

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.

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

IDEAS

Analyzing credit to create visualizations that enhance financial literacy.

Credit Canvas

- ✓ Credit Score Calculator
- ✓ Loan Approval Calculator
- ✓ Interest Rate Calculator
- ✓ Map of relative credit scores by state in the U.S.

Credit Score Calculator

Excellent, Good, Fair, Poor

Shade in from poor to calculated rating

Loan Approval Calculator

Loan Predictor, Interest Rate Calculator

Given credit score, loan amount, time to pay off, and type of loan, predict if loan will be approved (regardless of interest rate). Could this be rate?

Given credit score rating and loan amount/length, calculate expected interest rate

Give a good/bad/average rating depending on calculated interest rate relative to loan type (ex: auto, home, student)

Map of relative credit scores by state in the U.S.

- Heatmap (categorical coloring) or gradient?
- Information displayed when hovering over each state
 - State's average credit score
 - State's average income
- Compare each state with personal credit score/income
- Should any labels be added on heatmap?
- State initials?
- Number rating?

Data Sets and Models

Age, Income, Debt?

Age (after data set) 0 to 100

Income 0 to 100

Debt? 0 to 100

Loan type (ex: auto, home, student)

Interest Rate

Legend (Types): Auto, Home

Credit Score: Excellent, Good, Fair, Poor

Filter

- Datasets should be fake
- No way to get user's real credit info from online
- Use easy to obtain info from user (credit age, balance, etc) rather than utilization, debt to income ratios, etc.
- State map will have state's average credit score, debt to income ratio, and income after hovering on the state.
- When clicked, on a separate graph, user's info can be compared

Categorize

Credit Score Wheel

User Input Sliders

Loan Predictor Scale

Very unlikely, Unlikely, Likely, Very likely

OR

Very likely, Likely, Unlikely, Very unlikely

Interest Rate Scatterplot

Map

On click: State, Score, DIR, Income

1 Year Score

Combine and Refine

We can organize our multiple ideas into one large visualization space that shows multiple graphs, charts, interactions to provide whole credit overview

QUESTION

Our question was to understand how credit scores distributed across different demographics was influenced by age, credit lines, region, credit utilization, etc, and understanding loan approval. Our visualizations answer all of these questions.

Sheet 2: Initial Design #1

Credit Canvas

Your Credit Rating

Excellent

Poor

Great

Good

Fair

Good

(example of a good credit score)

YOUR INFO

What is my credit?

Based on your information, here is what may be contributing the most to your credit

• HEADER 1
Feedback 1

• HEADER 2
Feedback 2

You are here

Good 65 to 80

For your age, you lie in the 65 to 80 Percentile of credit

YOUR INFO

Age: 0 to 100

Annual Income: 0 to 1,000,000

Number of Cards: 0 to 5

Card 1: Year opened 1900 to 2025, Unpaid balance 0 to 100,000, Limit 0 to 100,000, Late Payments 0 to 100

Card 2: Year opened 1900 to 2025, Unpaid balance 0 to 100,000, Limit 0 to 100,000, Late Payments 0 to 100

Amount of Unpaid Card Balance: 0 to 10,000,000

Loans

Number of Loans: 0 to 5

Loan 1: Type (Auto, Home, Student, Cash), Length (years) 0 to 100, Amount 0 to 1,000,000, Year Started 1900 to 2025, Monthly Payment 0 to 100,000

Loan 2: Same sliders for rest of the loans.

LAYOUT

Title: Credit Canvas

Authors: Ruchi Gupta, Shubhrit Ravi, Suryam Gupta

Date: 2/26/2025

Sheet: 2 - Initial Design #1

Task: Credit Score Visualizer and Analysis.

OPERATIONS

- When each slider is moved, by clicking and dragging the mouse, the slider will store the new value. A text box will also appear under the slider box, where the exact desired value can be input, and the slider will adjust automatically.
- When the "What is my credit?" button is clicked, a loading page will appear while the credit model generates results, then:
 - The wheel chart fills in with the calculated rating and fills in the wheel with the specified color and amount.
 - The "Based on Your Information..." section is filled out with significant factors to the calculated rating from the model.
 - The percentile plot will show a range, in blue, based on the relative score for the specified age. The range will be shaded in blue, and hovering over it gives the exact percentile range.

