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?

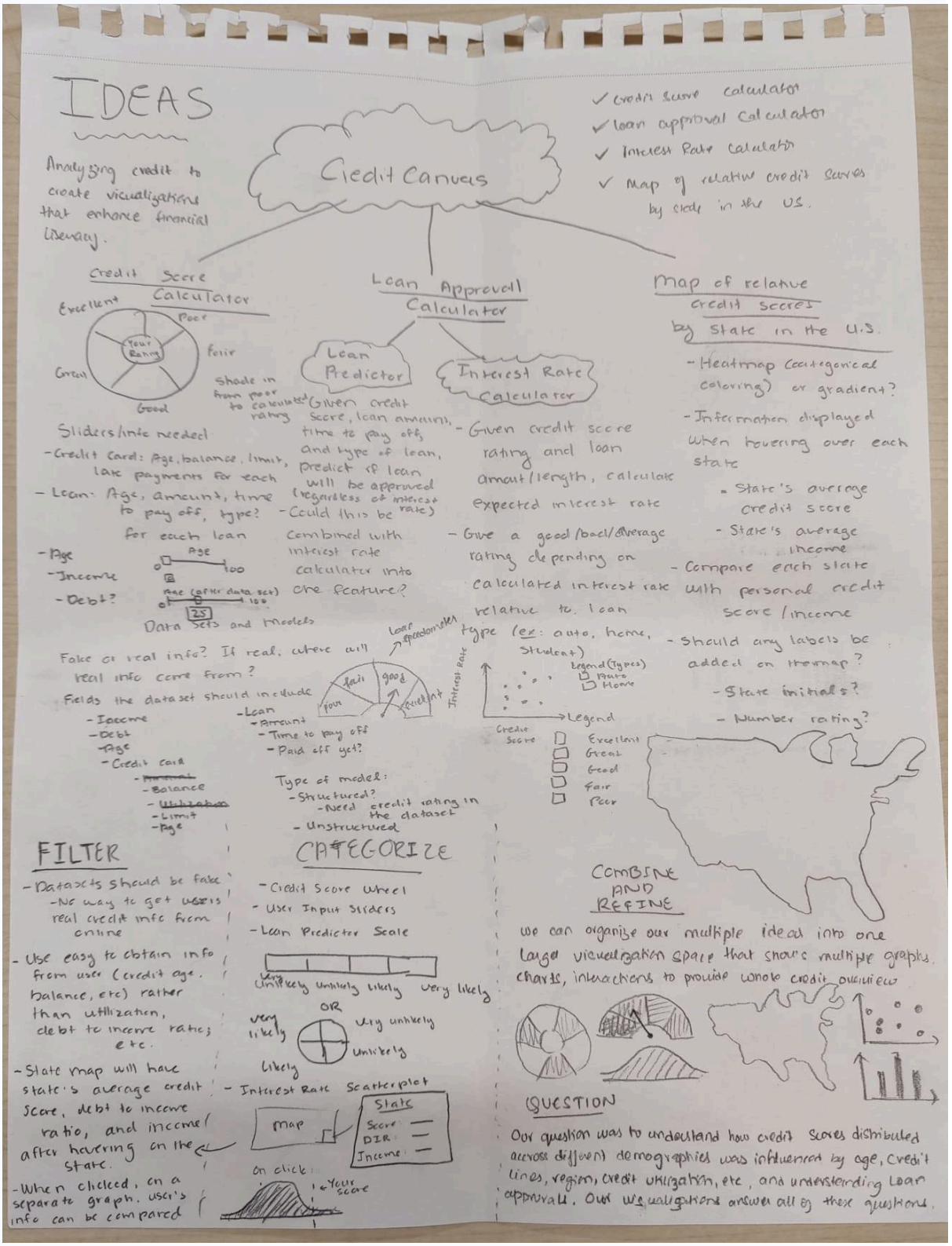
We will use **Python** for data preprocessing, leveraging libraries such as:

