Exploring Loan Default Patterns Through Advanced Visual Analytics and Data- Driven Insights

Date: 12/09/2024
Team: Vishnuvardhan Reddy Kollu, Santoshi Borra, Jaswanth Mandava
Course: Data Visualization (INFO – I 590) – Fall 2024

Contents:

Project Motivation

What if we could predict loan defaults just by looking at something before they happen?

Could a borrower's age, employment status, or credit history hold the key to unlocking their financial reliability?

What if we could visualize this data and uncover interesting trends within the vast dataset of people's bank accounts?

- ❖ We aimed to dive into these questions, use our visualizations to assess credit risk to revolutionize lending decisions and safeguard against financial losses.

Introduction

- Vehicle loans play a critical role in enabling people to purchase, but defaults on Equated Monthly Instalments (EMIs) pose significant financial risks for lenders.

- Lenders assess borrower reliability through factors like employment type, identity verification documents, collaterals, primary, secondary accounts information, and credit history details including credit scores and past loan defaults.

- Credit history evaluation includes critical details like credit scores, the number of active loans, past defaults, and total debt burden, all of which impact a borrower's perceived risk.

- Rising default rates cause lenders to tighten lending criteria, which directly affects loan approval rates and necessitates a balance between financial security and granting credit to deserving clients.

- A data-driven visualization dashboard is proposed to help individuals make informed decisions by picking out the key factors influencing loan outcomes.

Description of Data

#	Column	Non-Null Count	Dtype
---	------	--------------	-----
0	UniqueID	233154 non-null	int64
1	disbursed_amount	233154 non-null	int64
2	asset_cost	233154 non-null	int64
3	ltv	233154 non-null	float64
4	branch_id	233154 non-null	int64
5	supplier_id	233154 non-null	int64
6	manufacturer_id	233154 non-null	int64
7	Current_pincode_ID	233154 non-null	int64
8	Date.of.Birth	233154 non-null	object
9	Employment.Type	225493 non-null	object
10	DisbursalDate	233154 non-null	object
11	State_ID	233154 non-null	int64
12	Employee_code_ID	233154 non-null	int64
13	MobileNo_Avl_Flag	233154 non-null	int64
14	Aadhar_flag	233154 non-null	int64
15	PAN_flag	233154 non-null	int64
16	VoterID_flag	233154 non-null	int64
17	Driving_flag	233154 non-null	int64
18	Passport_flag	233154 non-null	int64
19	PERFORM_CNS.SCORE	233154 non-null	int64

20	PERFORM_CNS.SCORE.DESCRIPTION	233154 non-null	object
21	PRI.NO.OF.ACCTS	233154 non-null	int64
22	PRI.ACTIVE.ACCTS	233154 non-null	int64
23	PRI.OVERDUE.ACCTS	233154 non-null	int64
24	PRI.CURRENT.BALANCE	233154 non-null	int64
25	PRI.SANCTIONED.AMOUNT	233154 non-null	int64
26	PRI.DISBURSED.AMOUNT	233154 non-null	int64
27	SEC.NO.OF.ACCTS	233154 non-null	int64
28	SEC.ACTIVE.ACCTS	233154 non-null	int64
29	SEC.OVERDUE.ACCTS	233154 non-null	int64
30	SEC.CURRENT.BALANCE	233154 non-null	int64
31	SEC.SANCTIONED.AMOUNT	233154 non-null	int64
32	SEC.DISBURSED.AMOUNT	233154 non-null	int64
33	PRIMARY.INSTAL.AMT	233154 non-null	int64
34	SEC.INSTAL.AMT	233154 non-null	int64
35	NEW.ACCTS.IN.LAST.SIX.MONTHS	233154 non-null	int64
36	DELINQUENT.ACCTS.IN.LAST.SIX.MONTHS	233154 non-null	int64
37	AVERAGE.ACCT.AGE	233154 non-null	object
38	CREDIT.HISTORY.LENGTH	233154 non-null	object
39	NO.OF_INQUIRIES	233154 non-null	int64
40	loan_default	233154 non-null	int64

dtypes: float64(1), int64(34), object(6)

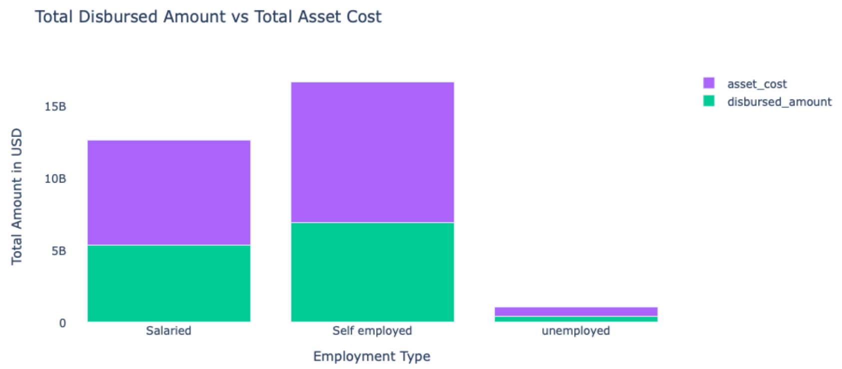
- The dataset we chose for this analysis comes from a Kaggle competition focused on vehicle loan default prediction. It includes detailed records for 233,154 loan applications, structured in a tabular format with 41 diverse attributes.

- We have considered important and unique columns that can reach a visualization easier to understand the common relations and its distributions across different branches in India to approve loans.