DISCUSSION

Advantages

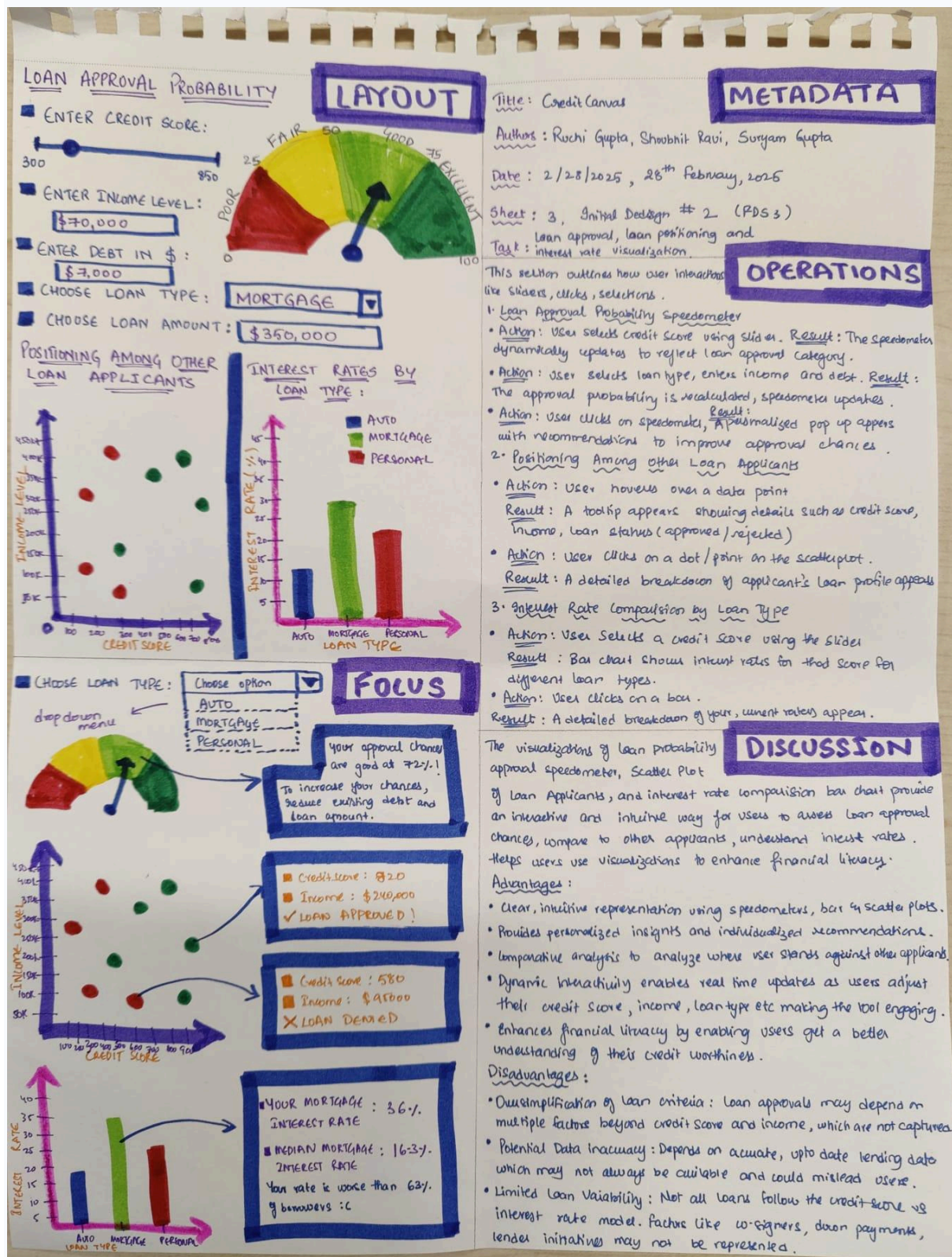
- The model produces simple results
- The Your INFO inputs are easily attainable by a user, while having a more detailed meaning to the model
- Sliders are divided into two sections (credit and loans) for easy distinguishing
- Simple layout, visualization is practical and not flashy

