CS732: Data Visualisation Assignment 3 Report

Akanksha
DT2023001
Akanksha@iiitb.ac.in

Ketki Bhatia
DT2023007
Ketki.Bhatia@iiitb.ac.in

Niharika Suri
DT2023015
Niharika.Suri@iiitb.ac.in

***Abstract*—This study analyzes fraudulent credit card transactions in the United States during 2019 and 2020, employing advanced data visualization techniques and statistical methods to uncover trends, patterns, and anomalies in fraudulent activities. The analysis examines fraud distribution across dimensions such as states, cities, merchant categories, job categories, and demographic groups, identifying geographical hotspots and high-fraud regions. A comparative analysis highlights evolving trends between 2019 and 2020. Spatial analysis, including heatmaps and bar charts, illustrates the geographical distribution of fraud. Demographic and behavioral insights are derived using treemaps, donut charts, and density scatter plots to examine factors like gender, generation, and job categories. Temporal analysis, through area charts and violin plots, uncovers time-based vulnerabilities, while monetary analysis provides a detailed perspective on transaction amounts in high-fraud regions.Machine learning models, enhanced by a feedback loop, addressed the challenges of imbalanced datasets through resampling, threshold optimization, and class-weight adjustments. Evaluation metrics included confusion matrices, ROC AUC curves, and Precision-Recall analysis to optimize model performance. Visualizations like time-series plots and boxplots identified limitations in ARIMA models and revealed fraud anomalies. These findings offer actionable insights to strengthen fraud detection systems.**

***Index Terms*—component, formatting, style, styling, insert**

I. Introduction

A. About the Dataset

The Credit Card Transactions Dataset provides comprehensive details on credit card transactions, capturing key attributes about transaction times, amounts, and associated personal, geographic, and merchant-specific information. It contains over 1.85 million rows, making it a rich resource for analysis, including fraud detection, customer behavior segmentation, and geospatial trends.

1) **Unnamed**: An automatically generated index column, likely created during data import. It does not hold analytical value and can be dropped.
2) **trans date trans time**: Records the exact date and time of each transaction. Useful for time-based trend analysis.
3) **cc num**: A tokenized representation of the credit card number, ensuring privacy while enabling customer-specific transaction tracking.
4) **merchant**: Identifier of the merchant where the transaction occurred, allowing for merchant-specific trend or fraud analysis.
5) **category**: The category of goods or services purchased (e.g., groceries, electronics). Useful for expenditure classification.
6) **amt** (created as rounded amt): The monetary value of the transaction. A rounded version has been created for further aggregation or comparison.
7) **first and last**: The first and last names of cardholders. These fields require anonymization to ensure privacy.
8) **gender**: The gender of the cardholder, valuable for demographic and behavioral analysis.
9) **street, city, state, zip**: Address details of the cardholder, enabling location-based analysis while respecting privacy considerations.
10) **lat and long**: The geographic latitude and longitude of the cardholder's residence. Useful for geospatial analysis.
11) **city pop**: Population of the cardholder's city, providing context for regional transaction patterns.
12) **job**: The profession of the cardholder. A derived field, job categories, groups jobs into broader classifications for sector-based analysis.
13) **dob and age**: The date of birth of the cardholder, with age computed for demographic segmentation.
14) **trans num**: A unique identifier for each transaction, ensuring traceability.
15) **unix time**: The transaction time in Unix timestamp format, standardized for machine learning models or time-series analysis.
16) **merch lat and merch long**: The geographic coordinates of the merchant's location, enabling geospatial clustering of transaction activity.
17) **is fraud**: A binary flag indicating whether the transaction was fraudulent (1) or legitimate (0).
18) **transaction date**: Extracted date of the transaction.
19) **transaction time**: Extracted time of the transaction.
20) **transaction time 24hr and transaction hour**: Provide time details in 24-hour format and rounded to the nearest hour.
21) **transaction month**: Categorizes transactions by the month, enabling seasonal trend analysis.
22) **generation**: Categorizes cardholders into generational cohorts based on their date of birth.
23) **state categories**: Groups states into broader regional categories for macroscopic geographic analysis.

NEW VARIABLES ADDED TO THE DATASET (K TO D):

24) **fraud victim count**: The number of fraudulent transactions associated with each credit card (cc num).
25) **fraud loss amount**: The total monetary loss due to fraud

per credit card.

26) **fraud per victim**: The ratio of fraudulent transactions to total transactions per victim, providing a measure of fraud prevalence for each cardholder.
27) **fraud per 1000**: The number of fraud cases per 1,000 people in each state, based on the population (city pop).

II. Methodology

A. Introduction

The methodology outlines a detailed approach to processing and analyzing a credit card transaction dataset using the Pandas library in Python. The focus of this process is on data cleaning, feature engineering, and transformation, which are essential steps in preparing the dataset for further analysis and visualization. These steps aim to clean the data and extract valuable features.

B. Workflow Overview

The workflow for processing the dataset involved multiple stages, beginning with the inspection of the dataset's structure, followed by subsetting the data and addressing any inconsistencies. Feature engineering was then performed to extract key attributes such as transaction date, time, and customer age. Additionally, various transformations were applied to ensure the data was clean, consistent, and formatted correctly for analysis and visualization.

C. Data Processing Steps

1) Initial Data Inspection: The dataset was initially inspected to understand its structure and characteristics. The shape of the dataset, which reveals the number of rows and columns, was examined using the shape attribute. This was followed by the info() function, which provided a detailed overview of the dataset's columns, data types, and the count of non-null values. This step helped to identify any missing or inconsistent data that needed to be addressed. A statistical summary was generated using the describe() function, providing key metrics such as mean, median, minimum, maximum, and quartiles, which were essential for understanding the distribution of the data.

2) Data Subsetting and Cleaning: Once the dataset was inspected, a subset of the data starting from index 924,850 was selected for further analysis using the .iloc[] function. This subset focused on the relevant portion of the data, ensuring that the analysis was performed on the required data points. The index was reset using the reset index() function, which re-established the continuity of the data while discarding the old index. The last column of the dataset, deemed unnecessary for the analysis, was dropped using the drop() function. The accuracy of this operation was confirmed by applying the head() function, ensuring the dataset was in the desired format.

3) Handling Missing Data and Duplicates: To ensure the dataset's integrity, any missing data was identified using the isnull().sum() function, which provided an overview of the missing values across all columns. The dataset was then checked for duplicate rows using the duplicated() function. Since no duplicates were found, no further action was required to handle duplicate records, ensuring that the dataset was free from redundancy.

4) Feature Engineering: Date and Time Processing: The trans date trans time column, which contains both transaction date and time, was split into two separate columns: transaction date and transaction time, using the str.split() method. The date was reformatted to a more readable dd-mm-yyyy format using pd.to datetime() and strftime() to ensure uniformity. Similarly, the transaction time was converted to a 24-hour format using pd.to datetime() with the errors='coerce' parameter to handle invalid time values by converting them to NaT (Not a Time). Additionally, the transaction hour was extracted from the transaction time column by converting it to a datetime format and then using the .dt.hour method to isolate the hour. The month of each transaction was extracted from the transaction date column by using the .dt.month method.

5) Feature Engineering: Amount Rounding and Categorization: To simplify the analysis of transaction amounts, the amt column was rounded to the nearest integer using the round() function and stored in a new column, rounded amt. The original amt column was then dropped to prevent redundancy in the dataset. Customer age was also categorized into generational groups (e.g., Baby Boomers, Millennials) using a custom function applied to the age column.

6) Data Transformation: 24-Hour Time Conversion and Filtering: The transaction time column, after being converted to the 24-hour format, was stored in a new column, transaction time 24hr, using strftime('H:M:S'). This ensured the time was in a consistent 24-hour format suitable for time-based analysis. The transaction time 24hr column was also converted into a time object by extracting only the time portion, removing the date. Additionally, rows corresponding to specific hours, such as 2:00 PM and 3:00 PM, were filtered using conditional statements to focus on specific time frames for further analysis.

7) Dropping Redundant Columns: To eliminate unnecessary variables, columns such as transaction time str and is 12 hour format were dropped, as they were deemed redundant after the transformation steps. Similarly, the transaction time 12hr column, which became unnecessary after the conversion to 24-hour time, was removed to streamline the dataset.

D. Calculated Fields of Tableau-

For creating customised visualisation depending upon the need we have created some calculated fields using the variables present in the dataset. The calculated variables are-

- Age- It is created using dob.
- Generation - It is created using age.
- Age in Range - It is also calculated using age
- Is Fraud(only 1) - It is calculated using Is Fraud.

III. Data Analysis and Visualisations

This report provides a comprehensive analysis of credit card fraud for the year 2019, with comparisons made to 2020. The analysis is divided into four main sections: Part 1 - Credit Card Fraud Analysis focuses on examining fraud patterns

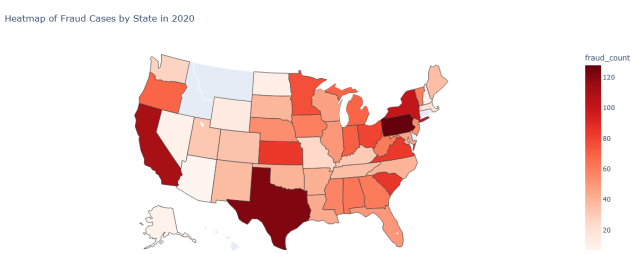

Fig. 1. Heatmap of Fraud Cases by States in 2020

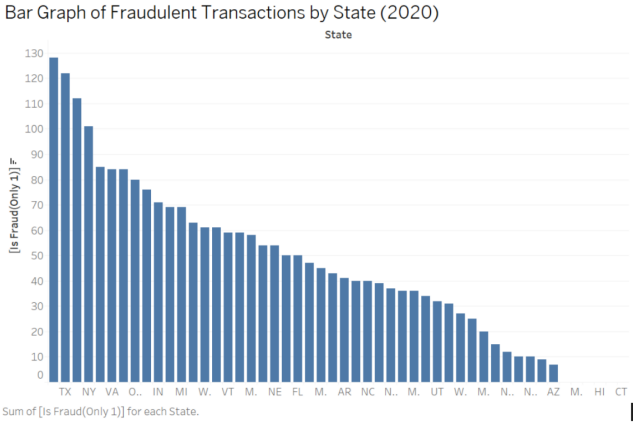

Fig. 2. Bar Graph of Fraudulent Transaction by States in 2020

in states with high fraud rates in 2019, including Texas, New York, Pennsylvania, Ohio, Florida, and Missouri. This section explores key variables such as job categories, merchant categories, and transaction hours using basic data exploration techniques, without employing advanced modeling. Part 2 - Comparative Analysis of Credit Card Fraud (2019 and 2020) builds on the findings from Part 1 by comparing fraud trends between the two years. It follows the Basic Visual Analytics loop, progressing from data visualization to modeling (with a model applied only in Exhibit 5) and then to knowledge, helping to identify trends and changes in fraud cases over 2019 and 2020. Part 3 - Knowledge to Data transitions from the insights gained in the earlier sections to more advanced data-driven models. This part aims to further investigate fraud patterns, enhance predictive capabilities, and provide actionable insights for future research and decision-making. Part 4 - This section of the report includes the Machine Learning Models applied (Forecasting, Classification) that provides a detailed analysis of our fraud detection approach, outlining the transition from time-series forecasting with SARIMA to classification-based methods for enhanced fraud detection.

A. Exhibit 1 - Fraudulent Credit Card Cases Across the United States

1) Exhibit 1.1 Heatmap of Fraud Cases (2020) : The heatmap illustrates the distribution of fraud cases across the United States in 2020. Darker shades of red indicate states with a higher incidence of fraud. Pennsylvania (PA) had the highest number of fraud cases, totaling 128. This was followed by Texas (TX) with 122 cases, California (CA) with 112 cases, New York (NY) with 101 cases, South Carolina (SC) with 85 cases, and Virginia (VA) with 84 cases. These states are represented with darker shades on the heatmap, highlighting their greater concentration of fraud cases compared to other regions.

2) Exhibit 1.2 Bar Chart - Fraudulent Transactions by State (2020): The bar chart illustrates the distribution of fraudulent transactions (is fraud = 1) across various states in 2020, highlighting the top six states with the highest number of fraud cases. The x-axis represents the states, sorted in descending order by the number of fraud cases, while the y-axis indicates the count of fraudulent transactions. At the top of the list are Pennsylvania (PA) and Texas (TX), which reported the highest number of fraudulent transactions, approximately 130 cases each. They are closely followed by New York (NY) and California (CA), with slightly over 120 cases each. Next are South Carolina (SC) and Virginia (VA), each with nearly 120 fraud cases, making them additional hotspots for fraudulent activity. These six states stand out as clear outliers with consistently high levels of fraud.

After Virginia, Ohio (OH), Indiana (IN), and Michigan (MI) contribute significantly to the overall fraud count. Moving further down the list, the number of fraud cases gradually decreases. States such as Florida (FL), North Carolina (NC), and Illinois (IL) also report relatively high counts, but there is a consistent decline compared to the leading states.

At the lower end of the chart, states like Arizona (AZ), Alaska (AK), and New Hampshire (NH), exhibit the least amount of fraudulent activity, with counts approaching zero. This steep decline in fraud cases indicates a highly skewed distribution, where a small number of states account for the majority of fraud incidents. Overall, this analysis highlights a clear disparity in the distribution of fraud across states, with a few high-risk states responsible for a disproportionate share of cases.

According to a Forbes report, several states in the U.S. have been identified for experiencing high levels of credit card fraud, with California, Texas, New York, Pennsylvania, and South Carolina standing out [2].

California, often regarded as the fraud capital of the U.S., had the highest number of fraud reports and total financial losses [2]. The state's large population and widespread online shopping and financial activities contributed to the elevated fraud rates [1]. Despite the high volume of incidents, California's fraud rate, when adjusted for population size, remained relatively moderate [2]. However, the financial impact on residents was significant, with many losing substantial amounts to fraudsters [2].

Texas also saw a large number of fraud incidents, ranking

high in both total reports and financial losses. The state's diverse and expansive population, along with its bustling economic sectors in major cities like Houston and Dallas, provided ample opportunities for fraudsters to target a wide variety of transactions [2]. The state's high levels of commercial activity further increased the likelihood of fraud across both individual and business sectors [2].

New York had a notable number of fraud cases due to its status as a major financial hub and densely populated areas. While the total number of fraud reports was high, the state's fraud rate per capita was more moderate [2]. The combination of affluent residents and vulnerable groups, coupled with New York's global financial influence, makes it a prime target for both sophisticated cyber fraud and street-level scams.

Pennsylvania also ranked high for fraud, driven by its major urban centres like Philadelphia and Pittsburgh [2]. As e-commerce continues to rise, Pennsylvania has become increasingly vulnerable to online fraud, including data breaches and identity theft [2]. The growing number of online transactions in the state, paired with gaps in consumer education and protection, has contributed to the surge in fraudulent activity.
Although not on the Forbes list in the analysis, South Carolina also ranked among the top five states with significant credit card fraud activity. The state's growth in tourism and population has led to more retail and online transactions, increasing opportunities for fraud [3]. Additionally, South Carolina has a relatively high proportion of retirees, who are often more susceptible to scams such as phishing and identity theft [4]. This demographic is sometimes less familiar with digital security practices, which makes them prime targets for fraudsters.

Alaska, North Dakota, and New Hampshire are ranked among the least populated states in the U.S., which naturally reduces the volume of financial transactions and, in turn, creates fewer opportunities for fraud [5]. In these states, smaller populations mean that financial activity is more localized and closely monitored, making unusual transactions more noticeable and likely to be detected quickly [5].

The smaller size also leads to stronger community connections, where residents are more likely to spot suspicious activities. In contrast, Nevada and Arizona, despite being larger and more urbanised, still experience lower fraud rates. Nevada benefits from strict regulatory measures in industries like gaming and finance, where high levels of scrutiny and robust security systems are standard [6]. Arizona, with its growing urban centers, emphasizes consumer education and awareness, helping residents recognize and avoid common fraud schemes.

These states, with their lower population density and focused fraud prevention efforts, are less susceptible to widespread fraudulent activity.

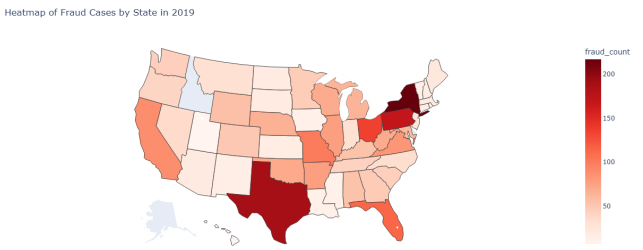

Fig. 3. Heatmap of Fraud Cases by States in 2019

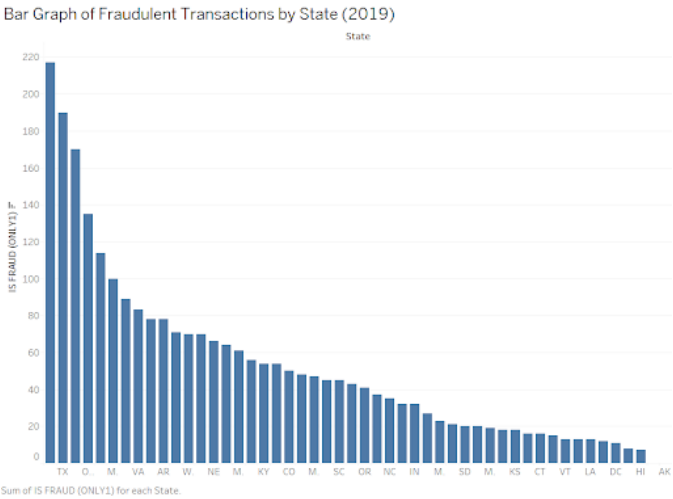

Fig. 4. Bar Graph of Fraudulent Transaction by States in 2019

3) Exhibit 1.3 Heatmap of Fraud Cases (2019): The geographic heat map illustrates the distribution of fraud cases across the United States in 2019. Darker shades of red indicate higher occurrences of fraud, while lighter shades represent fewer cases. The map clearly identifies the states with the highest levels of fraud activity. New York (NY) reported the most fraud cases, with 217 incidents, followed by Texas (TX) with 190 cases, Pennsylvania (PA) with 170 cases, Ohio (OH) with 135 cases, Florida (FL) with 114 cases, and Missouri (MO) with 100 cases. These states are prominently displayed on the map with darker shades, highlighting their greater vulnerability to fraudulent activities. In contrast, many rural and sparsely populated states show lighter shades, indicating minimal occurrences of fraud.

4) Exhibit 1.4 Fraudulent Transactions by State (2019): The bar chart illustrates the distribution of fraudulent transactions (is fraud = 1) across various states. It highlights the states with the highest number of fraud cases, while also showcasing a broader pattern of how fraud is distributed. The x-axis represents the states in descending order of fraud cases, while the y-axis denotes the count of fraudulent transactions.

The chart reveals significant disparities in fraud counts across states, with a few states demonstrating disproportionately high fraud activity, while others exhibit minimal occurrences. At the top of the chart, New york(NY), stands out as the state with the highest number of fraudulent transactions, reporting over 220 cases. This makes New York a clear outlier in fraud prevalence. Following New York comes Texas(TX) with about 190 cases and then PA with 170 cases. Following this Ohio (OH) ranks next with over 160 fraud cases, and Michigan (MI) closely follows with slightly fewer than 150 cases. These states also exhibit a significant concentration of fraud, though not to the extent observed in New York.

The next group of states includes Virginia (VA), Arkansas (AR), and Wisconsin (WI), each reporting over 100 fraud cases. As we move further down the list, states such as New Mexico (NM), Kentucky (KY), Colorado (CO), and South Carolina (SC) report moderate fraud levels, with cases ranging between 80 and 100. These states are not among the most fraud-prone but still show noteworthy activity. At the lower end of the chart, states like South Dakota (SD), Vermont (VT), District of Columbia (DC), Hawaii (HI), and Alaska (AK) report the lowest fraud counts, with some states approaching zero cases.

The discrepancies in fraud counts among states, particularly with New York being a notable outlier, can be explained by its unique characteristics. As a densely populated state with significant economic activity and a concentration of financial institutions, New York offers numerous opportunities for fraudulent activities. The state's diverse economy and high-value financial transactions may attract fraudsters seeking lucrative targets [10].

Similarly, other states such as Texas and Pennsylvania, which also report high fraud cases, share attributes like large populations and significant economic hubs. These factors facilitate a high volume of transactions, increasing the likelihood of fraud occurrences [8] [9].

In contrast, states with low fraud counts, such as Vermont or South Dakota, generally have smaller populations, less diverse economies, and fewer large-scale financial activities. These elements reduce the opportunities and incentives for fraud, contributing to their lower prevalence of fraud. Additionally, the District of Columbia (DC) shows low fraud counts, likely due to its unique status as a political hub with stringent financial regulations and strict enforcement of anti-fraud laws, which act as a deterrent to fraudulent activities [7].

PART 1 - CREDIT CARD FRAUD ANALYSIS FOR 2019 For the 2019 analysis, the six high-fraudulent states included Texas (TX), New York (NY), Pennsylvania (PA), Ohio (OH), Florida (FL), and Missouri (MO). The three common high-fraud states from both 2019 and 2020 were Texas (TX), New York (NY), and Pennsylvania (PA). This selection was made to provide a comprehensive overview and identify patterns distinguishing high-fraud states from others. As the dataset was new to us, we concentrated on acquiring fundamental insights on job categories, merchant categories, transaction hours, and other essential aspects. We applied a basic Data to Knowledge loop to explore these features in detail, without utilizing any models in Part 1 of the analysis. This approach allowed us to gain a better understanding of the dataset and identify initial patterns and trends that could guide further analysis.

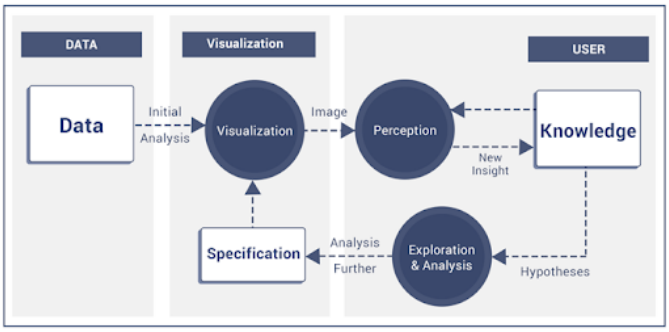

Fig. 5. Data to Knowledge Loop

B. Exhibit 2 - Synopsis of 2019 Analysis

The rationale behind creating the following visualizations of this exhibit is to provide a comprehensive analysis of fraud-related data across multiple dimensions. The dumbbell plot was designed to compare the number of unique credit card numbers categorized as fraudulent and non-fraudulent across high-fraud U.S. states in 2019, highlighting regional variations in fraudulent activity. The parallel coordinate plot of High fraud cities (NY, TX, PA) further explores the intricate relationships between fraud data, such as cities, transaction types, and job sectors, to identify patterns and vulnerabilities. Together, these visualizations offer insights into the nature and demographics of fraudulent activities, aiding in targeted fraud prevention strategies.

DATA

The primary data points considered in this analysis include: Is Fraud (whether a transaction is fraudulent or non-fraudulent), State (identifying high-fraud and common states), Merchant Categories (types of transactions), Job Categories (professions), Generations (Millennials, Generation Z, and Baby Boomers), and Cities (specific locations within high-fraud states).

The high-fraud states identified in Main Exhibit for 2019 include Texas (TX), Florida (FL), New York (NY), Pennsylvania (PA), Missouri (MO), and Ohio (OH), all of which exhibited elevated levels of fraud. The common states observed in both the 2019 and 2020 data include Texas (TX), New York (NY), and Pennsylvania (PA), which were also classified as moderate-fraud states.

The dataset contains a variety of merchant categories, such as grocery, shopping, entertainment, travel, and healthcare, which are used to analyze trends in fraud. It also includes job categories like healthcare, STEM, public service, law, and business and finance to evaluate the correlation between

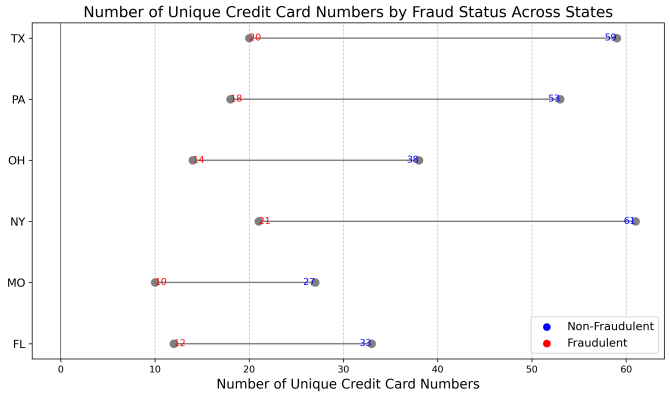

Fig. 6. Number of Unique Credit Card Numbers by Fraud Status Across States

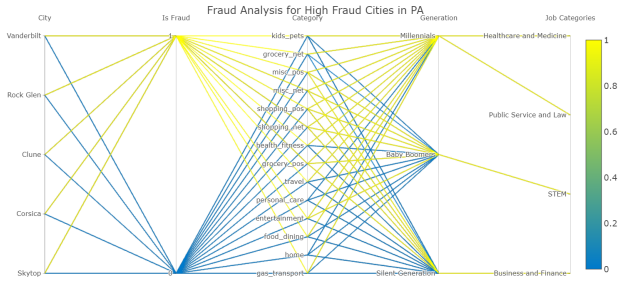

Fig. 7. Fraud Analysis for High Fraud Cities in PA

different professions and fraud occurrences. Additionally, the dataset encompasses generational cohorts, including Millennials, Generation Z, and Baby Boomers, to identify which age groups are more vulnerable to fraud. Furthermore, it features city-level data from high-fraud states, specifically Houston, Dallas, and New York City, in order to reveal city-specific trends.

VISUALIZATION

1) Exhibit 2.1: Unique CC Numbers by Fraud Across High Fraud States: The dumbbell plot compares the number of unique credit card numbers categorized as fraudulent and non-fraudulent across high fraudulent U.S. states in 2019. For each state, the blue marker represents the count of non-fraudulent credit card numbers, while the red marker represents the fraudulent ones.

From the plot, Texas (TX) shows the largest number of unique credit card numbers overall, with 50 non-fraudulent and 10 fraudulent. Similarly, Pennsylvania (PA) follows with 52 non-fraudulent and 8 fraudulent numbers. New York (NY) has the highest count of non-fraudulent cards (62), with only 1 fraudulent case, indicating strong security. Conversely, Missouri (MO) and Florida (FL) exhibit slightly higher ratios of fraudulent to non-fraudulent cases compared to other states, with Missouri reporting 27 non-fraudulent and 10 fraudulent cards, while Florida has 33 non-fraudulent and 12 fraudulent cards. Ohio (OH) reports 38 non-fraudulent and 4 fraudulent cards, maintaining a relatively low fraud count.

Overall, the plot highlights regional variations in fraudulent activities, with some states, like NY, having minimal fraud cases relative to their total credit card usage, while others, like MO and FL, show relatively higher fraudulent activity.

According to Forbes, which ranks states based on the number of fraud incidents, per capita rates, and financial losses, Florida, Texas, New York, and Pennsylvania occupy the second, fourth, eighth, and tenth positions, respectively [2]. These rankings reflect not only the frequency of fraud but also its financial impact on residents and businesses. They highlight systemic vulnerabilities in certain states while showcasing relatively better fraud management in others.

In the 2019 dataset, Florida demonstrated a high prevalence of fraud, with 33 non-fraudulent cases and 12 fraudulent cases [2]. The state's significant retiree population makes it particularly susceptible to scams targeting older individuals. This vulnerability contributes to Florida's high fraud rate per capita, as emphasized by Forbes, along with substantial financial losses amounting to 99.9 million dollar in a single quarter, marking it as one of the hardest-hit states for fraud [2].

Texas recorded 50 non-fraudulent cases and 10 fraudulent cases in the same dataset. As the second-most-populous state in the U.S., Texas naturally experiences a higher volume of financial transactions, leading to increased exposure to fraud [1]. Forbes reported significant financial losses from fraud in Texas, totaling 119.6 million dollar, reflecting the challenges posed by its vast economic scale and population [2].

New York, ranked eighth by Forbes, showcases effective fraud prevention measures. In the 2019 dataset, it had the highest number of non-fraudulent cases (62) while reporting just one fraudulent case, highlighting its strong security infrastructure [2]. Despite being densely populated and having extensive financial activity, New York's relatively moderate fraud losses of 64.9 million dollar suggest that effective mitigation strategies are in place, even in the face of frequent attempts [2].

Pennsylvania, ranked tenth by Forbes, recorded 52 non-fraudulent cases and 8 fraudulent cases in the 2019 dataset. While the state has a significant fraud rate per capita, its total losses of 38.8 million dollar are relatively contained compared to higher-ranked states. This suggests that, although fraud is common, the financial impact per incident tends to be lower, indicating some level of control over fraudulent activity [2].

Missouri and Ohio, despite emerging in our 2019 dataset with notable ratios of fraudulent to non-fraudulent cases, do not feature in Forbes' high-fraud rankings. Missouri's smaller population and lower economic activity reduce its overall exposure to fraud, while Ohio, with a relatively low fraud ratio (38 non-fraudulent and 4 fraudulent cases), experiences fewer incidents and financial losses. This might have kept both states outside the top-ranking fraud lists.

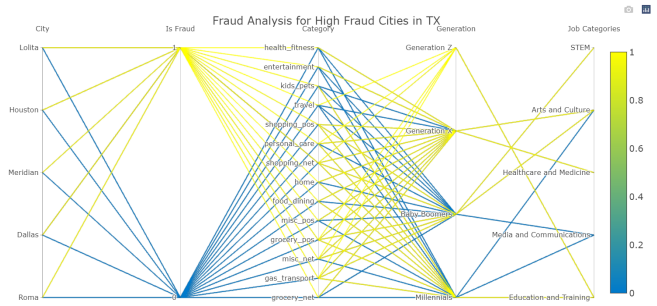

Fig. 8. Fraud Analysis for High Fraud Cities in TX

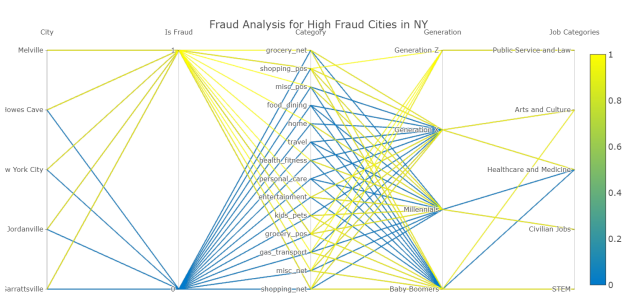

Fig. 9. Fraud Analysis for High Fraud Cities in NY

2) Exhibit 2.2: An Analysis of High Fraudulent Cities in Pennsylvania : The figure (7) visualizes connections between fraud-related data across several categories. The City axis displays high-fraud cities in Pennsylvania, including Rock Glen, Clune, and Skytop. The Is Fraud axis indicates whether a transaction was fraudulent, with a value of 1 representing fraud and 0 representing non-fraud. The Category axis represents different transaction types, such as ”grocery pos,” ”kids pets,” and ”shopping net,” which specify the types of purchases or activities involved. The Generation axis divides fraud cases by generational cohorts, including Millennials, the Silent Generation, and Baby Boomers. The Job Categories axis associates fraud cases with various professional sectors, such as Healthcare and Medicine, STEM, Public Service and Law, and Business and Finance.