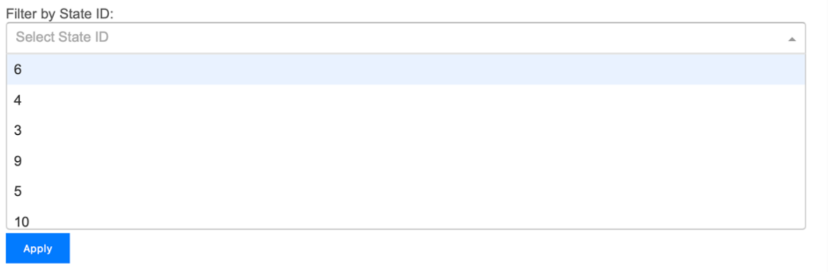
- Ages of the accounts, credit histories, loan defaults, geographical information like its state id's, pincodes and other columns are considered.

Visualizations of data(1 - 1/3) – Analyzing Risk

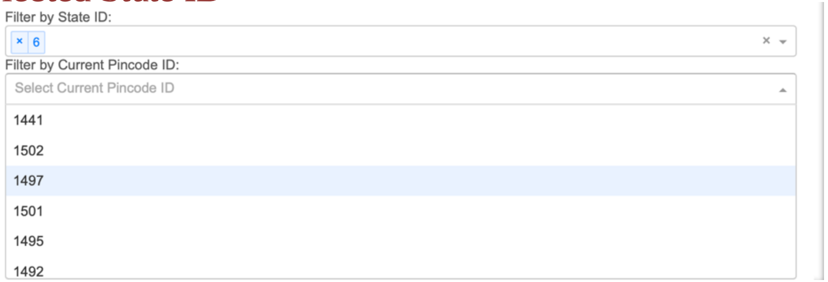
Dashboard



Step 1 – Selecting the required State ID

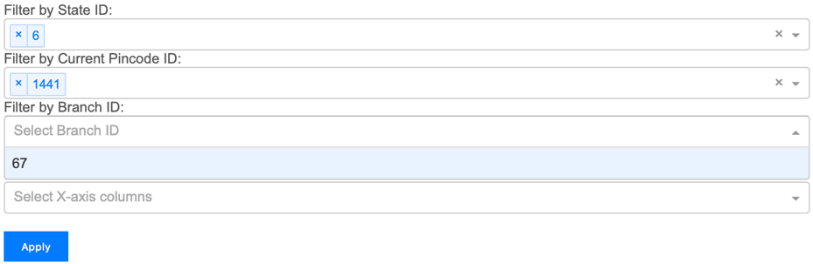


Step 2 – Drop down shows Pincode ID's available for selected State ID

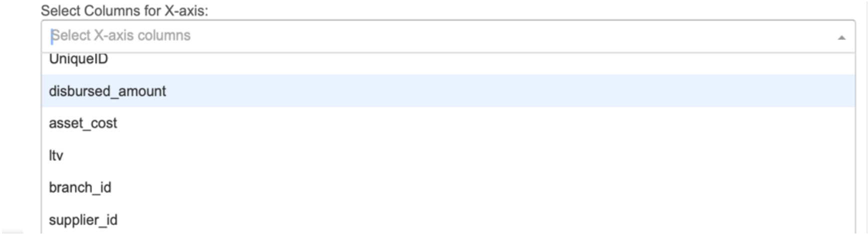


Visualizations of data(1 - 2/3) – Analyzing Risk

Step 3 – Drop down available Branch ID's for selected State ID and Pincode



Step 4 – Drop down showing available columns to select

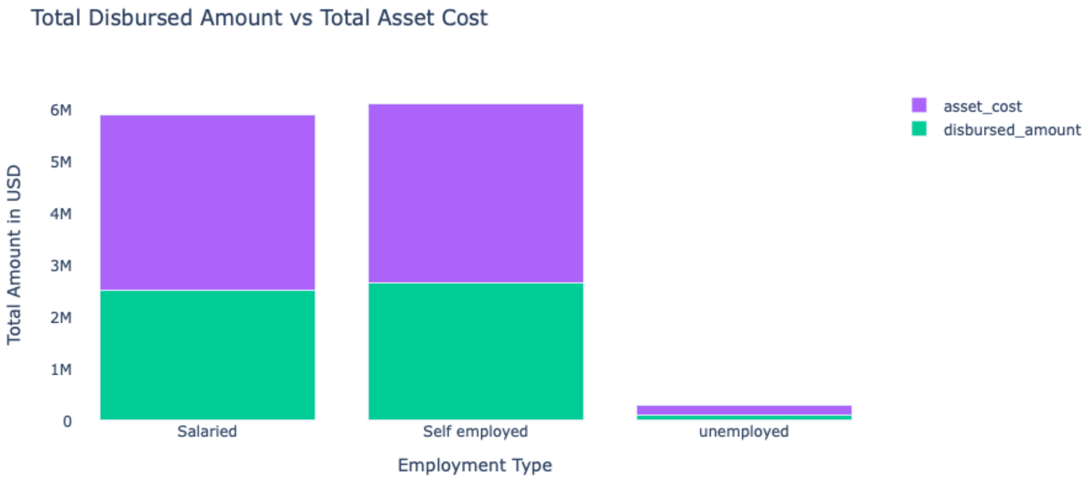


Demo visualization after selecting required items of interest

Above graph shows the total sum Disbursed Amount vs Asset Cost for State ID – 6, Current Pincode ID 1441, and Branch ID 67

Visualizations of data(1 - 3/3) – Analyzing Risk

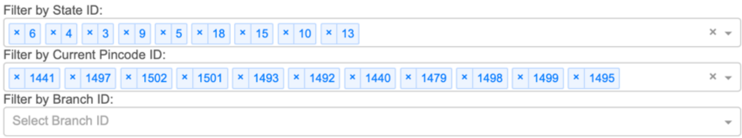
- Our interactive dashboard provide option to select multiple State ID’s, Pincode ID’s and Branch ID’s as well.

