Interactive InfoVis Using User Centric Design

A Design Study Approach Using Kaggle Dataset

Project Summary:

The goal of this project is to do Exploratory Data Analysis (EDA) and Visualizations (Infovis) to gain insights from the Credit Card transactions dataset from Kaggle. By doing an EDA on this dataset we will be able to identify the patterns of finding fraud in Credit card transactions. Furthermore, we will be able to build Fraud detection models and minimize the fraud losses for customers and the Credit card companies. This is a very interesting problem because we are dealing with an unbalanced dataset. We will follow the step-by-step process outlined in DTSA-5304.

Most of the theoretical concepts in this document are accompanied by InfoVis. 70% of the time was spent on data wrangling so that we can use the data for visual encodings needed to answer our questions. My notebook uses the Altair that we learnt, seaborn and other python imp libs. I also did a peer review with two of my close friends to complete the Validation section.

Project Files: The project is uploaded to GitHub.
https://github.com/rajesh4aiml/DTSA5304-Fundamentals-of-Data-Visualization

Our Users and Domain Info:

In Machine Learning, this is a Binary Classification problem, since there is one target or dependent variable named 'IS_FRAUD' (True or False). The **Data Scientist's,** who are our users, will build a **Supervised ML model** to detect Fraud. Our **Infovis** will help the Data Scientists to identify the important independent or predictor variables that are key

contributors to fraud. They will leverage the insights gained from our Infovis to do Feature Engineering and to train and build ML models.

The main question for this domain is - Which are the most important independent variables for Fraudulent Transactions?

Note - Not all independent variables are important. For e.g. Person's name is not important to identify if a transaction is fraud, however, Merchant name is important variable, because some Merchants may be bad actors.

Dataset Info:

Kaggle dataset: **Credit Card Fraud Prediction**

URL: https://www.kaggle.com/datasets/kelvinkelue/credit-card-fraud-prediction/data

Keynote: This is a large dataset for Exploratory Data Analysis. This dataset offers a variety of attributes valuable for comprehensive analysis. It contains 555,719 instances and 22 attributes, a mix of categorical and numerical data types. Several techniques such as strategic sampling and aggregations were used to produce our InfoVis.

Variables: As part of feature engineering to draw more insights, following variables were created-

Variable Type	Details
Original (Independent)	trans_date_trans_time', 'cc_num', 'merchant', 'category', 'amt', 'first', 'last', 'gender', 'street', 'city', 'state', 'zip', 'lat', 'long', 'city_pop', 'job', 'dob', 'trans_num', 'unix_time', 'merch_lat', 'merch_long',
New Features	tran_date: The transaction date in format yyyy-mm-dd tran_month: Numeric transaction month tran_month_nm: transaction month name. age: Age at the time of transaction. Derived using (tran_date – dob) lst_day_diff: Day difference between the current transaction and previous transaction lst_amt_pct_chng: Increase/decrease in amount between the current transaction and previous transaction.
Original (Dependent)	is_fraud

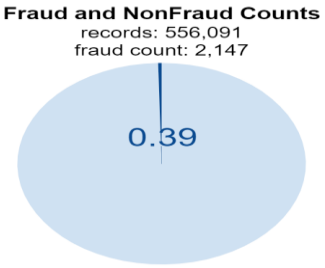
Supplementary datasets:

Spatial file used for maps: https://www.census.gov/geographies/mapping-files/time-series/geo/carto-boundary-file.html

Section 1: Goals & Tasks

S1.1: Why is the task pursed?

The Credit Card Fraud Transaction dataset is highly imbalanced. 0.39 percent (Less than 1 percent) are Fraudulent Transactions.
To build a fraud detection model, we want to understand the relationship between the **predictor**(independent) variables and **target** variable using Infovis. Our InfoVis, will provide insights to our Data Scientists in building a predictive model.



S1.2: How is a task conducted?

We will define the means by which people work with the data.

Navigation:

Our users are data analysts and data scientists. They will be visually processing the data.

We will create interactive charts for them. They will use interactive controls to identify new information. The visual encoding below will demonstrate the same.

Organizing:

We will perform various transformations on the data before applying the visual encodings. We will filter the data, aggregate the data, sort it. We will build new knowledge and new characteristics. The Visual encodings below will demonstrate the same.

Relation:

We were able to find the relations that emerged through our distributions & correlations. Please see details in the sections below.

S1.3: What does a task seek to learn about the data?

Relationship between dependent and independent variables. We want to be able to classify a transaction as fraud from a large dataset, so we need a deep understanding of the independent variables in this dataset.

S1.4 What characteristics, patterns, or properties matter?

We will present traditional statistics in this section.

Low level characteristics

Ideally, we can check basic stats such as min, max, mean, etc. for important independent variables.

In Sections 5.2, I have shared some distributions.

High Level characteristics:

Trends in the data: We created a time series chart (Fig 5.1A, 5.1B below) to identify the trend in the data.

Trend with different filters: We have also allowed users to click and find details.

Section 2: Target Data

S2.1 Where does the task operate?

Since this is a case study, we have downloaded the data from Kaggle. In real-life scenarios, we will have to collect data from different sources using data pipelines. We will store data in data lakes and apply various transformations to clean the data before starting the Analysis. The clean data will be sourced from Data Lake which has complete lineage. We can use Cloud Vendors such as Azure or AWS or Google.

S2.2 What target data do we need to work with?

We will need the Credit Card Fraud Transaction data. The data was available in Kaggle.

Kindly note, you will need to sign up to download the data. I have added a copy to my GitHub for testing this project.

For Geo spatial views, we can take the data from the US Census.

S2.3 What frames of reference are we working within?

Absolute reference frame:

We will be looking at different States in the USA and try to discover fraud patterns within USA.

Relative reference frame:

In some cases, we have to compare Fraud and Non-Fraud transactions to check if any new patterns emerge.

Section 3: Workflow

S3.1: Formulate Main Question: (Purpose of Analysis)

Which are the **most important independent variables** for Fraudulent Transactions?
Intuition: The motivation behind this generic question is: We are building a prediction model to detect fraud from an unbalanced dataset (Many nonfraud transactions vs less than 5% fraud transactions). In this dataset, there are many variables such as person-name, card-number, age, transaction-amount, etc. Person-name and address does not help us in our analysis; hence we need to pick the right predictor (independent) variables that will help us predict the target variable: **is-fraud.**

To answer this question, we had to answer many other questions for generating new insights.

The entire brainstorming processes led to snowballing effect.

S3.2: Preparing the data (Data Wrangling):

We need the data to be in the right shape and format to do the Visual Encodings.

The Data Cleanup tasks included:

A) Profiling on the dataset after cleaning the dataset.
B) Identify the variable types as categorical, Boolean, numerical – continuous.
C) Do univariate, bivariate and multivariate analysis on this dataset.
D) Null value imputation and cleaning the data.

The data analysis also led to intuition that resulted in asking more questions on the Domain.

S3.3: Snowballing Effect:

The main question above led to very detailed questions as below. The second column is Answered Y/N?