In Pennsylvania, the city of Vanderbilt exhibits a mix of fraudulent and non-fraudulent transactions, with notable connections to categories like ”kids pets” and ”travel.” The fraudulent activities in this area are largely associated with Healthcare and Medicine professions.

Rock Glen, on the other hand, displays a concentration of fraudulent activity in the ”grocery pos” category. Millennials dominate the fraudulent cases in this city, which are strongly linked to jobs in Public Service and Law, reflecting potential vulnerabilities in these sectors.

In Clune, fraudulent transactions are predominantly associated with the ”shopping net” and ”entertainment” categories. The Silent Generation forms a significant demographic for fraud cases in this area, with links primarily to jobs in Business and Finance, indicating the sector’s susceptibility to fraud.

The city of Corsica shows a balanced distribution of fraudulent transactions across the ”misc pos” and ”gas transport” categories. Again, Millennials are prominent in the fraud demographics, with a noticeable connection to STEM professions, highlighting potential exploitation within this field.

Skytop demonstrates a unique pattern, with significant fraud connections to the ”food dining” and ”travel” categories. Fraudulent activities here are associated with both Millennials and the Silent Generation, spanning job categories such as Public Service and Law and Business and Finance.

3) Exhibit 2.3: An Analysis of High Fraudulent Cities in Texas: The figure (8) visualizes connections between various categories of fraud-related data. The City axis lists high-fraud cities in Texas, which include Lolita, Houston, Meridian, Dallas, and Roma. The Is Fraud axis indicates whether a transaction was fraudulent, with a value of 1 representing fraud and 0 representing non-fraud. The Category axis represents different types of transactions, such as ”health fitness,” ”kids pets,” ”grocery net,” and ”entertainment,” specifying the nature of the purchases or activities involved. The Generation axis divides fraud cases by generational cohorts, including Millennials, Generation Z, and Baby Boomers. The Job Categories axis links fraud cases to various professional sectors, including STEM, Arts and Culture, Healthcare and Medicine, Media and Communication, and Education and Training.

In Texas, the city of Lolita shows no fraudulent transactions (Is Fraud = 1) but is connected to non-fraudulent activities (Is Fraud = 0), particularly in categories like ”health fitness” and ”kids pets.” These transactions are associated with individuals in the STEM and Arts and Culture sectors.

Houston exhibits fraudulent transactions in the ”travel” and ”shopping net” categories, with links to Generation Z and jobs in Arts and Culture.

Meridian demonstrates a concentration of fraudulent activity in the ”grocery pos” and ”misc pos” categories, primarily involving Millennials and individuals in the Healthcare and Medicine sector.

Dallas displays fraudulent activity across the ”entertainment” and ”food dining” categories, with Baby Boomers being the primary demographic. Fraudulent transactions in this city are strongly tied to jobs in Media and Communication.

Roma highlights fraudulent transactions in the ”gas transport” and ”grocery net” categories, with Millennials notably represented in these cases. These activities are linked to the Education and Training sector.

4) Exhibit 2.4: An Analysis of High Fraudulent Cities in New York: The figure (9) illustrates the connections between various categories of fraud-related data. The City axis highlights high-fraud cities in New York, including Melville, Howes Cave, New York City, Jordanville, and Garrettsville. The Is Fraud axis indicates whether a transaction was

fraudulent, with a value of 1 representing fraud and 0 indicating non-fraudulent transactions. The Category axis denotes different types of transactions, such as ”grocery net,” ”kids pets,” and ”shopping net,” which specify the nature of purchases or activities. The Generation axis categorizes fraud cases by generational cohorts, including Millennials, Generation Z, and Baby Boomers. Lastly, the Job Categories axis connects fraud cases to various professional sectors, including Public Service and Law, Arts and Culture, Healthcare and Medicine, Civilian Jobs, and STEM.

In New York, Melville shows a concentration of fraudulent transactions primarily in the ”grocery net” category, correlated with Public Service and Law jobs. Notably, non-fraudulent transactions have no associations (Is Fraud = 0), underscoring its significant fraudulent activity.

Howes Cave exhibits fraudulent activity in the ”kids pets” and ”food dining” categories, with Generation Z dominating the fraud cases in this city. These cases are mainly linked to jobs in Public Service and Law.

New York City reveals a wide distribution of fraudulent transactions across various categories, with Millennials being the primary demographic affected. Fraud in this area is connected to jobs in Healthcare and Medicine, as well as Civilian Jobs.

In Jordanville, fraudulent transactions are associated with the ”gas transport” and ”misc net” categories. Baby Boomers represent a significant demographic in fraud cases here, showing strong connections to STEM professions.

Garrettsville indicates a link between fraudulent activity and the ”shopping net” and ”entertainment” categories. Both Baby Boomers and Millennials are prominent demographics in fraud cases here, with connections to the Arts and Culture sector.

PERCEPTION

The visualizations from Exhibit 2, including the dumbbell chart and PCP diagrams, provide an understanding of fraud distribution across various high-fraud states and cities. These tools highlight patterns such as the concentration of fraud in specific states like Texas, Florida, and New York and the roles of demographic, transactional, and professional factors. States and cities exhibited varying ratios of fraudulent to non-fraudulent activities, reflecting localized vulnerabilities in financial security. Generational differences, along with correlations to job sectors like Healthcare, STEM, and Public Service, also emerged, pointing to specific areas of susceptibility.

KNOWLEDGE

The visualizations in Exhibit 2, which include the dumbbell chart and PCP diagrams, offer insights into the distribution of fraud across high-fraud states and cities. The charts reveal significant patterns of fraudulent activity in regions such as Texas, Florida, and New York, while also highlighting connections between demographic, transactional, and professional factors.

In Texas, the substantial number of fraudulent activities can be attributed to the state’s size and large population, with cities like Houston showing notable fraud linked to Generation Z and professions in Arts and Culture. Florida, with its higher proportion of retirees, demonstrates greater vulnerability to scams targeting older individuals. In contrast, New York has a relatively low fraud rate despite its dense population, suggesting effective fraud management strategies are in place.

The analysis of specific cities, such as Rock Glen in Pennsylvania, indicates that Millennials are particularly involved in fraudulent activities related to grocery point-of-sale purchases, often connected to jobs in Public Service and Law. Similarly, in New York, fraud cases are linked to Healthcare and Medicine professions, especially among Millennials.

Generationally, both Millennials and Baby Boomers are identified as the most vulnerable groups. Millennials are notably active in fraudulent activities in Texas and Pennsylvania, particularly in sectors such as Public Service, Law, and STEM. Baby Boomers also feature prominently in fraud cases, especially in Texas and New York, where they are associated with STEM professions.

The types of transactions involved in these fraud cases, including grocery point-of-sale, online shopping, and travel, suggest that these categories are particularly targeted for fraud, emphasizing the need to focus on them for prevention efforts.

EXPLORATION AND ANALYSIS

Hypothesis- Fraudulent transactions are more likely to occur in ”shopping net,” ”grocery pos,” and ”travel” transaction categories.

Fraudulent transactions commonly occur in the ”shopping net,” ”grocery pos,” and ”travel” categories, as shown by various PCP visualizations. These categories indicate systemic vulnerabilities, likely due to high transaction volumes and the prevalence of card-not-present activities. For instance, the ”shopping net” category is particularly susceptible to online fraud, including phishing and account takeovers. The ”grocery pos” category may be vulnerable to skimming at physical terminals. Additionally, the ”travel” category often involves booking scams and the fraudulent use of rewards points. These trends emphasize the need for a focused examination of merchant-level security practices and increased consumer awareness.

Hypothesis - Millennials and Baby Boomers are more likely to be involved in fraudulent transactions compared to other generational cohorts

Millennials and Baby Boomers are notable groups that are more likely to be involved in fraudulent transactions, as consistently shown in PCP diagrams. Millennials often face financial pressures, such as student loan debt, and are targeted by scams that exploit their digital habits [12]. Common threats include fake check schemes, job fraud, and phishing attempts. On the other hand, Baby Boomers, who grew up in a less digitally savvy era, are more susceptible to traditional scams,

such as lottery and prize fraud or investment scams, especially as they approach retirement [12] [11].

Hypothesis - Healthcare and STEM job categories are more frequently associated with fraudulent activities than other job sectors.

Healthcare and STEM job categories are common among all the PCP visualizations. Healthcare professionals, due to their high earnings and demanding schedules, are frequently targeted by schemes like medical equipment scams, which exploit crises to sell fake or overpriced supplies. Additionally, billing fraud schemes leverage their involvement with complex insurance systems. STEM professionals, often perceived as tech-savvy, are targeted for fraudulent schemes like tech support scams or credential theft, which capitalize on their frequent online interactions.

SPECIFICATION

Furthermore, we will analyze merchant categories, generational cohorts, and job categories, along with additional variables such as transaction hour, transaction month, and transaction amount. This approach will help us identify patterns and relationships that contribute to fraud and provide a deeper understanding of the key influencing factors.

C. Exhibit 3 - Exploration of Fraudulent by Job Categories

The rationale for the following visualizations was their effectiveness in highlighting key aspects of the data. The density scatter plot illustrates the relationship between job categories and transaction amounts, revealing clusters, outliers, and the concentration of fraudulent transactions within specific ranges. Its ability to overlay fraud status onto continuous variables makes it ideal for identifying high-risk areas.

The area chart clearly shows how fraudulent activity varies over a 24-hour period, highlighting peaks and declines across different job categories. Its layered design allows for easy comparison of overlapping trends, helping to identify when and where fraud occurs most frequently.

The treemap is effective in representing multidimensional data by combining variables such as state, gender, job category, and month. Its hierarchical layout emphasizes the prevalence of fraud across different combinations, while variations in size and color provide clarity and highlight significant clusters, such as patterns related to specific states or genders.

DATA

The data in Exhibit 3 includes various variables for analyzing fraudulent transactions across different job categories, transaction amounts, and temporal and geographical factors. Job categories include classifications such as ”Business and Finance,” ”Public Service and Law,” ”Healthcare and Medicine,” ”STEM,” ”Media and Communications,” among others, which categorize the occupation related to each transaction.

Transaction amount represents the monetary value of each transaction and is crucial for understanding the relationship between fraud and transaction size. Fraud status (is fraud) is a binary variable that indicates whether a transaction is fraudulent (1) or non-fraudulent (0). Transaction time categorizes transactions by hour (from 0 to 23), highlighting the times of day when fraud is most prevalent.

State indicates the geographical location of each transaction, focusing on high-fraud states identified in Exhibit 1, including Texas (TX), Pennsylvania (PA), and New York (NY). Gender distinguishes between transactions associated with males and females. Transaction month covers the months from July (7) to December (12).

Fig. 10. Density Scatter Plot of Job Category vs Transaction Amount by Fraud Status

VISUALIZATION

1) Exhibit 3.1: Density Scatter Plot of Job Categories vs Transaction Amount by Fraud Status: The scatter plot illustrates the relationship between job categories and transaction amounts, highlighting both fraudulent and non-fraudulent transactions in 2019. Fraudulent transactions, represented by yellow points, are sparse compared to the red points that indicate non-fraudulent transactions. Most fraudulent transactions occur at lower amounts, typically below dollar 5,000, across all job categories.

Fraud instances are primarily observed in categories such as ”Civilian Jobs,” ”Public Service and Law,” and ”Business and Finance.” In contrast, categories like ”Media and Communications,” ”STEM,” and ”Healthcare and Medicine” show a higher occurrence of larger monetary transactions, with notable outliers exceeding dollar 20,000. However, these outliers are predominantly legitimate.

Overall, the majority of transactions, regardless of job category, are concentrated at lower amounts, indicating a skewed distribution. The data suggests a low frequency

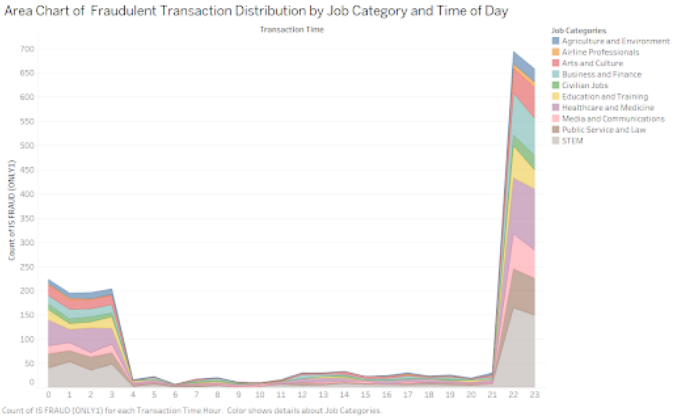

Fig. 11. Area Chart of Fraudulent Transaction Distribution by Job Category and Time of Day

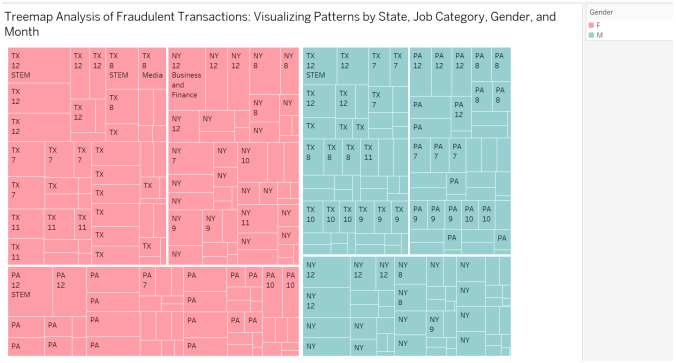

Fig. 12. Treemap Analysis of Fraudulent Transactions: Visualizing Patterns by State, Job Category, Gender and Month

of fraud, primarily in smaller transactions within specific categories, while high-value transactions across other categories remain largely legitimate.

2) Exhibit 3.2:Area Chart (Transactions Over 24 Hours for different Job Categories): Fraudulent activity between 00:00 and 06:00 is relatively high and steady, albeit slightly lower than the late-night peak. Job categories such as Business and Finance, Healthcare and Medicine, and Education and Training exhibit noticeable counts of fraud during this time. This trend could indicate that these job categories are commonly linked to transactions processed late at night, potentially due to the use of online systems for professional services or purchases.

On the other hand, job categories like Arts and Culture, Public Service and Law, and STEM experience lower levels of fraudulent activity during these hours. From 06:00 to 20:00, there is a sharp decline in fraud counts across all job categories, reflecting the trend illustrated in the area chart of merchant categories and transaction hours. During this period, instances of fraud remain minimal and uniformly low, suggesting that timing, rather than the specific job category, is a stronger determinant of fraudulent behavior. There is a dramatic surge in fraudulent transactions during late-night hours (21:00 to 23:59), peaking around midnight. Fraudulent activity during this period spans multiple job categories, with Business and Finance, Healthcare and Medicine, and Education and Training being particularly significant. Categories such as Media and Communications and STEM also show notable activity but are less prominent compared to others.

Business and Finance consistently ranks as a high-risk group across all time periods, especially during the late-night peak. Healthcare and Medicine shows consistent activity, particularly late at night. In contrast, Agriculture and Environment exhibits minimal fraudulent activity, possibly due to fewer online transactions or the nature of purchases in this sector being less targeted by fraudsters. Low fraud counts in Public Service and Law may reflect robust security measures or lower transaction volumes in this category. While STEM is present in the data, fraudulent transactions in this category remain comparatively low, suggesting that either the nature of their transactions involves higher security measures or they are less susceptible to fraud.

3) Exhibit 3.3: Treemap (Fraudulent Transactions by Common State, Gender, and Job Category): The treemap offers a multi-dimensional view of fraudulent transactions (is fraud = 1). The size of each rectangle reflects the count of fraud cases, while color, labels, and text overlays add layers of additional information. The variables represented in this visualization include gender (indicated by color), state, job categories, and transaction month (displayed as text labels). The size of the rectangles signifies the volume of fraud cases—larger rectangles indicate a higher number of fraudulent transactions tied to a specific combination of attributes.

Notably, the largest rectangles highlight states like Texas (TX), Pennsylvania (PA), and New York (NY), which are major contributors to fraudulent activities across various job categories and months. Within these states, specific job categories such as STEM, Media and Communication, Business and Finance, and Healthcare and Medicine show relatively larger rectangles, suggesting these sectors are more susceptible to fraud.

The color coding differentiates between fraudulent transactions involving females (F) (pink) and males (M) (blue). A visual examination reveals that fraud among females appears more prevalent in certain states and sectors, like Texas in the STEM and Healthcare categories, as indicated by the larger pink areas. Conversely, male-associated fraud is more pronounced in states such as Pennsylvania and New York, particularly in STEM and Business and Finance.

The treemap depicts a substantial number of fraud cases from Texas (TX), Pennsylvania (PA), and New York (NY). While fraud cases are distributed across different job categories in these states, certain states may dominate specific professions. For instance, Texas shows a prominent representation of fraud

in the STEM sector, suggesting a high count of fraud in this area for the state.

Job categories like STEM, Healthcare and Medicine, Business and Finance, and Media and Communication occupy a significant portion of the treemap, indicating that these sectors may be more vulnerable to fraud. In Texas, STEM fraud seems heavily female-dominated, as demonstrated by the larger pink rectangles in this category. Similarly, New York STEM shows notable male involvement, reflected by larger blue rectangles. Healthcare and Medicine and Business and Finance exhibit a more balanced distribution of gender contributions, although the prevalence varies from state to state.

The treemap also highlights transaction months, represented numerically within each segment (e.g.,12 for December). Among these, December (12) stands out as the month with the highest frequency of fraudulent transactions, as indicated by its presence across many large rectangles. This suggests a spike in fraudulent activity during the end-of-year period. Other notable months such as August (8), October (10), and July (7) also appear prominently, although to a lesser extent than December.

PERCEPTION

Fraudulent transactions tend to occur at lower amounts, with notable activity observed in job categories such as "Business and Finance," "Healthcare and Medicine," and "STEM." In contrast, high-value transactions in categories like "Media and Communications" and "Healthcare" are primarily legitimate, although there are some occasional outliers. Timing plays a crucial role as well, with spikes in fraudulent activity typically occurring during late-night hours (from 12:00 AM to 6:00 AM and from 9:00 PM to 11:59 PM). This suggests that fraud is more likely to happen during nighttime transactions. Additionally, the data indicates that specific months, particularly from July to December, experience higher rates of fraud.

KNOWLEDGE

The analysis reveals that timing significantly influences fraudulent transactions. Late-night hours, especially between midnight and early morning (12:00 AM to 6:00 AM), exhibit the highest frequency of fraud. Certain job categories, such as "Business and Finance" and "Healthcare and Medicine," are particularly vulnerable to these activities, likely due to their engagement with online transactions or supportive systems.

Geographical patterns indicate that states like Texas, Pennsylvania, and New York consistently report higher fraud rates. In these regions, job categories such as "STEM" (Science, Technology, Engineering, and Mathematics) and "Healthcare" show notable instances of fraud which align with the Exhibit 2 Hypothesis . Gender patterns also emerge, with fraud being more prevalent among females in Texas, particularly within STEM and Healthcare sectors, while Pennsylvania and New York tend to see a more male-dominated trend in fraudulent activities.

The distribution of transaction amounts tends to be skewed towards smaller transactions; however, larger amounts occasionally appear in legitimate sectors, highlighting outliers that deviate from the overall fraud trend. Additionally, there is a notable spike in fraudulent activity from July to December, emphasizing the seasonal nature of fraud.

EXPLORATION AND ANALYSIS

Hypothesis - Fraudulent transactions are more likely to happen in the Business and Finance, Healthcare and Medicine, and STEM job categories, especially in transactions under dollar 5,000.

Analysis of Exhibit 3.1 shows that most fraudulent transactions occur in smaller amounts, usually less than dollar 5,000, within the Business and Finance, Healthcare and Medicine, and STEM sectors. This indicates that these areas may be more at risk for fraud in lower-value transactions. Since fraud mainly happens with smaller amounts, it's important to monitor transactions above this threshold more closely. Fraud detection systems should also include alerts for higher-value transactions to reduce risks.

Hypothesis - Fraudulent transactions are more likely to occur late at night, particularly between 21:00 and 23:59, with the Business and Finance and Healthcare sectors being the most affected.

The area chart (Exhibit 3.2) illustrates a significant increase in fraudulent transactions during late-night hours, especially between 21:00 and 23:59. This time frame aligns with the extended working hours common in sectors such as Business and Finance and Healthcare, where professionals often work late or remotely. Many employees in these fields may be working nights due to shift schedules, remote work arrangements, or urgent job demands. As a result, there is a heightened risk of fraud, particularly in online or digital transactions, during these hours. The reduced level of monitoring for late-night transactions can create more opportunities for fraud to occur.

Hypothesis - Female involvement in fraudulent transactions is particularly notable in job categories such as STEM and Healthcare.

The treemap (Exhibit 3.3) illustrates that these sectors show a higher representation of females in fraudulent activities, as indicated by the larger pink areas. This suggests that there may be a disproportionate amount of fraud committed by women in these fields.

Hypothesis - Male involvement in fraudulent transactions is particularly noticeable in job categories such as Business and Finance.

The treemap (Exhibit 3.3) shows that these sectors experience a higher volume of fraud associated with males, as indicated by the larger blue areas. This illustrates a gender-specific trend in fraudulent activities within these job categories.

Hypothesis - Fraudulent transactions tend to increase toward the end of the year, with December showing the highest volume of fraud across various job categories.

The treemap (Exhibit 3.3) reveals that December has the most

recorded fraudulent transactions, indicating a seasonal peak in fraud rates. This spike may be attributed to heightened consumer spending, year-end financial pressures, bonuses, and opportunistic fraud, as individuals and businesses engage in larger transactions at year-end. Additionally, fraudsters might exploit the holiday shopping rush and the rise in online shopping, where payment security can be more compromised.

SPECIFICATION No further analysis or visualization for Job Categories.

D. Exhibit 4 - Exploration of Fraudulent Transactions by Merchant Categories

The reason for using violin plots in these exhibits is their effectiveness in visualizing the distribution, density, and variability of fraudulent transaction hours in high-fraud cities across each state. Violin plots provide a detailed representation of the range and concentration of fraud occurrences, highlighting differences in timing patterns between locations.

These plots reveal broader trends, such as the presence of concentrated peaks during specific times of the day or more evenly distributed fraud activity over extended hours. For instance, some cities exhibit sharp peaks, indicating consistent and focused fraud during limited time windows, while others display broader distributions, reflecting greater variability with activity spread over a wider range of hours.

DATA

The data presented in Exhibit 3 is similar to that in Exhibit 4, but instead of focusing on job categories, it emphasizes merchant category spending. These categories include ”entertainment,” ”gas transport,” ”grocery net,” ”shopping pos,” and ”travel.” They represent the different types of merchant transactions associated with either fraudulent or non-fraudulent activities.

VISUALIZATION

1) Exhibit 4.1: Density Scatter Plot of Merchant Category vs Transaction Amount by Fraud Status: The scatter plot illustrates the distribution of transaction amounts (rounded) across various spending categories in 2019, differentiating between fraudulent and non-fraudulent transactions. The x-axis represents spending categories such as ”entertainment,” ”gas transport,” ”grocery net,” ”shopping pos,” and ”travel,” while the y-axis displays transaction amounts up to 30,000 units. Fraudulent transactions are highlighted in yellow, while non-fraudulent ones are marked in red.

The data reveals certain patterns: higher transaction amounts, particularly in the ”travel” category, indicate some instances of fraud. Categories like ”shopping net,” ”grocery net,” and ”shopping pos” also show notable fraudulent activity, albeit at lower transaction amounts. Conversely, categories such as ”personal care” and ”health fitness” exhibit minimal or no signs of fraud.

Overall, while fraudulent transactions (yellow dots) are relatively sparse compared to non-fraudulent ones, they tend to occur in specific categories, often associated with higher transaction values. This suggests a possible correlation between transaction amount, category, and the likelihood of fraud.

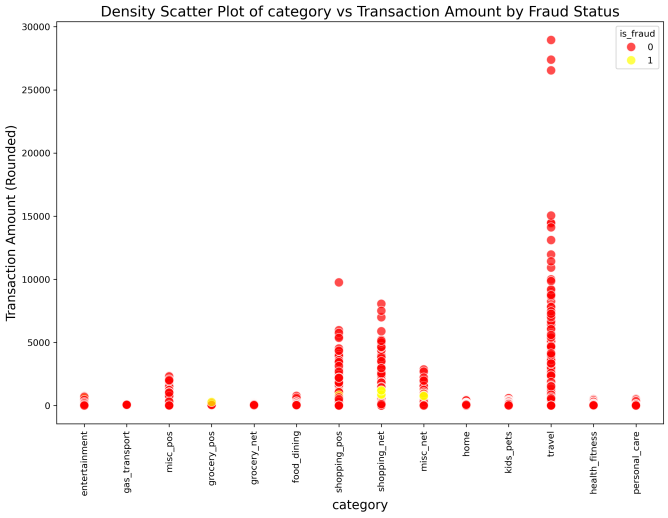

Fig. 13. Density Scatter Plot of category vs Transaction Amount by Fraud Status

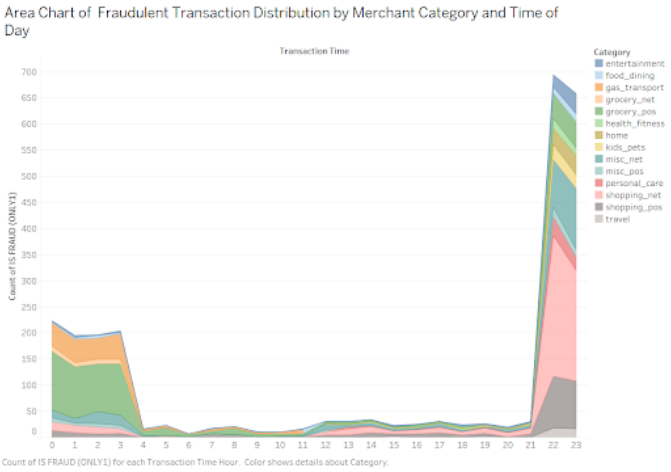

Fig. 14. Area Chart of Fraudulent Transaction Distribution by Merchant Category and Time of Day

2) Exhibit 4.2: Area Chart (Transactions Over 24 Hours for different Merchant Categories): The image depicts an area chart that illustrates the distribution of transactions over a 24-hour period, categorized by different merchant types. The chart shows a clear peak in transactions around the 21st hour, indicating a surge in activity during the evening. Notably, the ”entertainment” category dominates this peak, followed by ”food dining” and ”gas transport.” Other categories, such as ”grocery pos,” ”health fitness,” and ”home,” also show a significant presence during this time.

Additionally, fraudulent transactions are relatively high during

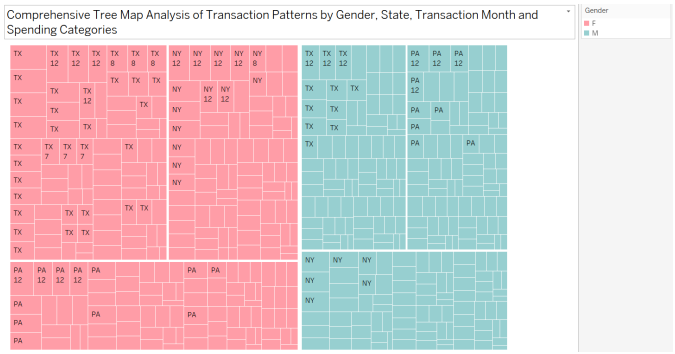

Fig. 15. Treemap Analysis of Transaction Patterns by Gender, State, Transaction Month and Spending Categories

the early morning hours (from 00:00 to 06:00). The categories of entertainment, food dining, and grocery net are the most affected during this period. However, fraudulent activity targeting the home and personal care categories remains minimal, even during peak activity times.

After 06:00, there is a sharp decline in fraudulent transactions, and their count remains low and consistent throughout the day (from 06:00 to 20:00). This decline suggests that fraudsters may avoid active hours when transactions are more likely to be noticed or scrutinized by cardholders and businesses. Point-of-sale categories, such as shopping pos and gas transport, reflect legitimate daytime activity but are not heavily targeted for fraud. Moreover, categories like home, personal care, and kids pets show minimal transaction activity throughout the 24-hour period.

3) Exhibit 4.3: Treemap (Fraudulent Transactions by State, Gender, and Merchant Category): The tree map offers a detailed analysis of transaction patterns categorized by gender and state, using pink to represent females (F) and blue to represent males (M). This color coding allows for a straightforward comparison of transaction counts. Larger sections labeled with state abbreviations (e.g., TX, NY, PA) indicate higher transaction volumes for specific states and genders, with numbers like 12 and 8 representing the number of transaction months.

States such as Texas (TX) and New York (NY) show a predominance of female-driven transactions, as evidenced by the larger pink sections. In contrast, Pennsylvania (PA) reflects a relatively balanced or male-leaning participation.

In terms of categories, larger rectangles represent significant areas such as Shopping POS, Gas Transport, Home, Shopping Net, Grocery POS, and Personal Care, which dominate the transaction volumes across states and highlight their essential nature. Smaller rectangles, representing categories like Food Dining, Entertainment, Grocery Net, and Travel, indicate lower transaction volumes and suggest discretionary spending. Categories dominated by females include Shopping POS, Shopping Net, and Personal Care, particularly in states like TX and NY. Meanwhile, male-dominated categories, such as Gas Transport, Home, and Grocery POS, are more pronounced in PA and TX.

At the state level, TX stands out for its significant transaction volumes across major categories. Females largely contribute to shopping-related transactions, while males focus on gas and transport. NY demonstrates strong female-driven contributions in Shopping Net and Personal Care, whereas PA shows a slight male dominance in Gas Transport and Home.

Overall, Shopping POS and Grocery POS emerge as the primary contributors to transaction counts, reflecting routine expenditures, while Gas Transport indicates commuting-related expenses with a higher male presence. Smaller categories such as Travel, Entertainment, and Food Dining show fewer transactions, spread across genders and states, suggesting they are associated with leisure or non-essential spending.

PERCEPTION

Fraudulent transactions are more likely to occur in categories such as ”travel,” ”shopping online,” and ”grocery online,” often involving larger amounts. In contrast, categories like ”personal care” and ”health and fitness” exhibit minimal fraud activity. Fraud tends to be concentrated during late-night and early-morning hours, specifically from 12:00 AM to 6:00 AM, with notable peaks in the ”entertainment,” ”food and dining,” and ”gas and transportation” categories. After 6:00 AM, the incidence of fraudulent activity sharply declines and remains low throughout the day. Additionally, fraud within routine spending categories like ”Shopping POS” and ”Grocery POS” is lower, while leisure or non-essential spending categories experience higher levels of fraud, particularly during certain times of the day.

KNOWLEDGE

Exhibit 4 illustrates the distribution of fraudulent transactions across various merchant categories, transaction times, and states. The data indicates that fraudulent transactions are primarily concentrated in categories with higher transaction amounts, such as ”travel.” Notably, fraud is also present in the ”shopping net” and ”grocery net” categories, even though these typically involve lower transaction amounts. Categories like ”personal care” and ”health fitness” show minimal or no fraud, suggesting they are less targeted by fraudsters.