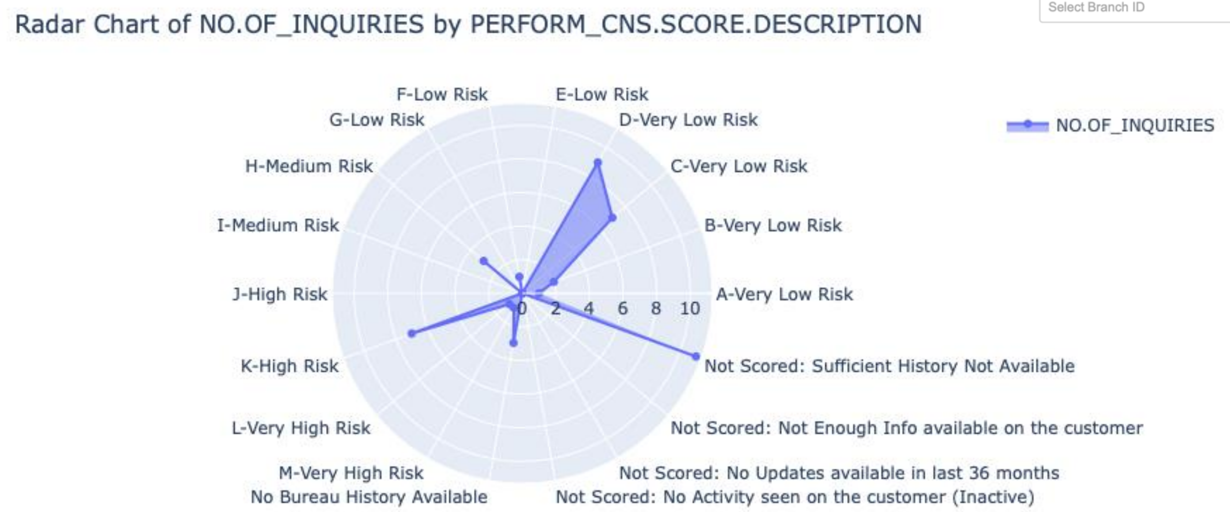


Note: This graphs also help us by showing values when we interact and hover on them.

Visualizations of data(2 - 1/3) - Analyzing Trend & Pattern

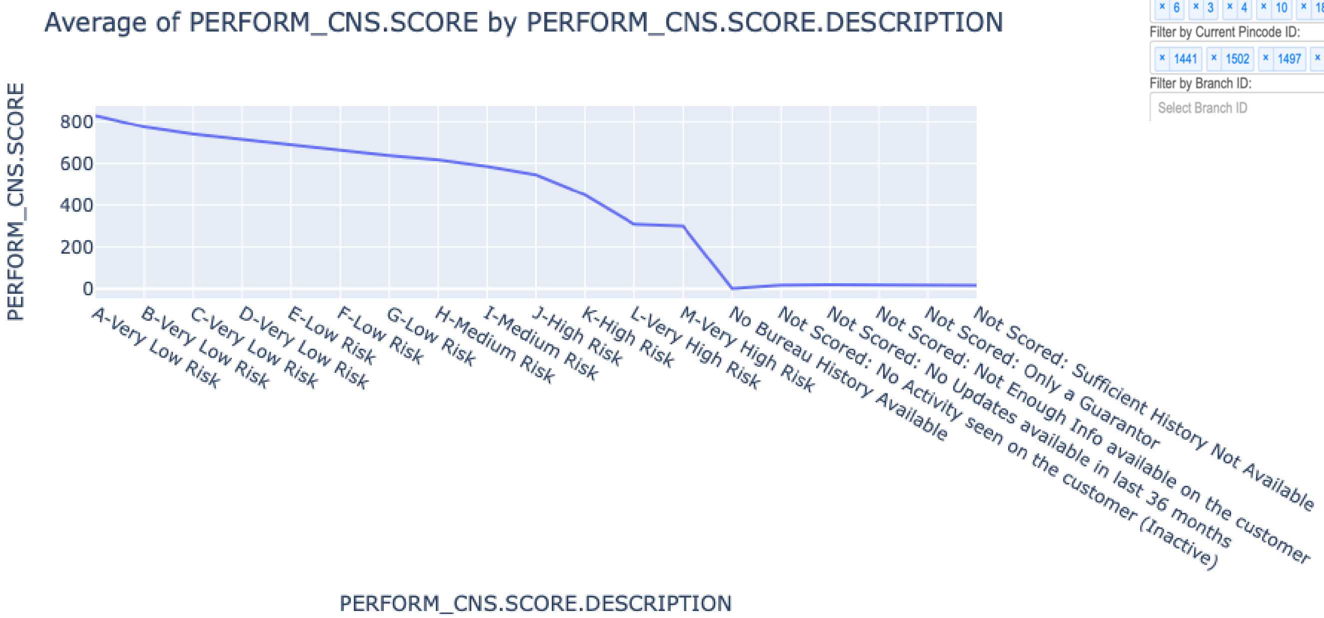
- This graph shows how many inquiries are encountering for eacl
  specific selections.





