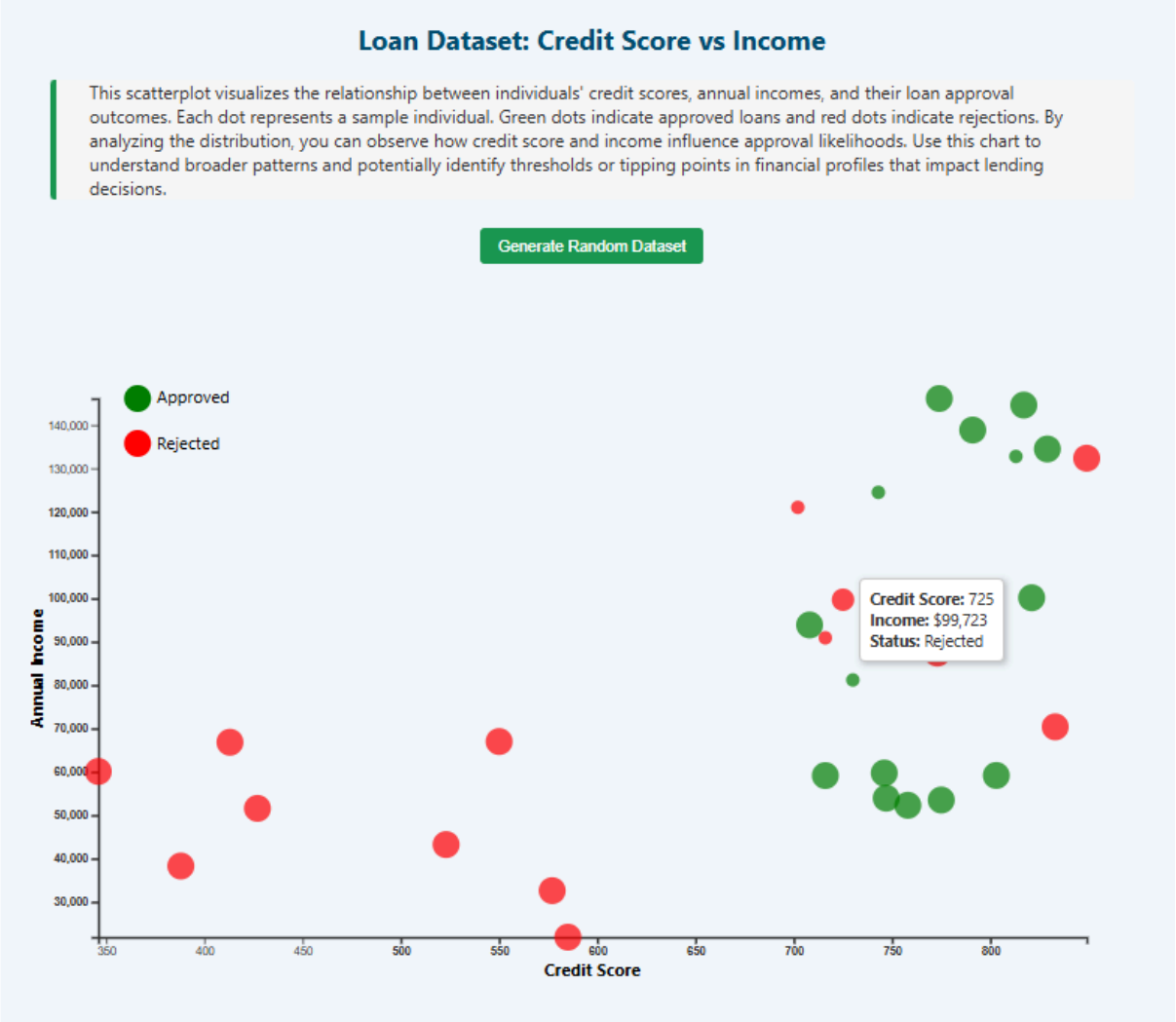
This scatterplot visualizes how **credit score and income correlate with loan approval outcomes**. When users click “Generate Random Dataset,” a GET request is made to the `/scatter-sample` Flask route, which returns 30 random records from the cleaned loan dataset. Each point on the plot represents an individual applicant.