The analysis of transaction times reveals a significant surge in fraud during the early morning hours (from 00:00 to 06:00), particularly in the ”entertainment,” ”food dining,” and ”grocery net” categories. This pattern implies that fraudsters tend to operate when transactions are less likely to be detected. After 06:00, the occurrence of fraudulent activity drops sharply and remains low throughout the day.

At the state level, the data shows a distinct geographic distribution of fraudulent transactions, with Texas and New York exhibiting a higher proportion of fraud driven by females. In contrast, Pennsylvania demonstrates a more balanced or male-dominated pattern of fraud. Additionally,

the data highlights that categories such as ”Shopping POS” and ”Grocery POS” have higher transaction volumes and are associated with routine spending, while ”Gas Transport” and ”Home” categories tend to be more gender-specific, with a predominance of male activity. Smaller categories like ”Travel,” ”Entertainment,” and ”Food Dining” show fewer transactions, indicating they are more related to discretionary or non-essential spending.

EXPLORATION AND ANALYSIS Hypothesis - Fraudulent transactions are more likely to occur in categories with higher spending values, such as ”travel” and ”shopping net,”

The scatter plot (Exhibit 4.1) shows that fraudulent transactions tend to happen more frequently in these higher-value categories, similar to the patterns observed in (Exhibit 2). While fraud is generally less common, it appears to be concentrated in transactions with larger amounts. This indicates that fraudsters may specifically target these higher-value transactions, where they are less likely to attract scrutiny.

Hypothesis - Female fraudulent transactions are more likely to occur in categories such as ”Shopping POS” and ”Personal Care.”

The treemap (Exhibit 4.3) indicates that females tend to commit fraudulent activities in these categories, particularly in Texas and New York. This trend may be attributed to consumer behaviors associated with shopping-related activities. These categories reflect common consumer behavior, highlighting that women are often the main purchasers, especially in routine, non-discretionary purchases, which makes them particularly vulnerable to fraudulent activities in these areas.

Hypothesis - Male fraudulent transactions are predominantly found in categories like ”Gas Transport” and ”Home.”

The treemap (Exhibit 4.3) shows that fraudulent activities driven by males are especially prevalent in these areas, particularly in Pennsylvania. This suggests that men are more inclined to target categories related to routine or commuting expenses. These categories encompass essential, regular expenses, such as commuting (Gas Transport) and home maintenance. Their high transaction volume and frequency may make them prime targets for fraud.

SPECIFICATION No further analysis or visualization for Merchant Categories.

E. Exhibit 5 - Exploration of Fraudulent Transactions by Gender

The rationale behind this visualization is to effectively combine geographical data with demographic insights to show the distribution of fraud by gender across states with high fraud rates. The donut charts placed over each state clearly represent the gender breakdown, making it easy to identify trends in different regions.

The use of color coding blue for females and red for males provides a straightforward visual distinction. Hover-over percentages offer precise numerical context without overwhelming the viewer. This geographic overlay enables viewers to link gender distribution trends directly to specific states, highlighting pronounced disparities where they exist.

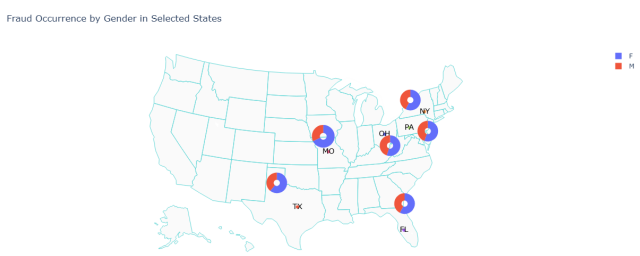

Fig. 16. Fraud Occurrence by Gender in Selected States (i.e. TX, MO, OH, PA, NY, FL)

DATA

Exhibit 5 uses data on fraudulent transactions in six U.S. states with the highest rates of fraud, as identified in Main Exhibit : Texas (TX), New York (NY), Pennsylvania (PA), Ohio (OH), Florida (FL), and Missouri (MO). It also uses information on the gender distribution of fraud cases, indicating the proportion of incidents involving females and males in each state. Furthermore, the data covers transactions that occurred from July to December.

VISUALIZATION

This plot visualizes the distribution of fraud occurrences by gender across six U.S. states: Texas (TX), New York (NY), Pennsylvania (PA), Ohio (OH), Florida (FL), and Missouri (MO). The donut charts, placed on each state’s location, represent the gender breakdown (male and female) of fraud incidents, with percentages displayed on hover. The color coding is straightforward, with blue representing female fraud occurrences and red representing male fraud occurrences, offering a clear visual distinction of gender distribution in each state.

The fraud distribution varies across states. In Texas (TX), female fraud occurrences constitute 61 percent, while males account for 39 percent. Missouri (MO) exhibits an even stronger skew, with 68 percent of fraud cases involving females and 32 percent involving males. Ohio (OH) shows a more balanced distribution, with 55.3 percent female and 44.7 percent male fraud cases. Florida (FL) follows a similar trend, where 57.6 percent of fraud occurrences involve females, and 42.4 percent involve males. Pennsylvania (PA) sees 55.4 percent female and 44.6 percent male fraud occurrences, while New York (NY) displays a similar distribution, with females constituting 57.6 percent of fraud cases and males making up 42.4 percent.

PERCEPTION

The donut charts on the map of the USA illustrate that female

fraud cases surpass male fraud cases in all six high-fraud states, with varying degrees of disparity.

KNOWLEDGE

Missouri shows the largest gender gap, with 68 percent of fraud cases involving females and 32 percent involving males. Texas also demonstrates a significant skew, with females making up 61 percent of fraud cases.

In contrast, states like Ohio and Pennsylvania exhibit a more balanced gender distribution, where females account for approximately 55-56 percentof cases and males for 44-45 percent. Similarly, Florida and New York show that females represent about 57-58 percent of fraud cases.

These findings emphasize a consistent trend of higher female involvement in fraud across these states, especially pronounced in Missouri and Texas.

EXPLORATION AND ANALYSIS

Hypothesis - The involvement of females in fraudulent transactions is more common than that of males in several high-fraud states across the U.S.

The donut chart shows that in many of these states, such as Texas, Missouri, and Florida, cases of fraud involving females account for a larger percentage of the total fraud incidents. This indicates that women may be disproportionately engaged in fraud in these regions. Specifically, the high rates of female-related fraud in states like Texas (61 percent) and Missouri (68 percent) suggest potential regional trends in gender-based fraudulent activity or be linked to lack of digital literacy among women.

SPECIFICATION

No further analysis or visualization for Gender in high fraud states

F. Exhibit 6 - Exploration of Fraudulent Transactions in Common States by Transaction Hour

The reason for using violin plots in these exhibits is their effectiveness in visualizing the distribution, density, and variability of fraudulent transaction hours in high-fraud cities across each state. Violin plots provide a detailed representation of the range and concentration of fraud occurrences, highlighting differences in timing patterns between locations.

These plots reveal broader trends, such as the presence of concentrated peaks during specific times of the day or more evenly distributed fraud activity over extended hours. For instance, some cities exhibit sharp peaks, indicating consistent and focused fraud during limited time windows, while others display broader distributions, reflecting greater variability with activity spread over a wider range of hours.

DATA

The data presented in Exhibit 6 focuses on fraudulent transaction patterns in key states highlighted in Exhibit 1, specifically Texas (TX), Pennsylvania (PA), and New York (NY). This analysis includes cities with high fraud rates from these states: in Texas, the cities are Dallas, Houston, Roma, Lolita, and Meridian; in Pennsylvania, they are Rock Glen, Vanderbilt, Clure, Corsica, and Skytop; and in New York, they include New York City, Howes Cave, Melville, Garrettsville, and Jordanville. The data tracks the timing of transactions using the transaction hour, which ranges from 0 to 23 hours.

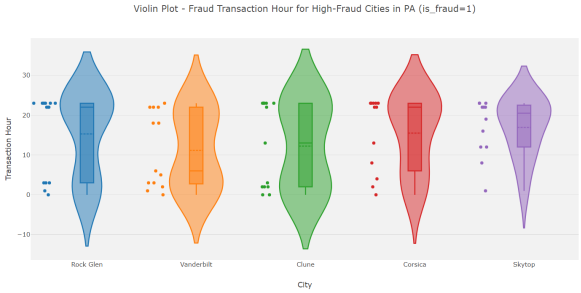

Fig. 17. Violin Plot - Fraud Transaction Hour for High-Fraud Cities in PA (is fraud=1)

VISUALIZATION

1) Exhibit 6.1: Violin Plot (Transaction Hour Distribution in High Fraud Cities - PA): The figure (17) visualizes the distribution of fraudulent transaction hours across five cities in Pennsylvania: Rock Glen, Vanderbilt, Clure, Corsica, and Skytop. It highlights distinct patterns and variations in the timing of fraudulent activities within these high-fraud areas, shedding light on city-specific behavioral trends.

Rock Glen exhibits a moderately broad and slightly asymmetric shape, indicating that fraudulent activities occur throughout the day, with a noticeable peak around midday. This suggests heightened fraud activity during those hours.

Vanderbilt has a wider and more symmetric shape, revealing significant fraud activity that spans a broad time range from early morning to late evening. This implies a steady distribution of fraud throughout the day.

Clure shows a tall and sharp peak centered around midday, reflecting a highly concentrated pattern of fraudulent transactions during this specific time window, with minimal variation in timing.

Corsica features a compact and narrower shape, where fraudulent activities are primarily clustered in the afternoon. This indicates limited variability and a shorter time range for fraudulent events.

Skytop presents a slightly wider and more uniform shape, indicating that fraud activities span a broader time range from morning into late evening hours. However, the density of occurrences is comparatively lower than in the other cities.

2) Exhibit 6.2: Violin Plot (Transaction Hour Distribution in High Fraud Cities - TX): The figure(18) illustrates the distinct timing patterns of fraudulent transactions in several cities in Texas, including Dallas, Lolita, Houston, Meridian, and Roma.

Dallas exhibits a concentrated pattern of fraud activity,

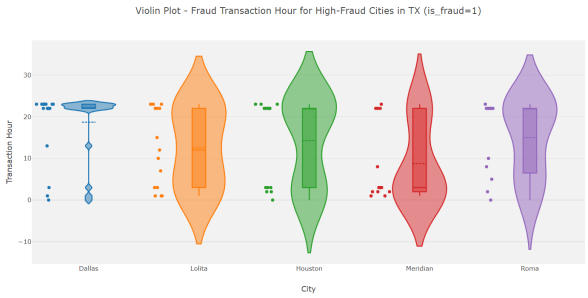

Fig. 18. Violin Plot - Fraud Transaction Hour for High-Fraud Cities in TX (is fraud=1)

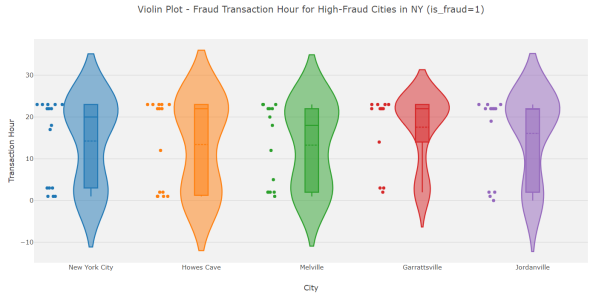

Fig. 19. Violin Plot - Fraud Transaction Hour for High-Fraud Cities in NY (is fraud=1)

primarily occurring during the afternoon hours. Its distribution has a narrow and symmetrical shape, indicating a specific and consistent time window for fraudulent behavior. In contrast, Lolita shows a broader distribution of fraudulent transaction hours, with significant activity observed late at night and early in the morning. This wider shape reflects greater variability in the timing of fraudulent transactions.

Houston experiences a peak in fraudulent transactions during daytime hours, characterized by a sharper and more defined shape that aligns closely with typical business hours. Meridian's fraudulent activities are mainly concentrated around midday, resulting in a compact shape that demonstrates less variability in timing compared to the other cities. Lastly, Roma shows a wide range of fraud transaction hours, with activity occurring from early morning to late evening. This broader shape indicates consistent occurrences of fraud throughout the day.

3) Exhibit 6.3: Violin Plot (Transaction Hour Distribution in High Fraud Cities - NY): The figure (19) visualizes the distribution of fraudulent transaction times across New York City, Howes Cave, Melville, Garrettsville, and Jordanville. It highlights the variations in the timing and density of fraudulent activities in these locations.

New York City exhibits a moderately broad shape, indicating that fraudulent activity occurs throughout the day, with a slight concentration during midday hours. Howes Cave has a wider and more symmetrical shape, reflecting significant fraud activity that persists from early morning to late evening.

Melville shows a sharp and narrow shape centered around midday, suggesting a consistent and concentrated pattern of fraudulent transactions. Garrettsville presents a compact and narrower shape, with fraudulent activities primarily occurring in the afternoon, indicating limited variability in timing.

Jordanville features a taller and thinner shape, with fraud activity spanning a wider time range, including early morning and late evening hours, although the density of activities is lower compared to the other cities. These patterns and shape variations provide valuable insights into the timing and consistency of fraudulent behavior across these high-fraud areas in New York.

PERCEPTION

Exhibit 6 presents a visualization of fraudulent transaction hours across five cities in Pennsylvania (PA), Texas (TX), and New York (NY). This distribution of fraud is illustrated using violin plots, which depict how fraudulent activity fluctuates throughout different times of the day in these high-fraud areas. Each city exhibits unique patterns, indicating specific time windows when fraudulent transactions peak.

KNOWLEDGE

The analysis of transaction hours reveals distinct patterns of fraud timing across various cities. In Pennsylvania, cities like Rock Glen experience a peak in fraudulent activity during midday, while others, such as Vanderbilt, show a consistent spread of fraud throughout the day. In Texas, cities like Dallas have a concentration of fraud in the afternoon, whereas Lolita exhibits more late-night to early-morning fraudulent activities. In New York, fraud activities also show variability; for instance, Melville has a sharp peak around midday, while Jordanville demonstrates a wider distribution of fraud throughout the day, although with lower density. These patterns suggest that fraudulent activities in different cities are shaped by unique local behaviors, likely influenced by times of day when there is less scrutiny or a higher likelihood of successful fraudulent transactions.

EXPLORATION AND ANALYSIS

Hypothesis: In cities in Pennsylvania with high fraud rates, fraudulent transactions often happen around midday. This shows that fraud occurs more at certain times of the day.

The data (see Exhibit 6.1) shows that places like Rock Glen and Clure experience noticeable peaks of fraud in the middle of the day. This means these areas see more fraud activity during these hours, suggesting concentrated fraud during this time.

Hypothesis: In Texas cities with high fraud levels, fraudulent transactions occur at various times, including late at night and early in the morning.

The violin plot for Texas cities (Exhibit 6.2) shows that fraud happens at different times of the day. For example, cities like Lolita and Roma have fraud activity in the early morning, afternoon, and late-night. This indicates that fraud behavior varies greatly in Texas cities, without specific times when it peaks. So, fraudulent activity here spreads across both day and

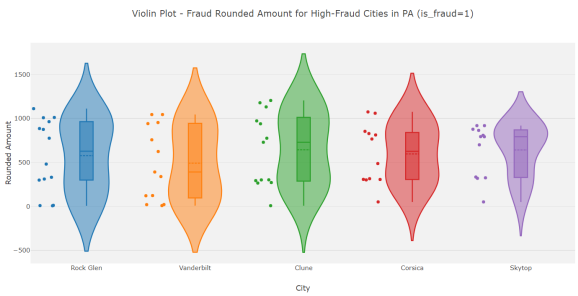

Fig. 20. Violin Plot - Fraud Rounded Amount for High - Fraud Cities in PA (is fraud=1)

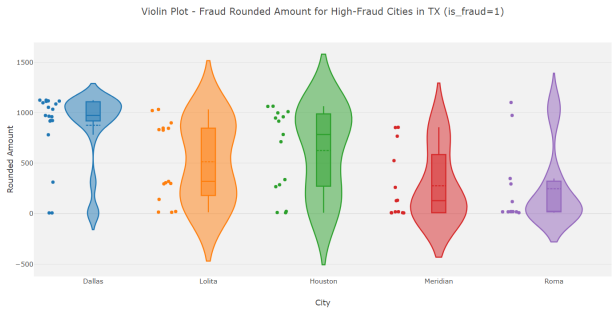

Fig. 21. Violin Plot - Fraud Rounded Amount for High - Fraud Cities in TX (is fraud=1)

night.

Hypothesis: In high-fraud cities in New York, fraudulent activities take place throughout the day. However, some cities have clearer peaks around midday, while others show a wider spread of fraud transactions.

According to the violin plot for these cities (Exhibit 6.3), fraud occurs all day long. Cities like Melville have a clear peak around midday, while places like New York City and Howes Cave show a broader range of fraud activity. This indicates that New York cities generally have fraud spread throughout the day, but some locations see more significant peaks at specific times.

SPECIFICATION

No further analysis or visualization on transaction hour

G. Exhibit 7 - Exploration of Transaction Amount Distributions in Common States

Plots used here are similar to those in Exhibit 6, with the same rationale applied. Violin plots effectively visualize the distribution, density, and variability of fraudulent transaction amounts, highlighting regional and temporal patterns without losing detail or clarity.

DATA

Exhibit 7 visualizes the amounts of fraudulent transactions across the same high-fraud cities in Pennsylvania (PA), Texas (TX), and New York (NY) as Exhibit 6, but instead of focusing on transaction hours, it highlights the distribution of transaction amounts.

VISUALIZATION

1) Exhibit 7.1: Violin Plot (Transaction Amount Distribution in High Fraud Cities - PA): The figure (20) illustrates the distribution of fraudulent transaction amounts across five cities in Pennsylvania known for high levels of fraud: Rock Glen, Vanderbilt, Clure, Corsica, and Skytop. It emphasizes city-specific variations in both the magnitude and concentration of fraud amounts, providing valuable insights into fraud patterns.

Rock Glen shows a moderately broad distribution, with fraudulent amounts varying from low to high. The highest density is observed near the median, but the presence of several outliers indicates occasional high-value fraudulent transactions.

Vanderbilt demonstrates a narrower and more symmetric distribution, suggesting that fraudulent amounts cluster consistently around the median, exhibiting limited variation. This indicates a steady and predictable pattern in the fraudulent amounts.

Clure features a tall and narrow shape, with fraud amounts tightly concentrated near the median. The limited range and absence of significant outliers suggest a uniform pattern in the fraudulent transaction amounts.

Corsica presents a slightly wider distribution, indicating moderate variability in fraudulent amounts. While the density is concentrated near the median, a noticeable spread towards mid-to-high values suggests that larger fraudulent transactions occur occasionally.

Skytop has a narrow and uniform distribution, with most fraudulent transaction amounts falling within a small range near the median. The lack of significant outliers highlights consistent, low-to-moderate fraudulent amounts.

2) Exhibit 7.2: Violin Plot (Transaction Amount Distribution in High Fraud Cities - TX): The figure (21) illustrates the distribution of fraudulent transaction amounts across five high-fraud cities in Texas: Dallas, Lolita, Houston, Meridian, and Roma. It highlights the variations in the concentration and spread of fraudulent transaction amounts in these locations.

Dallas has a moderately narrow and symmetrical distribution, with fraudulent amounts concentrated around the median. While most transactions fall within a moderate range, a few higher-value outliers indicate occasional large fraudulent transactions.

Lolita exhibits a broader and more symmetric distribution, reflecting significant variability in fraudulent amounts. Transactions range from low to high, suggesting a diverse range of fraudulent activity across the city.

Houston has a tall and narrow distribution, with fraudulent amounts closely clustered around the median. This pattern indicates limited variability and consistent fraudulent transaction values, with minimal outliers.

Meridian demonstrates a slightly narrower distribution, where fraudulent amounts are primarily centered around the median,

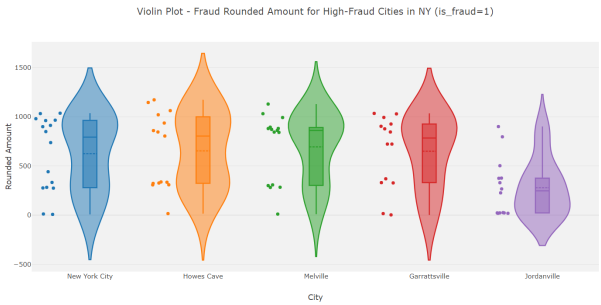

Fig. 22. Violin Plot - Fraud Rounded Amount for High - Fraud Cities in NY (is fraud=1)

with a few outliers extending to higher values. This suggests moderate variability in transaction amounts.
Roma features a distinct narrow and uniform distribution, with fraudulent amounts closely grouped around the median. There is minimal variability, and the lack of outliers indicates consistent low-to-moderate fraudulent transactions.

3) Exhibit 7.3: Violin Plot (Transaction Amount Distribution in High Fraud Cities - NY): The figure (22) illustrates the distribution of fraudulent transaction amounts across five high-fraud cities in New York: New York City, Howes Cave, Melville, Garrettsville, and Jordanville. It highlights the differences in the magnitude and density of these fraudulent transactions, offering insights into financial fraud patterns in these locations.
New York City demonstrates a moderately broad distribution of fraudulent amounts, spanning a wide range. The highest density is observed near the median, but the presence of outliers indicates occasional high-value fraudulent transactions.
Howes Cave exhibits a wider and more symmetric distribution, suggesting a diverse range of fraudulent amounts. A significant proportion falls across both low and high values, indicating consistent fraud activity with notable variability in amounts.
Melville features a tall and narrow distribution, with fraudulent amounts concentrated close to the median. This suggests limited variability and a more uniform distribution of fraudulent transactions throughout the city.
Garrettsville shows a slightly narrower and more compact distribution, where fraudulent amounts are primarily clustered around the median. A few outliers indicate sporadic high-value fraud transactions, but overall, the variability remains lower than in other cities.
Jordanville presents a unique distribution that is relatively narrow and symmetric. Fraudulent amounts are tightly clustered around the median, indicating consistent low-to-moderate transaction values with minimal outliers.

PERCEPTION

Exhibit 7 illustrates the distribution of fraudulent transaction amounts across high-fraud cities in Pennsylvania (PA), Texas (TX), and New York (NY). It highlights city-specific variations in both the magnitude and concentration of these fraudulent amounts. The violin plots demonstrate how fraudulent transaction values are distributed in each city, revealing differences in the spread and consistency of fraud over the course of the day.

KNOWLEDGE

The data provides insights into the distribution of fraudulent transaction amounts across cities with high rates of fraud. In Pennsylvania, cities like Rock Glen and Corsica show a wider range of fraudulent amounts, indicating that they occasionally experience high-value fraudulent transactions. In contrast, cities such as Clure and Skytop exhibit more consistent and lower fraudulent amounts.
In Texas, Lolita stands out with a broader distribution of amounts, suggesting a variety of fraudulent activities. Meanwhile, Dallas and Houston have narrower distributions, with fewer outlier transactions.
In New York, Howes Cave and New York City display significant variability in fraudulent amounts, with some transactions reaching higher values. On the other hand, cities like Melville and Garrettsville show more uniform distributions centered around the median.
These patterns indicate that the timing and nature of fraudulent activities differ across these cities, with some experiencing larger or more variable fraudulent amounts than others.

EXPLORATION AND ANALYSIS

Hypothesis: Cities with a wide range of fraudulent transaction amounts show different fraud patterns, including both small and high-value frauds.
Violin plots show that cities like Lolita (TX) and Howes Cave (NY) have wide and balanced distributions. This means there is a lot of variation in fraudulent transaction amounts, with many small frauds and occasional large ones. The differences in these fraud patterns may come from the specific demographics, economic activities, and spending habits of these areas.
Hypothesis: Cities with a narrow range of fraudulent transaction amounts have consistent fraud activity, with amounts clustered closely around the median and little variation.
Cities like Houston (TX), Melville (NY), and Clure (PA) have tight, narrow distributions in the violin plots, showing uniform fraud behavior. This consistency suggests that fraud patterns are predictable, possibly due to local fraud schemes or repeated behaviors related to specific types of transactions or community dynamics.
Hypothesis: Cities with large outliers in fraudulent transaction amounts tend to experience occasional high-value frauds, indicating more targeted or sophisticated fraud schemes.
The violin plots reveal outliers in cities such as Rock Glen (PA), Dallas (TX), and New York City (NY). These outliers indicate rare but significant instances of high-value fraud. Such patterns may come from targeting high-value accounts

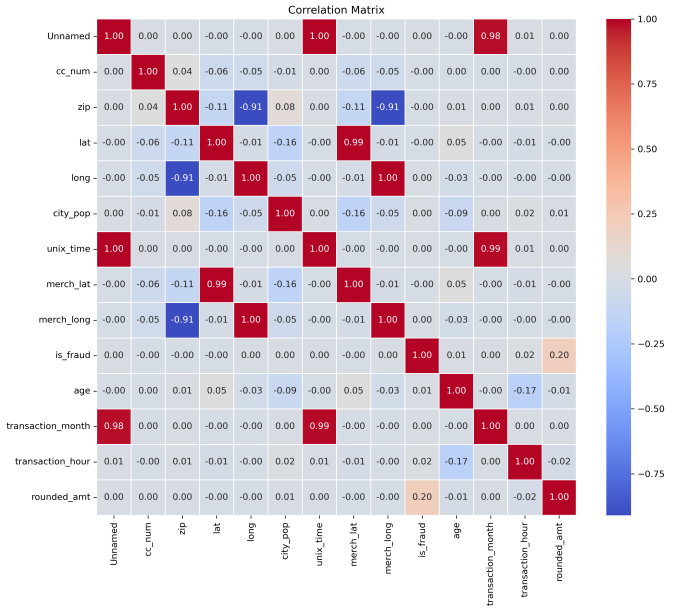

Fig. 23. Correlation Matrix of all variables

or taking advantage of weaknesses in processing high-value transactions in these urban areas.

SPECIFICATION

No further analysis or visualization on transaction amount.

H. Exhibit 8 - Statistical Insights and Relationships

Correlation and covariance heatmaps are valuable tools for revealing linear relationships among numerical features. They provide insights into how variables depend on each other and where redundancies may exist. These heatmaps help identify patterns, such as strong spatial or temporal alignments, while also emphasizing weak associations between the variable ”is fraud” and others.

DATA

The data used for this exhibit contains the following variables: is fraud, rounded amt, lat, long, merch lat, merch long, unix time, transaction date, transaction hour, transaction month, age, transaction year and others

For the analysis, we also focused on the following specific variables: is fraud, rounded amt, transaction hour, age, and transaction month. These variables were chosen for their potential to reveal insights into the patterns of fraudulent transactions. In particular, transaction hour, transaction amount, and age were selected for their possible connections to behavioral patterns that may help distinguish fraudulent transactions from legitimate ones.

VISUALIZATION

1) Exhibit 8.1: Correlation Matrix (For all variables): The correlation heatmap offers a detailed overview of the relationships between various numerical features in the dataset, with values ranging from -1 (strong negative correlation) to 1 (strong positive correlation). The diagonal values are all equal to 1, representing the self-correlation of each variable.

A prominent positive correlation is observed between geographical variables, such as latitude (lat) and merchant latitude (merch lat) at 0.99, as well as longitude (long) and merchant longitude (merch long) at 0.99. These high correlations are expected, indicating the geographical proximity between customers and merchants involved in transactions. Similarly, a high correlation of 0.99 exists between the Unix timestamp (unix time) and the transaction month (transaction month), reflecting their temporal alignment, as the Unix timestamp corresponds to specific months of the year.

Another significant positive correlation of 0.96 is found between transaction hour (transaction hour) and rounded amount (rounded amt), suggesting that certain transaction amounts tend to occur during specific hours, possibly indicating behavioral or operational patterns.

On the other hand, some strong negative correlations are also present. For example, zip code (zip) shows a negative correlation with both latitude (lat) and merchant latitude (merch lat) at -0.91. This suggests that the zip codes in the dataset are associated with decreasing latitudes, reflecting geographic or regional variations. These trends likely arise from the spatial distribution of locations within the dataset.

Interestingly, the target variable, is fraud, demonstrates weak or negligible correlations with most features. For instance, is fraud and transaction month show almost no correlation (0.00), implying that the occurrence of fraud is not seasonally dependent. Similarly, is fraud exhibits weak correlations with geographical variables like latitude and longitude, indicating that location alone does not strongly influence the likelihood of fraud. However, a moderate positive correlation of 0.22 exists between is fraud and transaction hour, suggesting that the time of the transaction may influence fraud detection, with certain hours displaying a slightly higher propensity for fraudulent activity.

The relationships between temporal variables, such as unix time and transaction month, and between latitude and longitude with their merchant counterparts, arise from their inherent conceptual connections rather than from statistical or causal relationships. These correlations emphasize the importance of careful feature selection to avoid redundancy and multicollinearity in predictive modeling.

Overall, the heatmap indicates that while some variables exhibit strong relationships, most features show weak or no correlation with is fraud. This suggests that linear relationships alone may not be sufficient to explain fraud patterns in the dataset.

2) Exhibit 8.2: Correlation Matrix (For specific variables): The correlation matrix visually represents the linear relationships between different variables in the dataset. The variable ”Is Fraud” shows a weak positive correlation with the ”Rounded Amount,” suggesting that larger transactions may be

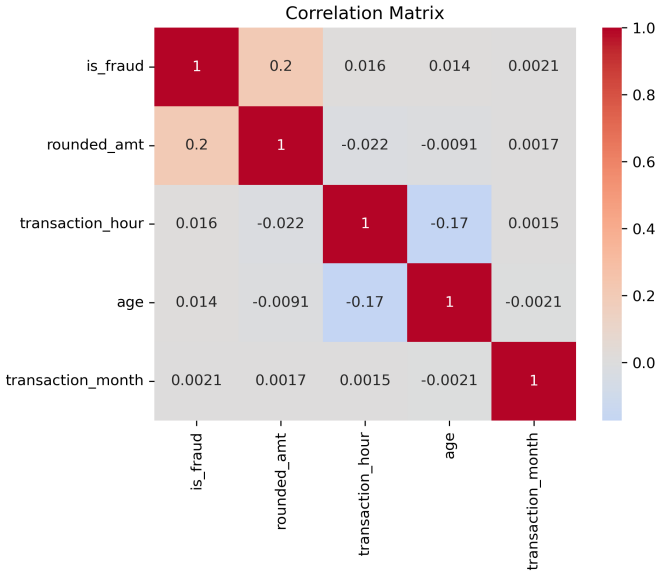

Fig. 24. Correlation Matrix of is fraud, rounded amount, transaction hour, age and transaction month

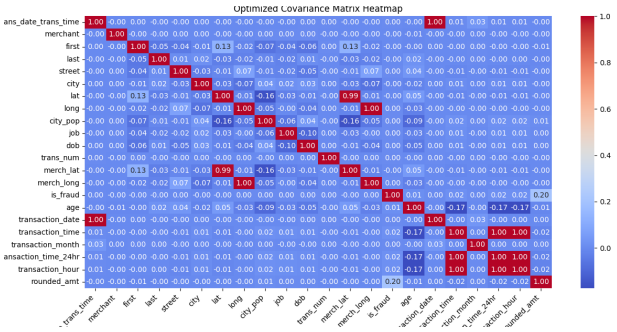

Fig. 25. Covariance heatmap of all variables

slightly more likely to be fraudulent. However, this correlation is weak, meaning it is not a strong predictor. Additionally, "Is Fraud" has negligible correlations with other variables, indicating that factors such as transaction time, age, and transaction month have little to no impact on the likelihood of a transaction being fraudulent.