Disadvantages

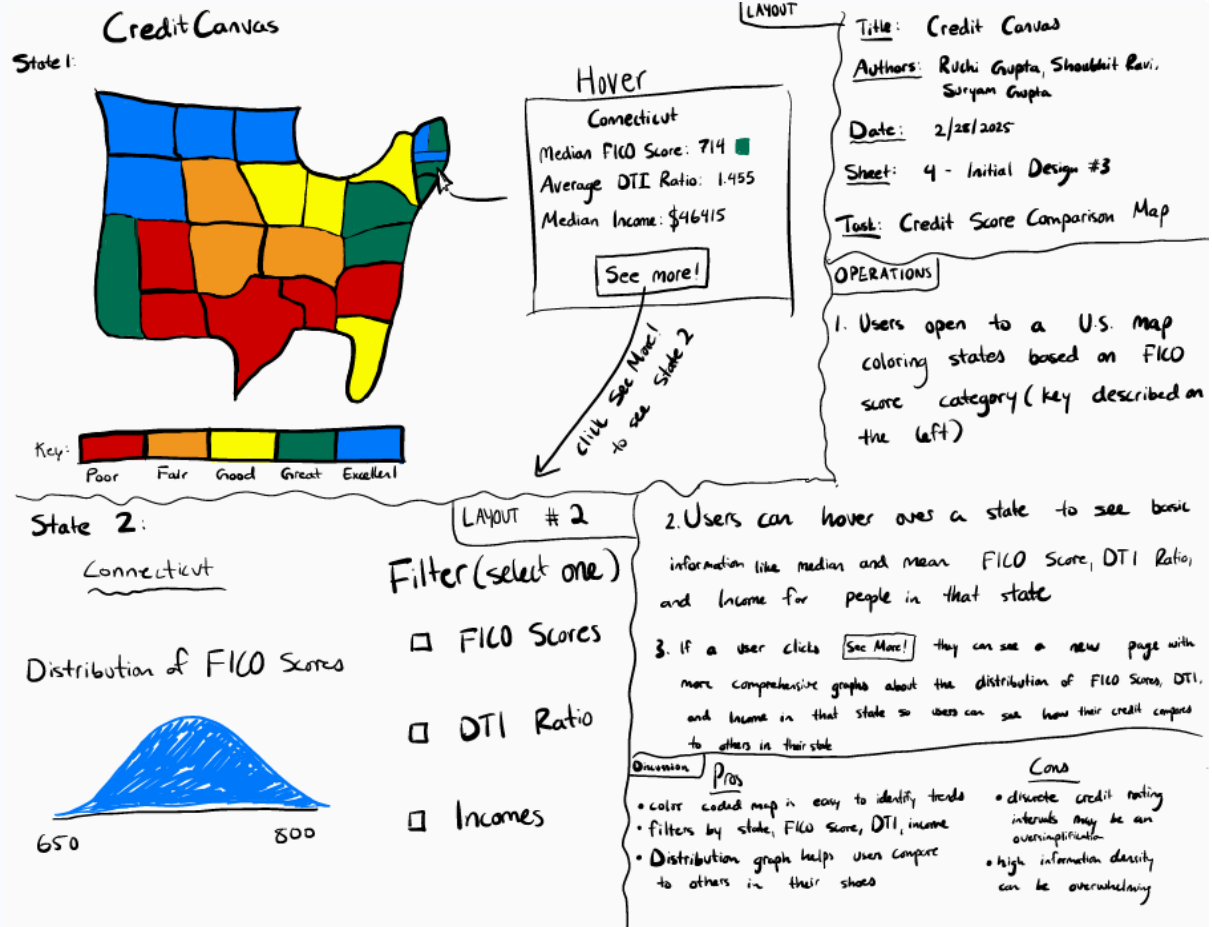
- A lot of text and sliders
- The model needs to be used to check all user inputs before running a calculation, AND/OR
- The visualization itself needs to be used to check all text box inputs (example: no strings are allowed in numerical fields)
- With the amount of inputs, a good model will require an abundance of hard-to-find credit data

As a whole, after inputs are filled in and results are generated, the design layout is clear, but before then, the design layout may appear unclear (for example, what is in the "Based on Your Info" box before calculations?)