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

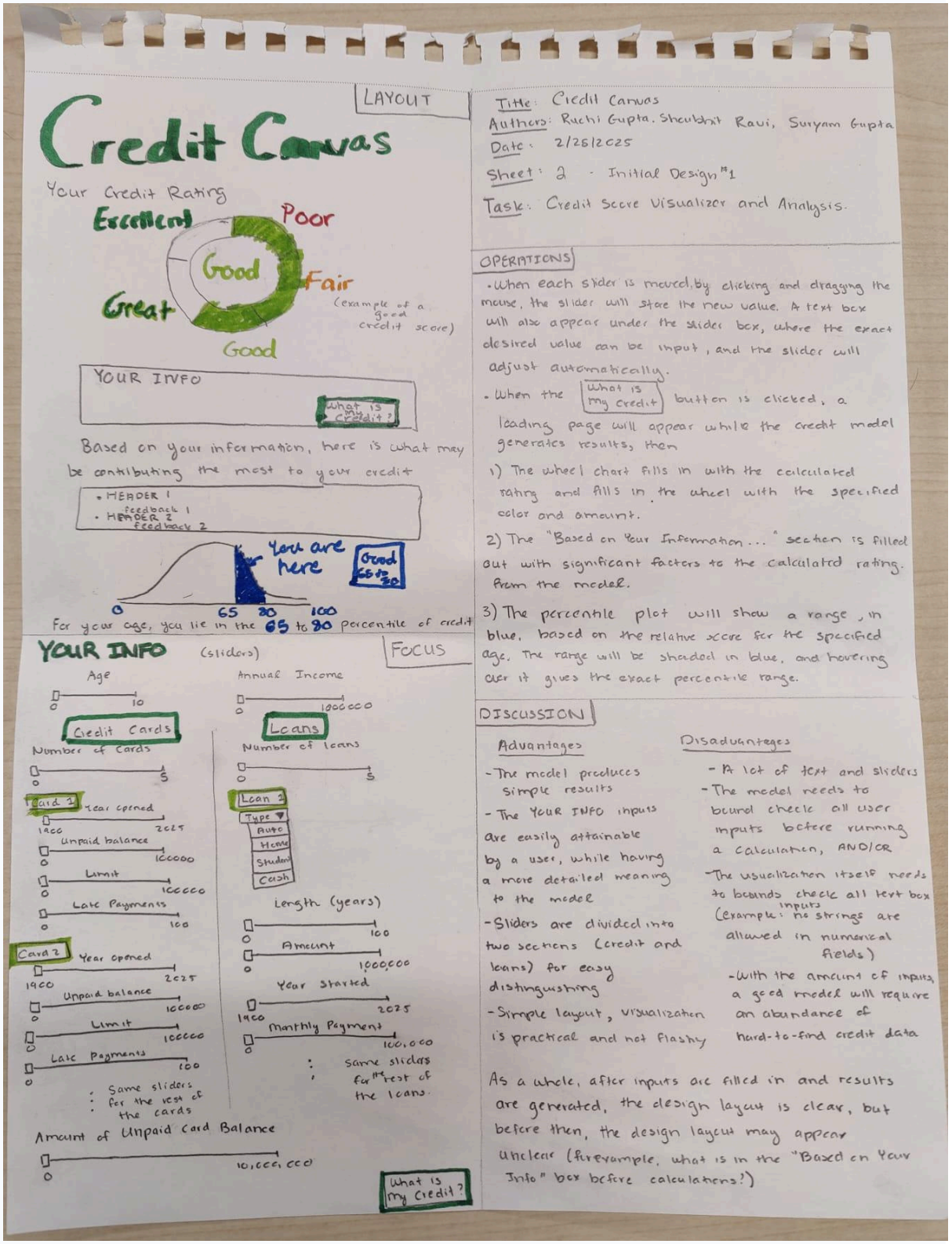
---