The "Rounded Amount" variable exhibits a strong self-correlation, which is expected, as any variable is perfectly correlated with itself. It also shows weak correlations with other variables, suggesting that rounded amounts are largely independent of factors like transaction time, age, and month of the transaction.

"Transaction Hour" displays a weak negative correlation with age, indicating that younger individuals may tend to make transactions at different times compared to older individuals. It also has negligible correlations with other variables, suggesting that transaction time is not strongly related to fraudulent activity or rounded amounts.

Age shows a weak negative correlation with transaction hour, which aligns with the observation regarding younger individuals and their transaction timings. This variable also exhibits negligible correlations with other factors, implying that age is not a significant predictor of fraud or transaction patterns.

Finally, "Transaction Month" has negligible correlations with all variables, suggesting that the month in which a transaction occurs has little to no impact on any other variables, including fraudulent activity.

The overall analysis reveals that most variables in this dataset are weakly correlated which means that predicting fraudulent activity based on these factors alone might be challenging.

3) Exhibit 8.3: Covariance Heatmap (Feature Relationships): The covariance heatmap offers insights into the relationships between different features in the dataset. The diagonal values, all equal to 1.00, reflect the self-covariance of each feature, indicating that a feature is always perfectly correlated with itself.

High positive covariance values, represented by red areas on the heatmap, indicate a strong positive linear relationship between certain features. Conversely, features with consistently low covariance may have less influence or may be independent of others in the dataset.

Each value in the matrix represents the covariance between a pair of features, with values ranging from -1 (strong negative relationship) to 1 (strong positive relationship).

It's important to note that some feature connections—such as those among time variables (like transaction time, transaction time 24hr, and transaction date) or geographic coordinates (like merchant lat and merchant long)—are not indicative of statistical relationships. Instead, they arise from their inherent conceptual linkages. For instance, transaction time, transaction time 24hr, and transaction date demonstrate high covariance due to their shared temporal characteristics, as they describe different aspects of time related to the same transactions.

Similarly, merchant lat and merchant long may show non-zero covariance because they represent paired coordinates for geographic locations. Their connection is rooted in their role in specifying locations rather than reflecting a statistical relationship. Additionally, the feature is fraud may exhibit varying levels of covariance with other features, which could assist in identifying potential predictors for fraud detection.

4) Exhibit 8.4: Covariance Heatmap (For specific variables): The heatmap illustrates a covariance matrix that offers insights into the linear relationships among the variables: is fraud, rounded amt, transaction hour, age, and transaction month. The diagonal elements of the matrix, each with a value of 1.00, indicate the variance of each variable with itself, which signifies perfect correlation. The off-diagonal values reflect the strength and direction of the relationships between different pairs of variables.

Starting with is fraud, we observe a mild positive covariance (0.20) with rounded amt, suggesting a slight association between fraudulent transactions and the amounts involved. However, is fraud shows very weak or negligible covariance

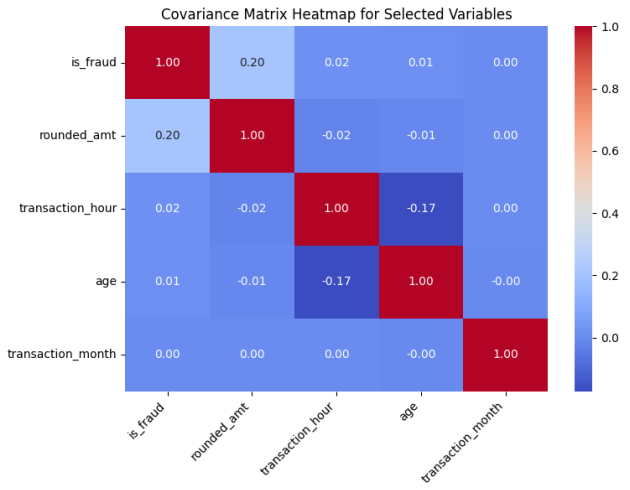

Fig. 26. Covariance heatmap of is fraud, rounded amount, transaction hour, age and transaction month

with other variables, such as transaction hour (0.02), age (0.01), and transaction month (0.00), indicating that these variables have little to no linear relationship with the occurrence of fraud.

The variable rounded amt demonstrates strong self-correlation (1.00), as expected, while its relationships with most other variables are weak. The covariance values for rounded amt with transaction hour, age, and transaction month are close to zero, suggesting that rounded amounts are largely independent of the timing of transactions, the age of the individual, or the month in which the transaction occurs.

Interestingly, transaction hour and age exhibit a mild negative covariance (-0.17), implying that certain transaction times may inversely correlate with individuals' ages. This could suggest that younger individuals tend to transact at specific hours that older individuals do not, or vice versa. Beyond this, transaction hour shows minimal relationships with other variables, while age has weak covariance with other variables, apart from its relationship with transaction hour.

Lastly, transaction month displays near-zero covariance with all other variables, indicating no meaningful relationship with is fraud, rounded amt, transaction hour, or age. This suggests that the month of a transaction is likely independent of the other variables in this dataset.

Overall, the analysis reveals that most variables in this dataset exhibit weak correlations.

PERCEPTION

The correlation and covariance heatmaps provide a comprehensive overview of the relationships between various features. These visualizations indicate that geographical variables, such as latitude and longitude, exhibit very strong correlations, which reflect the expected spatial proximity between customers and merchants. Temporal variables, like Unix timestamp and transaction month, also show high correlation, consistent with the time-based nature of the data. In contrast, many features display weak or negligible correlations with the fraud variable (is fraud), suggesting that location, transaction time, and amount are not strong indicators of fraud. These findings imply that while some variables are strongly related, their linear associations may not be sufficient for accurate fraud prediction.

KNOWLEDGE The correlation matrix and covariance heatmaps provide several key insights. Firstly, the correlation values reveal significant relationships among certain features, as expected. For instance, there is a high correlation of 0.99 between latitude and merchant latitude, as well as between longitude and merchant longitude. This indicates a strong geographical relationship between customers and merchants. Additionally, there is a strong correlation of 0.99 between the Unix timestamp and transaction month, demonstrating their temporal alignment. The correlation of 0.96 between transaction hour and rounded amount suggests that transaction amounts may be linked to specific times of the day, hinting at potential patterns in transaction behavior.

However, the target variable, is fraud, shows weak correlations with most features. This is particularly noteworthy for variables like transaction month, age, and geographical location, which exhibit negligible relationships with the likelihood of fraud. The only notable correlation is with transaction hour (0.22), indicating that certain times of the day may experience slightly higher instances of fraud. The absence of strong linear relationships with fraud emphasizes that other factors beyond these variables must be considered for more accurate fraud detection.

In the covariance analysis, similar patterns are observed. The covariance matrix indicates a mild positive covariance of 0.20 between is fraud and rounded amount, but negligible relationships with transaction hour, age, and transaction month. The rounded amount variable, as expected, demonstrates strong self-correlation (1.00) but weak covariance with other features, suggesting that transaction amounts are largely independent of time or age.

Moreover, transaction hour and age show a mild negative covariance of -0.17, implying that younger individuals may transact at different times than older individuals, though this relationship is weak. Transaction month exhibits near-zero covariance with other variables, reinforcing the idea that the timing of a transaction does not significantly correlate with fraud or other transactional attributes.

Overall, the correlation and covariance analyses suggest that fraud detection may not heavily depend on linear relationships with features such as transaction hour, amount, or geographical location. The weak correlations and covariances observed indicate that predictive models for fraud detection will need to explore non-linear relationships and consider additional factors beyond those currently captured in the dataset.

EXPLORATION AND ANALYSIS

Hypothesis - The occurrence of fraud is not strongly influenced by individual variables such as geographical location, transaction time, or age, as these features show weak or

negligible correlations with the target variable, "is fraud."
The correlation matrix indicates minimal relationships between "is fraud" and most variables, including transaction month (0.00), latitude (near zero), and age (negligible). The strongest observed correlation with "is fraud" is 0.22 for transaction hour, suggesting only a slight tendency for fraud to occur at specific times. This finding implies that linear dependencies between fraud and these features are weak, highlighting the need for non-linear models or additional variables to improve fraud detection. Furthermore, cities like Houston (TX) and New York City (NY), with their diverse transaction patterns, may underscore the importance of contextual factors that are not captured by these variables.
Hypothesis - The covariance analysis between certain features, such as transaction hour and age, as well as transaction amount and fraud occurrence, indicates that while there are some mild relationships, they are not strong enough to serve as reliable predictors for fraud detection.
The covariance heatmap shows a mild negative covariance of -0.17 between transaction hour and age, and a slight positive covariance of 0.20 between the rounded transaction amount and fraud occurrence. These values suggest weak correlations between these variables. Although transaction time and amount may have some relationship with fraud incidence, their predictive power remains limited.
In cities like New York (NY) or Houston (TX), where there are large volumes of transactions, this weak covariance further reinforces the notion that effective fraud detection requires more complex, non-linear models or additional factors to accurately capture these subtle relationships.

SPECIFICATION

Since the initial analysis revealed no significant correlation or covariance, we did not pursue further exploration of the existing variables. Instead, we will adopt a Visual Analytics Model approach to shift our focus from knowledge to data. This approach will involve introducing new variables, such as (new variable 1) and (new variable 2), which recent research papers have identified as potentially relevant to fraud activity. We expect these new variables to establish stronger correlations with fraud and provide actionable insights.

PART 2- COMPARATIVE ANALYSIS FOR CREDIT CARD FRAUD (2020 and 2019)

In 2020, since we were building on previous data and needed to compare fraud cases between 2019 and 2020, we applied the Basic Visual Analytics loop. This process involved moving from Data to Visualization, then to Modeling (although a model was only applied in Exhibit 5, as the other sections did not require one), and finally to Knowledge. This approach facilitated a clear comparison and provided valuable insights into the trends and changes in fraud cases over the two years.

I. Exhibit 1 - Fraud Distribution and Trends across Common States and Cities 2020 and 2019

The rationale for using bubble charts in these visualizations is their ability to effectively convey the distribution of fraud across multiple dimensions specifically city, frequency, and intensity at a glance. By encoding the number of fraud cases in both the size and color intensity of the bubbles, the charts illustrate patterns and disparities across different cities within states. Larger, darker bubbles signify cities with significant fraud activity, while smaller, lighter bubbles indicate areas with minimal incidents.
This dual representation offers an intuitive way to compare dynamics between urban and rural areas, highlighting how metropolitan regions such as Dallas, Houston, and New York City often have the highest fraud counts due to greater transaction volumes or population density. In contrast, smaller towns with less activity are easily identifiable through their distinct visual characteristics.

DATA

Two datasets are used 2020 and 2019. The data used in this exhibit is the fraud distribution for common states as identified in main Texas, New York, and Pennsylvania during 2019 and 2020 includes information on fraudulent transactions across various cities within these states. This dataset features a binary fraud variable (0 for non-fraudulent transactions and 1 for fraudulent transactions), along with information about high-fraud cities in each state.

VISUALIZATION

1) Exhibit 1.1 - Bubble Chart (2019 TX - Fraud Distribution across Cities): The figure (27) chart illustrates the distribution of fraud cases across various Texas cities in 2019. Each bubble's color intensity and size provide insights into the total number of fraud cases in each city. The color intensity reflects the total instances where fraud is indicated (is fraud = 1), with darker blue shades representing a higher number of fraud cases. In contrast, lighter blue shades indicate cities with fewer fraud incidents.

The size of each bubble is also proportional to the total number of fraud cases, so larger bubbles such as those for Dallas and Houston represent cities that experienced the highest levels of fraud in Texas. Conversely, smaller bubbles, like those for Pecos and Harper, indicate these locations have lower levels of fraud activity.
Overall, this chart offers a clear visual representation of fraud distribution across Texas. It highlights that larger cities like Dallas and Houston are significant contributors to the state's fraud cases, as evidenced by both their large size and dark blue color. In contrast, smaller cities such as Pecos and Harper exhibit much smaller bubbles and lighter blue shades, signaling notably lower levels of fraud.

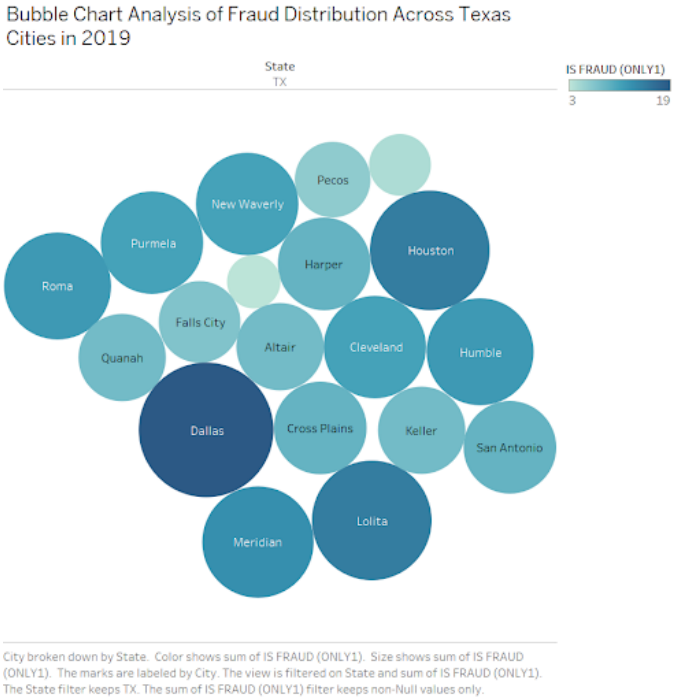

Fig. 27. Bubble Chart (2019 TX - Fraud Distribution across Cities)

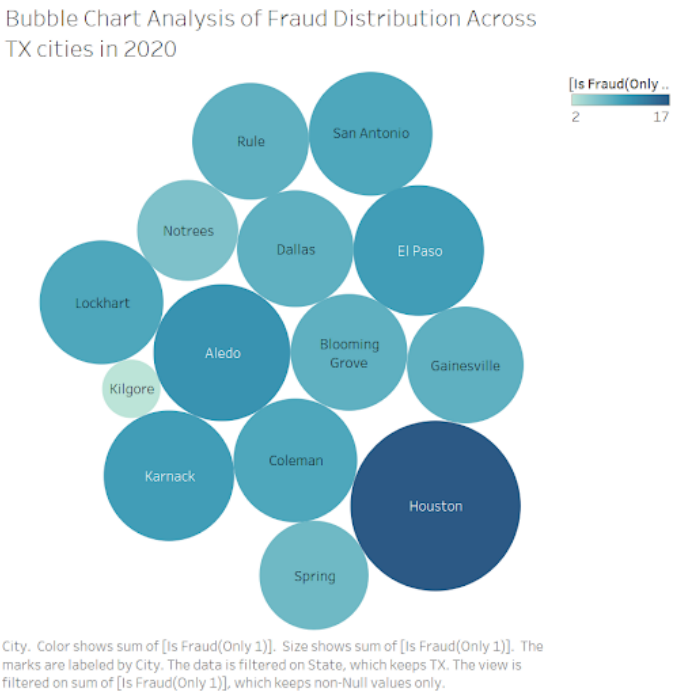

Fig. 28. Fraud Distribution across Cities of TX (2020)

2) Exhibit 1.2 - Bubble Chart (2020 TX - Fraud Distribution across Cities): The bubble chart illustrates the distribution of fraudulent cases (where Is Fraud = 1) across various cities in Texas in 2020. Each bubble represents a city, with the size of the bubble indicating the number of fraud cases and the color intensity showing the frequency of these cases darker shades indicate higher counts. Houston stands out as the most significant outlier, depicted with the largest bubble and the darkest color, indicating it had the highest number of fraud cases among the cities listed. Other cities, such as Dallas, San Antonio, and El Paso, also have relatively large bubbles, suggesting moderately high levels of fraud.

The trend indicates that larger bubbles appear in major metropolitan areas (like Houston, Dallas, and San Antonio), while smaller bubbles are found in rural or less populated cities, hinting at a possible correlation between fraud occurrences and factors such as population density and economic activity. In contrast, smaller cities like Kilgore, Notrees, Lockhart, and Karnack are represented by smaller, lighter-colored bubbles, reflecting minimal fraudulent activity.

3) Exhibit 1.3 - Bubble Chart (2020 NY - Fraud Distribution across Cities): The bubble chart illustrates the distribution of fraudulent cases (where Is Fraud = 1) across various cities in New York in 2020. Each bubble represents a city, with the size and color intensity of the bubble indicating the number of fraud cases.

Margaretville stands out as the city with the highest number of fraud cases, evident from its large bubble size and dark color. This indicates that Margaretville experienced a disproportionately high level of fraudulent activity compared to other cities in the state.

Cities such as Montrose, Rock Tavern, and Brooklyn have moderately sized bubbles, suggesting a significant, but smaller, number of fraud cases relative to Margaretville. In contrast, East Rochester, Cowlesville, Camden, and Palmyra are represented with smaller bubble sizes and lighter colors, reflecting relatively low levels of fraud.

Lastly, cities like Hudson, Stittville, and Saint Bonaventure show the smallest and lightest bubbles, indicating minimal fraud cases. Overall, the distribution pattern in this chart suggests that fraud in New York during 2020 was concentrated in specific cities, particularly Margaretville, while smaller and less populous towns experienced comparatively fewer cases.

4) Exhibit 1.4 - Bubble Chart (2019 NY - Fraud Distribution across Cities): The figure (30) shows the distribution of fraud cases across various cities in New York in 2019. Similar to the Texas chart, this visualization uses size and color to represent the number of fraud cases. Larger bubbles and darker shades of blue indicate a higher number of fraud incidents, while smaller bubbles and lighter shades signify fewer cases.

New York City and Garrattsville stand out with their larger, darker bubbles, highlighting their significant contribution to the state's fraud activity. This suggests that these cities are hotspots for fraud, potentially due to their larger populations or higher transaction volumes. In contrast, cities like Newark Valley and Oakdale are depicted with smaller, lighter bubbles, indicating much lower levels of fraud activity in these areas.

Fig. 29. Fraud Distribution across Cities of NY (2020)

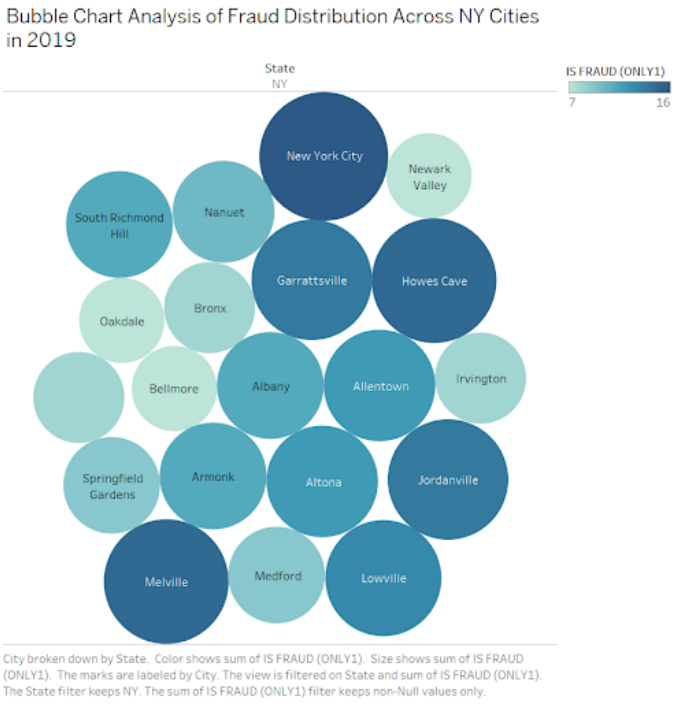

Fig. 30. Fraud Distribution across Cities of NY (2019)

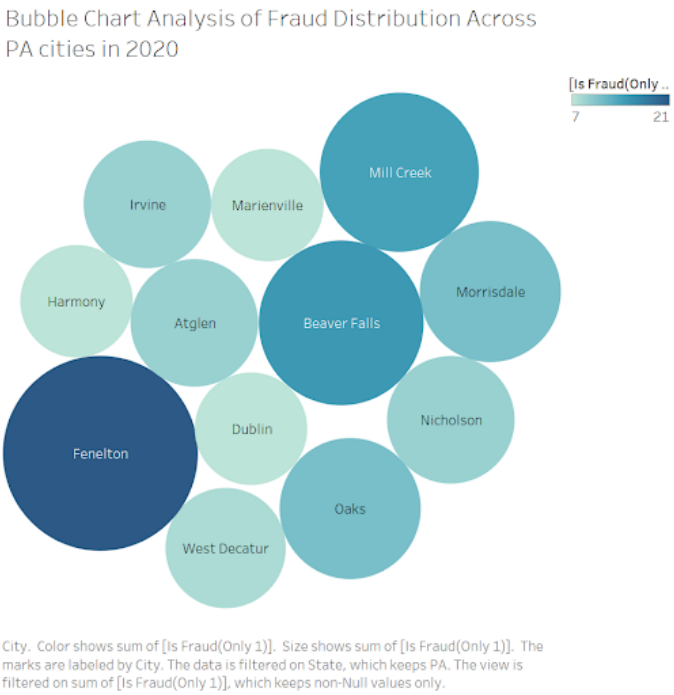

Fig. 31. Fraud Distribution across Cities of PA (2020)

These differences effectively illustrate the disparity in fraud cases between larger metropolitan areas and smaller towns.

Exhibit 1.5 - Bubble Chart (2020 PA - Fraud Distribution across Cities) The bubble chart illustrates the distribution of fraudulent cases (where Is Fraud = 1) across various cities in Pennsylvania in 2020. Each bubble represents a city, with its size reflecting the number of fraud cases and its color intensity indicating the magnitude, with darker shades corresponding to higher counts.

Among the cities, Fenelton stands out with the largest bubble and the darkest color, indicating it had the highest number of reported fraud cases. This highlights Fenelton as a significant hotspot for fraudulent activity in Pennsylvania during this period. Beaver Falls, with a slightly smaller and less intense bubble, also shows a relatively high number of fraud cases.

Other cities, such as Mill Creek, Atglen, and Dublin, have moderate bubble sizes, reflecting fewer fraud cases compared to Fenelton and Beaver Falls, but still notable within the dataset. In contrast, cities like Harmony, Marienville, Nicholson, and West Decatur are represented by smaller and lighter-colored bubbles, indicating minimal fraudulent activity in these areas.

The variation in bubble sizes and colors suggests a clear disparity in fraud prevalence, with larger and more active cities experiencing higher levels of fraud. This distribution implies that fraud occurrences in Pennsylvania are concentrated in a few key areas, while smaller or rural towns report lower levels of fraudulent activity.

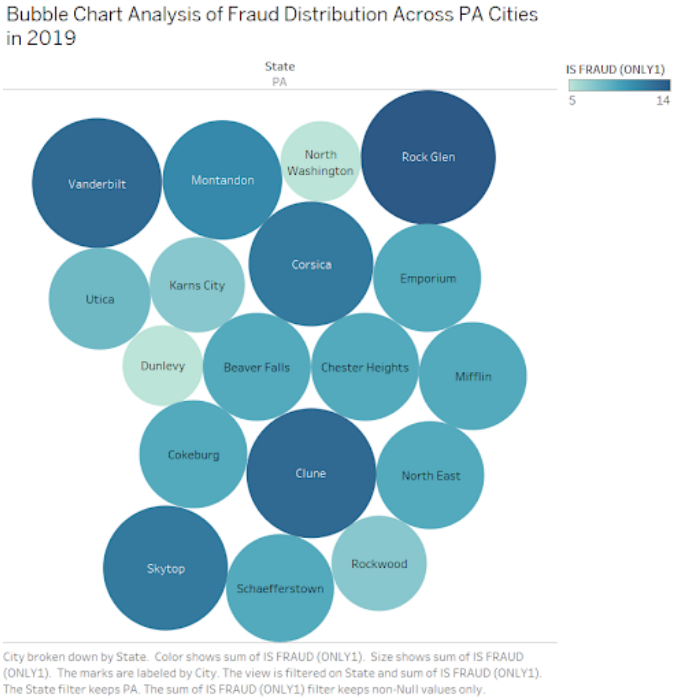

Fig. 32. Fraud Distribution across Cities of PA (2019)

5) Exhibit 1.6 - Bubble Chart (2019 PA - Fraud Distribution across Cities): The figure (32) depicts the distribution of fraud cases across various cities in Pennsylvania (PA) in 2019. Similar to the charts for Texas and New York, this visualization uses bubble size and color intensity to represent the frequency of fraud cases. Larger bubbles and darker blue shades indicate a higher number of fraud incidents, while smaller bubbles and lighter shades signify fewer cases.

In particular, Rock Glen and Clune are highlighted with prominent, darker-colored bubbles, reflecting their higher levels of fraud activity. This suggests that these areas may have a notable concentration of fraudulent activities. In contrast, cities like Dunlevy and North Washington are represented by smaller, lighter-colored bubbles, indicating relatively low levels of fraud activity in these locations.

MODEL

No models have been used for this.

KNOWLEDGE

In 2019 and 2020, several cities in Texas, New York, and Pennsylvania experienced high levels of fraud. Common high-fraud cities during this period included Houston and Dallas in Texas, as well as New York City in New York. These cities remained significant hotspots for fraudulent activity due to their large populations and substantial economic activity.

On the other hand, some uncommon high-fraud cities exhibited notable changes between the two years. Margaretville in New York and Fenelton in Pennsylvania emerged as fraud hotspots in 2020, even though they were less significant in 2019. These changes may be influenced by limitations in the dataset, including inconsistent reporting, differences in data collection methods, or missing records for certain areas, all of which can skew the perceived significance of specific cities. Additionally, shifts in regional demographics and economic activity such as population growth or an increase in financial transactions may create new opportunities for fraudulent behavior. Smaller cities might see a rise in fraud incidents due to vulnerabilities like less sophisticated prevention measures or the emergence of new businesses, which are often easier targets for fraudsters.

J. Exhibit 2 - Comparison of Fraud by Generation 2020 and 2019

The rationale for using stacked bar charts in these visualizations is their effectiveness in clearly displaying the distribution of fraud cases across different generations while also segmenting the data by state. This format enables a comparative analysis of fraud prevalence among generational cohorts and highlights specific trends within each state.

By stacking generational data for each state, the charts provide an intuitive understanding of both total fraud counts and the relative contribution of each generation.

DATA

The data presented in this exhibit includes fraud cases from high-fraud states during 2019 and 2020, categorized by generation (Baby Boomers, Generation X, Millennials, and the Silent Generation) to illustrate generational trends in occurrences of fraud.

VISUALIZATION

1) Exhibit 2.1 - Stacked Bar Chart (2019 Fraud by Generation in High Fraud States): The stacked bar chart from 2019 illustrates the distribution of fraud cases across different generations: Baby Boomers, Generation X, Millennials, and the Silent Generation. These cases are segmented by state (FL, MO, NY, OH, PA, TX). Generation X exhibits the highest number of fraud cases overall, closely followed by Millennials, indicating that younger generations are more susceptible to fraud. Baby Boomers account for a moderate number of fraud cases, while the Silent Generation consistently has the lowest fraud counts across all states.

Among the states, Texas (green) stands out as a significant contributor to fraud cases across all generations, particularly for Generation X and Millennials. Florida (blue) also has a substantial share, especially among Baby Boomers and Generation X, signifying a high vulnerability in these groups. New York (orange) shows a consistent contribution across generations, with a particularly strong presence in Millennials. Missouri (brown) and Ohio (light blue) make moderate contributions, while Pennsylvania (red) shows slightly lower fraud counts compared to other states.

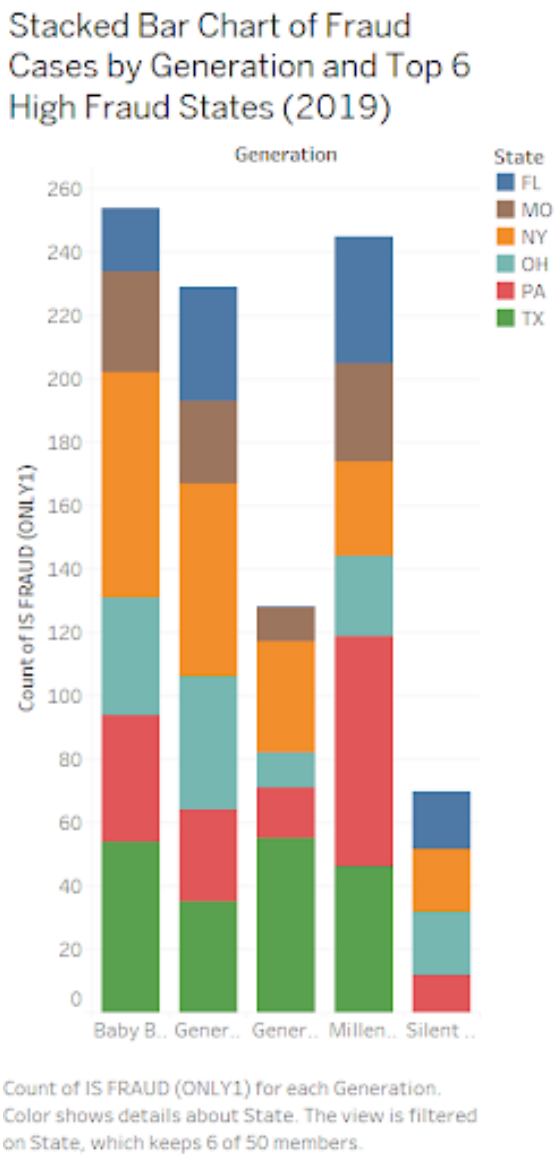

Fig. 33. Stacked Bar Chart (2019 Fraud by Generation in High Fraud States)

Generation X displays a widespread distribution of fraud across all states, with Texas, Florida, and New York as the primary contributors. Millennials follow a similar trend, although New York accounts for a relatively larger share in this group. Baby Boomers exhibit moderate levels of fraud, with Florida and Texas leading in counts for this generation. The Silent Generation's cases are minimal, with Texas and Florida continuing to be the dominant contributors.

2) Exhibit 2.2 - Stacked Bar Chart (2020 Fraud by Generation in High Fraud States): The stacked bar chart illustrates the occurrence of fraud by generation in 2020 across six states with the highest rates of fraud: California (CA), New York (NY), Pennsylvania (PA), South Carolina (SC), Texas (TX), and Virginia (VA).