#	Question	Ans	Reference
1	Timeseries - Did we observe any trends in the fraud transactions? Are we on an increasing or decreasing fraud trend?	Y	Fig 5.1A, 5.1B
2	In timeseries, we noticed there were spikes in amounts – the related question is: Will change in amounts (spike in Customer spending) lead to Fraud?	Y	Fig 5.2A, 5.2B, 5.2C
3	Which states have the most fraud?	Y	Fig 5.5A, 5.5B
4	Which fraud merchants (bad actors who deceive customers)?	Y	Fig 5.6A, 5.6B, 5.6C
5	Which age group are more victims of Fraud?	Y	Fig 5.4B
6	Which categories have the most Fraud?	Y	Fig 5.4A
7	Is Gender contributing to the fraud trends?	Y	Fig 7.1A

S3.4 Tasks

Provenance –

Here we define the history of tasks being conducted. The list of tasks include:

Data Wrangling:

Based on 4Vs of data, Data comes from various sources, shape, form and volume. The sub-tasks are:

- Clean up the data,
- Missing value/null value imputation.

- Review the different dimensions in the data.

Exploratory Data Analysis

The EDA Tasks include

#	Tasks	Reference
1	**Timeseries Analysis**: Checking how the distributions for different variables across an absolute reference frame	S5.1A, S5.1B
2	**Univariate Analysis**: Understanding one variable at a time to gain deeper insights about it	S5.2
3	**Bivariate Analysis:** Understanding the relationships between two variables.	S5.3
4	**Multivariate Analysis**: Understanding relationships between multiple variables. We can remove strongly correlated variables during this analysis. For e.g. MileageInKm, MileageInMiles is same.	S5.4

Section 4: User Centered Design Process

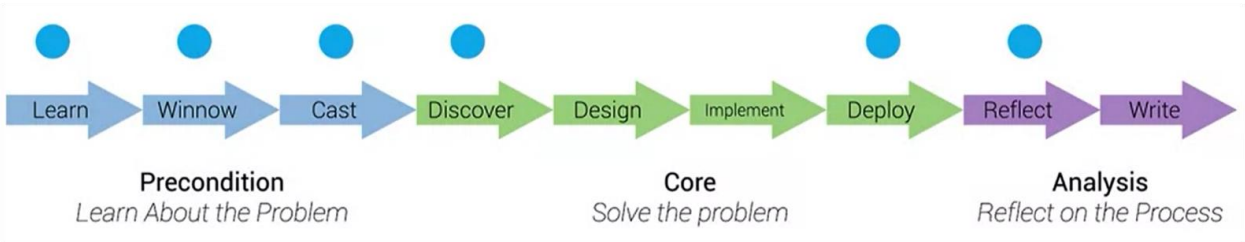
We will follow the iterative process for designing this InfoVis.

Design Studies Approach: A project that analyzes a real-world problem in order to design a visualization system that contains a validated design where designers reflect about lessons learned.

Wicked problem:

In our case, we don't have a wicked problem.

Linear model: determinate problems have definite conditions – We should identify conditions and design solution.



Learn: We need to gain insights into this domain. For e.g. How often does Fraud happen? What are the characteristics of Fraud Transactions?

Winnowing: We need to collaborate with the experts in this field. We need to spend time with them to check what techniques they used and what were their pain points. We need to connect with Customer Service Agents who is servicing fraud and gain some insights about the customers reporting fraud.

Cast: Next we form the team to execute our Tasks. Here we need critiques and friends. The team should be respective and goal oriented and not subjective and personal.

Discover: In the workflow above, we have characterized our problem. We need to keep doing analysis to answer all the questions and possibly come up with a new set of questions.

Design: Detail design for each iteration is provided below. These are low fidelity prototypes. Jupyter notebook and python libs such as Altair, Seaborn and Geopandas was used.

Implement & Deploy: Are not in scope for this case-study. In this part, we plan to productionalize the application.

Reflect: Confirm Hypothesis

Here we check if we learnt something new by checking the questions we asked and reflecting on it by checking if our questions were answered and if new questions need to be asked.

In the last few weeks, I refined the questions many times to come up with more streamlined questions.

Write:

This document lists the summary, findings, approach, EDA, design choices made and conclusions that will help the stakeholders.

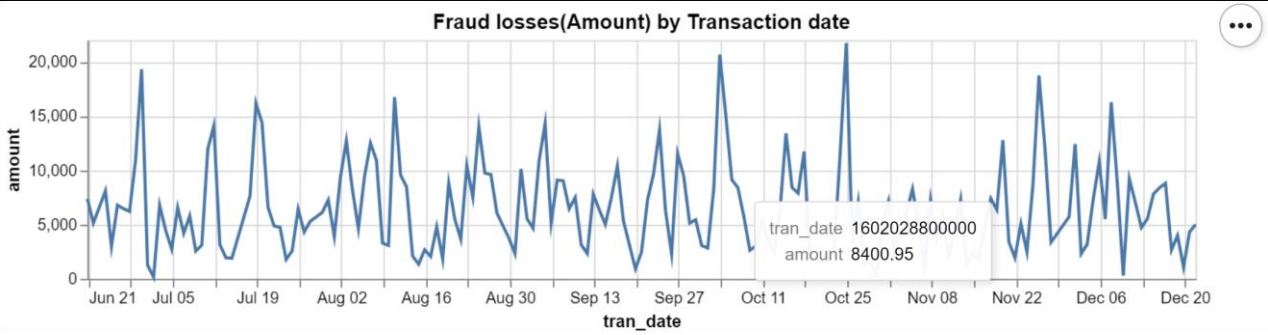
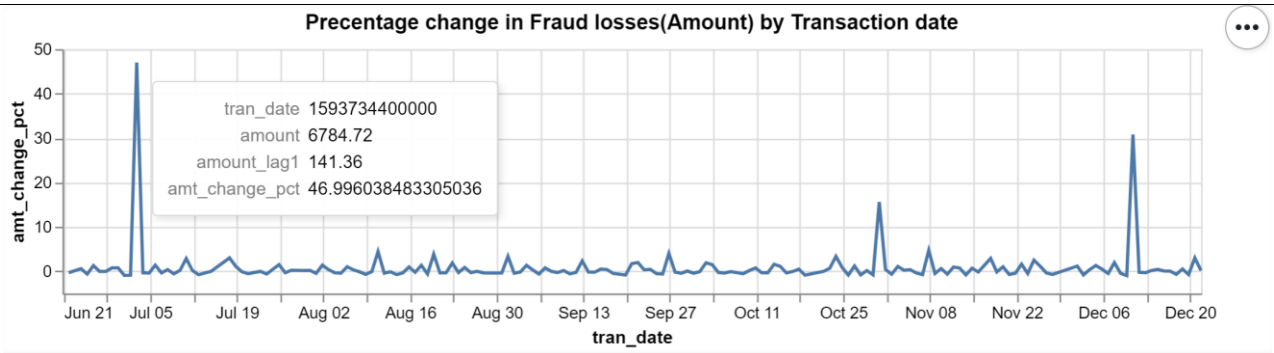
Section 5: Exploratory Data Analysis using Altair, Seaborn, Geopandas

S5.1A - Timeseries Analysis – What are the daily trends in fraud?