Visualization Design

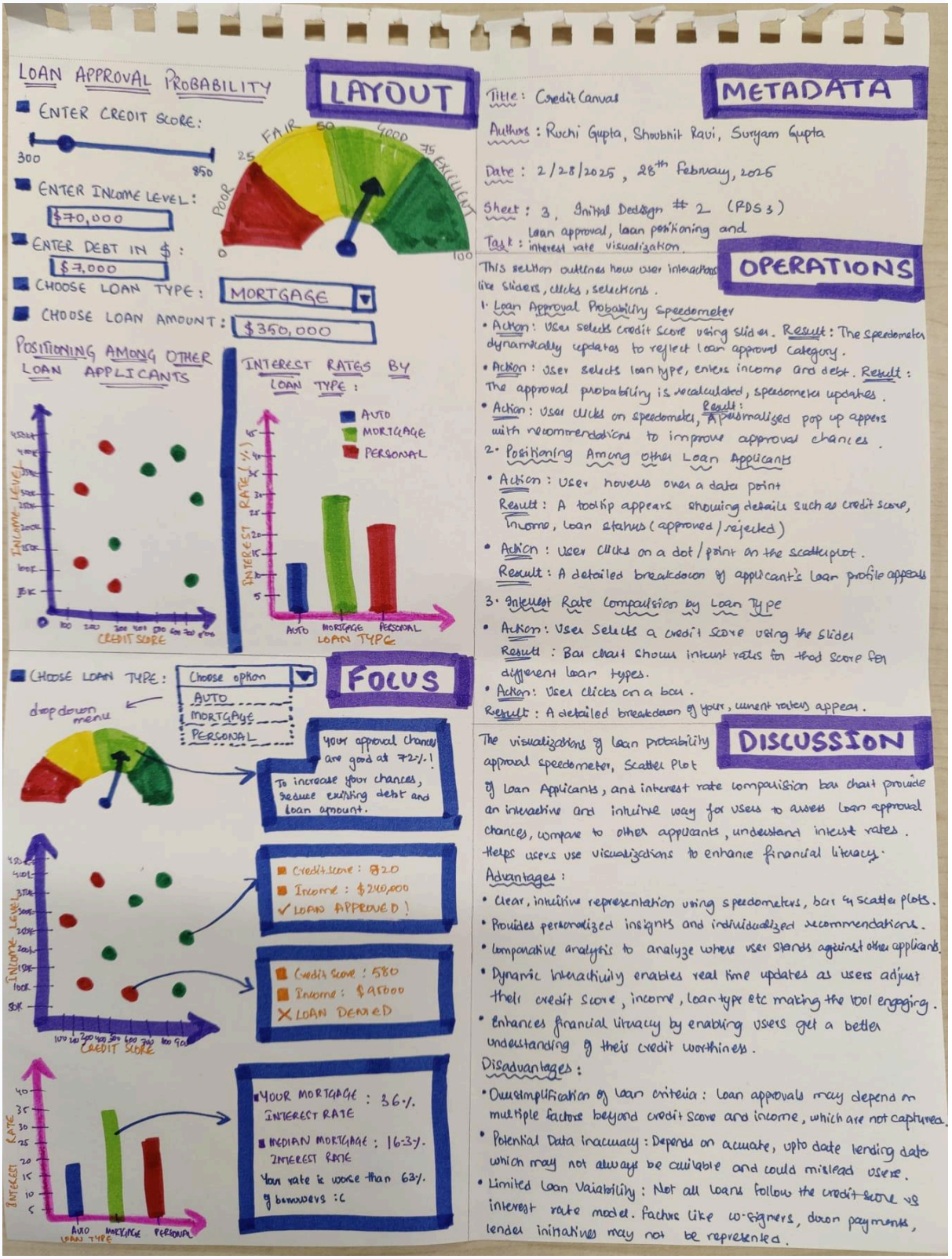
Sheet 1: Brainstorm



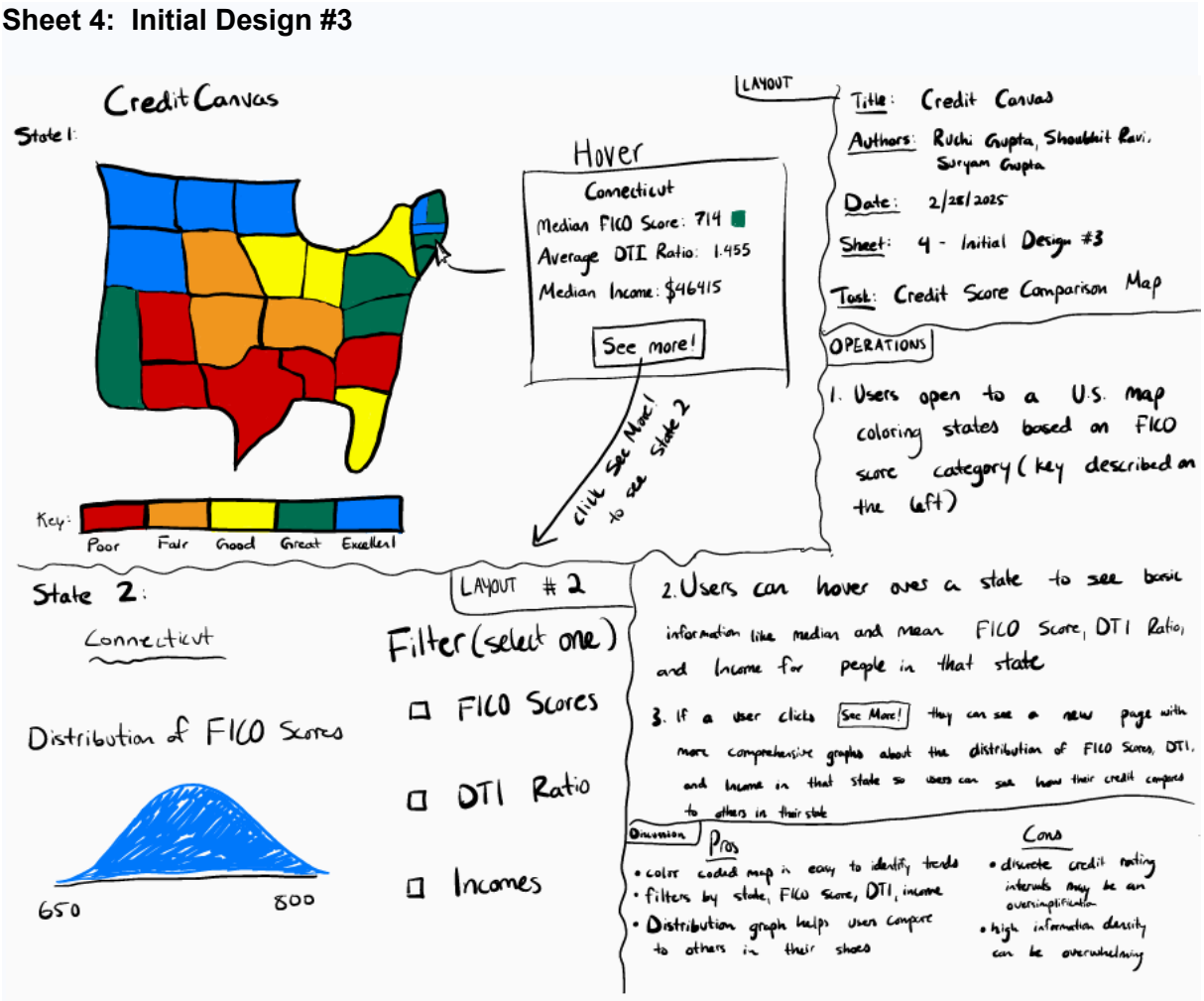