In California(Gray), fraud cases are most prevalent among

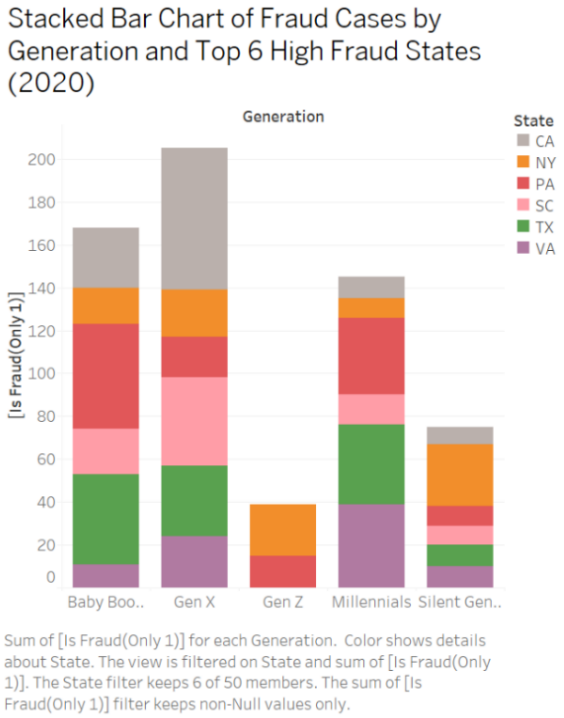

Fig. 34. Stacked Bar Chart (2020 Fraud by Generation in High Fraud States)

Generation X (Gen X), followed by Baby Boomers. Generation Z and Millennials have relatively lower fraud cases, with Gen Z being the least affected. In New York, Gen X again leads in fraud occurrences, closely followed by Baby Boomers. Millennials exhibit a moderate number of fraud cases, while Gen Z experiences fewer cases, mirroring the trend seen in California.

Pennsylvania(Red) shows a high number of fraud cases among Baby Boomers, with Gen X also contributing significantly. Millennials and Gen Z have fewer cases in this state, and the Silent Generation reports the least number of fraud incidents.

In South Carolina (Pink), Baby Boomers are the most affected generation, followed by Gen X. The Silent Generation also shows some instances of fraud, but Millennials and Gen Z have the least occurrences in this state.

Texas (Green) reports a significant number of fraud cases among Gen X, followed by Baby Boomers. Millennials contribute moderately to the total fraud occurrences, while both Gen Z and the Silent Generation show fewer cases.

Virginia (Lavender) experiences notable fraud cases among Baby Boomers and Gen X. The Silent Generation also reports some incidents, while Millennials and Gen Z experience lower frequencies of fraud compared to other generations.

Across the six states, Gen X consistently reports the highest number of fraud occurrences, followed closely by Baby Boomers. Millennials and Gen Z experience relatively fewer fraud cases, although there is some variation in the distribution across different states. Each state exhibits slightly different patterns, with California and Texas having the highest overall

fraud rates, primarily impacting Gen X and Baby Boomers.

MODEL

No models have been used for this.

KNOWLEDGE

The insights gained from the stacked bar charts in Exhibits 2.1 and 2.2 reveal significant patterns in the distribution of fraud across different generational cohorts in high-fraud states. In 2019, Generation X consistently exhibited the highest number of fraud cases, followed closely by Millennials. This suggests that younger generations, particularly in states like Texas and Florida, are particularly vulnerable to fraud.

Several factors contribute to this vulnerability. For instance, Generation X is often targeted by tech support scams, healthcare frauds, and scams related to managing technology for their families, all of which exploit their busy, multitasking lifestyles [12]. Millennials face similar risks, frequently targeted by job scams, student loan fraud, and fake check scams, which often take advantage of their financial pressures and extensive online interactions. Both generations are active online, making them prime targets for scams that exploit their tech-savviness and trust in digital communication.

In contrast, the Silent Generation showed minimal fraud cases in 2019, largely due to their limited internet usage, which reduces their exposure to online fraud. However, they remain susceptible to traditional scams like lottery fraud and investment scams, which prey on their retirement savings and potential isolation [12].

In 2020, the trend of Generation X being the most affected continued, with Baby Boomers also experiencing a significant number of fraud cases, particularly in states like California, Pennsylvania, and Texas. Baby Boomers are often targeted by investment scams and romance fraud, which exploit their financial security and, in some cases, loneliness [12]. Fraud targeting Millennials persisted, especially concerning job scams and student loan fraud. Meanwhile, Generation Z, the youngest cohort, was the least affected but remained vulnerable to online shopping scams, investment fraud (especially through social media and cryptocurrency schemes), and employment scams, which exploit their familiarity with technology and heavy social media use [12]. These insights illustrate that fraud is more prevalent among middle-aged to older generations (Generation X and Baby Boomers), but younger generations like Millennials and Generation Z also face significant risks. The patterns of fraud distribution across states highlight how regional factors, such as economic conditions and demographic compositions, influence the susceptibility of different generations to fraud.

K. Exhibit 3 - Comparison of Fraud by Generation 2020 and 2019

The reason for using word clouds in these visualizations is that they offer a quick and clear summary of transaction frequency across different merchant categories. The size

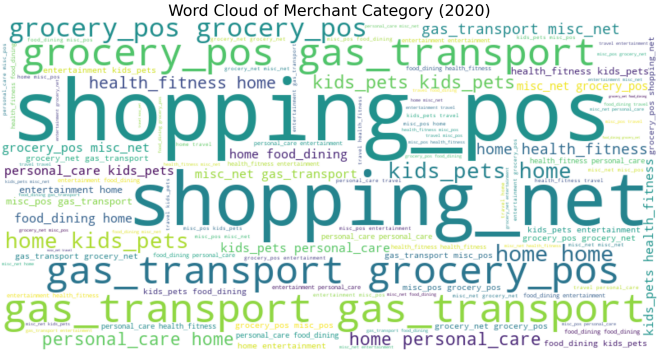

Fig. 35. Word Cloud of Merchant Category (2020)

of each word directly reflects the volume of transactions, allowing for an easy identification of the most and least common categories at a glance.

DATA

This dataset features transactions within merchant categories such as shopping pos (point-of-sale shopping), shopping net (online shopping), grocery pos (grocery purchases at physical stores), gas transport (gas and transportation-related transactions), and home (home-related expenses).

VISUALIZATION

1) Exhibit 3.1 - Word Cloud of Merchant Category (2020): The word cloud illustrates the merchant categories for transactions in 2020, with the size of each word indicating the frequency of transactions within that category. The most prominent categories in the word cloud include shopping pos (point-of-sale shopping), shopping net (online shopping), grocery pos (grocery purchases at physical stores), gas transport (gas and transportation-related transactions), and home (home-related expenses). This suggests that these were the most common areas of spending in 2020.

Smaller categories such as personal care, health fitness, kid's pets, entertainment, and food dining show less frequent but still notable spending. The dominance of categories like shopping net may reflect a shift toward online shopping, likely influenced by the COVID-19 pandemic. At the same time, the significance of gas transport and grocery pos indicates essential expenditures during that year.

2) Exhibit 3.2 - Word Cloud of Merchant Category (2019): The word cloud for merchant categories in 2019 highlights the most frequent transaction types for that year. The dominant categories include shopping at point-of-sale (shopping pos), gas and transportation-related expenses (gas transport), grocery purchases at physical stores (grocery pos), and online shopping (shopping net). These categories indicate a strong emphasis on in-person retail shopping, essential spending on groceries and gas, and some degree of online shopping.

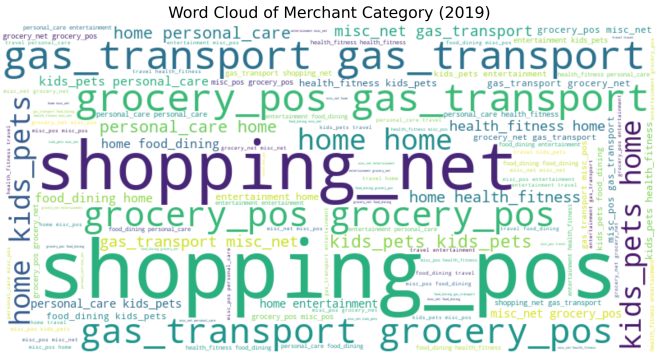

Fig. 36. Word Cloud of Merchant Category (2019)

Smaller but notable categories include home, personal care, health and fitness, kids and pets, food and dining, and entertainment. These categories suggest a mix of discretionary and necessity-driven spending. The prominence of gas transport and shopping pos indicates that in-person transactions and travel-related expenses were significant in 2019, likely reflecting consumer behavior before the pandemic.

MODEL

No models have been used for this.

KNOWLEDGE

The insights gained from the word clouds in Exhibits 3.1 and 3.2 highlight significant trends in consumer spending across different merchant categories in 2019 and 2020. In 2020, there was a noticeable shift in consumer behavior, likely influenced by the COVID-19 pandemic, as indicated by the increased prominence of categories such as online shopping (shopping net), grocery purchases (grocery pos), and transportation expenses (gas transport). Online shopping saw a substantial rise, while spending on essentials like groceries and transportation remained crucial. This suggests that, despite economic disruptions, basic needs and online shopping emerged as the primary drivers of consumer spending.

In contrast, the word cloud for 2019 reveals a stronger focus on in-person shopping (shopping pos) and transportation costs (gas transport), reflecting consumer behavior prior to the pandemic. Although online shopping (shopping net) was still relevant, the prevailing trend was centered around physical retail and travel-related expenditures, indicating a more traditional spending pattern. Smaller categories, such as personal care and entertainment, represented discretionary spending but were less prevalent compared to essential items. These findings underscore the pandemic's impact in 2020, illustrating a shift toward online shopping and essential spending, while 2019's habits were more consistent with pre-pandemic norms.

L. Exhibit 4 - Comparison of Fraud with Job Categories by State and City 2020 and 2019

The purpose of the visualization is to provide a clear and organized analysis of fraud occurrences across different job categories, cities, and states for the years 2019 and 2020. The sunburst chart is specifically designed to illustrate the relationships among states, cities, and job categories, revealing patterns and concentrations of fraud over these years.

These charts facilitate the identification of the most fraud-prone sectors by visually highlighting job categories such as STEM, healthcare, arts and culture, and business and finance. By organizing the data at the state level and drilling down into individual cities and job categories, we can identify the areas that are most affected by fraud. This approach allows for a comprehensive understanding of how various regions and sectors contribute to overall fraud rates.

DATA

The data presented in Exhibits 4.1 and 4.2 includes instances of fraud reported across six states: New York, Texas, Pennsylvania, California, South Carolina, and Virginia for the year 2020, as well as New York, Texas, Ohio, Pennsylvania, Florida, and Missouri for 2019. The data is categorized by job sectors such as STEM, healthcare, business and finance, civilian jobs, arts and culture, education, and public service. A binary variable is used to indicate fraud, where 0 represents non-fraudulent transactions and 1 indicates fraudulent transactions. This approach helps track the occurrence of fraud in various locations and sectors.

VISUALIZATION

1) Exhibit 4.1 - Sunburst Chart (Job Categories by State, City (2020) : The sunburst chart illustrates the occurrences of fraud in 2020 across six states: New York (NY), Texas (TX), Pennsylvania (PA), California (CA), South Carolina (SC), and Virginia (VA). It highlights cities with significant fraud cases and organizes the information by job categories. At the center of the chart are the states, which branch out into individual cities and then into specific job sectors. The thickness of the segments represents the volume of fraud cases.

In Texas, the cities of Houston and Aledo showcase significant fraud occurrences, particularly in the STEM (Science, Technology, Engineering, and Mathematics) category, which dominates the state. Other notable areas of fraud include arts and culture, business and finance, and education and training.

In California, cities such as Los Angeles, San Diego, Oakland, Corona, and Indian Wells report widespread fraud, especially in STEM jobs. Civilian jobs, business and finance, and education-related sectors also exhibit a high volume of fraud, with STEM leading across most cities.

In New York, fraud is concentrated in cities like New York

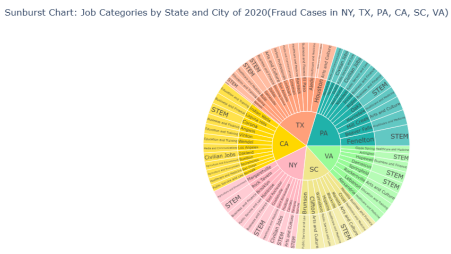

Fig. 37. Sunburst Chart (Job Categories by State, City (2020)

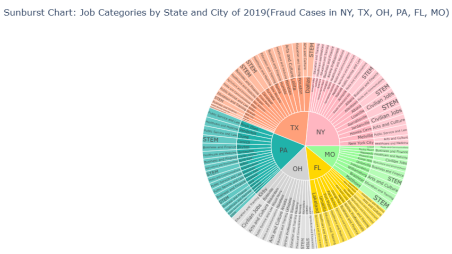

Fig. 38. Sunburst Chart (Job Categories by State, City (2019)

City, Brooklyn, Rochester, and Margaretville. Once again, STEM jobs are a major area of concern, followed by civilian jobs and arts and culture. Pennsylvania displays a high volume of fraud in STEM, particularly in Fenelton, Beaver Falls, and Altoona. Other significant categories include healthcare, arts and culture, business and finance, and public service and law, particularly in cities such as Morrisdale, Mill Creek, and Oaks.

In South Carolina, the cities of Brunson, Winnsboro, and Clifton demonstrate fraud cases primarily in the STEM category, followed by arts and culture and education and training. Virginia shows widespread fraud in STEM across cities like Arlington, Alexandria, and Springfield. Other sectors contributing to the fraud cases in cities such as Hopewell, Glade Spring, and Ruckersville include healthcare and medicine, business and finance, and arts and culture.

Overall, the chart indicates that STEM is the most fraud-prone job category across all six states, prominently affecting cities like Houston (TX), Los Angeles (CA), New York City (NY), and Fenelton (PA). While STEM stands out as the leading category, other sectors such as healthcare, civilian jobs, and arts and culture also experience significant fraud, particularly in Pennsylvania, California, and Virginia. This analysis reveals the prevalence of job-related fraud and highlights the sectors most impacted in each state.

2) Exhibit 4.2 - Sunburst Chart (Job Categories by State, City (2019) : The sunburst chart provides a comprehensive analysis of fraud occurrences in 2019 across six states: New York (NY), Texas (TX), Ohio (OH), Pennsylvania (PA), Florida (FL), and Missouri (MO). It focuses on cities with significant fraud occurrences and categorizes them based on job sectors. The chart is structured hierarchically, starting with states at the core, followed by cities, and finally their associated job categories. The thickness of each segment signifies the relative volume of fraud cases within that category and location.

In Texas (TX), the cities of Houston and Dallas show significant fraud occurrences, particularly in STEM-related jobs, which dominate the fraud cases in the state. Other notable categories include arts and culture, healthcare and medicine, and public service and law, with public service and law trailing closely behind STEM in terms of prominence. In New York (NY), fraud is concentrated in New York City and Melville. The highest occurrences are observed in healthcare and civilian jobs, indicating widespread fraud in these sectors. STEM jobs also feature prominently, alongside arts and culture.

Ohio (OH) exhibits notable fraud cases in the cities of Kirby and Barton. STEM and education and training are the dominant categories, with significant fraud also observed in agriculture and environment, as well as business and finance. Pennsylvania (PA) displays a broad spread of fraud across multiple cities, including Rock Glen, Vandergrift, Clune, Corsica, Skytop, and Montandon. STEM is the leading category in Pennsylvania, with business and finance, public service and law, and healthcare and medicine also featuring prominently in fraud cases. In Florida (FL), the cities of Lake Alfred and Matlacha stand out for significant fraud occurrences, particularly in STEM-related sectors. Fraud is also observed in arts and culture, education and training, and healthcare and medicine. STEM is clearly the dominant category in Florida. Missouri (MO) includes multiple smaller cities with significant fraud, including categories such as STEM, civilian jobs, public service and law, and arts and culture. STEM remains the leading category, followed by civilian jobs and arts and culture.

Overall, STEM is the most prominent category across all six states, with cities like Houston, Dallas, Rock Glen, and New York City demonstrating significant fraud cases in this field. Healthcare and medicine emerges as another notable category, especially in New York, Pennsylvania, and Florida. Civilian jobs are primarily observed in Missouri and New York, while public service and law is prominent in Texas, Pennsylvania, and Missouri. Arts and culture appears in Missouri, Florida, and Texas but is less widespread compared to the other categories. The chart underscores the pervasive nature of STEM-related fraud across all the states while also highlighting state-specific trends in other sectors.

MODEL

No models have been used for this.

KNOWLEDGE

In 2019 and 2020, sunburst charts illustrate the prevalence

of fraud across various job categories in multiple states and cities. In both years, STEM (Science, Technology, Engineering, and Mathematics) jobs consistently emerge as the most fraud-prone sector. This is particularly evident in cities like Houston (TX), Los Angeles (CA), and New York City (NY), where STEM-related fraud cases dominate. The prominence of STEM in these locations reflects the growing vulnerability of highly skilled technical sectors to fraudulent activities, likely due to the increasing reliance on technology and online platforms in these fields.

In 2020, alongside STEM, other sectors such as arts and culture, healthcare, business and finance, and education and training also show notable occurrences of fraud, though they remain secondary to STEM. In Texas cities like Houston and Aledo, significant fraud is concentrated in the STEM category, while arts and culture and business and finance also contribute to fraud cases. In California, cities such as Los Angeles, San Diego, and Oakland show a similar trend, with STEM jobs leading, followed by civilian jobs, business and finance, and education sectors.

New York (NY) in 2020 reveals a concentration of fraud in STEM, but healthcare and civilian jobs also experience considerable fraud, especially in cities like New York City, Brooklyn, and Rochester. Pennsylvania shows high fraud rates in STEM jobs, but other categories, including healthcare, arts and culture, and public service and law, also feature prominently in cities like Morrisdale, Mill Creek, and Oaks. South Carolina and Virginia follow the same pattern, with STEM jobs being the most fraud-prone, though other sectors such as arts and culture and healthcare also demonstrate significant fraud in cities like Brunson, Winnsboro, Clifton (SC), and Arlington, Alexandria, and Springfield (VA).

In 2019, overall trends remained consistent, with STEM jobs continuing to dominate across all states. Texas again stands out, with cities like Houston and Dallas showing significant fraud in STEM jobs. New York, particularly in cities like New York City and Melville, reveals high fraud rates in healthcare and civilian jobs, while Ohio shows a concentration of fraud in STEM and education sectors. Pennsylvania in 2019 exhibits similar trends, with STEM leading, followed by business and finance, healthcare, and public service. In Florida, STEM-related fraud is particularly prominent in cities like Lake Alfred and Matlacha, followed by fraud in arts and culture, education, and healthcare. Missouri sees significant fraud across STEM, civilian jobs, and public service sectors, further reinforcing the dominance of STEM across the states. Overall, data from both 2019 and 2020 emphasize the widespread nature of fraud in the STEM sector, alongside a notable presence in sectors such as healthcare, civilian jobs, arts and culture, and business and finance. The year 2020 also highlights the growing impact of sectors like education and training, reflecting the changing nature of fraud during the pandemic, particularly as more people shifted to online environments and remote work.

M. Exhibit 5 - Comparison of KDE of Fraudulent Transaction 2020 and 2019

The rationale for using Kernel Density Estimation (KDE) in Exhibit 5 lies in its ability to provide a smooth and continuous representation of the distribution of fraudulent transaction amounts. This technique highlights patterns that may be hidden in other types of visualizations. KDE is particularly effective at identifying clusters and outliers, as it smooths the data while maintaining the underlying trends.

The visualizations reveal multiple peaks in fraudulent transactions, indicating that fraud is not concentrated around a single value; rather, it spans a range from small test transactions to occasional high-value outliers. This method enables a clearer understanding of the variability in fraudulent activity over time.

DATA

The analysis utilizes data from 2019 and 2020, specifically focusing on fraudulent transactions. It includes the amounts of these transactions, which represent the value of fraud, as well as a fraud status indicator that shows whether a transaction is fraudulent (is raud = 1). Additionally, the analysis calculates statistical measures such as the mean, median, and mode of the fraudulent transaction amounts to better understand the central tendencies and distribution of these transactions over the two-year period.

MODEL

The Kernel Density Estimation (KDE) model was used for the analysis. This non-parametric technique allowed for the creation of a smooth curve based on a set of data. KDE works by placing a kernel function at each data point and then smoothing these functions to obtain a density estimate.

VISUALIZATION

1) Exhibit 5.1 - Kernel Density Estimation of Fraudulent Transactions (2020): This graph illustrates the distribution of transaction amounts for fraudulent transactions (is fraud = 1), using kernel density estimation (KDE) for visualization. It also marks key statistical measures: mean, median, and mode for the dataset. The blue KDE curve represents the probability density of transaction amounts.

The distribution is multimodal, displaying several peaks that indicate distinct clusters of fraudulent transaction amounts. The curve covers a wide range, suggesting substantial variability in these fraudulent values.

The average transaction amount for fraudulent activities is dollar 533.80, indicating a central tendency towards mid-range values. The median is slightly lower than the mean, which suggests a right-skewed distribution. This skewness implies that a small number of high-value transactions are influencing the overall average. The mode is notably low at dollar 7.25, reflecting a significant number of small fraudulent transactions, which creates a peak at the lower end of the

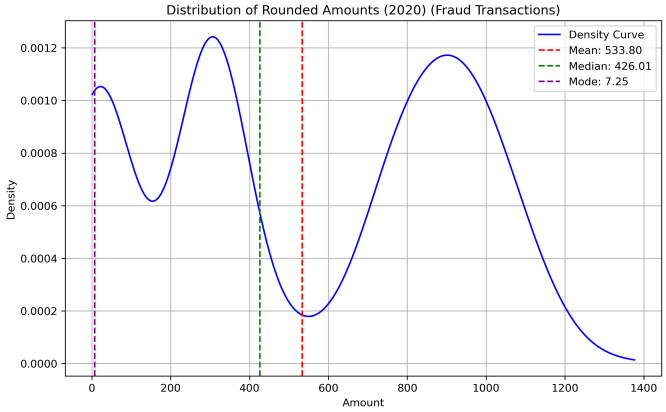

Fig. 39. Kernel Density Estimation of Fraudulent Transactions (2020)

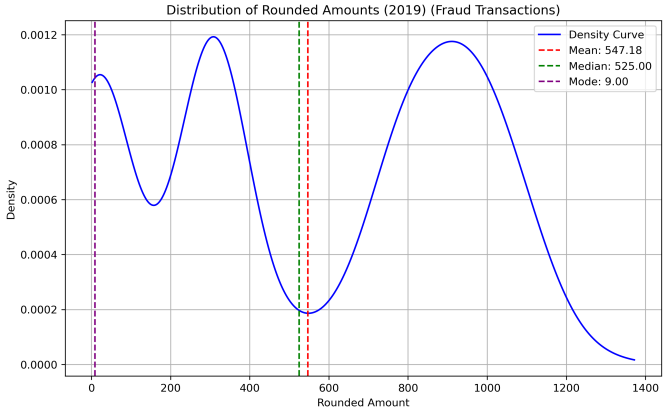

Fig. 40. Kernel Density Estimation of Fraudulent Transactions (2019)

distribution.

A prominent peak near the mode highlights a cluster of low-value fraudulent transactions, likely associated with testing or probing activities by fraudsters. The second peak occurs between dollar 400 and dollar 600, aligning with the median and mean, indicating that many fraudulent transactions frequently fall within this range. Another smaller peak is observed near dollar 1,200, representing high-value fraudulent transactions.

Overall, the curve spans from approximately dollar 0 to dollar 1,400, indicating that fraudulent transactions occur at both low and high values, although high-value fraud is less common. The right skew and peak near dollar 1,200 suggest the existence of high-value outliers, which may correspond to targeted, high-impact fraud.

2) Exhibit 5.2 - Kernel Density Estimation of Fraudulent Transactions (2019): This graph illustrates the distribution of rounded transaction amounts for fraudulent transactions (is fraud = 1) using kernel density estimation (KDE). It also marks the mean, median, and mode for reference. The KDE curve (in blue) represents the probability density of transaction amounts associated with fraudulent activities.

The distribution displays multiple peaks, indicating potential clusters of transaction amounts that are frequently targeted by fraud. The curve does not have a perfect bell shape, suggesting that fraudulent amounts do not follow a strict normal distribution and may be influenced by specific patterns. On average, fraudulent transactions amount to approximately dollar 547.18, indicating a tendency towards relatively higher amounts. The median is close to the mean, suggesting a somewhat symmetric distribution in this range, with fewer extreme outliers. In contrast, the mode is significantly lower at dollar 9.00, implying that numerous fraudulent transactions occur at very low amounts, creating a separate peak at the lower end of the curve.

The first peak around dollar 9 aligns with the mode, indicating a significant cluster of small fraudulent transactions, which may represent test transactions or low-value fraud strategies. A second, more prominent peak is observed between dollar 500 and dollar 600, corresponding with the mean and median values, signifying a higher frequency of mid-range fraudulent transactions. The third peak, approaching dollar 1,000, indicates a smaller but still notable cluster of high-value fraudulent transactions.

The KDE curve spans a wide range of values, approximately from dollar 0 to dollar 1,400, highlighting the variability in fraudulent transaction amounts. The decline at higher amounts suggests that extremely high fraudulent transactions are less common. The sharp rise near the mode (low values) and the distribution's spread towards higher amounts hint at the presence of outliers, particularly at both ends of the distribution.

KNOWLEDGE

Exhibit 5.1 and Exhibit 5.2 provide insights into the distribution of fraudulent transaction amounts for 2020 and 2019, respectively, using kernel density estimation (KDE). Both distributions display multimodal patterns, indicating distinct clusters of transaction amounts that fraudsters target, with variability across different values.

In 2020, the average fraudulent transaction amount was dollar 533.80, while the median was slightly lower, suggesting a right-skewed distribution. This skew is attributed to a small number of high-value transactions that elevate the mean. The mode, located at dollar 7.25, indicates a peak at lower values, likely driven by testing or probing activities. Additionally, there are significant peaks around dollar 7.25 (low-value transactions), dollar 400 to dollar 600 (mid-range transactions), and around dollar 1,200 (high-value transactions). This pattern suggests that fraudulent transactions occur at both low and high values, although high-value fraud is less frequent.

For 2019, the average fraudulent transaction amount was dollar 547.18, and the median was close to the mean, implying a more symmetrical distribution compared to 2020. Similar to the 2020 data, the mode is low at dollar 9.00, reflecting the frequency of small fraudulent transactions. The peaks in this distribution are at dollar 9 (low-value fraud), dollar

500 to dollar 600 (mid-range fraud), and around dollar 1,000 (high-value fraud). The range spans from dollar 0 to dollar 1,400, with a steep rise near the mode and a decline at higher amounts, indicating fewer extreme high-value fraudulent transactions.
Both distributions highlight the presence of small fraudulent transactions, which may relate to testing, as well as larger fraudulent activities, which are less common but still noteworthy. The distinct peaks observed in both years emphasize that fraud can occur across a range of transaction values, targeting both low and high amounts, albeit with lower-frequency high-value fraud.

PART 3- KNOWLEDGE TO DATA In Part A1, the initial analysis of fraud across all U.S. states was cluttered and difficult to interpret. To improve clarity, the states were categorized into five standard regions: Capital, Northwest, South, Midwest, and West. This organization streamlined the visualization for a more comprehensive analysis. To further enhance the understanding of fraud, key metrics such as total fraud loss, fraud per victim, and fraud victim count were included. Each of these metrics was calculated using Python, providing a clearer insight into the fraud patterns across the different regions.

DATA

In this analysis, we utilized newly calculated fields total fraud loss, fraud per victim, and fraud victim count along with state categories (Capital, Northwest, South, Midwest, West) and transaction hours.

VISUALIZATION

N. Exhibit 1 - Density Scatter Plot for Fraud Loss Amount and Fraud Victim Count

The density scatter plot illustrates the relationship between the number of fraud victims (on the x-axis) and the amount of financial loss due to fraud (on the y-axis). The visualization shows a positive trend: as the number of victims increases, the fraud loss amount generally increases as well. For each victim count, there is a range of loss amounts, indicating variability in the severity or scale of the fraud incidents.
A denser clustering of points is observed for victim counts between 5 and 15, indicating that most fraud cases are concentrated within this range. Although the overall trend is positive, there is notable variability in loss amounts for similar victim counts, highlighting differences in the nature or extent of the fraud cases.

Some data points deviate significantly from the main trend, particularly at higher victim counts and greater fraud loss amounts. These outliers may represent exceptional or extreme cases, such as large-scale fraud schemes or unique instances resulting in significant financial impact. The clustering of points and the presence of outliers together suggest that

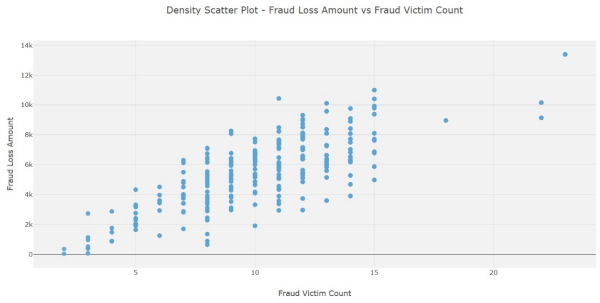

Fig. 41. Density Scatter Plot for Fraud Loss Amount and Fraud Victim Count

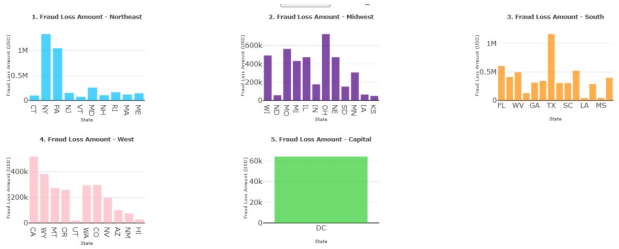

Fig. 42. Regional Analysis of Fraud Loss Amounts Across States

mid-sized fraud cases are more common, while larger-scale incidents, although rarer, often result in considerably higher losses.

O. Exhibit 2 - Regional Analysis of Fraud Loss Amounts Across States

The bar plots provide a clear overview of fraud loss amounts across states, categorized by five regions: Northeast, Midwest, South, West, and the Capital (D.C.). Each plot emphasizes the financial impact of fraud in USD, highlighting variations among states within each region.
In the Northeast, Pennsylvania and New York report the highest fraud losses, with Pennsylvania being the most affected state in this region. In contrast, states like Connecticut, Vermont, and Maine demonstrate relatively lower fraud losses, indicating a lesser financial impact from fraud in these areas.

In the Midwest, Illinois and Ohio show the highest levels of fraud losses, signifying significant financial impacts in these states. Other states, such as Michigan and Missouri, follow with moderate fraud losses, while Kansas and Iowa exhibit comparatively lower amounts.
The South region presents a striking pattern, with Texas recording the highest fraud loss amount by a considerable margin, indicating that Texas is disproportionately affected compared to other states in the region. States like Georgia, South Carolina, and Florida report moderate fraud losses, while Mississippi and West Virginia have lower values.
In the West, California leads with the highest fraud loss amount, underscoring its significant financial impact from

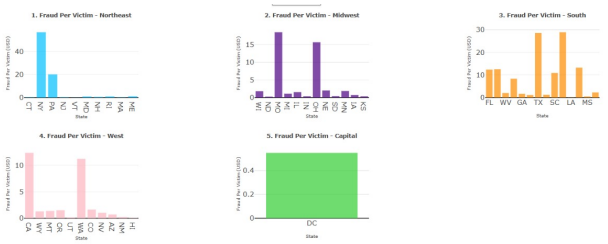

Fig. 43. Regional Analysis of Fraud Per Victim Across States

fraud in this region. Other Western states show a gradual decline in fraud losses, with Washington and Oregon displaying moderate amounts and Nevada and Arizona reporting smaller losses.

The Capital region, represented solely by Washington, D.C., shows moderate fraud loss amounts that are relatively smaller compared to the most affected states in other regions. Nonetheless, its losses are notable given its unique status and size.