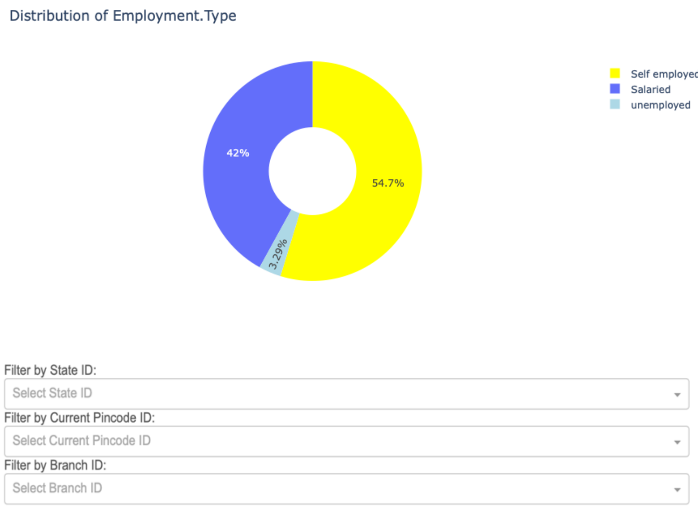
Visualizations of data(2 - 2/3) - Analyzing Trend & Pattern

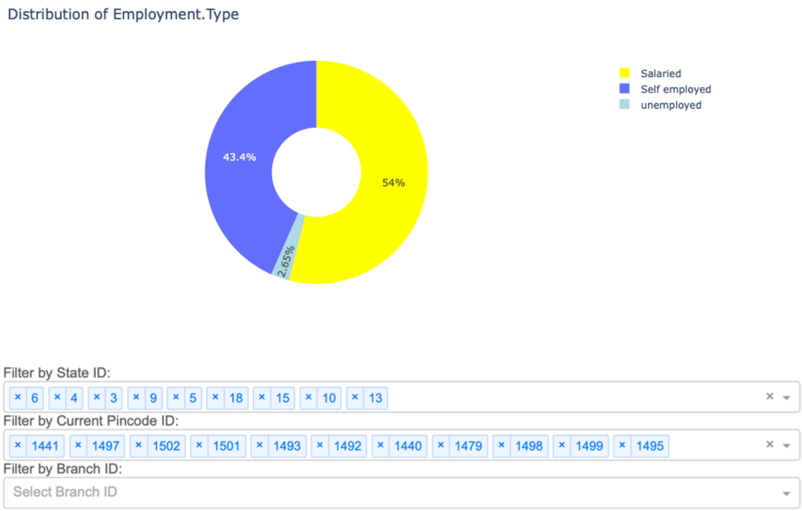
- This graph shows what is the average CNS score for each CNS score description after making specific



Visualizations of data(2 - 3/3) - Analyzing Trend & Pattern

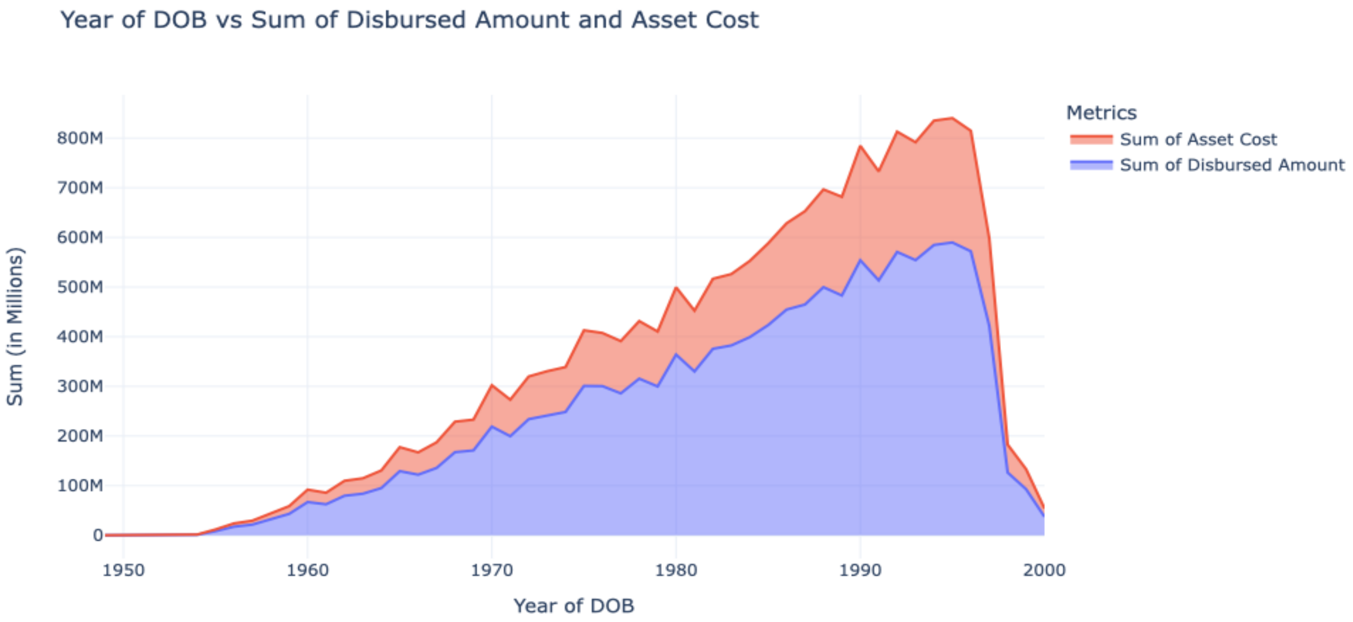
- Distribution of Employment Type by default and when selecting specific State_ID, Pincode and Branch_ID