- **X-axis**: Credit Score
- **Y-axis**: Annual Income
- **Color**: Green (approved), Red (rejected)
- **Size**: Uniform for now, but potential for expansion based on loan amount or debt

Hovering over a dot reveals a tooltip with the exact:

- Credit score
- Income
- Approval decision

To help with visual feedback, once a point has been viewed, it **animates and shrinks slightly**, giving users a way to track which data points they've already explored. This small touch adds a bit of fun and improves user experience without requiring extra clicks. This will help them visually compare their profile with others.



4. Credit Score Prediction Wheel

This gauge-based visualization estimates a user's credit score based on key financial and personal inputs. Upon submitting the form, a POST request is sent to the `/credit` Flask endpoint, which returns a predicted credit score using a Random Forest model trained on historical data.

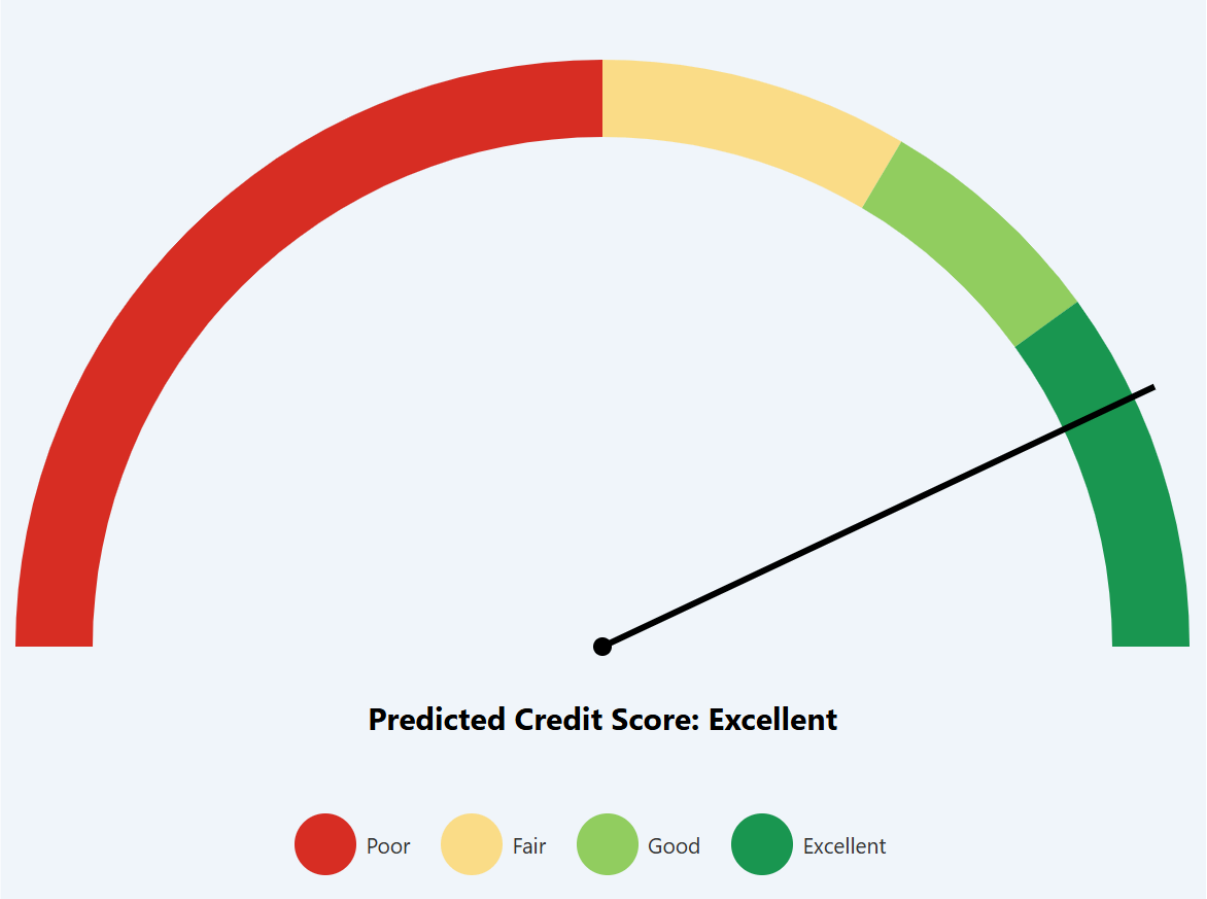
- **Needle Position**: Indicates predicted credit score