Observations: Fig 5.1A below has fraudulent transactions by date and amount. The users can clearly see the fraud loses across absolute reference frame. There were periodic spikes in the amounts, hence we would like to understand the percentage increase in fraud losses per day.

S5.1B - Timeseries Analysis – What would the trends look like if we changed the Y-axis to display the increase in amount per day?

Observations: Fig 5.1B provides more insights into the increase in fraud losses. This generates more insights for our case study. We would like to zoom into specific transaction dates and explore details to find more insights.

Fig 5.1A	 These chart shows only Fraudulent transactions by date and amount. X axis: Transaction date Y axis: The total transaction amount.
Fig 5.1B	 Describes only Fraudulent transactions by date and percentage increase in amount. X axis: Transaction date Y axis: The change in amount from previous day (prev_amt – curr_amt)/prev_amt

S5.2 Univariate Analysis - Plotting distributions for Amount and Change in Amount

Related question: Will change in amounts (spike in Customer spending) lead to Fraud?

We need to understand how the amount is distributed for Fraud Transactions.

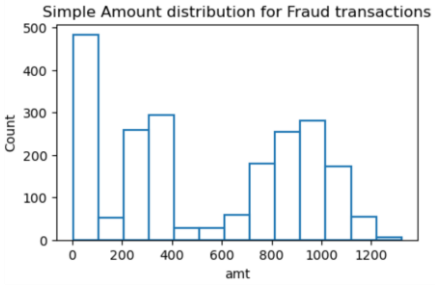
Change in

Observations: Visual encodings 5.2A and 5.2B are similar except that 5.2B shows the distribution of amounts by Categories.

5.2B shows Groceries category has more fraud than other. DS can choose to do deeper analysis on this category.

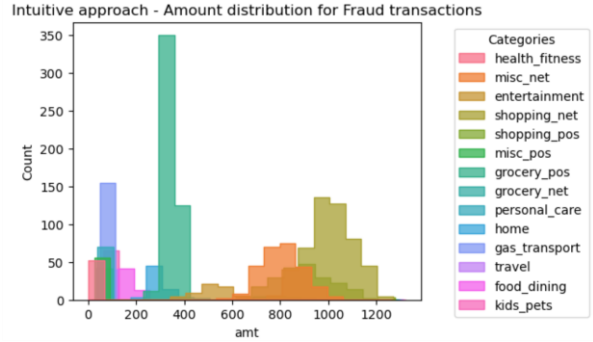
5.2C Displays distribution for variable: lst_amt_pct_chng. This variable is the percentage change between the current and last transaction. We noticed that most of the customers have between 0 to 20 percent change in amounts.

Fig 5.2A
F 5.2B



X: count of fraud transactions
Y: amount bins.

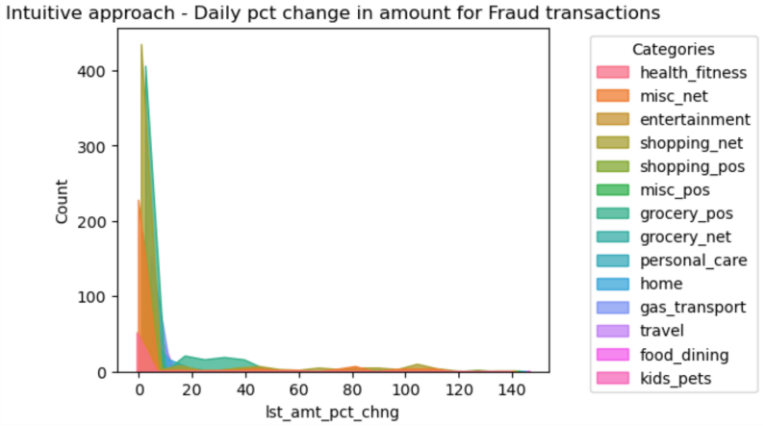
Fig 5.2A
Traditional Seaborn chart



X: count of fraud transactions
Y: amount bins.

Fig 5.2B
Altair Layered Distribution

Fig 5.2C



S5.3 Bivariate Analysis – Comparing two variables.

A correlation matrix is simply a table which displays the correlation coefficients for different variables. The matrix depicts the correlation between all the possible pairs of values in a table. It is a powerful tool to summarize a large dataset and to identify and visualize patterns in the given data.

A correlation matrix consists of rows and columns that show the variables. Each cell in a table contains the correlation coefficient.

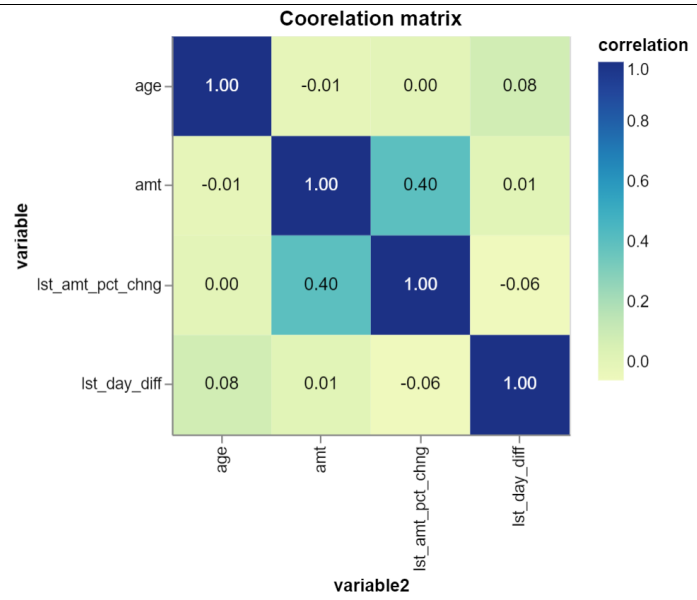
Value	Meaning
-1	Two variables have an inverse linear relationship: when X increases, Y decreases
0	No linear correlation between X and Y
1	The two variables have a linear relationship: when X increases, Y increases too.

Note: Correlation does not imply causation. When correlation between X and Y is close to 1, we cannot say that a change in X implies a subsequent change in Y.

Observations: The variables compared were age, amount, last-amount-pct-change (Customers amount spending change percent from previous transaction), last-day-diff (Days since last transaction).