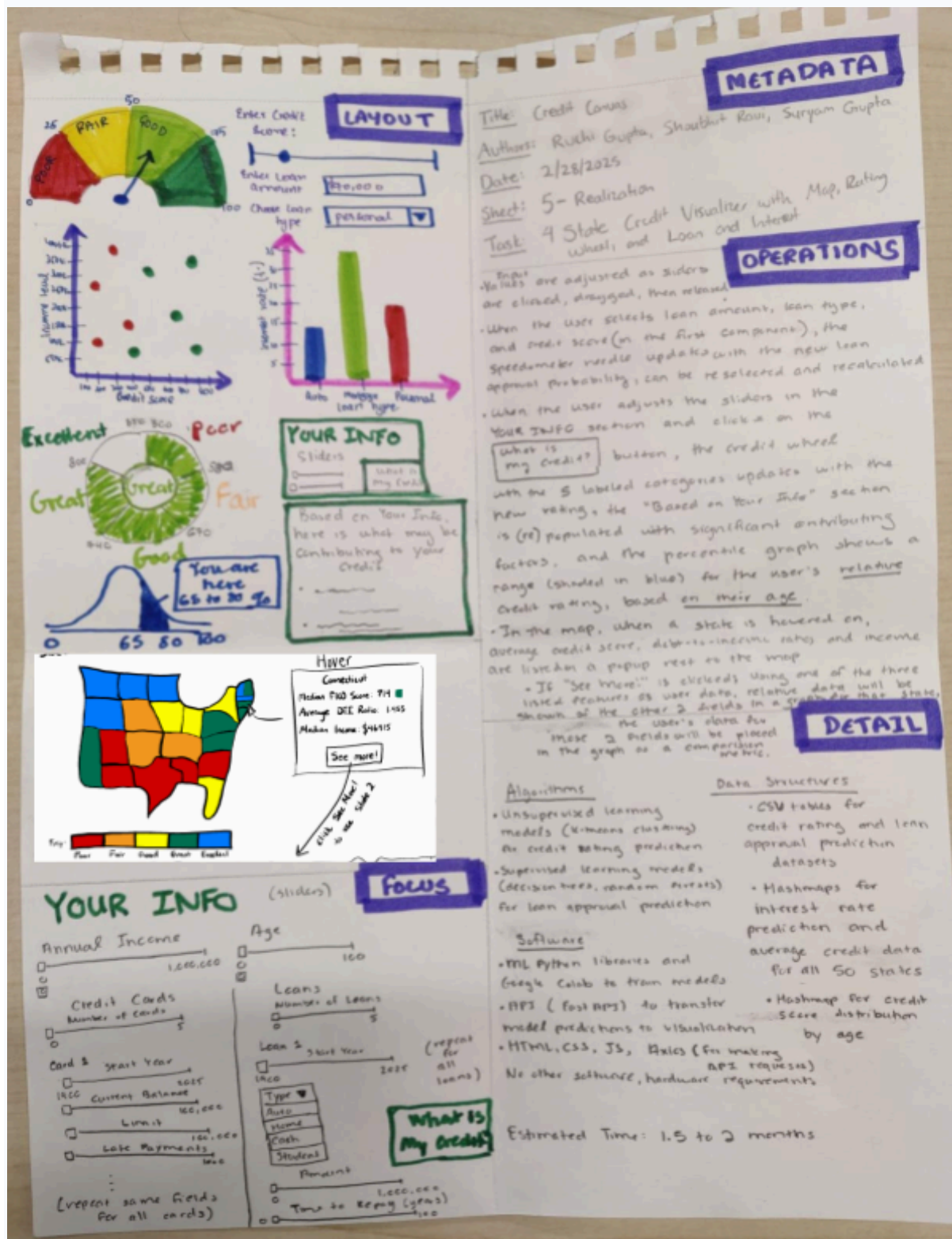
Sheet 3: Initial Design #2



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.

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](#)