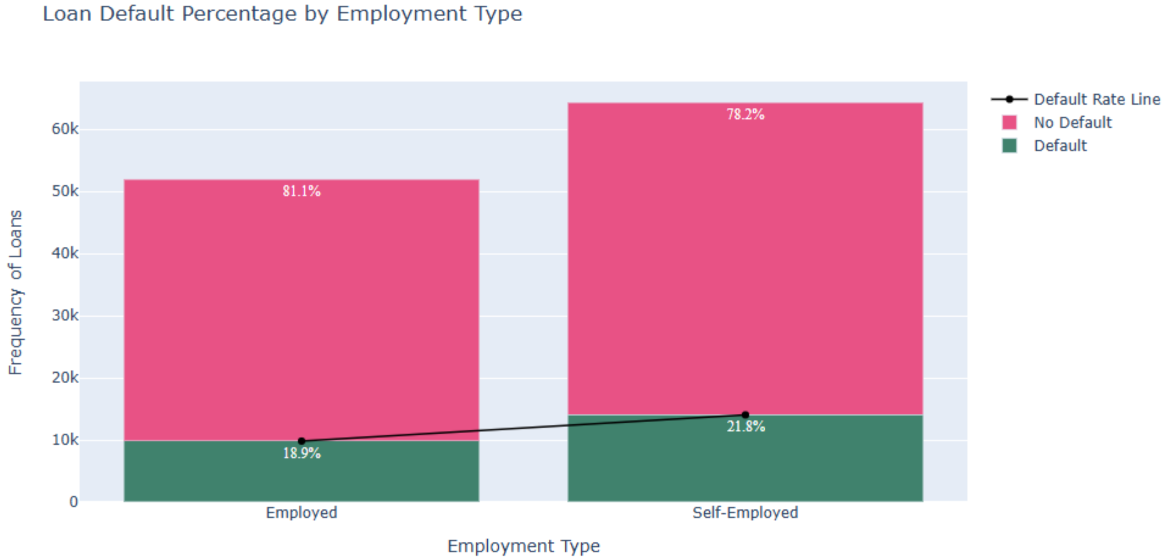


Visualizations of data(3)

Our interactive area graph helps in understanding pattern between Asset cost and amount disbursed every year



Visualizations of data(4) - Analyzing Loan default



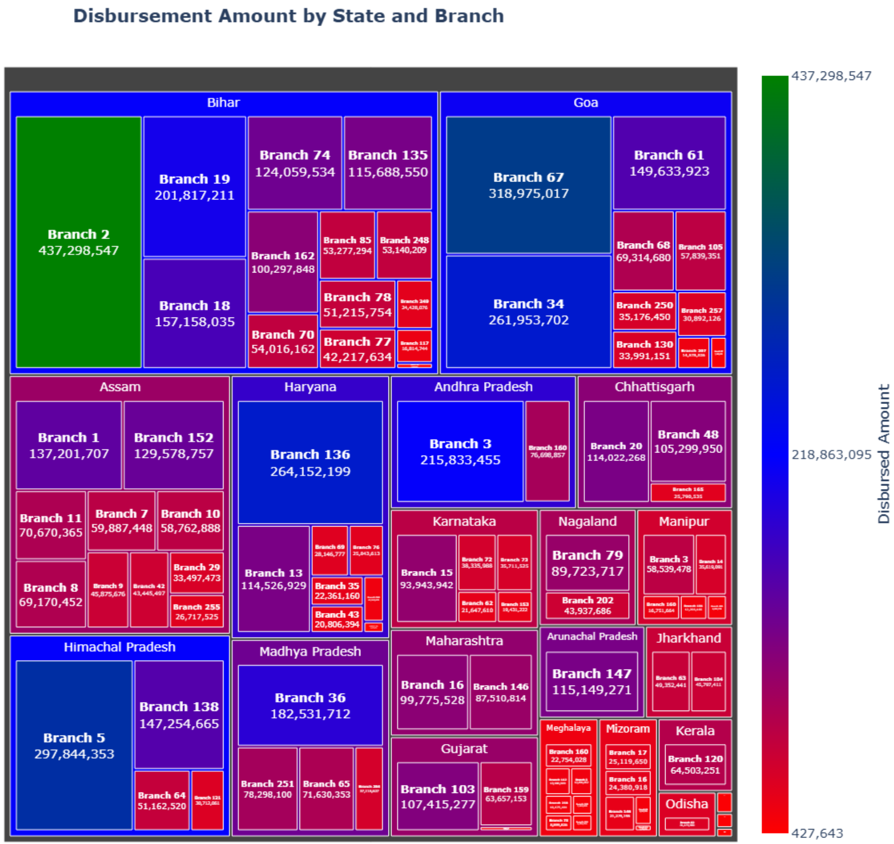
This chart presents the default rates for loans issued to Employed and Self-Employed individuals, illustrating variations in repayment dependability.

Employed Borrowers exhibit more reliable repayment behavior with a lower default rate of 18.9%.

Self-Employed Borrowers show a higher risk with a 21.8% default rate.

The trend line indicates self-employed individuals are riskier for lenders. This information could guide financial institutions in refining their risk assessment and lending strategies.