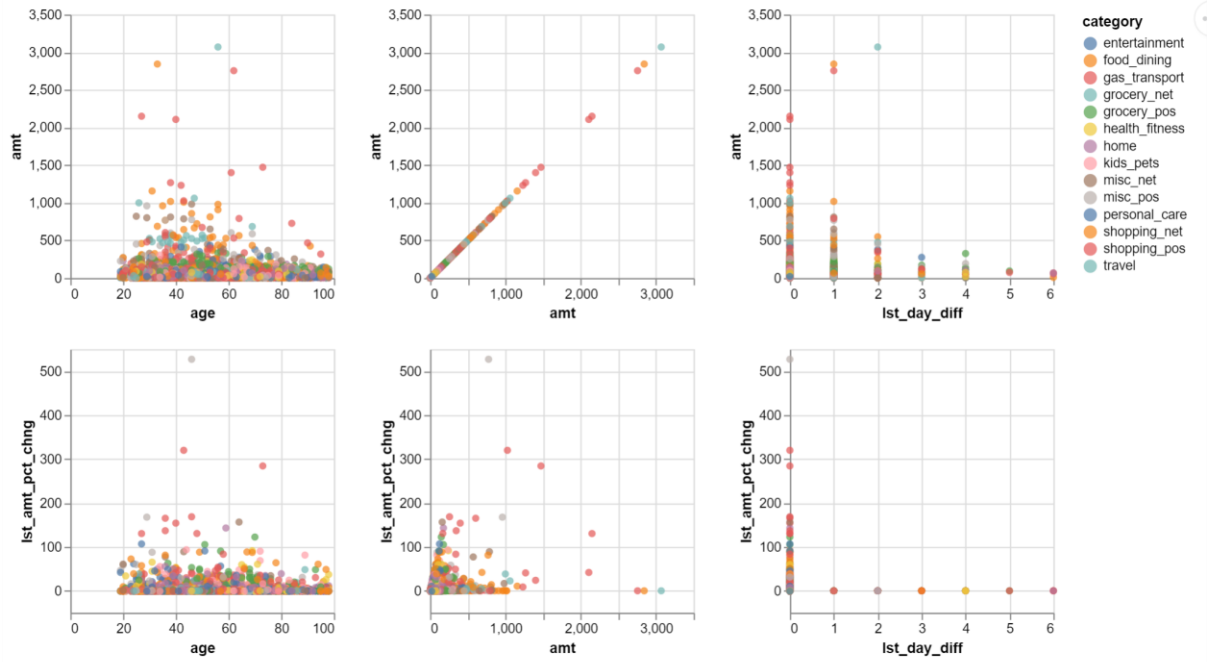
We used two visual encodings. Traditional approach (correlation matrix) displayed as Altair rectangles and Graphical approach. The Graphical approach with hue provides more insights into the variables.

Fig 5.3A



Relationship between variables: age, amount, last-amount-pct-change, last-day-diff

Fig 5.3B



Relationship between variables: age, amount, last-amount-pct-change, last-day-diff

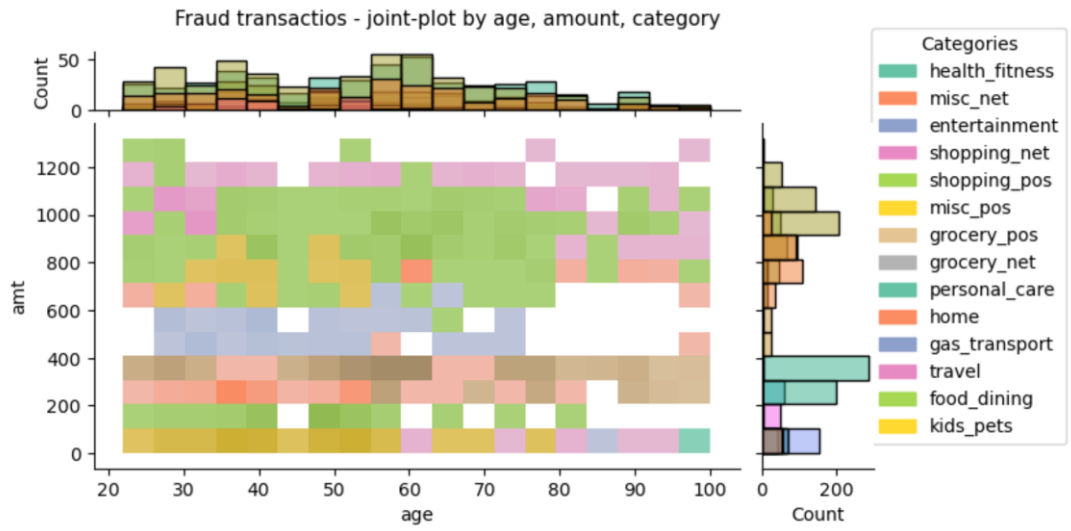
S5.4 Multivariate Analysis – Comparing two variables.

We compare multiple independent variables and their relationships to one another using joint distribution charts.

Observations: The visual encoding in 5.4 A is **intuitive** because we can see the distributions for both variables.

Users can figure out the categories that lead to most fraud. In this case shopping_pos has more transactions and shopping_net has a higher amount but less transaction. We can get deeper insights in Fig 5.4B where we compare three variables using an interactive chart. When clicked on a category it will display rectangular plots. This was very useful to users.

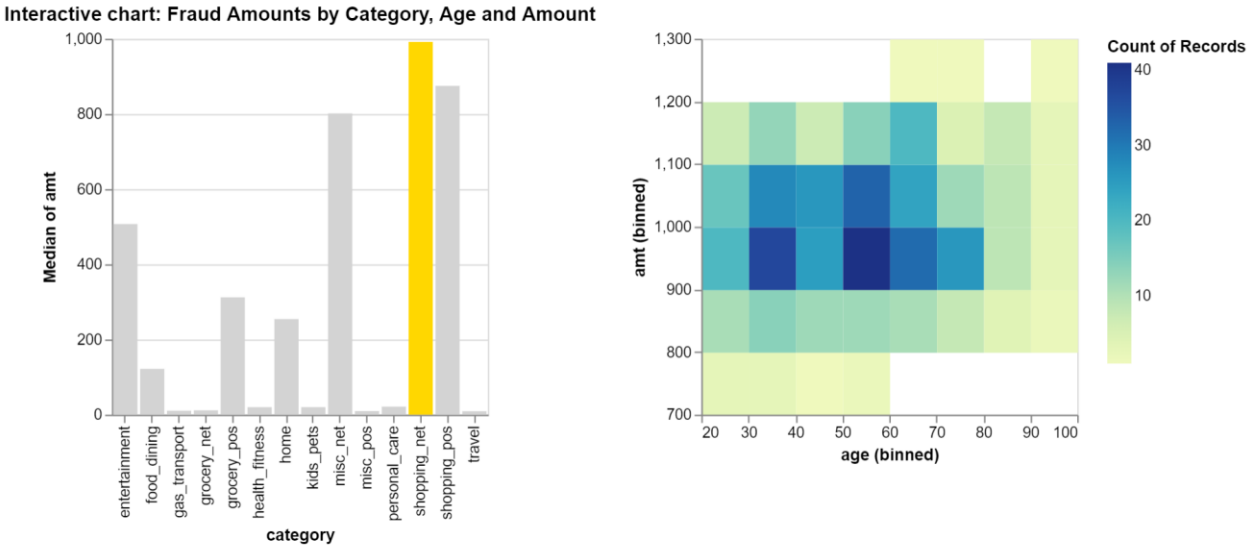
Fig 5.4A



X axis: age

Y axis: amount

Fig 5.4B



Clicking on a category will show the age, amount and count of transaction in color.

Interactive Left chart: X-axis: category, Y-axis: median amount

Right chart: X-axis: age binned, Y-axis: amount binned

S5.5 Interactive Charts – Analysis of Variables, State, Amount, Categories, Transaction date.

Related question: Which states have the most fraud?

In this chart, we want to gain insights into variables: amount, date, category and transaction month.

We allow the user to pick a state and further explore details about the state, how the monthly fraud amount looks like for each category.