- **Color Zones**:
    - Red: “Poor” (< 580)
    - Yellow: “Fair” (580–669)
    - Light Green: “Good” (670–739)
    - Dark Green: “Excellent” (740+)

The score range is normalized and mapped to an angular position on a semicircular speedometer. Hovering over each zone displays tailored credit-building suggestions. Unlike raw numerical outputs, this wheel presents creditworthiness in a categorical and intuitive format.

User inputs include:
Age, marital status, employment, residential status, income, expenses, loan history (flag), number of existing loans, debt, and bank account tenure.

We decided not to show the exact score to avoid misinterpretation, especially since the model is trained on a synthetic dataset. Instead, the visualization displays the rating category and a dynamic pointer, offering users insight into their financial standing with contextual advice.



Addressing Midpoint Feedback

We took the midpoint feedback seriously and made sure to address each point with thoughtful improvements.

1. **Linking Views with User Data**
   We removed repetitive input sections to make the visualizations cohesive and more intuitive for users to understand.
2. **Input Validation and Range Clamping**
   We've added robust **form validations** to ensure inputs fall within realistic ranges. Credit scores are limited to the 300–850 range, and all numerical fields (like debt or income) are restricted to positive integers. Dropdowns and sliders reduce the chance of entering invalid values.
3. **Annotations and Narrative Support**
   Each visualization now includes **supporting text, tooltips, and legends**. These aren't just explanations, they guide the user, helping them understand what to focus on and how to interpret what they see. We made sure the visualizations don't just present data but tell a story.
4. **Navigation and Usability**
   We added a **persistent navigation bar** at the top of the page. This helps users seamlessly move between the U.S. map, loan predictor, and scatterplot without needing to scroll endlessly. It's a small addition, but it significantly improves usability.

Evaluation

Using our visualizations, we were able to uncover and understand several key trends and challenges in the credit and loan approval landscape. Each view offered unique insights into the data, while also highlighting important limitations and areas for improvement in our project.

What We Learned from the Visualizations

1. From the U.S. State Map:
We saw clear regional variations in financial health. States with higher per capita income generally had better average credit scores and lower debt-to-income ratios. However, this wasn't always consistent. Some states had surprisingly high debt loads relative to income, despite decent FICO scores. This prompted us to consider the role of cost of living, education, and healthcare in shaping credit health, factors that weren't in our dataset but could explain outliers. Hover-based exploration made this pattern discovery much more fluid than a table ever could.

2. From the Loan Approval Predictor (Speedometer):
This model revealed just how **sensitive loan approval can be** to certain inputs. Small changes in credit score or debt level could result in large swings in approval probability. We also observed that certain categorical features (like marital status or education level) had an outsized influence on the model's behavior, possibly reflecting bias or imbalance in the original training data.

This raised a larger question about fairness in credit modeling. Are users being judged more on easily manipulable labels rather than their full financial story? We couldn't solve this in one project, but we now better understand the challenge of building equitable models.

3. From the Scatterplot (Income vs Credit Score):
We discovered that income alone does not guarantee approval. There were many high-income individuals with poor credit scores who were still rejected, and conversely, some moderate-income users with strong credit scores were approved. This visualization helped reinforce our assumption that **credit behavior is a stronger predictor than income** alone. The scatterplot also made these exceptions easier to spot and investigate.