P. Exhibit 3 - Regional Analysis of Fraud Per Victim Across States

The bar plots illustrate the financial impact of fraud on victims in USD across various states, categorized into five regions: Northeast, Midwest, South, West, and the Capital (Washington, D.C.). Each region highlights differences in the average fraud loss per individual victim.

In the Northeast, Pennsylvania reports the highest fraud loss per victim, significantly surpassing other states such as New York and Vermont, which demonstrate much lower values. This indicates that fraud cases in Pennsylvania are particularly severe in terms of financial impact on individuals.

The Midwest reveals a similar trend, with Illinois and Missouri showcasing high fraud loss per victim amounts. In contrast, states like Ohio and Michigan exhibit much lower figures, underscoring disparities in individual fraud impacts within the region.

In the South, Texas and Georgia emerge as leaders with notably high fraud loss per victim, suggesting a substantial financial burden per individual. Other states, including South Carolina and Florida, display moderate amounts, while Mississippi and West Virginia report lower values.

The West region is primarily dominated by California, where the fraud loss per victim is the highest, followed closely by Wyoming. Other states, such as Nevada and Arizona, show significantly lower values, highlighting California's exceptional status in this region.

Washington, D.C., classified under the Capital category, displays a moderate fraud loss per victim, reflecting a relatively consistent per capita financial impact compared to the other states.

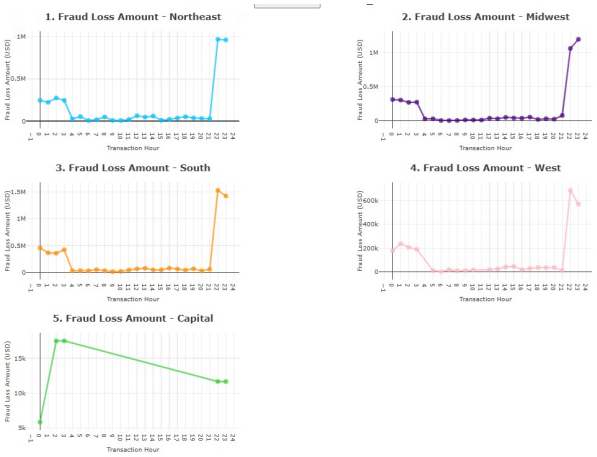

Fig. 44. Transaction Time Analysis of Fraud Loss Amounts Across States

Q. Exhibit 4 - Transaction Time Analysis of Fraud Loss Amounts Across States

The line charts illustrate the amounts of fraud losses (in USD) across various regions—Northeast, Midwest, South, West, and Capital—over different transaction hours. Each region displays distinct patterns in the timing of these fraud losses.

In the Northeast, fraud losses are moderate during most hours but sharply escalate around the 22nd hour, peaking at over 1 million dollar. This indicates a concentration of fraudulent activity during the late evening hours.

The Midwest follows a similar trend, with minimal losses for most of the day and a dramatic spike after the 21st hour, reaching over 1 million dollar. This highlights significant late-hour fraud occurrences.

In the South, the pattern shows relatively stable losses during the earlier hours, followed by the highest peak among all regions, exceeding 1.5 million dollar near the 22nd hour. This suggests particularly severe fraud activity during this period.

The West experiences moderate losses throughout the day, with a noticeable rise around the 22nd hour, peaking at approximately 600,000 dollar. Although this is lower than in other regions, it still underscores the vulnerability to fraud during late hours.

In the Capital, Washington, D.C., fraud losses remain consistently low compared to other regions. A small peak occurs during the early hours, but it lacks the dramatic spikes seen elsewhere, indicating a relatively lower impact of fraud in this area.

R. Exhibit 5 - Transaction Time Analysis of Fraud Per Victim Across States

The line charts illustrate fraud losses per victim (in USD) across different transaction hours in five regions: Northeast, Midwest, South, West, and Capital. Each region shows a unique distribution, with notable peaks indicating increased

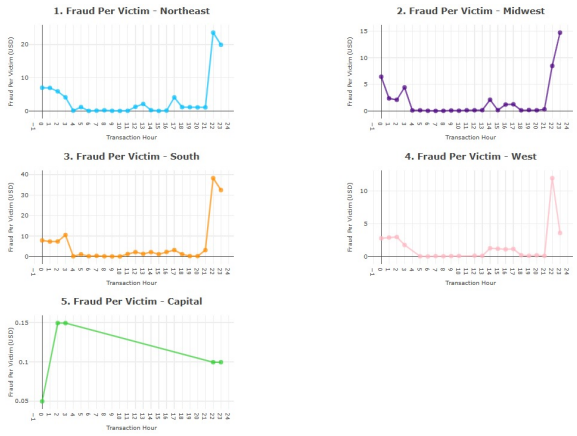

Fig. 45. Transaction Time Analysis of Fraud Per Victim Across States

vulnerabilities during late hours.

In the Northeast, fraud losses per victim remain low during most hours, with a sharp increase after the 21st hour, peaking at approximately 30 dollar per victim. This suggests that concentrated fraudulent activities occur in the late evening.

The Midwest exhibits a similar trend, with consistently low losses earlier in the day followed by a significant spike after the 21st hour, exceeding 15 dollar per victim. This pattern underscores the prevalence of fraud during late hours.

The South stands out with the highest losses per victim among all regions, experiencing a dramatic spike near the 22nd hour to almost 40 dollar per victim. This indicates a particularly high financial burden on individual victims during this time.

The West experiences moderate losses per victim throughout the day but shows a noticeable peak near the 22nd hour, reaching over 10 dollar per victim. Although less severe than the South, the late-hour spike is still evident.

The Capital region exhibits minimal losses per victim compared to the other regions. While a small peak occurs in the early hours, the overall impact remains negligible, with losses not exceeding 0.15 dollar per victim.

MODEL No models were used here

KNOWLEDGE

The analysis provides valuable insights into the patterns and impacts of fraud across different regions in the U.S. It offers a detailed examination of fraud loss amounts, fraud per victim, and transaction timing. By organizing states into five standard regions Capital, Northwest, South, Midwest, and West the clarity of the data visualization improved, allowing for more focused observations of regional trends. The introduction of calculated metrics, such as total fraud loss, fraud per victim, and fraud victim count, enabled a more nuanced understanding of both the financial and individual effects of fraud.

The density scatter plot revealed a positive correlation between the number of fraud victims and the amount of financial loss, with a concentration of cases involving 5 to 15 victims.

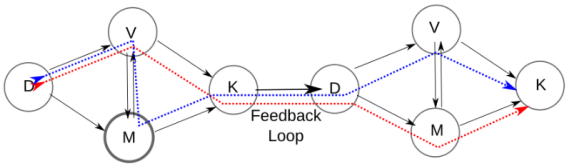

Fig. 46. Feedback Loop

Outliers in the data indicated large-scale fraud schemes or extreme incidents, highlighting the variability in the scale and impact of fraud. This variability underscores the need for tailored approaches to address both common and exceptional fraud cases.

Regional analysis using bar plots illustrated distinct patterns of fraud across the U.S. Pennsylvania, Texas, and California emerged as states with the highest fraud losses in their respective regions, reflecting a disproportionate financial burden on these areas. Meanwhile, states such as Vermont, Mississippi, and Arizona experienced lower fraud losses, indicating less severe impacts. The analysis of fraud per victim further emphasized regional disparities, with Pennsylvania, Texas, and California standing out for their high financial impacts on individual victims, while other states showed more moderate or minimal losses.

The temporal analysis of fraud losses and fraud per victim across transaction hours revealed that late evenings, particularly around the 9 PM and 10 PM hours, are critical periods for fraudulent activity. The South region experienced the most pronounced spikes in both overall losses and per-victim impact, suggesting heightened vulnerability during this time. In contrast, the Capital region displayed consistently lower losses, pointing to regional differences in fraud dynamics.

PART 4- Machine Learning Model - Forecasting and Classification

S. Exhibit 1 - Predictive Forecasting of fraud in The state OH using ARIMA model

The feedback loop was used to identify discrepancies between actual and predicted fraudulent transactions, enabling iterative improvements in the model. The analysis went from data to knowledge and then knowledge to data. The time series plot highlighted the ARIMA model's inability to capture sudden spikes, while the boxplot revealed temporal patterns and anomalies in fraud activity.

DATA

Data preprocessing involved loading and filtering the dataset to include only fraudulent transactions, specifically those where the variable 'is fraud' was equal to 1. The 'transaction date' column was parsed as a datetime object, and any potential parsing errors were addressed to ensure data integrity. The focus was on transactions from the state of Ohio (OH), and the data was grouped by transaction dates to count the number of fraudulent transactions occurring each day. To

```
Mean Absolute Error (MAE): 2.1115527947653456
Root Mean Squared Error (RMSE): 3.8516862343919924
```

Fig. 47. ARIMA Model Performance

maintain continuity, the dataset was resampled to a daily frequency, filling in any missing dates with a count of zero. For the train-test split, the data was divided into training and testing sets, with the last 30 days reserved for testing purposes.

MODEL

The process of selecting the ARIMA model involved using the pmdarima library's auto arima function to automate the identification of the best parameters for the ARIMA model, specifically (p, d, q). The chosen model order was (1, 0, 1), which minimized the Akaike Information Criterion (AIC). AIC is a metric that evaluates the quality of a model by balancing its fit to the data and its complexity.

The ARIMA(1, 0, 1) model was then fitted to the training data, allowing the model to learn the underlying patterns and trends associated with fraudulent transaction occurrences. After training, the ARIMA model was utilized to predict fraudulent transactions over a 30-day test period.

The forecast followed the baseline pattern observed in the training data; however, it did not effectively capture the spikes evident in the test data. This discrepancy indicates that the ARIMA model may not be well-suited to detect abrupt changes or extreme outliers in fraudulent activity.

The model's performance was evaluated using Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE). The MAE was 2.11, suggesting that the model's predictions deviate by approximately 2.11 fraudulent transactions from the actual observed values. The RMSE was 3.85, which gives more weight to larger errors due to the squaring process, suggesting a larger overall prediction error compared to MAE and reflecting greater penalties for substantial discrepancies.

Both the MAE and RMSE highlight errors in the model's predictions, which may result from missing important features or the model's inability to capture relevant patterns in the data. The higher value of RMSE indicates that larger prediction errors significantly impact the model's overall performance.

VISUALIZATION

1) Exhibit 1.1 - Plotted the training data, test data, and forecasted values for visual comparison.: The visualization illustrates a time series forecast of fraudulent transactions in Ohio (OH) from August 2019 to December 2019, using three color-coded elements to convey information.

The blue line represents the training data, showcasing the historical fraudulent transactions observed up until the model was trained. This line reveals a fluctuating pattern with peaks and valleys, indicating that fraudulent activities in Ohio occur sporadically, at varying frequencies throughout the year.

The green dotted line signifies the forecasted fraudulent

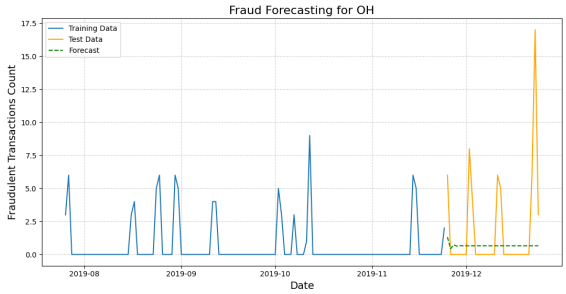

Fig. 48. Fraud forecasting using ARIMA model for state OH

transactions for the period following the training data, extending into late 2019. The forecast remains close to zero, suggesting that the model predicts either a minimal occurrence of fraud or fails to account for potential future fraudulent activity in the area.

The orange line denotes the actual test data, displaying the real counts of fraudulent transactions observed after the training period. There is a noticeable discrepancy between the forecast and the actual data, as the test data reveals several significant spikes in fraudulent transactions that the model did not predict. These spikes highlight periods of higher-than-expected fraudulent activity that were overlooked by the model.

KNOWLEDGE

The time series forecast of fraudulent transactions in Ohio from August 2019 to December 2019 reveals discrepancies between the predicted and actual levels of fraud activity. The blue line, which represents the training data, displays fluctuating patterns with irregular peaks and valleys, indicating sporadic fraudulent incidents throughout the year. These fluctuations imply that fraud occurrences in Ohio may happen with some periodicity or randomness, although they do not follow a consistent trend.

The green dotted line represents the forecasted fraudulent transactions, remaining close to zero. This suggests that the ARIMA model predicts very few future fraud incidents or struggles to account for potential increases. In contrast, the orange line illustrates the actual test data, which reveals significant spikes in fraudulent transactions, particularly in December 2019 an event the model failed to predict. These unexpected spikes underscore that the model may be too simplistic to capture sudden surges in fraudulent activity or changes in underlying patterns.

The discrepancy between the forecast and the actual test data, especially regarding the December surge, indicates a potential limitation of the ARIMA model in handling irregular and sudden increases in fraudulent behavior. This suggests a need for improvements to the model to better predict such anomalies.

DATA

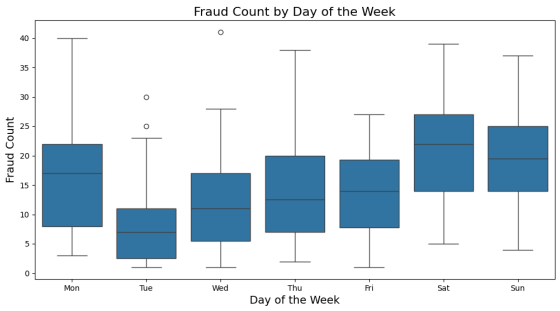

Fig. 49. Box plot showing Fraud Count by Day of the Week

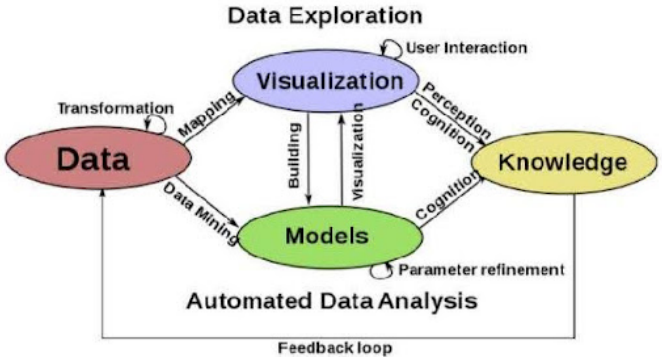

Fig. 50. Diamond Feedback loop

Feature engineering was performed to create new, meaningful features (such as the day of the week, month, etc.) that can improve the model's ability to identify trends or seasonal patterns in the data. For instance, fraudulent activity may be influenced by time-based trends (like weekends versus weekdays), which raw transaction counts alone might not capture.

VISUALIZATION

2) Exhibit 1.2 - Feature Engineering: This boxplot illustrates the distribution of fraudulent transactions for each day of the week, highlighting how fraud counts vary over time.

The boxes represent the interquartile range (IQR), encompassing the middle 50 percent of the fraud counts for each day. The horizontal line inside each box marks the median, providing a central tendency for fraud activity. Whiskers extend from the boxes, showing the minimum and maximum values within a reasonable range, while any individual circles represent outliers, indicating unusually high or low fraud counts.

Fraud counts are not evenly distributed across the days of the week. Monday and Saturday have higher median fraud counts compared to other days, suggesting these days experience relatively more fraudulent activity. Additionally, the variability (spread of data) is higher on these days, as indicated by the larger IQR and longer whiskers.

Tuesday and Thursday have notable outliers, showing occasional spikes in fraudulent activity that deviate significantly from the regular pattern. In contrast, midweek days such as Wednesday and Friday exhibit narrower IQRs and less variability, suggesting more stable and predictable fraud activity on these days. Weekends (Saturday and Sunday) display wider distributions, indicating greater fluctuation in fraud counts.

MODEL

Here no model was used.

KNOWLEDGE

The boxplot illustrates distinct patterns in fraud activity over time, highlighting days with more frequent or intense fraudulent behavior. Mondays and Saturdays stand out due to higher median fraud counts and greater variability, as indicated by their larger interquartile ranges (IQR) and longer whiskers. These features suggest increased fraudulent activity on these days, possibly influenced by behavioral or transactional factors such as the beginning of the workweek or heightened weekend spending.

Notable outliers on Tuesdays and Thursdays indicate isolated spikes in fraudulent transactions, suggesting specific events or campaigns that deviate from the typical pattern. These anomalies may require focused investigation to identify underlying triggers or vulnerabilities. In contrast, midweek days like Wednesdays and Fridays exhibit narrower IQRs and less variability, reflecting more stable and predictable fraud activity.

The broader distribution of fraud counts during weekends, particularly on Saturdays, may be attributed to increased consumer activity and a variety of transaction types such as online shopping, entertainment, and travel. Weekends often involve discretionary spending and leisure activities, creating diverse opportunities for fraudulent behavior. Additionally, relaxed vigilance or operational limitations in monitoring systems during weekends could further contribute to the fluctuations in fraud activity.

T. Exhibit 2 - Forecasting of fraud in The state OH using SARIMA model

Rationale-The feedback loop was used to evaluate the SARIMA model's performance in forecasting fraudulent transactions, enabling iterative refinements. The time series plot highlighted the model's effectiveness in capturing general trends while exposing its limitations in predicting extreme spikes, and the residual histogram revealed inconsistencies in handling anomalies, suggesting areas for improvement.

DATA

Data preprocessing involved loading and filtering the dataset to include only fraudulent transactions, specifically those where the variable 'is fraud' was equal to 1. The 'transaction date'

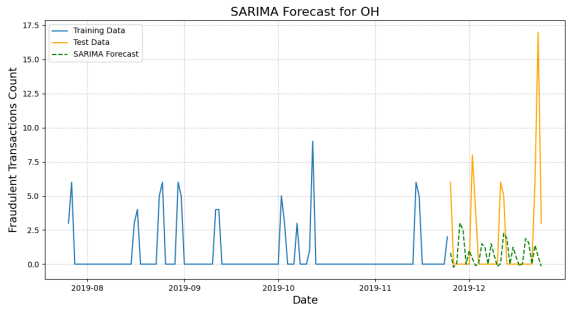

Fig. 51. SARIMA Forecasting of OH

column was parsed as a datetime object, and any potential parsing errors were addressed to ensure data integrity. The focus was on transactions from the state of Ohio (OH), and the data was grouped by transaction dates to count the number of fraudulent transactions occurring each day. To maintain continuity, the dataset was resampled to a daily frequency, filling in any missing dates with a count of zero. For the train-test split, the data was divided into training and testing sets, with the last 30 days reserved for testing purposes.

MODEL

SARIMA (Seasonal AutoRegressive Integrated Moving Average) is an effective model for forecasting time-series data that shows both seasonal patterns and trends. Its capacity to specifically account for seasonality makes it an excellent choice for analyzing trends in fraudulent transactions.

Feature engineering, such as extracting attributes like the day of the week or month, helps to identify trends in the data. For instance, there may be an increase in fraudulent activity during weekends or a recurring weekly or monthly cycle. SARIMA is particularly adept at modeling these seasonal and trend-based variations, making it an ideal tool for capturing these dynamics.

Once time-based features are identified and incorporated, SARIMA can forecast future fraud patterns based on observed historical trends. This capability supports proactive decision-making by providing insights into when fraudulent activity is likely to occur.

Neglecting time-based features could lead to suboptimal results, as the model may miss key seasonal trends. Including these features ensures that SARIMA effectively identifies and models complex seasonality, thereby enhancing the accuracy of its predictions.

Visualizing the predictions of the SARIMA model alongside actual data offers insights into how well the model captures seasonal patterns and trends. These visual diagnostics can validate the model's ability to track and forecast occurrences of fraud over time, highlighting its practical utility in managing seasonal variations.

VISUALIZATION

```
SARIMA Model Performance:
Mean Absolute Error (MAE): 2.2993170149890956
Root Mean Squared Error (RMSE): 4.001721272229995
```

Fig. 52. SARIMA Model Performance

1) Exhibit 2.1 - Forecasting of fraud in The state OH using SARIMA model: The training data, shown by the blue line, reveals sporadic patterns in the counts of fraudulent transactions, characterized by occasional peaks and troughs. Most values hover around zero, but there are periodic spikes. This indicates that fraudulent transactions are not evenly distributed over time; instead, they occur in bursts, likely triggered by specific times, external factors, or systemic issues.

The test data, represented by the orange line and covering the last 30 days, displays greater variability, including a significant spike. This trend may suggest an increase or erratic nature in fraudulent activity, possibly due to external influences such as holidays, new campaigns, or system vulnerabilities during this period.

The SARIMA forecast, depicted as the green dashed line, aligns with the general trend of the historical data, predicting small variations along with occasional spikes. While it captures the sporadic nature of fraudulent transactions, the model may have difficulty predicting extreme spikes, as seen in the highest peak of the test data. The forecast implies that fraudulent activity will continue to fluctuate, with potential spikes in the near future.

The forecast aligns well with the lower and moderate counts of the test data but tends to underestimate extreme peaks. Although the SARIMA model provides solid baseline predictions, it may require additional features or a different modeling approach to more effectively predict rare, extreme anomalies.

To validate the model, we examine the performance metrics of the SARIMA model. The Mean Absolute Error (MAE) is calculated to be 2.299, indicating that, on average, the forecasted values deviate from the actual counts of fraudulent transactions by approximately 2.3 transactions per day. Additionally, the Root Mean Squared Error (RMSE) is 4.002, which signifies that the typical error magnitude is around 4 transactions per day. It is important to note that the RMSE penalizes larger errors more heavily than the MAE.

The SARIMA model incorporates adjustments for seasonality, enabling it to capture patterns that occur at regular intervals, such as weekly or monthly trends. However, its performance is somewhat inferior to that of the ARIMA model, which recorded a Mean Absolute Error (MAE) of 2.11 and a Root Mean Square Error (RMSE) of 3.85.

2) Exhibit 2.2 - Residual Error : This image illustrates the distribution of residuals from a SARIMA (Seasonal AutoRegressive Integrated Moving Average) model. The following interpretations can be made from the histogram: The x-axis represents the residual values, which are the

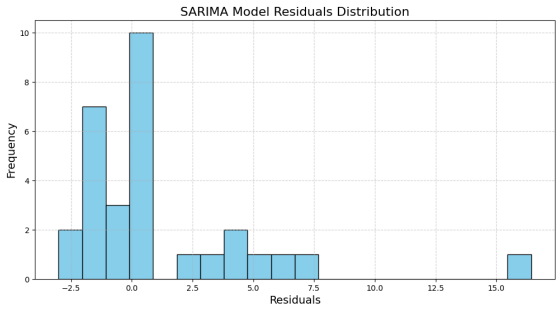

Fig. 53. SARIMA Model Residuals Distribution

differences between the observed and predicted values. The residuals are concentrated around 0, indicating that the SARIMA model generally makes accurate predictions for most test cases. However, there are outliers, particularly on the far right, where residuals reach approximately 15. This suggests that the model struggles to predict extreme spikes in fraudulent transactions. Positive residuals occur when the model underpredicts fraud, while negative residuals arise when the model overpredicts fraud. The presence of these outliers, especially at higher residual values, highlights the areas where the model fails to capture unusual spikes in fraudulent activity.

The y-axis shows the frequency of residuals for each bin. Most residuals fall between approximately -2.5 and 5, with the highest frequency near 0. This indicates that for the majority of data points, the model's predictions are close to the actual values. The frequency distribution suggests that although the model performs well overall, it is less consistent in extreme cases. This indicates a need for further tuning or feature engineering to enhance performance for these edge cases.

The residuals somewhat resemble a normal distribution centered around 0, but they exhibit visible skewness. The presence of positive outliers suggests that the model tends to underpredict fraud during certain periods. Additionally, the residuals deviate slightly from a perfect normal distribution, indicating that the model may not capture all patterns in the data, particularly during periods of high fraudulent activity. In summary, while the SARIMA model effectively captures the general trend, it struggles to predict extreme or unusual spikes in fraudulent transactions.

KNOWLEDGE

The image illustrates the distribution of residuals from a SARIMA (Seasonal AutoRegressive Integrated Moving Average) model. The x-axis represents the residual values, which are the differences between the observed and predicted values. Most residuals are concentrated around 0, indicating that the SARIMA model makes accurate predictions for the majority of the test cases. However, there are some outliers, particularly on the far right, where residuals reach up to around 15. This indicates that the model has difficulty predicting extreme spikes in fraudulent transactions.

Positive residuals occur when the model underpredicts fraud, while negative residuals indicate overprediction. The presence of outliers, especially at higher residual values, shows that the model may not fully capture unusual spikes in fraudulent activity. Most residuals fall between approximately -2.5 and 5, with the highest frequency near 0. This suggests that, for the majority of the data points, the model's predictions closely align with the actual values.

The frequency distribution indicates that while the model performs well overall, it is less consistent in extreme cases. This points to a need for further tuning or feature engineering to enhance performance for these edge cases. The residuals exhibit a shape somewhat resembling a normal distribution, centered around 0, but there is noticeable skewness. The presence of positive outliers suggests that the model tends to underpredict fraud during certain periods. Overall, the residuals show slight deviations from a perfect normal distribution, indicating that the model may not capture all patterns in the data, particularly during periods of high fraudulent activity. Thus, while the SARIMA model effectively captures the general trend, it struggles to predict extreme or unusual spikes in fraudulent transactions.

During our forecasting phase with the SARIMA model, we learned important things about fraudulent transaction trends. The model captured seasonal patterns well, but it had trouble with extreme variations, like significant outliers in the data. This showed us that seasonality and time-series trends alone were not enough to understand the full complexity of fraudulent activity.

To improve our analysis, we shifted to a classification approach. We included a wider range of variables such as transaction amount, transaction hour, customer age, state, job category, and gender. This method helped us better understand patterns of fraudulent behavior.

The classification model worked alongside the forecasting insights by analyzing individual transactions in more detail, which improved our ability to detect fraud. It also demonstrated its effectiveness by predicting potential fraudulent transactions in unseen data based on what it learned.

U. Exhibit 3 - Model Training, Building and Evaluation through Random Classifier

The feedback loop in this fraud detection system was essential because it addresses the iterative nature of building effective models for highly imbalanced datasets. The feedback loop ensures that steps like resampling, threshold optimization and class-weight adjustments are implemented iteratively to improve the model's performance on the minority class (fraud).

To provide a detailed view of the model's prediction performance,the confusion matrix visualisation were made and moreover to evaluate the model's diagnostic ability across different thresholds the ROC AUC curve was made by plotting the true positive rate against the false positive rate.

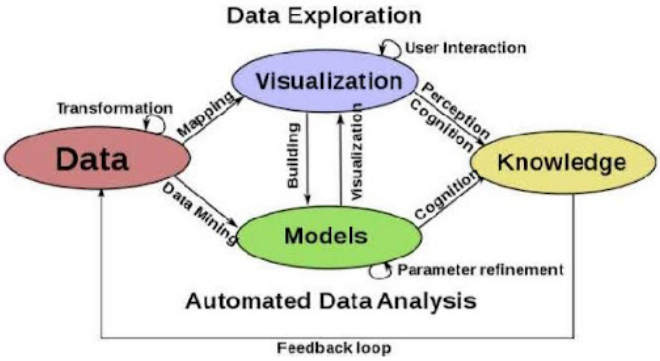

Fig. 54. Diamond Feedback loop

Also during threshold optimisation the Precision-Recall curve (Fig 56) analysis was used to identify the optimal threshold for balancing precision and recall.

DATA

The feedback loop in the fraud detection system begins with the preprocessing of raw data to ensure that the model receives high-quality inputs. This process includes tasks such as removing outliers that could distort the model, addressing missing values to prevent disruptions in calculations, and balancing the dataset using techniques like Synthetic Minority Oversampling Technique (SMOTE). SMOTE is crucial for tackling the class imbalance problem, which is a common challenge in fraud detection due to the significantly lower number of fraudulent transactions compared to legitimate ones. By generating synthetic samples of the minority class, SMOTE helps ensure that the model does not become biased toward predicting non-fraudulent transactions.

Data Preparation : Structuring the Dataset for Analysis

Subset selection involved extracting a representative sample to facilitate quicker experimentation while maintaining the diversity and distribution of the data. The dataset was divided into training (80 percent) and testing (20 percent) subsets using the ‘train test split‘ function. The training data allowed the model to learn patterns, while the testing data evaluated its ability to generalize. This approach ensured that the model did not simply memorize the data, promoting better generalization to unseen samples and reducing the risk of overfitting.

Data Preprocessing: Transforming Data into Machine-Learning-Ready Formats

Feature scaling was performed using ‘StandardScaler‘ to standardize numerical data with varying ranges, such as transaction amounts. This standardization ensured that all features were centered around a mean of 0 with a standard deviation of 1, which helped algorithms converge faster during training. Categorical features, such as merchant location and transaction type, were encoded to be compatible with machine learning. One-Hot Encoding converted categories into binary columns (e.g., ”Location A” = 1, others = 0), while Label Encoding assigned numerical labels to ordinal categories (e.g., High = 3, Medium = 2, Low = 1).

MODEL

Once the data is prepared, the model is trained on this processed dataset using suitable algorithms. After the training phase, the model’s performance is rigorously evaluated through a combination of metrics to gain a clear understanding of its strengths and weaknesses. Key evaluation metrics include accuracy, which measures the overall correctness of the model; precision, which assesses how effectively the model identifies actual fraud cases among those predicted as fraudulent; recall, which focuses on the model’s ability to detect fraud cases out of all actual fraud instances; and F1-score, which is the harmonic mean of precision and recall, balancing their trade-offs.

Model Training and Evaluation

After preparing the data, the model is trained using suitable algorithms on the processed dataset. Following the training phase, the model’s performance is assessed through a combination of metrics to obtain a comprehensive understanding of its strengths and weaknesses.

Key Evaluation Metrics

Accuracy measures the overall correctness of the model. Precision evaluates the model’s ability to accurately identify actual fraud cases among those predicted as fraudulent. Recall focuses on the model’s capability to detect fraud cases from the total number of actual fraud instances. F1-score represents the harmonic mean of precision and recall, balancing their trade-offs.

Model Building: Random Forest Classifier

A Random Forest Classifier was selected as the initial model due to its robustness in handling both numerical and categorical features. The initial performance on the imbalanced dataset resulted in an accuracy of 99.56 percent. However, metrics specific to the fraud class, such as precision, recall, and F1-score, were notably low. This outcome indicated that while the model performed well for the majority (non-fraud) class, it struggled to effectively identify the minority (fraud) class.

Refining the Model to Address Class Imbalance

Class imbalance presented a significant challenge, leading to the implementation of refinement strategies aimed at enhancing the model’s ability to detect fraudulent transactions. To address this, class weights were adjusted in the Random Forest Classifier using class weight=’balanced’, which penalized misclassification of fraudulent transactions more heavily than those that were non-fraudulent. This adjustment encouraged the model to concentrate more on accurately predicting the fraud class.