Observation: In Fig 5.5A and 5.5B, NY was an outlier. In 5.5 B when NY was selected, the chart on the right with tooltip displays the spending for an outlier point.

This data helps Data Scientists to track **anomalies**.

Fig 5.5A

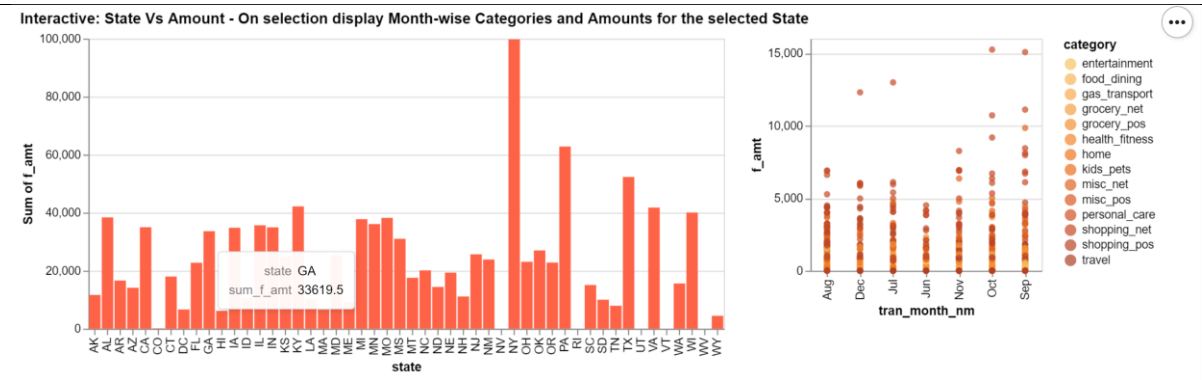
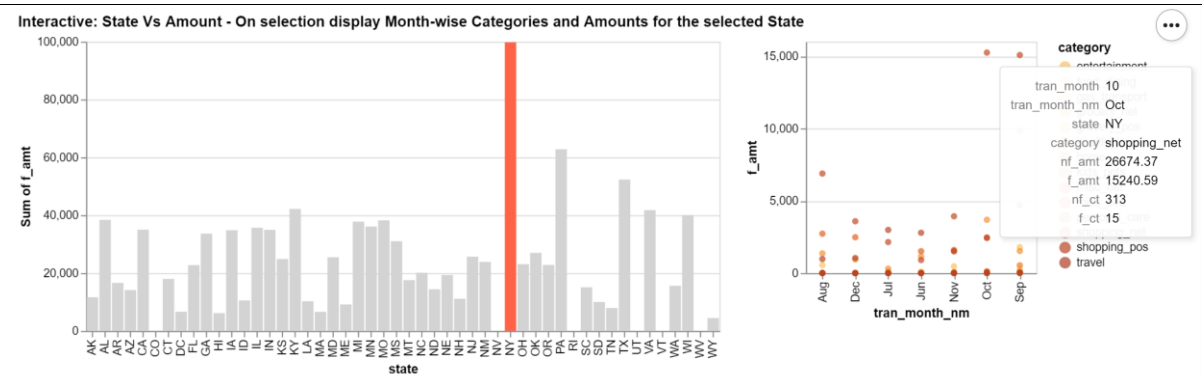


Fig 5.5B



Left chart: X-axis: state, Y-axis: fraud amount

Right chart: X-axis: month, Y-axis: fraud amount

The right chart displays data for the state selected on left.

S5.6 Interactive Charts – Analysis of Merchants.

Related question: Which fraud merchants (bad actors who deceive customers)?

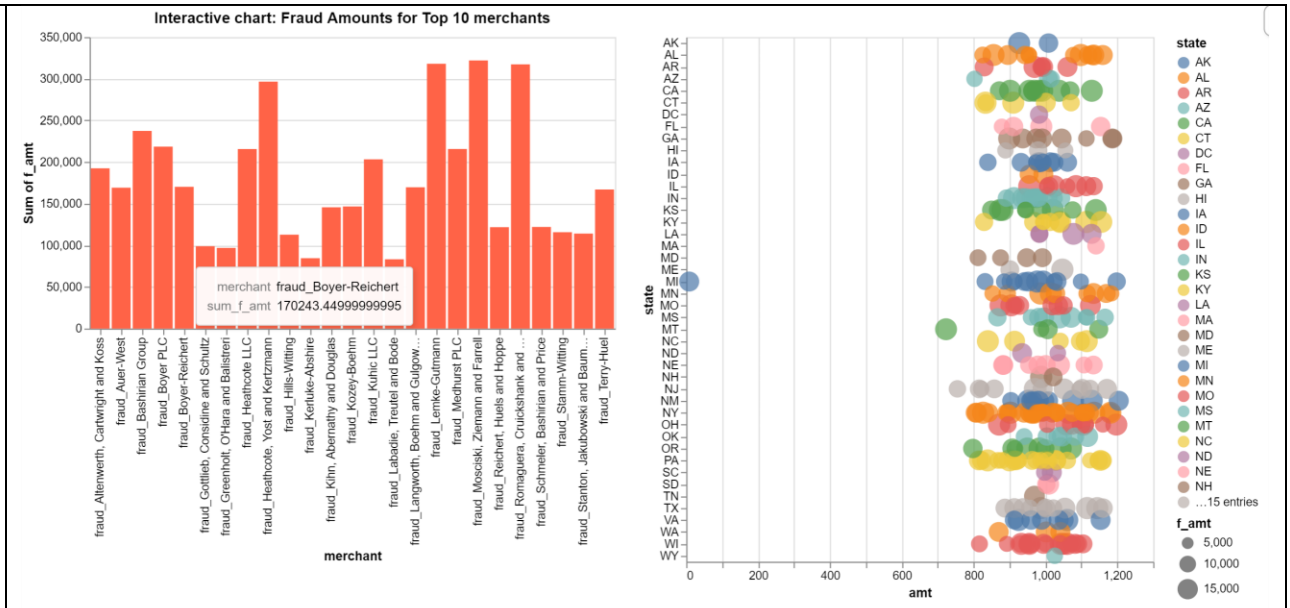
Here we took the Top 25 merchants and created a subset dataset. The left chart shows the Merchant and the total amount in fraud. When the Merchant is selected, the right chart will show the state where the merchant operates, the transaction amount and the categories for which fraud was reported.

Fig 5.6A and 5.6B are color coded by state; In Fig 5.6B When we select one or more merchants, we can see where that merchant operates and their transaction amounts.

Fig 5.6C is color coded by Category and shows the amount for category on the right.

Note: If we considered all transactions instead of Top 50 bad merchants, we would have got different bubble sizes.

Fig 5.6A



Left chart: X-axis: Total fraud amount, Y-axis: Merchant name

Right chart: X-axis: Amount, Y-axis: state.