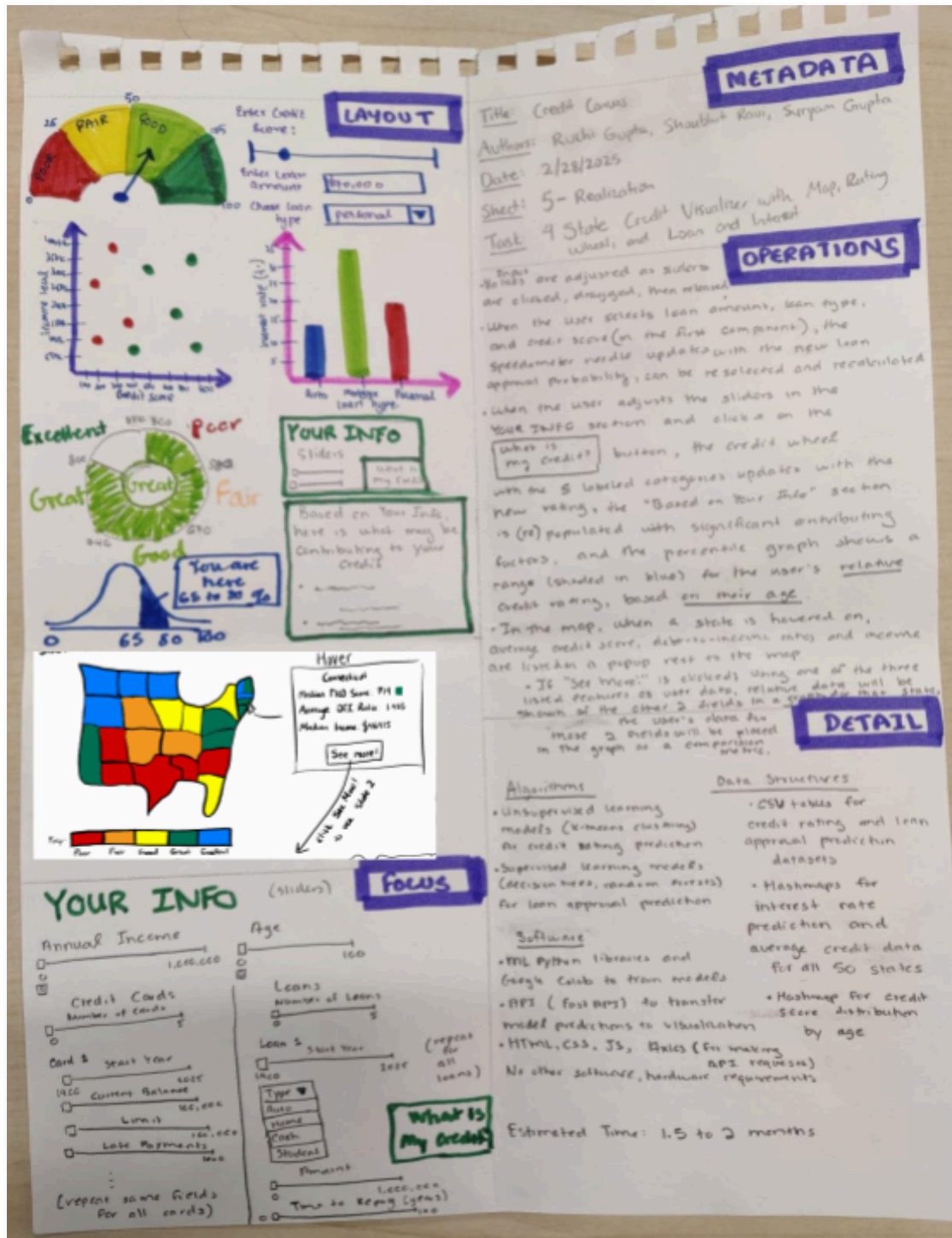
We split this dataset into two clean, labeled subsets:

- 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

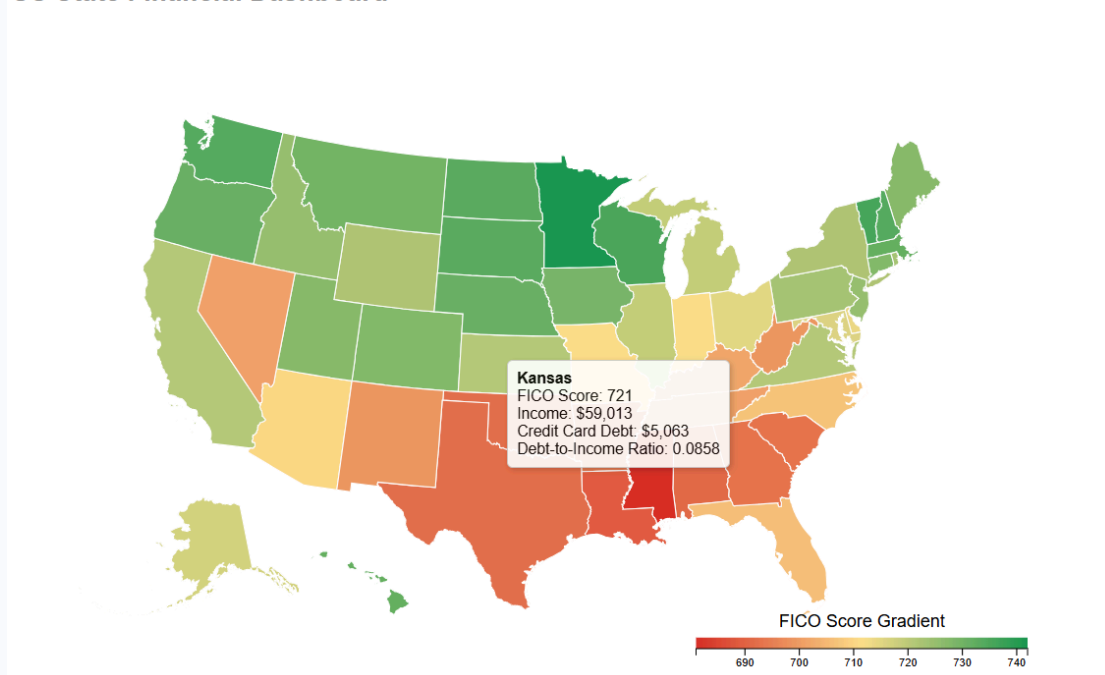
This section of the code creates a choropleth map of the United States using D3 and TopoJSON. Each state is filled with a color that corresponds to its **average FICO score**, scaled using a linear gradient from red (low scores) to green (high scores).

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.

US State Financial Dashboard



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.

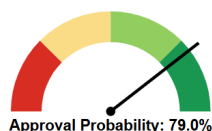
Gauge features:

- 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.

Loan Approval Predictor

Age: Dependents: Annual Income: Credit Score: Total Existing Loan Amount:
Outstanding Debt:
Marital Status: Education: Residential Status:



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.

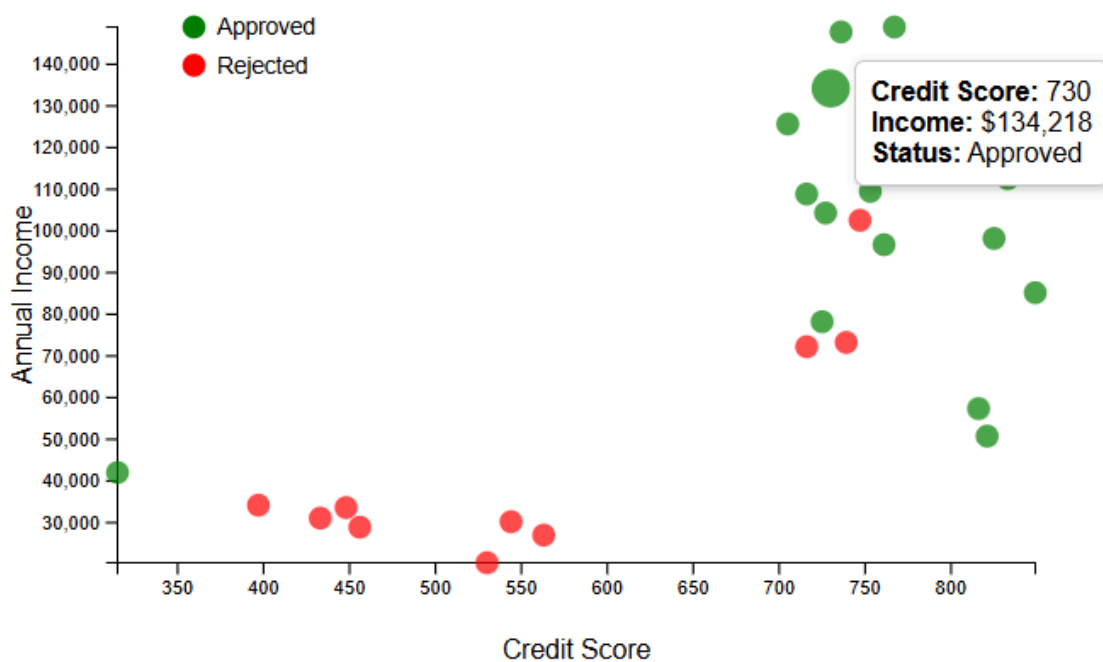
Interactivity:

- **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.