Additionally, threshold optimization was conducted by analyzing the Precision-Recall Curve to determine an optimal

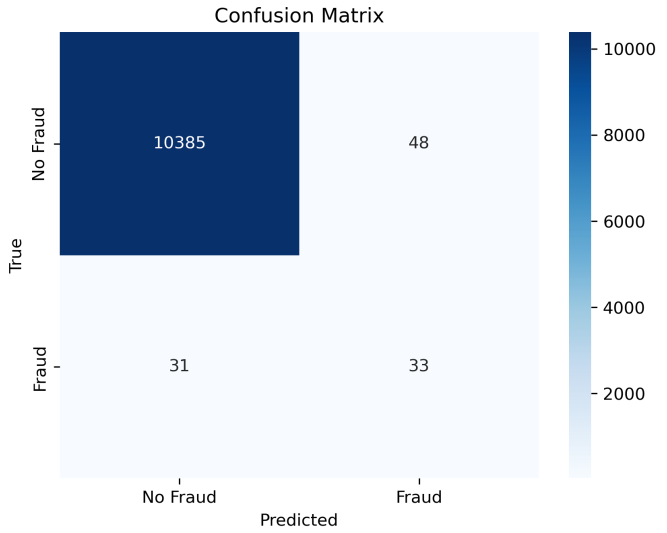

Fig. 55. Confusion Matrix displaying true and false positives, true and false negatives

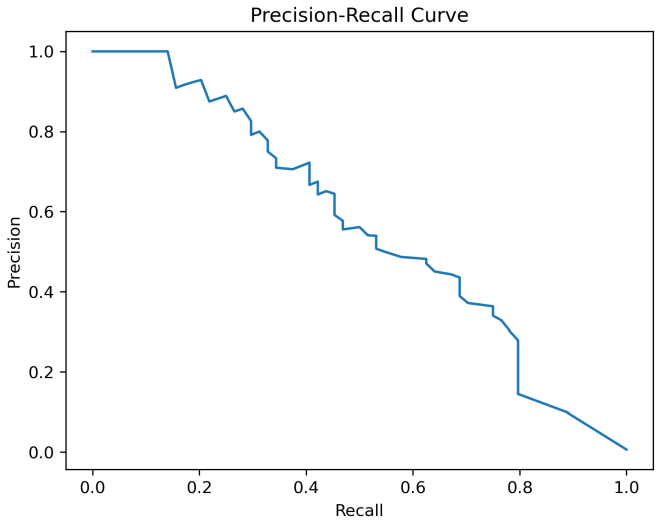

Fig. 56. Precision- Recall Curve

threshold that balanced precision and recall. After these optimizations, the model demonstrated improved fraud detection capabilities. Specifically, the precision for the fraud class rose to 0.64, recall improved to 0.53, and the F1-score increased to 0.52, reflecting better performance in detecting fraudulent transactions, especially regarding recall.

VISUALIZATION

The ROC AUC (Receiver Operating Characteristic - Area Under Curve) score is an important metric for evaluating a model's ability to distinguish between fraudulent and legitimate transactions across different thresholds. It provides a comprehensive overview of the model's diagnostic capabilities. Additionally, the confusion matrix is a crucial tool for visualizing the distribution of true positives, true negatives, false positives, and false negatives, which offers a deeper understanding of the model's errors.

1) Exhibit 3.1 : Confusion Matrix: The confusion matrix displays the counts of true positives (fraud cases correctly identified as fraud), false positives (non-fraud cases incorrectly identified as fraud), true negatives (non-fraud cases correctly identified as non-fraud), and false negatives (fraud cases incorrectly identified as non-fraud). This visualization highlights specific misclassification trends and provides insights for refining model strategies.

2) Exhibit 3.2 : ROC AUC Curve: The ROC AUC curve illustrates the trade-off between the true positive rate (recall) and the false positive rate, showcasing the model's ability to discriminate between fraudulent and non-fraudulent transactions. The model achieved an ROC AUC score of 0.9368, indicating a strong capacity to distinguish between these transaction types effectively.

MODEL

To enhance the model's performance and address class imbalance, various resampling techniques were implemented to balance the dataset.

1) **SMOTE (Synthetic Minority Oversampling Technique)**: SMOTE generated new fraudulent samples by interpolating existing fraud cases, effectively increasing the representation of the fraud class. This approach balanced the dataset, allowing the model to focus more on patterns from the minority class.

2) **Undersampling**: Undersampling reduced the number of non-fraudulent samples to match the size of the fraud class. By limiting the majority class, ensured that crucial patterns in the data were preserved.
Post-Resampling Performance: The resampling techniques led to significant changes in the model's performance metrics. Recall for the fraud class improved dramatically to 0.86, indicating a much higher detection rate for fraudulent transactions. However, precision dropped to 0.07, signaling an increase in false positives. This trade-off highlighted the focus on maximizing fraud detection while accepting a higher rate of false alarms.

Earlier Model Performance

In comparison, the earlier model achieved a precision of 51 percent for fraudulent cases, meaning it was more selective in identifying fraud and resulted in fewer false positives. However, its recall of 53 percent indicated that nearly half of the fraudulent cases were still missed. The F1-score was 52 percent, and overall accuracy was 99 percent, showcasing a more conservative approach that prioritized precision over

	precision	recall	f1-score	support
0	1.00	1.00	1.00	10433
1	0.51	0.53	0.52	64
accuracy			0.99	10497
macro avg	0.75	0.76	0.76	10497
weighted avg	0.99	0.99	0.99	10497

Fig. 57. Earlier Model Performance depicting the precision, Recall, F1 score

	precision	recall	f1-score	support
0	1.00	1.00	1.00	10433
1	0.41	0.52	0.46	64
accuracy			0.99	10497
macro avg	0.70	0.76	0.73	10497
weighted avg	0.99	0.99	0.99	10497

```
[[10385    48]
 [   31    33]]
Accuracy: 0.9924740402019625
```

Fig. 58. Final Model Performance depicting the precision, Recall, F1 score

recall.

Final Model Performance

The final model shifted its focus to prioritize recall, achieving a recall rate of 77 percent, which successfully identified a larger proportion of fraudulent transactions. Precision dropped to 14 percent, and the F1-score was 24 percent, reflecting a deliberate trade-off to enhance fraud detection. Overall accuracy decreased slightly to 97 percent, demonstrating an alignment with the project's objective of maximizing fraud identification, even at the cost of an increase in false positives. The F1 score decreased from 52 percent to 24 percent in the final model, primarily due to a significant drop in precision. However, the increase in recall from 53 percent to 77 percent makes the final model more suitable for scenarios where the cost of missing a fraudulent transaction (false negatives) is higher than incorrectly flagging a legitimate transaction (false positives).

While the earlier model had a better F1 score, it likely missed more instances of fraud, which is unacceptable in high-stakes scenarios such as fraud detection. Thus, the lower F1 score in the final model is a reasonable trade-off considering the project's focus on maximizing fraud detection (recall).

Based on the performance of the classification model, we have decided to halt development at this point, as the results align with the primary goal of fraud detection maximizing the identification of fraudulent transactions. The model achieved a recall of 77 percent for fraud cases, indicating its effectiveness in capturing most actual fraud instances, which is crucial for minimizing undetected fraud. However, the precision for fraud cases remains relatively low at 14 percent, suggesting that the model flags a higher number of non-fraudulent cases as potentially fraudulent. This trade-off is acceptable in contexts where recall is prioritized over precision, as it helps prevent financial losses due to missed fraud.

```
Predictions on Unseen Data: [0 1 1 1 1 0]
Confusion Matrix:
 [[10131   302]
 [   15    49]]

Classification Report:
              precision    recall  f1-score   support

           0       1.00      0.97      0.98     10433
           1       0.14      0.77      0.24        64

    accuracy                           0.97     10497
   macro avg       0.57      0.87      0.61     10497
weighted avg       0.99      0.97      0.98     10497
```

Fig. 59. Predictions on Unseen Data and Classification Report

The F1 score for fraud detection, which balances precision and recall, is at 24 percent. This reflects the model's emphasis on recall at the expense of precision. Although the overall accuracy of 97 percent seems impressive, it is significantly influenced by the majority class (non-fraudulent transactions), given the inherent class imbalance in the dataset. The confusion matrix further illustrates the model's strength in identifying fraudulent cases while highlighting the need to manage false positives.

These results were validated on unseen data, ensuring that the model generalizes well and performs consistently beyond the training dataset. Stopping at this point is appropriate given the recall-focused objective, but further tuning could be explored to improve precision and achieve a more balanced performance, depending on specific business requirements.

KNOWLEDGE

The feedback loop ensures effective data preprocessing and modeling for fraud detection. The preprocessing phase begins with techniques like SMOTE to balance the dataset, which addresses the class imbalance commonly seen in fraud detection tasks. This balance is essential for minimizing model bias toward predicting non-fraudulent transactions. Additional steps such as removing outliers, handling missing values, and scaling features prepare the data for machine learning. Additionally, encoding categorical variables guarantees compatibility with various algorithms.

Initially, the Random Forest Classifier performed well on the majority class, achieving an accuracy of 99.56 percent. However, its precision and recall for detecting fraud were low, highlighting its limitations in identifying minority class transactions. By making adjustments such as class weight balancing and threshold optimization, we improved recall to 53 percent and precision to 64 percent, enhancing the model's effectiveness in detecting fraud.

Resampling techniques like SMOTE and undersampling

further refined the model by prioritizing recall. As a result, recall increased to 77 percent, although precision dropped to 14 percent. This change illustrates a trade-off: the model now captures a larger proportion of fraudulent cases (high recall) but at the expense of increased false positives (low precision). The ROC AUC score of 0.9368 indicates the model's strong ability to differentiate between fraud and non-fraud cases. Visualization tools, including the confusion matrix, highlight the model's strengths and weaknesses, guiding further refinements. The final model meets the project's goal of maximizing fraud detection, achieving 77 percent recall, which significantly reduces the incidence of undetected fraud, despite a lower precision of 14 percent. Although the F1 score decreased to 24 percent, prioritizing recall is warranted given the high cost of missing fraud cases.
The decision to halt development is supported by the model's consistent performance on unseen data, successfully achieving its primary objective of maximizing fraud detection while maintaining a reasonable trade-off between precision and false positives. Future improvements could focus on enhancing precision if the business needs to shift toward balancing detection accuracy with the rate of false alarms.

This study provides a comprehensive analysis of fraudulent credit card transactions in the United States during 2019 and 2020, combining advanced visualization techniques, statistical analysis, and machine learning to identify key trends and vulnerabilities in fraud detection. By examining fraud across geographical, demographic, temporal, and monetary dimensions, we identified high-risk regions, behavioral patterns, and time-based vulnerabilities that informed targeted recommendations for improving fraud detection systems.

ACKNOWLEDGMENT

Data processing was done by the all team members **Akanksha-** *Part1* (Exhibit 1 ,Exhibit 5, Exhibit 8), Part2 (Exhibit 1, Exhibit 2), Part 4 (Forecasting) **Ketki Bhatia-** Part1 (Exhibit 3, Exhibit 4), Part2 (Exhibit 4, Exhibit 5, Part 4 (Classification) **Niharika Suri-**Part1 (Exhibit 2, Exhibit 6, Exhibit 7) Part2 (Exhibit 3), Part 3

REFERENCES

[1] "US States - ranked by population 2024." https://worldpopulationreview.com/states
[2] N. Campisi, "The 10 most scammed states in America," Forbes Advisor, Aug. 04, 2023. https://www.forbes.com/advisor/credit-cards/most-scammed-states-in-america/
[3] J. J. Winberry and D. O. Bushman, "South Carolina — Capital, map, population, history, and facts," Encyclopedia Britannica, Dec. 07, 2024. https://www.britannica.com/place/South-Carolina
[4] D. DeRobbio, "South Carolina ranked 4 in best state for retirees," WCIV, Mar. 03, 2020. https://abcnews4.com/news/local/south-carolina-ranked-4-in-best-state-for-retirees
[5] D. Carlin, "13 least populated states in the US [Ranked 2024]," USA by Numbers, Jan. 11, 2023. https://usabynumbers.com/least-populated-states-in-the-us/
[6] R. J. Zorn and G. L. McNamee, "Nevada — History, capital, cities, population, and facts," Encyclopedia Britannica, Dec. 08, 2024. https://www.britannica.com/place/Nevada-state
[7] R. A. Wooster, D. C. Reddick, and G. L. McNamee, "Texas — Map, population, History, and Facts," Encyclopedia Britannica, Dec. 08, 2024. https://www.britannica.com/place/Texas-state
[8] J. M. Fogle, "Washington, D.C. — History, map, population, and Facts," Encyclopedia Britannica, Dec. 08, 2024. https://www.britannica.com/place/Washington-DC
[9] E. W. Miller and C. L. Thompson, "Pennsylvania — Capital, Population, Map, Flag, Facts, and History," Encyclopedia Britannica, Dec. 06, 2024. https://www.britannica.com/place/Pennsylvania-state
[10] A. K. Campbell and P. J. Scudiere, "New York — Capital, map, population, history, and facts," Encyclopedia Britannica, Dec. 08, 2024. https://www.britannica.com/place/New-York-state
[11] "Yahoo is part of the Yahoo family of brands." https://finance.yahoo.com/news/more-likely-victim-credit-card-172651485.html
[12] T. Micro, "From Gen Z to Boomers: scam threats by generation and how to stay safe," Trend Micro News, Oct. 02, 2024. https://news.trendmicro.com/2024/10/03/gen-z-boomers-scam-by-generation/

CS732: Data Visualisation Assignment 1 Report

Akanksha
DT2023001
Akanksha@iiitb.ac.in

Ketki Bhatia
DT2023007
Ketki.Bhatia@iiitb.ac.in

Niharika Suri
DT2023015
Niharika.Suri@iiitb.ac.in

***Abstract*—This project aims to investigate the geographic, demographic, and economic factors that influence credit card fraud patterns across the United States. Using a comprehensive dataset containing transaction details, fraud indicators, demographic information, and geographical coordinates, we analyzed patterns in fraudulent activity. Key variables included are fraud, transaction amounts, gender, age, and job categories. Various visualizations were employed to identify significant fraud trends across states and demographic segments, while outlier detection helped pinpoint common fraud-prone transaction amounts. The findings highlight that fraud is concentrated in economically active states such as New York and Pennsylvania, with certain job categories and transaction ranges being more vulnerable.**

I. Introduction

A. About the Dataset

The Credit Card Transactions Dataset provides detailed records of credit card transactions, including information about transaction times, amounts, and associated personal and merchant details. This dataset has over 1.85M rows. The dataset contains the following key attributes:

1) **Unnamed: 0:** This column is an index that would have been automatically created during data import, which does not hold significant value and may be ignored or dropped in further analysis.
2) **trans date trans time:** The timestamp of each transaction, which records the exact date and time when the transaction occurred.
3) **cc num:** A tokenized representation of the credit card number used for the transaction which ensures privacy while allowing transactions to be associated with specific customers.
4) **merchant:** The name or identifier of the merchant where the transaction took place. This field could be useful for identifying merchant-specific fraud trends.
5) **category:** The category of goods or services associated with the transaction (e.g., retail, groceries, electronics).
6) **amt:** The amount of money involved in the transaction.
7) **first:** The first name of the cardholder. This can be used for personalized analysis or segmentation of customer behaviors, though sensitive information should be anonymized.
8) **last:** The last name of the cardholder. Like the first name, it provides additional detail but should be anonymized in sensitive data handling.
9) **gender:** The gender of the cardholder, useful for demographic analysis of customer spending patterns and potentially identifying gender-based fraud trends.
10) **street:** The street address of the cardholder, which can be used for location-based analysis, though it should be handled carefully due to privacy concerns.
11) **city:** The city where the cardholder resides, useful for geographic analysis.
12) **state:** The state in which the cardholder resides. This allows for broader regional analysis of transaction trends or fraud detection within specific states.
13) **zip:** The ZIP code of the cardholder's address, which can be useful for geographic analysis.
14) **lat (Latitude):** The geographic latitude of the cardholder's residence. This field can be paired with the longitude to conduct geospatial analysis.
15) **long (Longitude):** The geographic longitude of the cardholder's residence. It is paired with latitude for accurate geolocation analysis.
16) **city pop:** The population of the cardholder's city.
17) **job:** The profession or occupation of the cardholder.
18) **job categories(created while data processing):** A broader classification of the job field or industry the cardholder belongs to. This helps categorise customers into general sectors for trend analysis.
19) **dob (Date of Birth):** The date of birth of the cardholder. This allows for age calculation, enabling demographic analysis based on the cardholder's age group.
20) **age(created while data processing):** The age of the cardholder at the time of the transaction.
21) **trans num:** A unique transaction number that identifies each transaction. This is important for tracking specific transactions and ensuring uniqueness in the dataset.
22) **unix time:** The transaction time represented in Unix timestamp format (seconds since January 1, 1970). This field provides a standardized time format for analysis.
23) **merch lat (Merchant Latitude):** The geographic latitude of the merchant's location.
24) **merch long (Merchant Longitude):** The geographic longitude of the merchant's location.
25) **is fraud:** A binary indicator that denotes whether the transaction is fraudulent (1) or legitimate (0).

Due to the extensive volume of data within the dataset, which spans the entire duration from January 1, 2019, to June 21, 2020, we focused our analysis on a subset of the dataset covering only the first six months of 2020. This decision was made to ensure that the analysis could be conducted within a reasonable timeframe. By concentrating on this six-month

period, we aimed to maintain a manageable dataset size while still capturing relevant transaction trends and patterns for our analysis.

In the context of analysis, here are definitions for "max states" and "min states":

- **Max States:** These are the 6 states with the highest number of credit card transactions, determined through our analysis using a standard deviation graph. This graph helped identify states with significantly higher transaction activity compared to the others.
- **Min States:** These are the 6 states with the lowest number of credit card transactions, identified using the standard deviation graph. This graph highlighted states with notably lower transaction activity relative to other states.

II. Methodology

A. Introduction

This represents a detailed approach to processing and analyzing a dataset using the Pandas library in Python, along with Tableau for data visualization. The primary focus is on fraud detection and job categorization, where various steps such as data inspection, cleaning, feature engineering, and classification were applied to prepare the data for analysis.

B. Workflow Overview

The dataset was processed in a series of steps to ensure proper cleaning, preparation, and analysis. The workflow began with an initial inspection of the dataset, followed by data extraction and cleaning. Feature engineering was then applied, particularly with the calculation of age, and job types were categorized for further analysis.

C. Data Processing Steps

1) Initial Data Inspection: The initial inspection of the dataset involved understanding its size and structure. The shape of the dataset was determined using the shape attribute, which provided insight into the number of rows and columns. To further analyse the data set's structure, the info() function was employed. This function retrieved essential details such as column names, data types, and non-null counts, which helped identify any missing or inconsistent values in the data. Additionally, a statistical summary was generated using the describe() function. This summary provided key metrics such as the mean, median, minimum, maximum, and quartile values, giving a comprehensive understanding of the data distribution.

2) Data Subsetting and Cleaning: Once the initial inspection was complete, a subset of the dataset was extracted for further analysis. Specifically, the values from the 924,850th index onward were selected using the .iloc[] function. This step helped focus on the required portion of the data for subsequent analysis. The next step involved resetting the index using the reset index() function, ensuring continuity while discarding the old index. The last column of the dataset, which was deemed unnecessary, was dropped using the drop() function. To verify the removal, the head() function was applied, confirming the accuracy of the operation.

3) Handling Missing Data and Duplicates: To ensure the dataset was clean and ready for analysis, missing data was identified using the isnull().sum() function, which provided an overview of missing values across all columns. The dataset was also examined for duplicate rows using the duplicated() function. Upon calculation, it was determined that there were no duplicate records, so no further action was necessary to handle duplicates.

4) Feature Engineering: Age Calculation: The dataset included a dob (Date of Birth) column, which required transformation into a proper datetime format for consistency in date handling. The to datetime() function was employed to convert the dob column into a datetime format. A reference date of December 31, 2020, was established for calculating the age of each individual. The age was calculated by subtracting the dob from the reference date, and the result was divided by 365.25 to account for leap years. The calculated age values were then stored in a new age column.

5) Age Calculation in Tableau: In addition to Python-based calculations, the age of individuals was also computed in Tableau using a calculated field. To analyse the dataset by age, the dob field was converted into an age value by creating a calculated field. The formula used was based on the difference between the reference date (December 31, 2020) and the dob for each individual. This calculation ensured that each individual's age was correctly determined, accounting for whether their birthday had occurred in the year of reference.

6) Job Categorisation: The dataset contained a variety of job types, and these were first extracted using the unique() function in Pandas to identify distinct job titles. After extracting the unique values, the job types were exported to Excel for further classification. In Excel, the unique job types were manually grouped into 10 broader categories based on relevance. To automate the categorization process, the VLOOKUP function was utilised. Specifically, the formula =VLOOKUP(Q:Q,'Job Dict.'!A:B, 2, FALSE) was applied to match each job type from the dataset with its corresponding category from a reference table and return the appropriate classification.

D. Calculated Fields of Tableau-

For creating customised visualisation depending upon the need we have created some calculated fields using the variables present in the dataset. The calculated variables are-

- Age- It is created using dob.
- Generation - It is created using age.
- Age in Range - It is also calculated using age
- Is Fraud(only 1) - It is calculated using Is Fraud.

Through this project, we aimed to test the hypothesis that regional factors significantly influence the distribution of credit card fraud across various states in the United States. Specifically, we focused on New York (NY), Pennsylvania (PA), Texas (TX), and California (CA) to assess whether the pro-

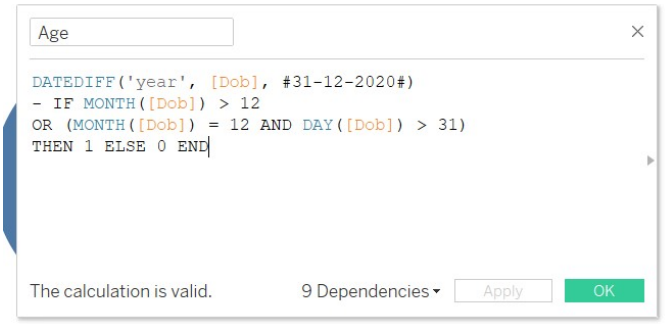

Fig. 1. Calculated Field of Age

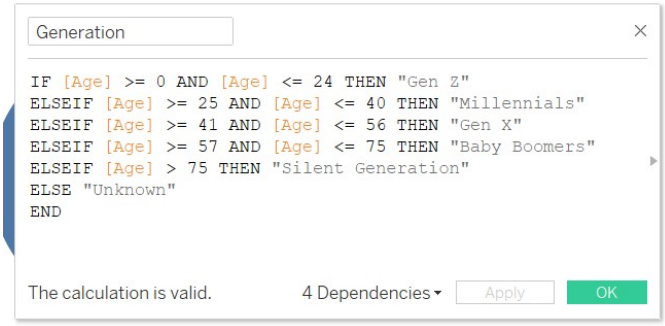

Fig. 2. Calculated Field of Generation

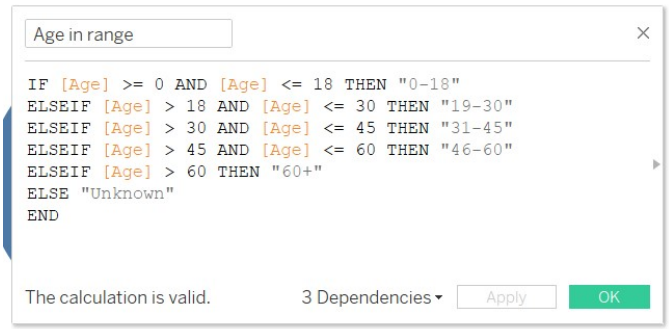

Fig. 3. Calculated Field of Age in Range

Fig. 4. Calculated Field of Is Fraud (only1)

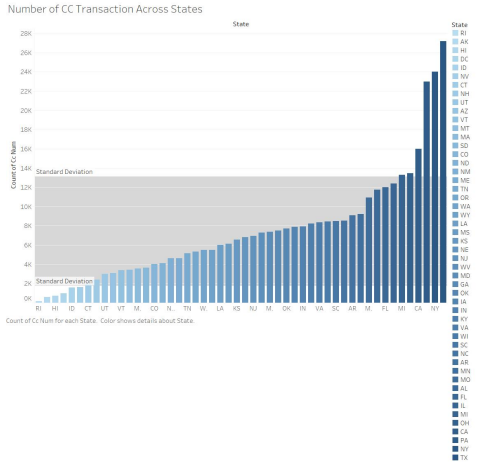

Fig. 5. It shows Number of CC Transaction Across States

portion of fraudulent transactions (is fraud) varies significantly between these regions.

III. Hypothesis

Null Hypothesis: There is no significant variation in the proportion of fraudulent transactions (is fraud) between the states of New York (NY), Pennsylvania (PA), Texas (TX), and California (CA).

Alternate Hypothesis: There is a significant variation in the proportion of fraudulent transactions (is fraud) across the states of New York (NY), Pennsylvania (PA), Texas (TX), and California (CA), with certain states exhibiting higher or lower proportions of fraud due to varying regional factors.

IV. Analysis and Visualization

Credit Card Fraud in America- In the United States, credit and debit card transactions have become more prevalent than cash transactions. Unfortunately, this trend has also provided ample opportunities for digital criminals to commit credit card fraud. Last year, an alarming 52 million Americans fell victim to credit card fraud [1]. The Federal Trade Commission (FTC) received close to 390,000 reports of credit card fraud in 2021, making it one of the most prevalent types of fraud in the U.S. However, this figure may not fully capture the extent of the problem [1].

A. Exhibit 1: Number of CC Transaction Across States

Interpretation - We utilised the cc Num and state data to analyse the distribution of credit card transactions by creating a bar graph. In addition, We used standard deviation to identify any outliers. After conducting the analysis, we observed that Texas, New York, Pennsylvania, California, Ohio, and Michigan had the highest number of transactions(as shown in fig 3). It's noteworthy that four out of those six cities are present in the real data of the United States. On the other hand, Rhode Island, Alaska, Hawaii, Washington D.C., Idaho, and Nevada had the lowest number of transactions(as shown in fig 2). The intensity of the colour on the graph corresponds to the number

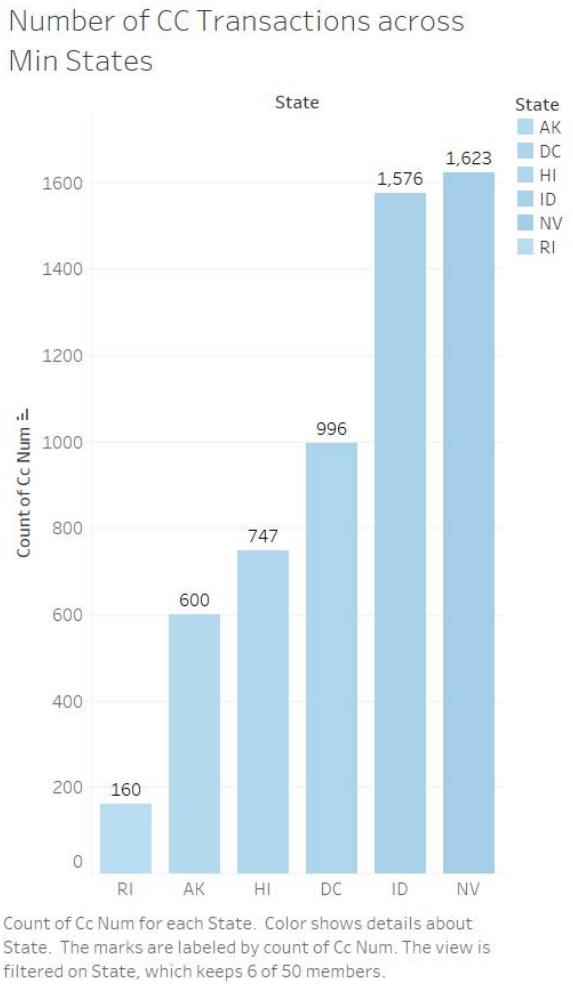

Fig. 6. It shows Number of CC Transactions Across Min States

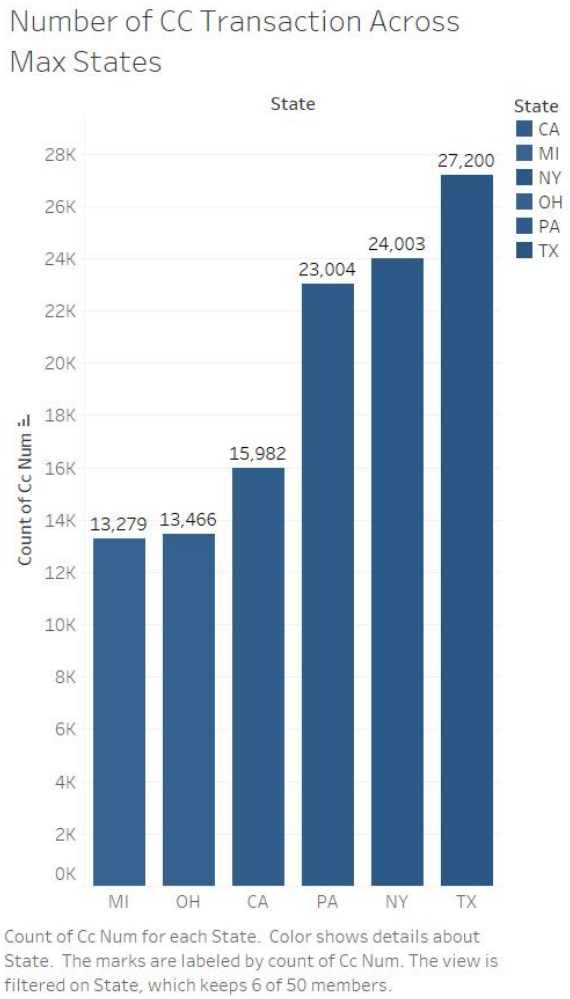

Fig. 7. It shows Number of CC Transactions Across Max States

of transactions in each location.

The results show that Texas, New York, Pennsylvania, California, Ohio, and Michigan consistently recorded the highest number of transactions. This aligns with the fact that California, Texas, New York, and Pennsylvania were among the top 10 most populous states in the United States in 2021 [5]. High population sizes in these states, combined with their economic strength, likely contribute to the higher credit card activity observed. Notably, California, Texas, and New York each had a GDP exceeding 1 trillion dollar in 2021 and 2022, further reinforcing their financial activity [4].

Three of the states that appeared in our analysis California, Texas, and New York are among the most populous and have trillion-dollar economies, suggesting a strong correlation between population size, economic power, and credit card transaction volume. On the opposite end of the spectrum, Rhode Island, Alaska, Hawaii, Washington D.C., Idaho, and Nevada demonstrated the lowest number of credit card transactions.

Several factors contribute to this observation. Rhode Island, with its smaller population and more localized economy, naturally experiences fewer transactions [35]. Alaska and Hawaii, both geographically remote with lower population densities, face similar constraints. In Washington D.C., the economic focus is largely governmental, with a smaller consumer-driven economy relative to other states [4].