Fig 5.6B

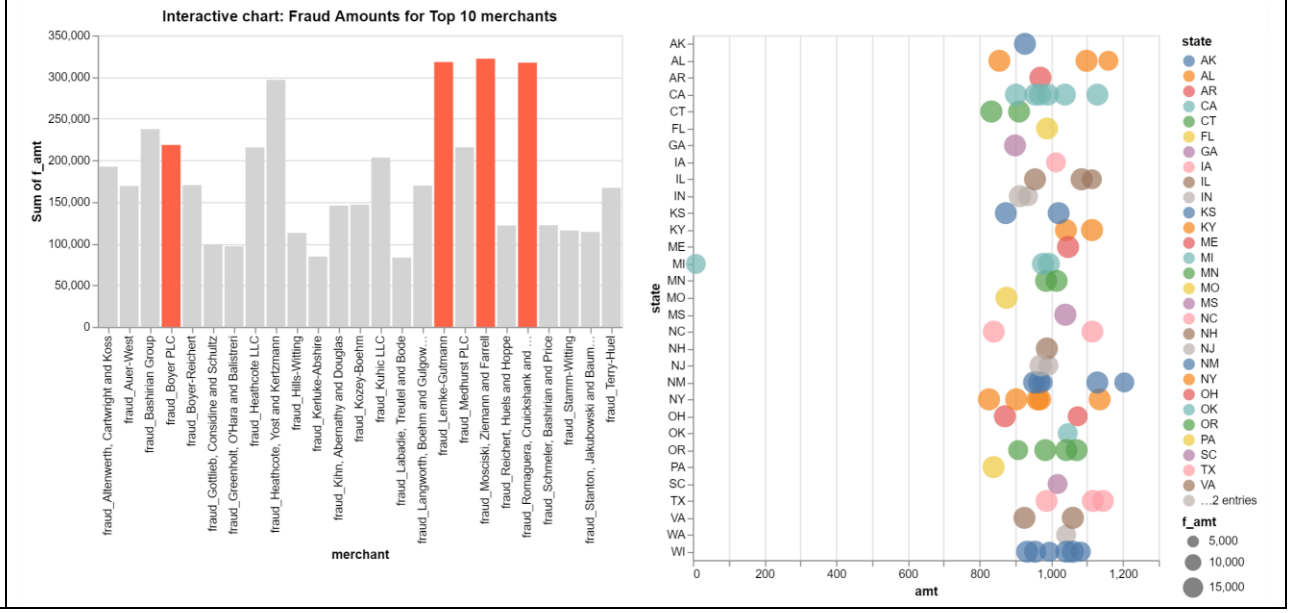
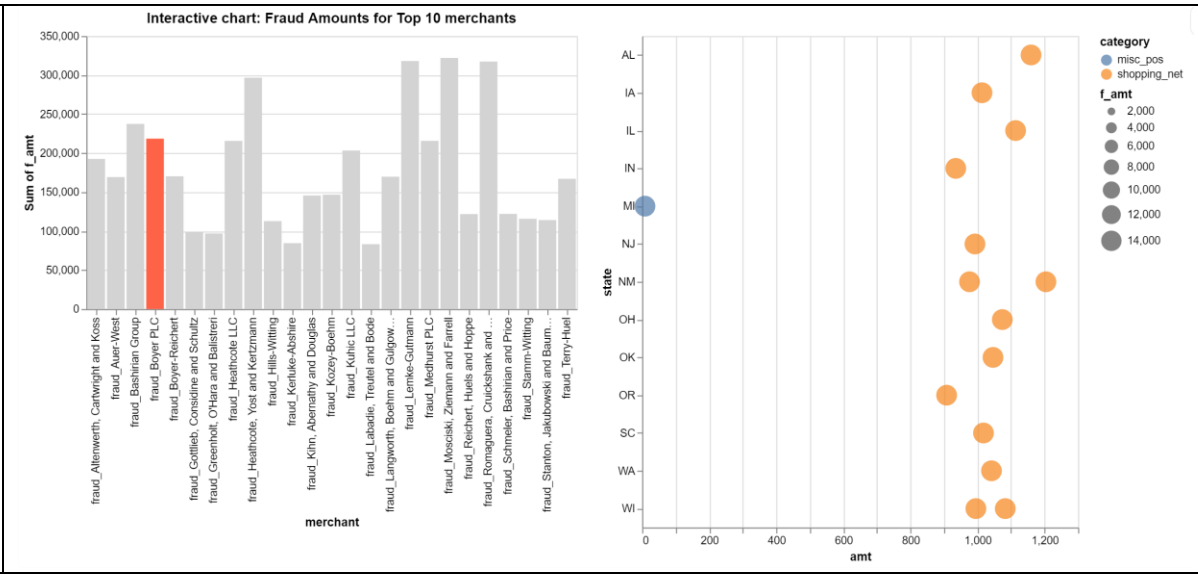


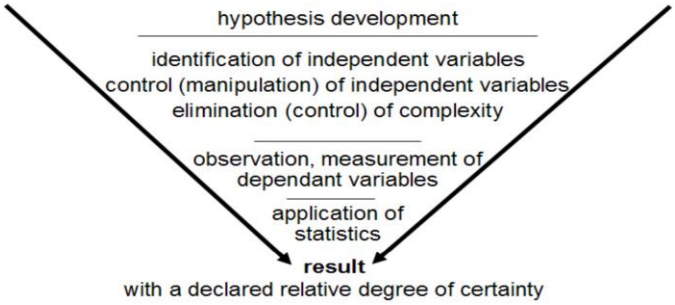
Fig 5.6C


Interactive chart: Fraud Amounts for Top 10 merchants
Sum of f_amt
350,000
300,000
250,000
200,000
150,000
100,000
50,000
0
fraud_Altenwerth, Cartwright and Koss
fraud_Auer-West
fraud_Bashirian Group
fraud_Boyer PLC
fraud_Boyer-Reichert
fraud_Gottlieb, Considine and Schultz
fraud_Greenholt, O'Hara and Balistreri
fraud_Heathcote LLC
fraud_Heathcote, Yost and Kertzmann
fraud_Hills-Witting
fraud_Kerluke-Abshire
fraud_Kihn, Abernathy and Douglas
fraud_Kozey-Boehm
fraud_Kuhic LLC
fraud_Labadie, Treutel and Bode
fraud_Langworth, Boehm and Gulgow...
fraud_Lemke-Gutmann
fraud_Medhurst PLC
fraud_Mosciski, Ziemann and Farrell
fraud_Reichert, Huels and Hoppe
fraud_Romaguera, Cruickshank and ...
fraud_Schmeler, Bashirian and Price
fraud_Stamm-Witting
fraud_Stanton, Jakubowski and Baum...
fraud_Terry-Huel
merchant
state
AL
IA
IL
IN
MI
NJ
NM
OH
OK
OR
SC
WA
WI
0
200
400
600
800
1,000
1,200
amt
category
misc_pos
shopping_net
f_amt
2,000
4,000
6,000
8,000
10,000
12,000
14,000

Section 6: InfoVis Evaluation Techniques

We used Experimental Evaluation approach for this case study.

The experiments or studies involve a rigorous process of hypothesis development, identification and control of the independent variables, observation and measurement of the dependent variables, and application of statistics which enable the declaration of the confidence with which the results can be taken.