4. Credit Score Predictor (Speedometer)

This view gave users a way to simulate their potential credit standing based on realistic financial inputs. One major insight was how factors like employment status, loan history, and debt levels cumulatively shaped the predicted credit range. The categorical inputs (e.g., residential status, marital status) often introduced sharp shifts in the outcome, hinting at potential biases in how creditworthiness is assessed in real-world data.

The visualization's design, merging "Very Good" and "Excellent" into a single final band, reflected a limitation in our dataset distribution. Still, the simplified range made it easier for users to interpret their results. Rather than showing an exact number, we chose to display only the rating label ("Good," "Fair," etc.) to protect privacy and avoid overprecision. The accompanying tooltips provided tailored suggestions, reinforcing the tool's educational value and helping users reflect on actionable financial habits.

The interactivity, especially shrinking visited points, added an extra layer of usability, encouraging users to explore further without feeling overwhelmed by too much clutter.

How We Answered Our Research Questions

Our visualizations addressed several of our core questions:

- **Q: How do credit scores vary across geography?**
  A: The map clearly showed geographic disparities in credit health. Users could quickly compare states and understand where they stand in a national context.

- **Q: What factors most strongly influence loan approval?**
  A: Our approval model and visualizations highlighted key drivers such as debt-to-income ratio, credit score range, and existing loan amounts. The speedometer made this information digestible, even for users without a technical background.

- **Q: Are there disparities in loan outcomes?**
  A: While our dataset lacked demographic features like race or ZIP code, we saw enough variation across categorical features (like employment type or education level) to suggest that further work in this direction would be valuable.

- **Q: How do income and debt interact with credit outcomes?**
  A: The scatterplot helped users see that credit score and approval outcomes don't always scale with income. It reinforced the idea that behavioral factors (repayment, debt ratio) often outweigh pure earnings.

How Well the Visualizations Worked

Overall, our visualizations were successful in making complex financial patterns more accessible. Each view was interactive, intuitive, and tied closely to user-driven questions. We received positive informal feedback about the clarity of the visuals and the responsiveness of the speedometer and map.

The design choices, like color encoding, tooltips, and modular input forms, helped users focus on insights without needing to interpret raw data tables or charts.

Limitations and Areas for Improvement

While we're proud of the results, there were definite challenges and things we would do differently next time:

1. **Loan Approval Model Limitations**
   The model was quite sensitive to certain fields. For instance, a minor change in income could sometimes swing the approval rating drastically, especially when debt levels were borderline. This reflects real-world volatility but also shows that the model could benefit from more robust feature engineering or regularization.

2. **Bias in the Training Data**
   Since the dataset came from Kaggle, we didn't have much control over how it was collected. Some features may be biased or outdated. For example, the approval outcomes might reflect historical inequalities in credit access that our model could unintentionally replicate.

3. **Data Sparsity in the Map**
   Our state-level data relied on public statistics from 2021. Ideally, we would have used more recent or real-time datasets, but reliable, free sources were limited.

4. **User Feedback and Guidance**
   While we added tooltips and annotations, we could still improve narrative scaffolding across the site. For example, highlighting "key insights" or providing an onboarding overlay might help first-time users navigate the dashboard more confidently.

5. **Input Validation and UX Edge Cases**
   While we added some input constraints, there's still room for improvement. For instance, preventing the user from entering nonsensical combinations (like extremely high income and very low debt yet still receiving a low approval) requires more thoughtful conditional logic.

Final Thoughts

CreditCanvas gave us a deeper appreciation for the complexity of financial data, and the importance of translating it into a form that people can actually understand. Our project isn’t perfect, but it’s a strong step toward helping users explore credit health in a more meaningful, engaging way. We’re excited about the potential to build on it in the future, with better data, deeper fairness considerations, and more personalized narratives.