Visualizations of data(5) - Tree Map

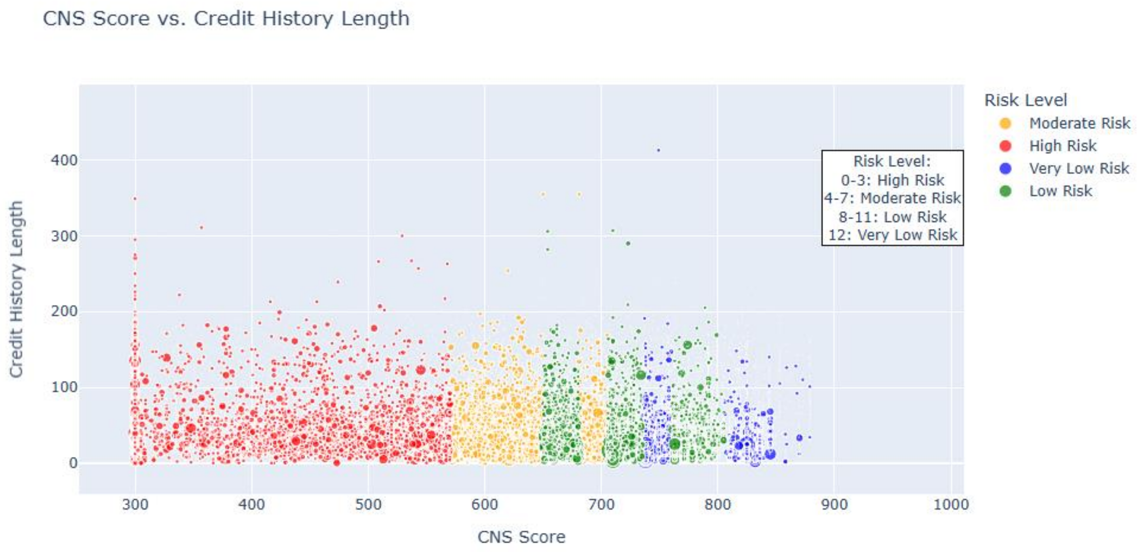


The visualization is a tree map telling disbursement amounts by state and branch.

This is particularly useful for displaying hierarchical, tree-structured data and part to whole relationships, using nested rectangles.

- **Hierarchy**: Organized by state, then by individual branches within each state.
- **Size of Rectangles**: Represents the disbursement amounts. Bigger rectangles indicate larger disbursements.
- **Color Gradient**: Green to red color scale, where green shows lower disbursements and red shows higher disbursements.

Visualizations of data(6) - Analyzing credit score



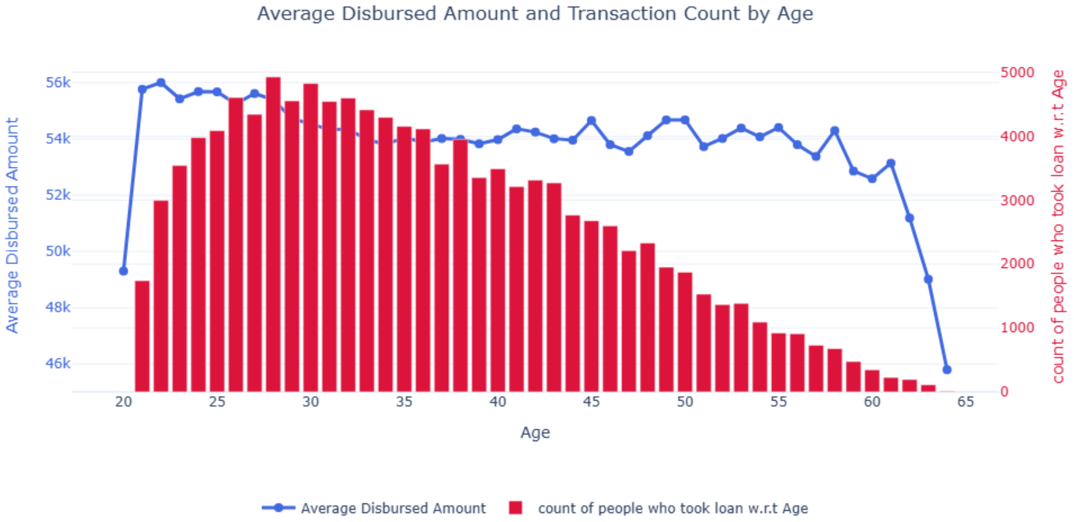
This visualization, CNS Score vs. Credit History Length, provides a detailed overview of the relationship between individuals' credit scores and the length of their credit history, segmented by the level of credit risk.

X-Axis (CNS Score): This axis quantifies credit scores on a scale from 0 to 1000. A CNS Score is a numerical representation of an individual's creditworthiness, derived from their credit history. Higher scores are indicative of more reliable borrowers who have historically been more consistent in repaying debts.

Y-Axis (Credit History Length): The length of credit history in months appears on this axis. This measure indicates the duration over which an individual has been active in the credit market. A longer credit history generally provides more data points, helping to ascertain a borrower's behavior with greater accuracy.

CNS scores into four risk levels: High Risk (0-3, red), Moderate Risk (4-7, orange), Low Risk (8-11, yellow), and Very Low Risk (12, blue), allowing easy identification of borrower reliability based on their credit score and history length.