Sheet 2: Initial Design #1



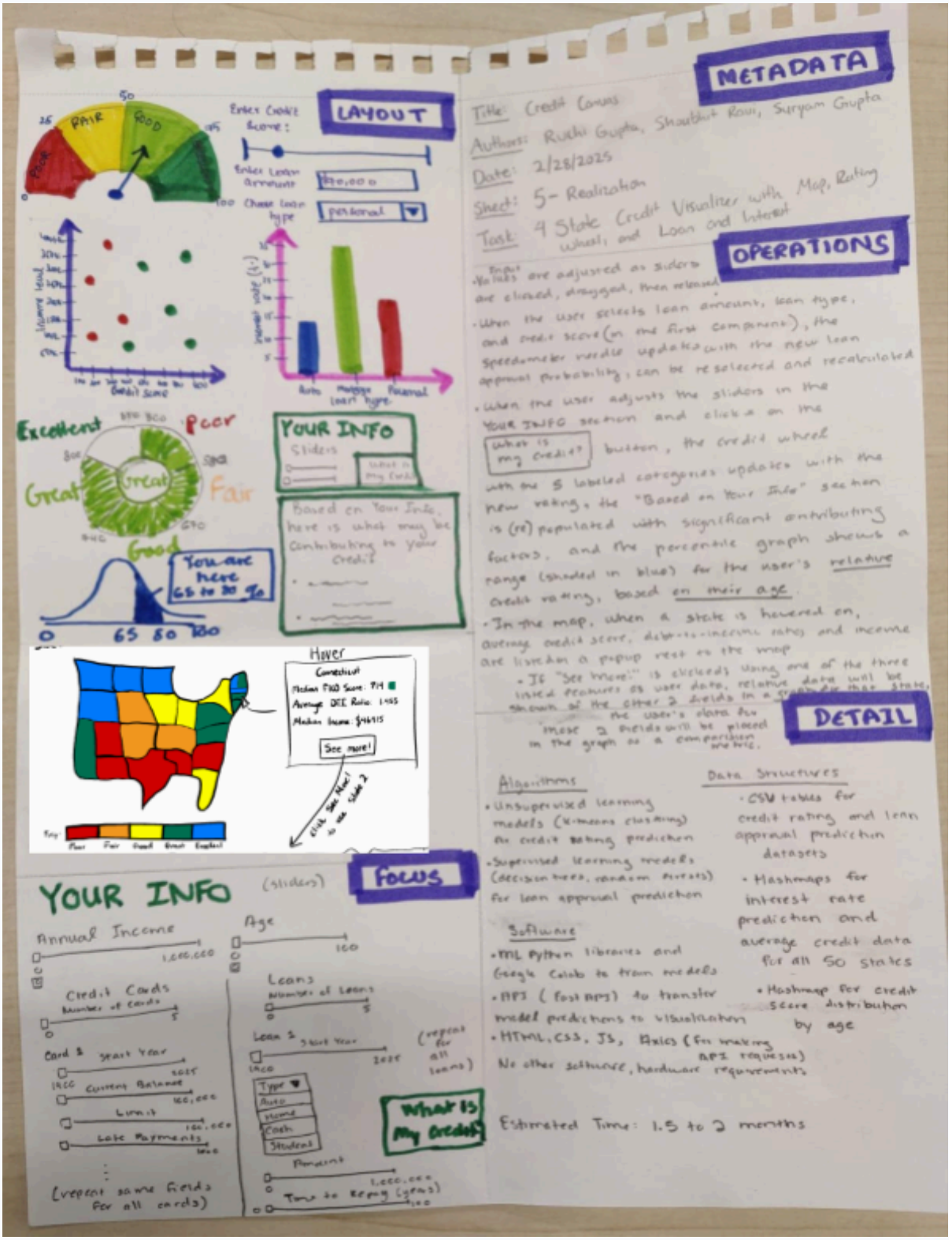
Sheet 3: Initial Design #2



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.

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

---