Loan Dataset: Credit Score vs Income

Generate Random Dataset




Credit Score Predictor

Age: Annual Income: Monthly Expenses: Existing Loans: Total Existing Loan Amount:

Outstanding Debt: Loan Term (months): Default Risk (0-1):

Education: Loan History: Bank Account History:



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 `/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

- **`train_credit_score_model.py`**

- Trains a random forest model to predict credit score as a regression task
- Evaluates performance with MSE and R^2 metrics
- 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

You can install them using pip:

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

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#>

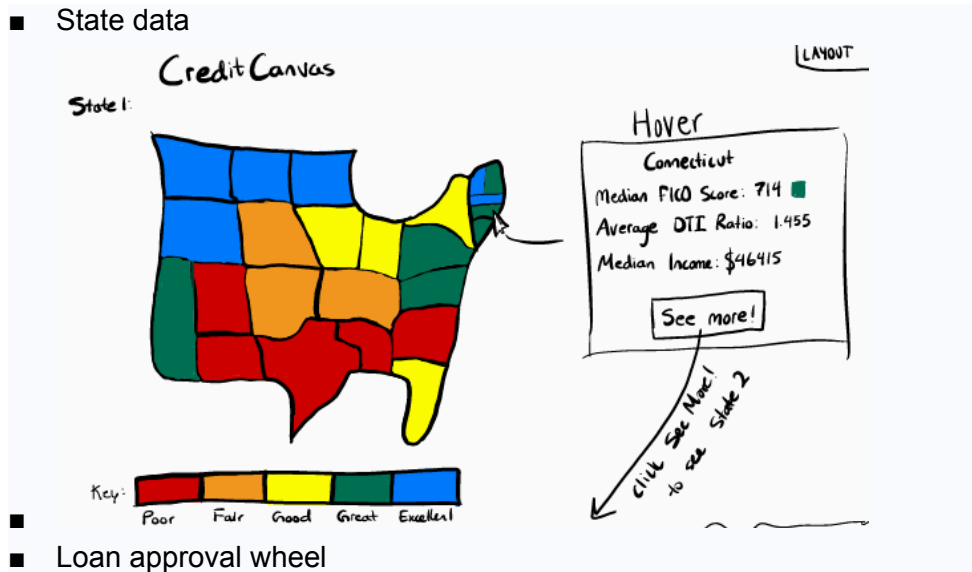
We did recognize that the state income data may be inaccurate, since it is four years outdated. However, we didn't address it immediately, as our main priorities were to start data cleaning, model training, and website skeleton building.

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):