Visualizations of data(7) -Analyzing Age Patterns

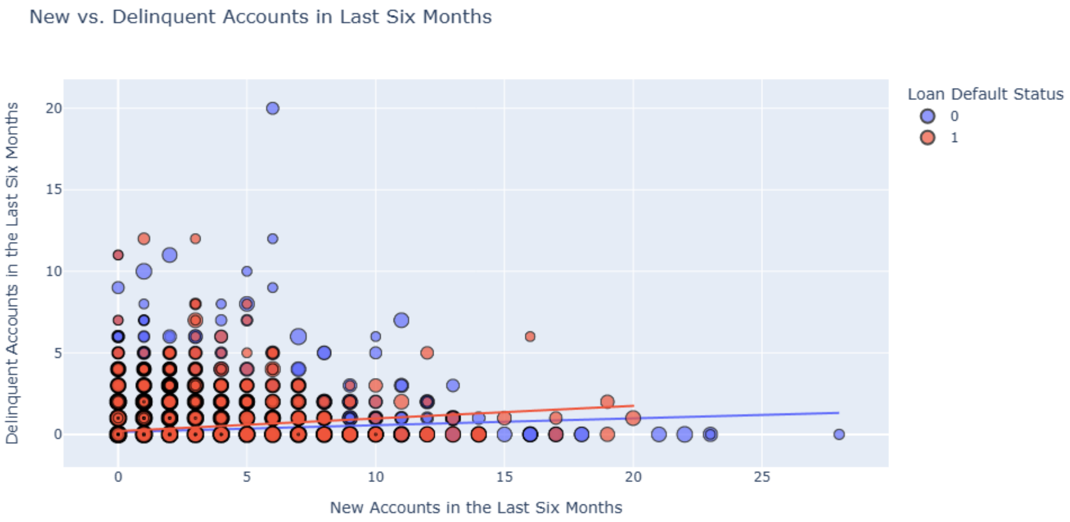


This graph shows two related trends: the average disbursed amount of loans and the number of loan transactions, both segmented by borrower's age.

Bar Chart - Red Bars: Represent the average disbursed amount for each age group. This value starts higher for the younger age groups and shows a slight initial increase before beginning a gradual decline as age increases.

Scatter plot - Blue Line: Indicates the count of people who took out loans, categorized by age. The highest number of transactions occurs in the younger age groups, with a sharp decrease as age increases.

Visualizations of data(8) - Analyzing the risk



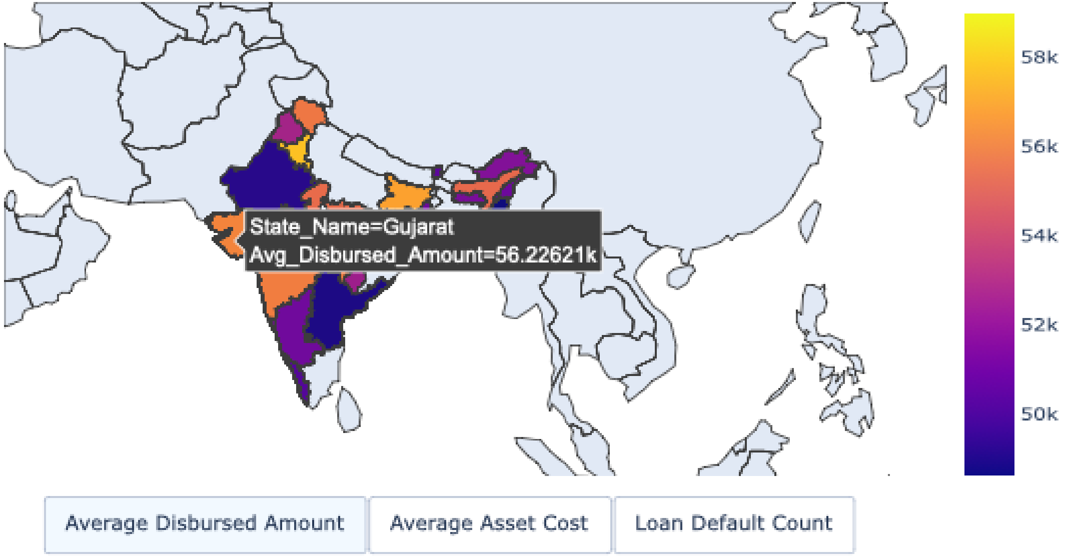
Colors Indicating Loan Default Status:

Red Circles: Represents borrowers who did not default on their loans.

Blue Circles: Represents borrowers who defaulted on their loans.

This chart illustrates that while new accounts are regularly opened, a smaller proportion of them turn delinquent or default within the first six months. This indicates effective identification by the lender to manage potential default risks.

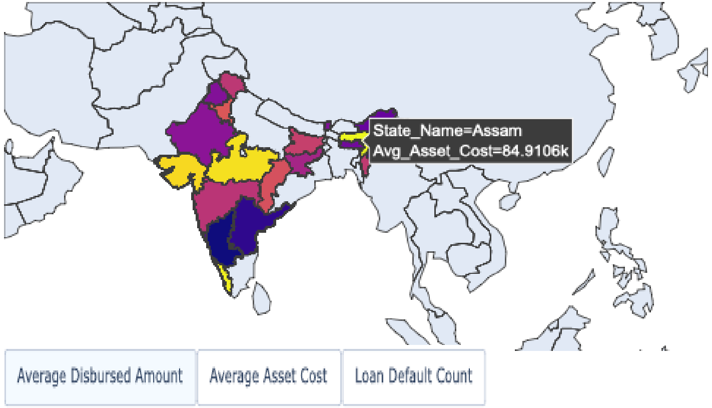
Visualizations of data(9 – 1/3) – Geographical Analysis