S6.1 Survey

Based on **Experimental Evaluation**, our findings are as follows

Survey Metric	Our Findings
Assess the domain problem the visualization solves Do the stakeholders concur with your assumptions?	Yes
Determine the data and tasks your users have Can people build the knowledge they need?	Yes
Choose the right encodings for your domain, data, and tasks Can people see the patterns they need correctly?	Yes
Build an algorithm to map the data to the encodings Does the algorithm perform correctly?	Yes
Design interactions to explore and refine data Can people quickly and intuitively interact with the data?	Yes

S6.2: Approach for Validations:

Insight-Based Evaluation Qualitative Evaluation	Experimental Evaluation Quantitative Evaluation
Knowledge: What new knowledge can my users gain? Complex deep understanding	Performance: - Does my visualization work better than other methods? controlled approach, measure performance
+ Holistic input + Grounded in real-world use + Richer data - Less controlled - Slower - Less precise	+ Highly precise + More generalizable + Easier to understand - Less specific - Less detailed - More abstract

S6.3 Experimental Evaluation-1

Hypothesis - Spatial data is the best.

For this experiment, I considered "**within the participant**" evaluation approach. This is geared toward user centric design and testing.

Independent variables -

1. Interactive maps with plots: (Fig 6.1 and 6.2). The VAL1.1 section below has the geo-spatial visual encodings that can help users to understand the fraud rates across United states for absolute reference frame (USA, 2020)
2. State Vs Amount Interactive bar chart: (Fig 6.3 and 6.4). The VAL1.2 section below has a State selection option by which the user can display month-wise Categories and Amounts for the selected State.

Dependent Variables:

1. **Accuracy** - The users were able to understand the infovis and draw the right conclusions.
   From the Observations sighted in S6.4, we were able to answer the question asked by the user in the most effective way.
2. **Time**- The time taken to navigate was short since a demo was given and then the user was able to use interactive controls to navigate and discover new relations. In our case, the time cannot be justified since our dataset size is small.

Control Variables:

SME's - We rely on subject matter experts (SME) who have deeper knowledge about Credit card networks such as Visa and Master cards and they have in-depth knowledge about such applications. They are field experts, and they are familiar with patterns of Fraudulent transactions.

In my case, I knew few SME's who have worked in Financial Services.

Dataset - We are working with synthetic data so the descriptive and Inferential Stats was useful.

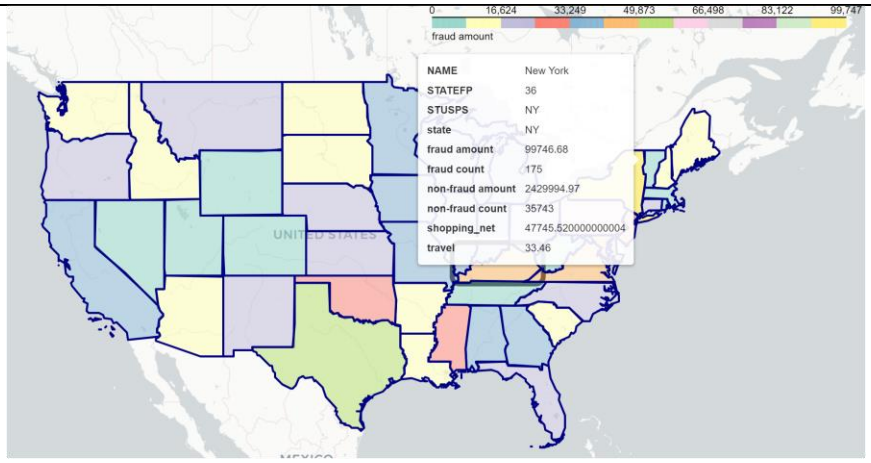
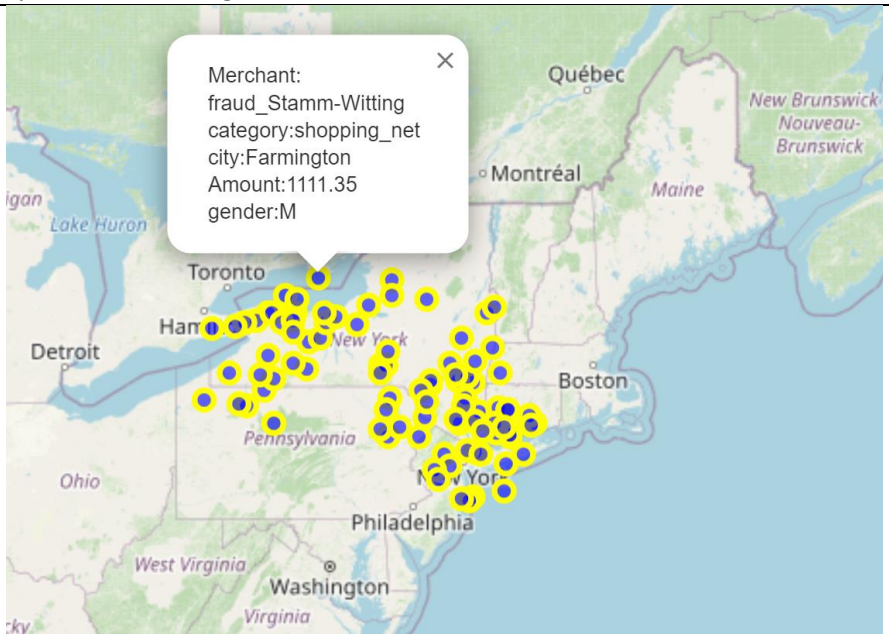
S6.4 Observations: I presented two types of Visual Encodings for Validation a) VAL1.1 b) VAL1.2. Based on the feedback from my users, it was considered that the interactive chart **Fig 6.4C and Fig 6.4D** was more useful to draw a conclusion. It was **accurate** and answered the question. The interactive bar chart also provided more insights into the state, its categories and amount per category for each month. Thus, the Hypothesis that Spatial data is the best was incorrect. Regardless, we learnt about our design from the user's perspective.

Furthermore, the following validations were done, and the results selected by user with reasons is as follows:

#	Encoding compared	Results/Selected	Reason
1	Fig 5.1A Vs Fig 5.1B Timeseries plots	5.1 B	Percent change was important
2	Fig 5.2A & Fig 5.2B Distributions	5.2B	The color encoding helped with insights
3	Fig 5.3A & Fig 5.3B Bivariate Analysis	5.3B	Visual stimuli are better than tabular info.
4	VAL1.1 Vs VAL1.2	VAL1.2	Provided better insights.

VAL1.1: Aggregate Analysis– Which states have the most fraud?