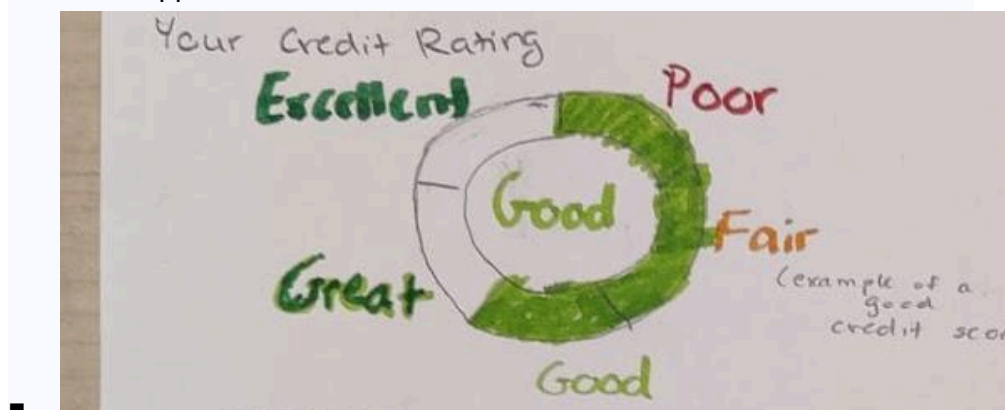
- State data



- Loan approval wheel

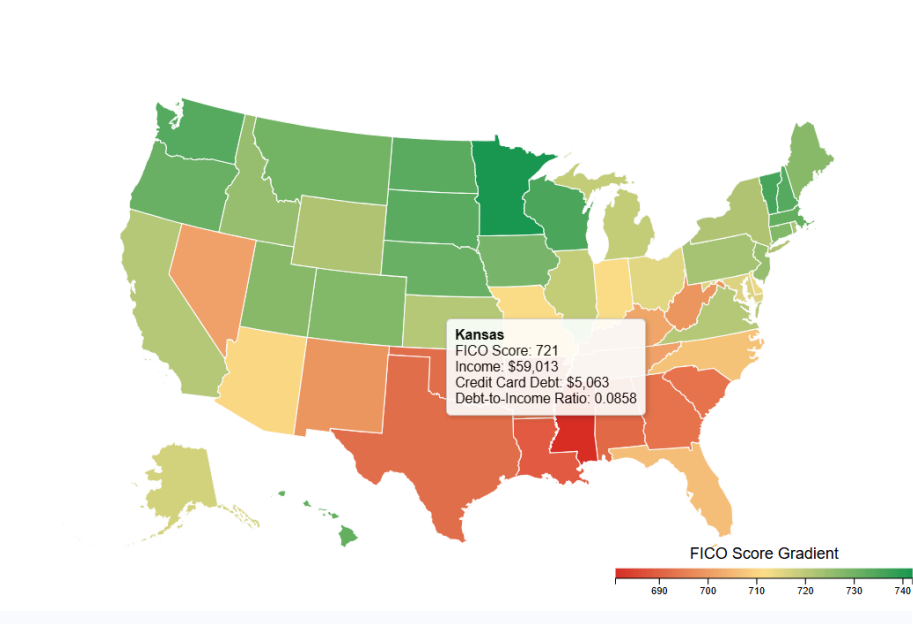


- Credit approval wheel



- Initial implementations (from the website)

US State Financial Dashboard

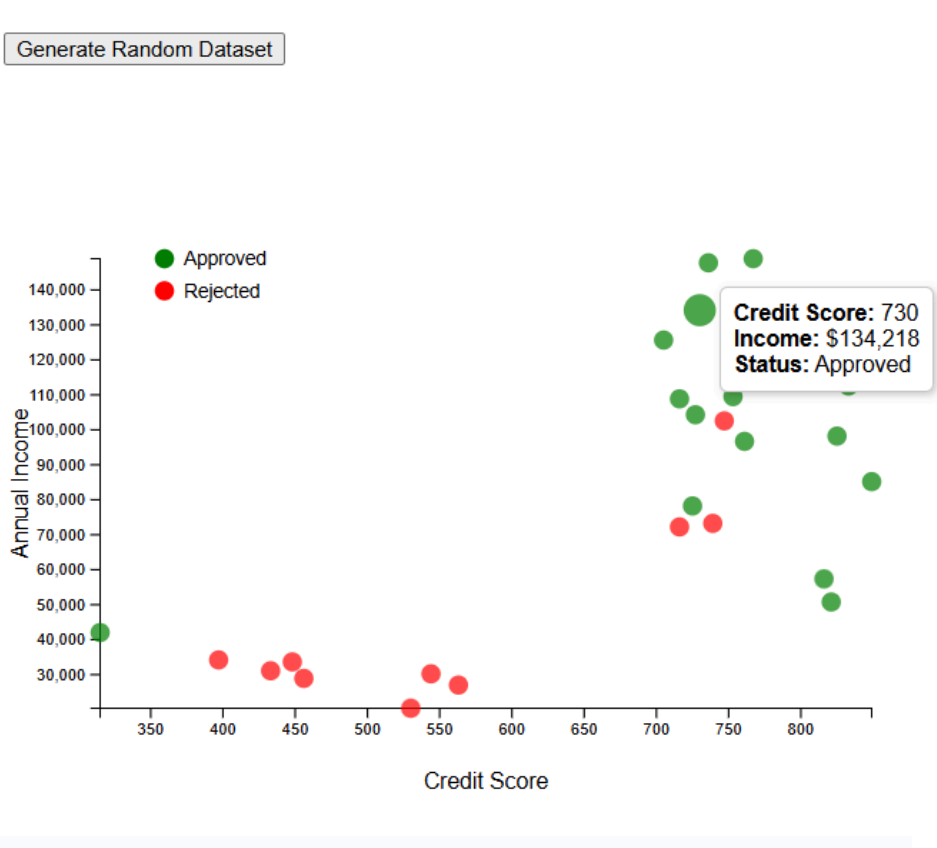


Loan Approval Predictor

Age: Dependents: Annual Income: Credit Score: Total Existing Loan Amount:
Outstanding Debt:
Marital Status: Education: Residential Status:

Approval Probability: 79.0%

Loan Dataset: Credit Score vs Income



- 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