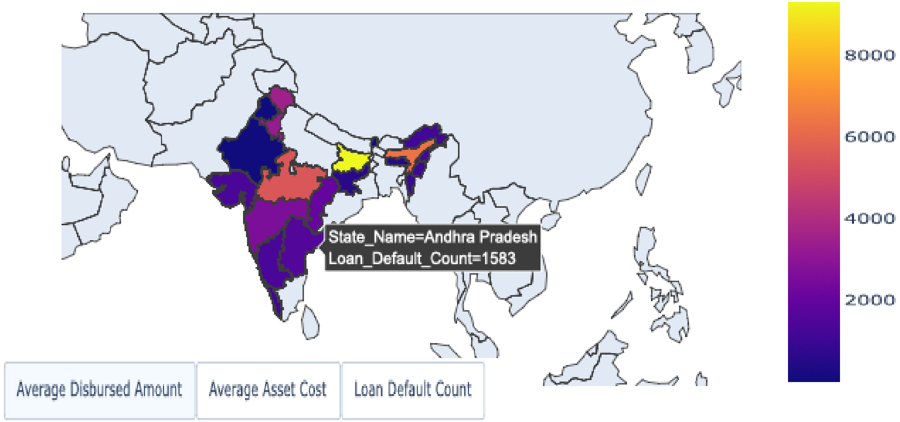
A choropleth map to show over the states of India.
It show the average distributions of disbursed amounts, asset costs, loan default counts when we select any of the 3 radio buttons below.
This will help a common person to understand the risk associated with each state for a general idea.

This visualization particularly shows average disbursed amount of Gujarat when the cursor is hovered on the outline of that state.

Visualizations of data(9 – 2/3) – Geographical Analysis

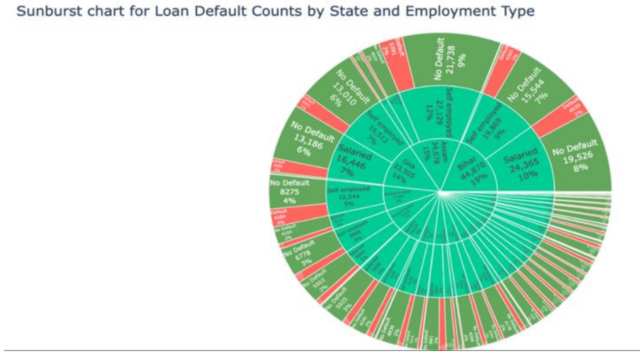


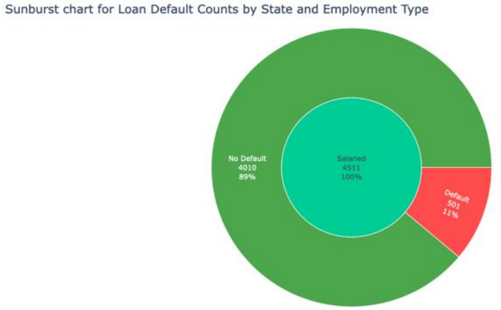
This image shows average asset cost of Assam when the cursor is pointed on the outline of Assam state and the Average_asset_cost radio button has been selected.

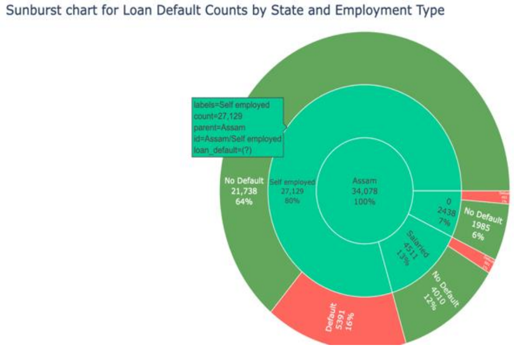


This visualization is when loan_default_count of Andhra pradesh is selected to see, and it displays the exact count which is 1583.

Visualizations of data(10)- **Sunburst Map: Hierarchical Visualization**







a) This sunburst chart extends the analysis nationwide, presenting loan default counts across all states and employment types. It reveals diverse patterns of loan defaults and employment type distributions across states. States are arranged in a hierarchical way that show larger counts of loan accounts, with a significant proportion marked as "No Default."

b) This chart focuses on a single state, Assam, and its loan default distribution by employment type. The chart highlights that 80% of loan accounts belong to self-employed individuals, followed by salaried individuals (13%) and others (7%). Among these, most accounts exhibit "No Default," with only a small proportion marked as "Default."

c) The third chart narrows the focus to a specific employment type: salaried individuals. The sunburst chart depicts that 89% of loan accounts for salaried employees have "No Default," while only 11% are classified as "Default."

Conclusion

- We have analyzed important factors influencing loan defaults, data on borrower demographics, loan details, financial history, and identification metrics.
- We identified trends and correlations within the dataset, shedding light on high-risk groups and potential warning signs. These tools enabled us to assess credit risk intuitively, paving the way for more informed and proactive lending decisions.
- Choropleth maps visualized geographical trends, identifying regions with higher default rates and favorable lending conditions.
- Radar charts, stacked graphs, donut charts, and scatter plots, to assess credit card scores and the loan default percentages for a broader insight are used.
- We have observed many trends across the data that seems very reliable when it comes to loan approvals.

Future Scope

- We could incorporate machine learning models to predict loan defaults more accurately based on real time data and emerging trends present it in a dashboard with much interactive features.
- Geospatial analysis can be visualised more accurately when we can include area-wise codes, incorporating demographic segmentation. Time series visualizations can be used for temporal analysis.
- We could include market trends or economic indicators, for a more comprehensive risk evaluation frameworks.

Questions

THANK YOU