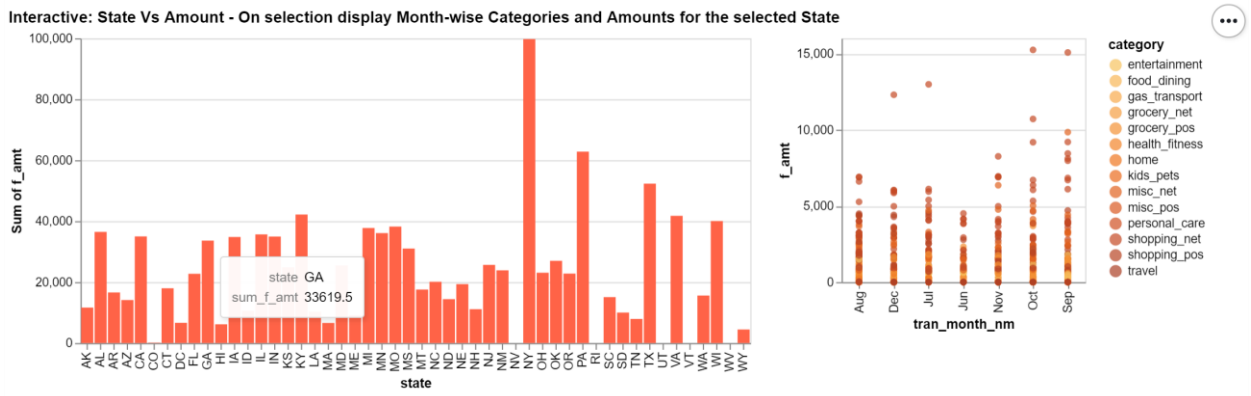
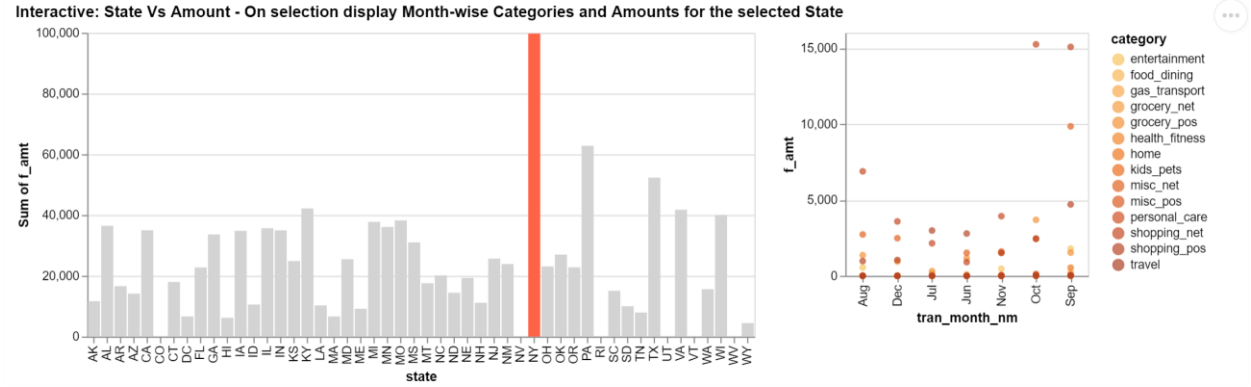
Intuition: This chart displays Geo-spatial data. The colors are based on the fraud amounts.

Fig 6.4A Default View	 Note: The Axis is hidden. When hovered over New York, it shows a drop down with additional columns such as Fraud amount, Non-Fraud amount, categories, etc. User does not have to spend time clicking; thus, time is saved.
Fig 6.4B Zoom	 Users can zoom-in into NY and check the ctiy and category that contributed to the fraud. The chart above shows the tooltip for one of the selected cities in NY.

VAL1.2: Aggregate Analysis– Which states have the most fraud?

Intuition: Here is alternate view of similar info. This is an interactive altair chart.

Left chart shows state on X axis and Amt on Y.

Fig 6.4C Default View	 Interactive Left Chart: X axis: State abbreviation. Y axis: Total fraud amount. Right Chart: X axis: Tran Month. Y axis: Total Fraud Amount. The Left chart displays Fraud Amount distribution across all the States. Right chart, by default, displays the Fraud amount for USA for each month for each category that is displayed in Color.
Fig 6.4D User selected **NY**	 Left/Right Chart: Axis same as Fig 6.4C Interaction: User clicked on state: **NY** which seems to be an outlier. The Left chart displays Fraud Amount distribution for **NY** with other states faded. The right chart displays the Fraud amount for **NY** for each month for each category that is displayed in Color.

Section 7: Uncertainly & Inferential Statistics

In many cases where there are large volumes of data due to which sampling techniques are used.

When the data volumes or datasets for analysis are large, sampling techniques are used. While sampling we want to ensure that the sample represents the true data. Biased samples will not help us in solving the problem.

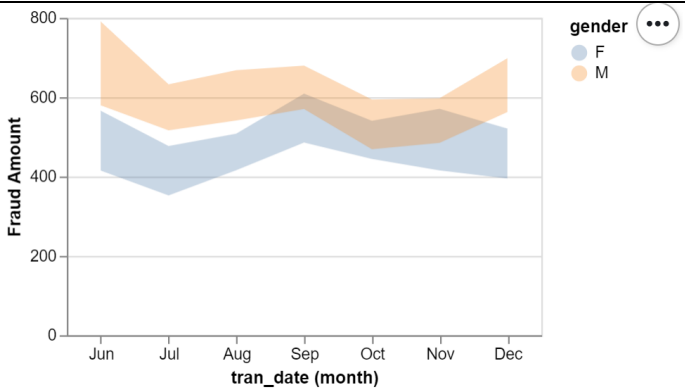
A good approach to address uncertainty is to use **Inferential Statistics (**Confidence Intervals, Z-Score, -tests, Violin plots).

S7.1: Aggregate Analysis– Is Gender contributing to the fraud trends?

Ans: If we can't use the entire dataset for our InfoVis, we have to use the charts like below to model Uncertainty for answering our questions.

Instead of showing the actual line, we show areas for Gender as below.

The 95% confidence interval bands help us to deal with uncertainty in visual encoding.

Fig 7.1A	 X-axis: The Fraud Transaction date, Y-axis: The Fraud Amount, Hue: Gender Encoding used: ci0: The lower boundary of the bootstrapped 95% confidence interval of the mean. ci1: The upper boundary of the bootstrapped 95% confidence interval of the mean. Reference: Encodings — Vega-Altair 5.3.0 documentation (altair-viz.github.io)

Section 8: Conclusion

We helped our Stakeholders and Data Scientists by generating new insights through our Visual Encodings.

We used Design study methodology to create workflows, define tasks, run experiments and validate our encodings.

During the iterative cycle, we asked important questions which led to a snowballing effect. We revised our InfoVis and we compared the effectiveness of different visual encodings and we built the most effective InfoVis that **generated new insights** each time.

Through the user centric design, we ran several experiments and validated those experiments with our users. We documented our learnings and approach so that we can be improve further using new tool sets.

Overall, our users were able to quickly **interact with the data and generate new insights**.

Section 9: Project Artifacts

#	Type	Description
1	My github project	https://github.com/rajesh4aiml/DTSA5304-Fundamentals-of-Data-Visualization
2	Dataset1: Fraud transactions	Kaggle: https://www.kaggle.com/datasets/kelvinkelue/credit-card-fraud-prediction/data
3	Dataset2: Geo spatial data	https://www.census.gov/geographies/mapping-files/time-series/geo/carto-boundary-file.html
4	Python Version	Python 3.11.8
5	Python Libs	altair 5.0.1; geopandas 0.14.3; pandas 2.1.4; numpy 1.26.4; seaborn 0.12.2; matplotlib 3.8.0; folium 0.16.0;

Section 10: References

Fundamentals of Data Visualization | Coursera

Vega-Altair 5.3.0

https://geopandas.org

seaborn 0.13.2 (pydata.org)