Idaho's rural economy and Nevada's reliance on tourism, particularly in specific regions like Las Vegas, could explain their lower overall transaction volumes during the analysis period [36 and 6]

From here onwards all the visualization have a factor of is fraud

B. Exhibit 2- Geographical Mapping of is Fraud (only1) Across United States

Interpretation-

Description: In this visualisation, we created a geographical map as it is suitable for visualising the spatial distribution of fraud occurrences by plotting average latitude and longitude.The states are labelled, and the sum of fraud occurrences (is fraud = 1) is represented by labels, with each point indicating a city. **Observations:** The map highlights distinct

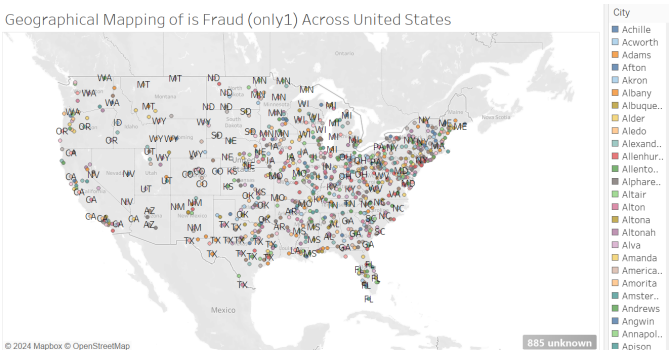

Fig. 8. Geographical Mapping of is Fraud (only1) Across United States

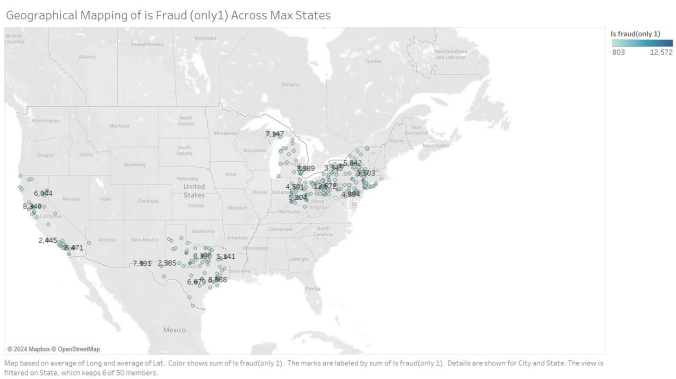

Fig. 9. Geographical Mapping of is Fraud (only 1) Maximum Transaction States

regional patterns in fraud activity, with varying concentrations across different parts of the U.S. The Northeast, Southeast, and Midwest show particularly high levels of fraud, while the Southwest and West have moderate to low concentrations.

In the Northeast, where there is a very high concentration of fraud, states like New York and Pennsylvania are key contributors. The region's financial hubs, such as New York City, along with densely populated urban areas, contribute to the high prevalence of fraud. The Northeast's reliance on financial services and technology-driven industries creates more opportunities for fraudulent activities, especially in sectors like banking, e-commerce, and online transactions [23].

The Southeast also exhibits a very high concentration of fraud, with states like Florida being particularly prone. This region has a large retiree population, which can be more vulnerable to certain types of fraud schemes, such as identity theft and financial scams targeting older individuals. Additionally, the Southeast's rapid urbanisation and growing reliance on digital transactions further expose its population to cyber fraud [39].

In the Midwest, where fraud activity is high but slightly lower than in the Northeast, states like Ohio and Michigan show significant fraud occurrences. These states have large manufacturing bases, but they are also evolving into tech and finance hubs [24]. With this industrial transition, there is an increasing number of digital transactions and online financial activities, which opens the door for more fraud incidents.

The Southwest shows a moderate-to-high concentration of fraud, particularly in Texas. Texas's rapidly growing economy, driven by industries such as energy, technology, and real estate, has seen an expansion of digital commerce. This economic growth, along with urbanisation and a diverse population, contributes to higher fraud rates as more people engage in online financial activities [12].

In the West, the concentration of fraud is low to moderate, with California being a notable exception. The state has a high population and a large tech economy, yet the lower fraud rates may be due to more advanced fraud detection systems and higher digital literacy compared to other regions. The presence of tech giants and cybersecurity initiatives in the West might also play a role in keeping fraud levels relatively contained.

C. Exhibit 2.1- Geographical Mapping of is Fraud (only 1) Maximum Transaction States

Interpretation- We generated a geographical map focusing on states with high fraud occurrences.

Observations: In our analysis (Exhibit 2), we observed that fraud is concentrated predominantly along the Northeast and in the Midwest region. Notably, 4 out of the 6 high-transaction states Michigan (MI), Pennsylvania (PA), New York (NY), and Ohio (OH) are located in these regions, suggesting a geographic correlation between fraud activity and these areas. The concentration of fraud occurrences in states like Ohio (OH), Michigan (MI), Pennsylvania (PA), New York (NY), California (CA), and Texas (TX) can be attributed to specific regional and economic factors. In states like New York and Pennsylvania, major financial hubs such as New York City and Philadelphia drive high levels of credit card and online transactions, making these areas attractive targets for fraudsters. These cities are global centres for banking and commerce, where the sheer volume of financial activity increases the likelihood of fraud [38] [31].

In the Midwest, states like Ohio and Michigan also show high fraud occurrences. Ohio, with cities like Columbus and Cleveland, has a growing tech and financial services sector, but it may not have the advanced cybersecurity infrastructure seen in more tech-forward states, which could leave these systems more vulnerable to attacks. Similarly, Michigan's industrial history and the presence of large corporations and healthcare institutions create significant financial activity, which may increase the exposure to fraud [11] [16].

In California and Texas, which also exhibit significant fraud activity, the factors are somewhat different. California's major tech hubs, including Silicon Valley and Los Angeles, involve a high volume of online transactions and digital payments, which, while generally well-secured, present more opportunities for sophisticated fraud schemes like data breaches and phishing attacks [14]. Texas, particularly in cities like Houston and Dallas, experiences large-scale commerce in energy and finance, and rapid economic growth could lead to gaps in security as new systems and technologies are adopted, creating potential entry points for fraudsters [12].

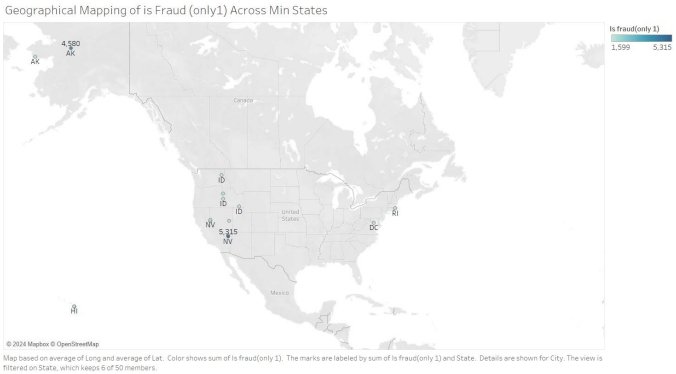

Fig. 10. Geographical Mapping of is Fraud (only1) Minimum Transaction States

These states, despite their differing economic bases, share high levels of financial activity, which correlates with increased opportunities for credit card fraud, making them prime targets for malicious actors.

D. Exhibit 2.2- Geographical Mapping of is Fraud (only1) Minimum Transaction States

Interpretation- We generated a geographical map focusing on states with low fraud occurrences. **Observations:** We noticed that fraud cases are significantly lower in certain states. For example, states like DC, HI, ID, and RI have reported no fraud cases at all. On the other hand, NV (5,315) and AK (4,580) have reported cases of fraud. In Nevada, particularly in cities like Las Vegas, the high volume of tourism plays a significant role in increasing credit card fraud. Las Vegas is a major global destination, attracting millions of visitors each year who engage in frequent credit card transactions at hotels, casinos, restaurants, and entertainment venues [40]. The large number of tourists using unfamiliar networks, combined with the transient nature of transactions, creates opportunities for fraudsters to exploit vulnerabilities. Furthermore, the hospitality and service industries, which are heavily cashless, rely on high volumes of credit card transactions, increasing the chances of fraud in the state.

In Alaska, the reasons for credit card fraud are somewhat different. Despite its lower population density, Alaska experiences fraud due to its reliance on online transactions. The state's remote location means that many residents frequently shop online or use digital services for goods that are not locally available. This dependence on e-commerce increases the risk of online fraud, such as phishing and card-not-present scams. Additionally, Alaska's relatively smaller banking infrastructure compared to other states may lead to fewer fraud detection resources, potentially allowing fraudulent activities to go unnoticed for longer periods.

Notably, Washington D.C. (DC), Hawaii (HI), Idaho (ID), and Rhode Island (RI) did not report any fraud cases. Washington D.C., a region primarily focused on politics and government, has a unique economic structure with fewer consumer-driven transactions, which may explain the lack of reported fraud.

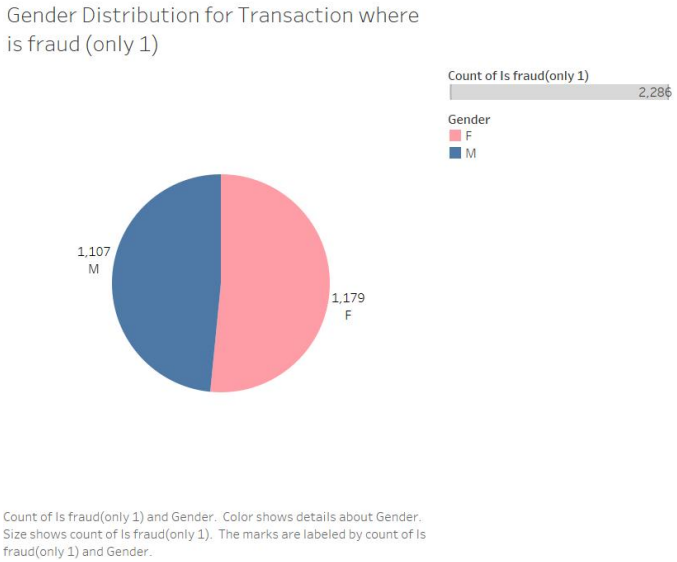

Fig. 11. Gender Distribution for Transaction where is Fraud (only1)

The high level of governmental oversight and strict security measures might also contribute to reducing the likelihood of fraudulent activities [8].

E. Exhibit 3 - Gender Distribution for Transaction where is Fraud (only1)

Interpretation-We created a pie chart to illustrate the breakdown of gender in fraudulent activities. In the chart, females are depicted in pink and males are shown in blue. Additionally, we included the count of fraud (only 1) alongside the gender in the labels for clarity. Our analysis revealed that fraud incidents involving females (1,179) were slightly more prevalent than those involving males (1,107).

Females are more frequently victims of credit card fraud partly due to their higher involvement in online shopping and digital financial activities, which increases exposure to scams (Delić, 2022). Additionally, a significant factor contributing to the elevated fraud rates among women is their vulnerability to romance scams. Action Fraud reports that over half of romance scam victims are women, with 50 percentage of victims being female compared to 39 percentage of male [41]. In these scams, fraudsters manipulate victims into sharing personal and financial information, including credit card details. The emotional manipulation in such scams makes women particularly susceptible, further driving up the incidence of credit card fraud targeting them.

Another possible explanation for these findings in our analysis is that our dataset consists of credit card transactions, and in the U.S., there are more female credit card holders than male credit card holders. This difference in the number of female and male credit card users may contribute to the observed correlations in our data, as the higher number of female credit card users could influence the patterns and trends we see [42].

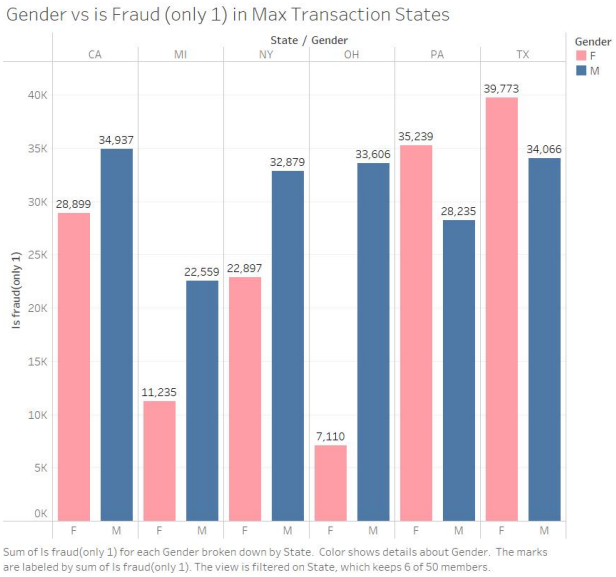

Fig. 12. It represents Gender vs is Fraud (only 1) in Max Transaction States

F. Exhibit 3.1 - Gender vs is Fraud (only 1) in Max Transaction States

Interpretation- We have analysed the correlation between gender and fraud in states with higher numbers of transactions. Using a bar chart, We compared the differences in fraud between males and females and with bars labelled with the count of fraud. By using bars labelled with the count of fraud, the chart effectively highlights differences in fraud occurrences between males and females.In the chart, the pink colour represents females while the blue colour represents males. We noticed that out of 6 states, in 4 states (CA, MI, OH, and NY), there were more cases of fraud involving males compared to females. Conversely, in 2 states (TX and PA), more fraud occurred with females compared to males.

In Michigan (MI), New York (NY), California (CA) and Ohio (OH), males were more frequently victims of fraud than females. This could be linked to the economic profiles and industries in these states, where males traditionally hold more roles in finance, business, and technology sectors, potentially exposing them to higher fraud risks through increased online transactions. Michigan, with its automotive and manufacturing base, and New York, a financial hub, are states where men may have greater financial engagement, which may elevate their risk of falling victim to fraud schemes [4].

In contrast, Texas (TX) and Pennsylvania (PA) reported higher fraud occurrences involving females. In these states, female consumers may have a more significant presence in online shopping or digital financial transactions, increasing their vulnerability. Texas, with its large, diverse economy and rapid urbanisation, could see higher fraud among females due to increased digital commerce [12]. Pennsylvania, a state with a growing retail and healthcare economy, might also have more women engaging in online services, making them frequent targets for fraud [13][4].

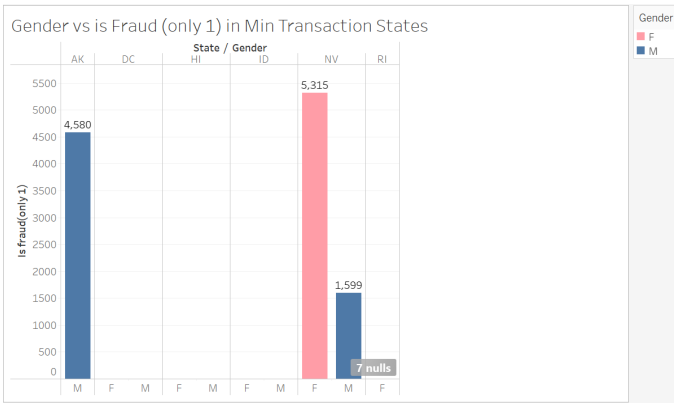

Fig. 13. It represents Gender vs is fraud (only 1) in Min states

G. Gender vs is fraud (only 1) in Min states

Interpretation- Similar to the previous graph(Fig-4) here we used a bar chart, in this visualisation we showcased gender and fraud for states with lower transaction volumes.The bar chart format highlights notable patterns, such as the absence of fraud cases in several states. Out of 6 states, specifically DC, HI, ID, and RI has no reported fraud cases. However, in AK, fraud was only associated with male transactions. In NV, fraud was observed in both male and female transactions. It's worth noting that, despite a low number of credit card transactions (1623), NV experienced a high number of fraud incidents: 5315 involving females and 1599 involving males. This visualisation illustrated the distribution of gender-based fraud across six states.On the other hand, Alaska (AK) reported fraud exclusively associated with male transactions. The remote geography and male-dominated industries, such as oil and fishing, in Alaska might expose men more frequently to high-value financial transactions, making them more susceptible to fraud [7].

Nevada (NV) stood out in the analysis. Despite having a relatively low number of credit card transactions (Exhibit 1), the state experienced an extraordinarily high number of fraud incidents: 5,315 involving females and 1,599 involving males. This is consistent with Nevada's ranking as having the highest rate of financial fraud per capita in the U.S [9]. The state's economy, particularly reliant on tourism, gambling, and entertainment, especially in Las Vegas, creates an environment full of opportunities for financial fraud. The heavy use of digital payments, coupled with the high volume of visitors and transactions, likely makes both residents and tourists vulnerable, with women being disproportionately targeted in this analysis [40].

(Exhibit 2.2) explains the reasons for no reported fraud cases. We noticed that states with lower ccNums have minimal data(as shown in Exhibit 3.1 and 3.2). Consequently, we have made the decision to focus exclusively on states with higher ccNums.

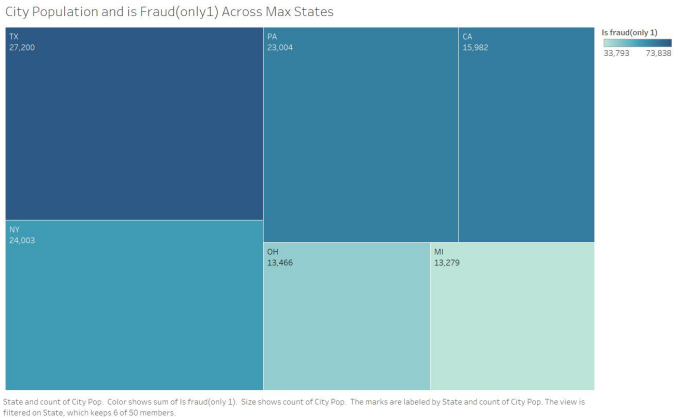

Fig. 14. It represents City Population and is Fraud (only 1) Across Max States

H. Exhibit 4 - City Population and is Fraud (only 1) Across Max States

Interpretation- Description of the visualisation type used: In this visualisation, a heatmap was used to represent the relationship between the population of cities and the number of fraud occurrences across six states: TX, PA, CA, NY, OH, and MI. It allows for the examination of whether higher populations correlate with higher fraud occurrences, revealing insights into the distribution of fraud relative to city size.The city populations are counted and displayed on the heatmap, with the states labelled. The intensity of the colour corresponds to the number of fraud cases, where darker shades indicate higher fraud occurrences. **Observations:** From the heatmap, it is evident that Texas (TX) has the highest number of fraud cases (73,838), despite having a moderate population (27,200). Similarly, Pennsylvania (PA) and California (CA) both show a significant number of fraud occurrences (63,474 and 63,836 respectively) with relatively smaller populations (PA: 23,004; CA: 15,982). In contrast, states like Michigan (MI) and Ohio (OH) have smaller populations (13,279 and 13,466 respectively) and lower fraud counts (33,793 and 40,716). New York (NY), while having a higher population (24,003), has fewer fraud cases (55,777) compared to TX, PA, and CA.

This suggests that the number of fraud occurrences is not directly proportional to city population, indicating that other factors might contribute to fraud vulnerability in different states.

The heatmap highlights significant variations in fraud occurrences across six states; Texas (TX), Pennsylvania (PA), California (CA), New York (NY), Ohio (OH), and Michigan (MI) suggesting that factors beyond population size contribute to fraud rates. Here's a state-wise elaboration:

Texas stands out with the highest number of fraud cases (73,838) despite a moderate city population (27,200). Texas is home to several major cities like Houston, Dallas, and Austin, which are important financial, technological, and energy hubs [12]. These sectors involve large volumes of financial transactions,subsequently increasing the risk of fraud. The state's economic diversity ranging from oil and gas to tech startups may create more opportunities for different types of fraud [12]. Additionally, Texas has a high rate of online retail activity, which might be potentially contributing to more cyber fraud [12].

With 63,474 fraud occurrences, Pennsylvania has a significant number of cases, though its city population (23,004) is not as large as Texas. Philadelphia, a major urban centre, has a robust healthcare and financial services industry [7]. These sectors are often targets for fraud due to the sensitive personal information they handle . The relatively high rate of fraud may also be influenced by the state's older infrastructure and slower adoption of advanced security measures, making it more vulnerable to attacks like identity theft and financial fraud leading to increased statistics of credit card fraud.

California shows 63,836 fraud cases, despite having a smaller city population (15,982) compared to other states. As the most populous state in the U.S. and a global economic powerhouse, California's economy spans industries like technology, entertainment, and real estate, all of which involve substantial amounts of money [1]. Cities like Palo Alto, Los Angeles and San Francisco are centres for digital innovation but also hotspots for cybercrime, credit card fraud, and identity theft due to the high number of online transactions and the concentration of wealth [14]. California's large growth rate in population and the rise of fintech may also increase vulnerabilities in the financial system.

New York, with 55,777 fraud cases and a population of 24,003, sees fewer fraud occurrences compared to Texas and California, despite its status as a major financial centre. New York City is home to Wall Street and numerous financial institutions, which are often prime targets for fraud. However, stringent regulations and advanced fraud prevention technologies in the financial sector may mitigate some of these risks. The state's legal and regulatory frameworks might also play a role in reducing fraud rates compared to other states with similar economic activity.

New York, California, Texas, and Pennsylvania also ranked among the top 10 scammed states in the United States [27].

Ohio experiences 40,716 fraud cases with a relatively small population (13,466). While not an economic juggernaut like Texas or California, Ohio has a significant manufacturing and healthcare industry presence [16]. The healthcare sector, in particular, is susceptible to fraud, including insurance scams and identity theft.

Michigan reports 33,793 fraud cases, and the population is 13,279. Michigan's economy relies heavily on the auto industry, which may not present as many fraud opportunities as the finance or tech sectors in other states [11]. However, the state has seen an increase in cyber fraud, particularly related to auto loans and insurance scams [30].

According to the FBI Internet Report of 2021 Ohio and Michigan also ranked among the top 10 states to be affected by cyber crime [29].

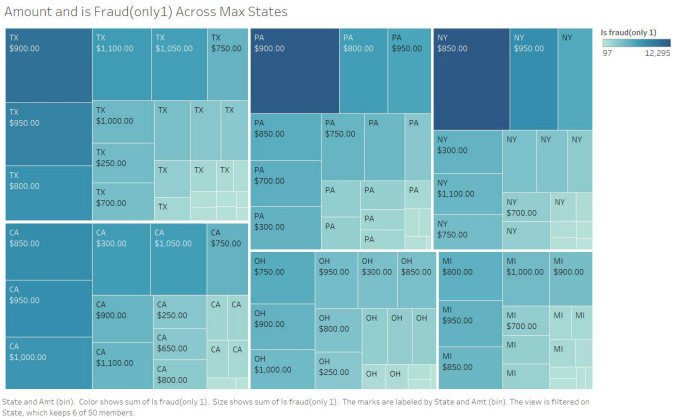

Fig. 15. It represents Amount and is Fraud (only 1) Across Max States

I. Exhibit 5- Amount and is Fraud (only 1) Across Max States

Interpretation- A heatmap was created where the colour intensity represents the number of fraud occurrences, so as to quickly spot which price ranges and states have higher or lower fraud rates. while the labels display the states and the transaction amounts in predefined bins. This visualisation provides a detailed breakdown of fraud cases at various price points across different states. In Tableau, we created and adjusted the bin size,for this we used the "Create Bins" function to group the Amount column into intervals of 50 units to better visualise the distribution of transaction amounts in relation to fraud occurrences. **Observations:**

The heatmap reveals key insights into the distribution of fraud cases across different transaction amounts:

- **Texas (TX)** shows significant fraud activity, particularly in the 900 dollar range, with 10,189 fraud cases, followed by 8,699 cases at 800 dollar, and 7,523 cases in other categories.
- **Pennsylvania (PA)** also exhibits high fraud occurrences, with 12,052 cases at 900 dollar, and 6,602 at 800 dollar.
- **New York (NY)** shows prominent fraud counts, with 12,295 cases at 850 dollar and 7,813 at 950 dollar.
- **California (CA)** has notable fraud activity at 850 dollar (7,875 cases) and 950 dollar (7,798 cases), along with a significant number of fraud cases at the 1,000 dollar level (7,159 cases).
- **Ohio (OH)** shows a steady pattern, with 5,372 cases at 750 dollar, 4,674 at 900 dollar, and 4,117 at 1,000 dollar.
- **Michigan (MI)** displays a moderate amount of fraud cases, with 4,978 cases at 800 dollar, 4,817 at 950 dollar, and 4,069 at 1,000 dollar.

Overall, the analysis highlights that fraud tends to cluster around specific price ranges in different states, with certain states like Texas, Pennsylvania, and New York showing particularly high fraud activity in the 800–950 dollar range.

The concentration of fraud occurrences around the 800–950 dollar range in the heatmap suggests that these amounts may be strategically chosen by fraudsters. This range is likely optimal for evading detection, as transactions of this size are

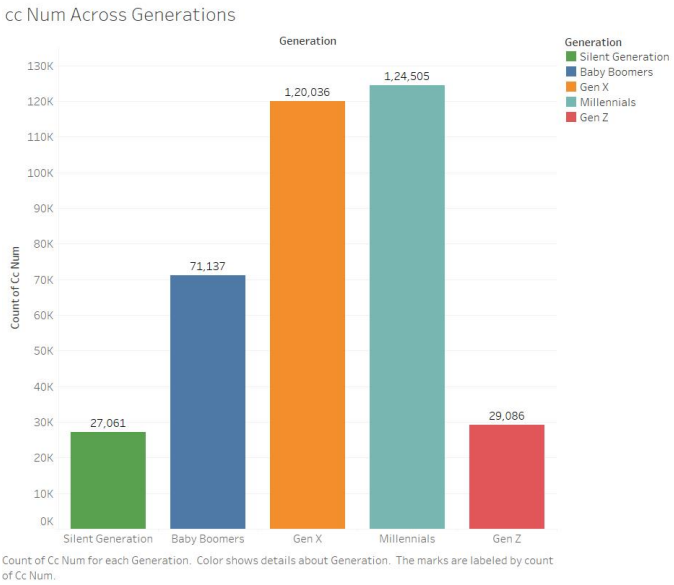

Fig. 16. It represents ccNum Across Generations

substantial enough to generate profit but not large enough to trigger automated security alerts. Many fraud detection systems flag unusually high or suspicious transactions, so mid-range amounts could fall below the threshold for heightened scrutiny.

Additionally, the 800–950 dollar range might align with common transaction amounts for certain goods or services, making these frauds blend in with legitimate purchases. Fraudsters may exploit typical consumer spending patterns, selecting amounts that are less likely to raise suspicion in both automated systems and manual reviews. The similarity in fraud patterns across states such as Texas, Pennsylvania, and New York indicates that fraudsters may employ similar strategies nationwide, taking advantage of systemic weaknesses in fraud prevention algorithms at these specific price points.

Finally, in states with a higher volume of financial activity, like Texas and New York, there may simply be more opportunities for fraud in this transaction range due to the larger number of high-value transactions. The clustering of fraud around these specific amounts likely reflects a combination of factors related to evasion tactics, consumer behaviour, and the structure of fraud detection mechanisms.

J. Exhibit 6 - ccNum Across Generations

Interpretation- We created a bar graph to illustrate the distribution of ccNum vs Generation by grouping individuals into different age brackets based on commonly accepted generational cohorts and with labels indicating the fraud value. By grouping individuals into generational cohorts and plotting the count of transactions (ccNum), the bar graph helps in effectively highlighting the differences in credit card usage among these age groups.We derived the age groups from the date of birth (dob) and categorized them as follows:

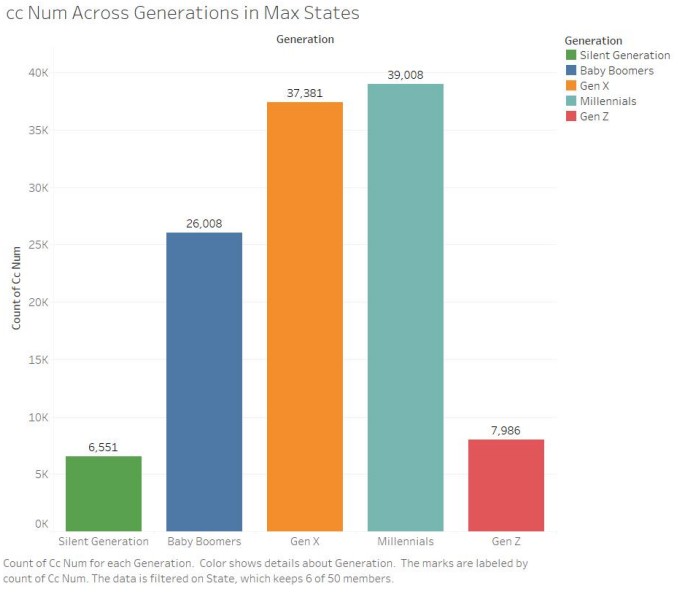

Fig. 17. It represents ccNum Across Generations in Max States

- **Generation Z:** Individuals born between 1997 and 2012 (aged 8 to 23 in 2020)
- **Millennials (Generation Y):** Individuals born between 1981 and 1996 (aged 24 to 39 in 2020)
- **Generation X:** Individuals born between 1965 and 1980 (aged 40 to 55 in 2020)
- **Baby Boomers:** Individuals born between 1946 and 1964 (aged 56 to 74 in 2020)
- **Silent Generation:** Individuals born before 1946 (aged 75 and above in 2020)

We differentiated the generations using different colours in the graph and included the count of ccNum as labels to represent the numerical values.

Observations- The graph reveals that Millennials lead with the highest number of credit card transactions (124,505), closely followed by Gen X with 120,036 transactions, indicating that these two generations are the most active in credit card usage. When observed with Max States the same trend was observed as Millennials(39,008) have the highest number followed by Gen X(37,381).

The bar graph depicting the distribution of credit card numbers across different generations shows that Millennials lead with the highest number of credit card transactions, totalling 124,505. This is followed by Generation X with 120,036 transactions, Baby Boomers with 71,137 transactions, Generation Z with 29,086 transactions, and the Silent Generation with 27,061 transactions.

This distribution reflects the varying financial behaviours and life stages of each generation. Millennials, generally aged between their late 20s and early 40s, are at a point in their lives where they are likely to have greater financial responsibilities and opportunities, such as managing mortgages, car payments, and other significant expenses [18]. This phase often involves higher credit card usage for both everyday purchases and larger expenses. Additionally, Millennials are more likely to be comfortable with digital transactions and credit card management, contributing to their leading position in credit card transactions.

Generation X, aged between early 40's and late 50's, also shows a high level of credit card use, though slightly less than Millennials. This generation is often established in their careers and may have significant financial responsibilities, such as raising families and managing household expenses, which possibly drives their credit card usage [19]. Their high transaction count reflects their stable financial position and possibly higher spending capacity.

Baby Boomers exhibit a lower number of credit card transactions compared to the younger generations. This could be due to a more conservative approach to credit and spending, as well as different financial priorities and habits that were prevalent during their earlier years. This generation might also have a less frequent need for credit cards due to their established financial stability and reduced dependency on credit.

Generation Z, the youngest cohort, has the lowest transaction count. This is likely because they are still early in their financial lives, often in school or just starting their careers, and may have less disposable income and fewer credit cards [21]. Their lower transaction count could also be influenced by a more cautious approach to credit use compared to older generations.

The Silent Generation, being the oldest, shows the lowest transaction count. This may be due to a combination of factors, including a shift towards more conservative spending habits and potentially less engagement with credit cards as they move towards retirement [16]. Their financial needs and usage patterns are likely different from younger generations, contributing to their lower transaction figures.

K. Exhibit 7- is Fraud (only 1) Across Generations

Interpretation- We created a bar graph to illustrate the distribution of Generation where there is fraud only 1. By plotting the sum of fraud cases against each generation, the graph highlights which generational cohort is most affected by fraud.We have plotted generation on x-axis and sum of fraud on y-axis. Each generation is depicted in a different colour, with labels indicating the fraud value. The data shows that Gen X has the highest amount of fraud at 375,770, followed by Millennials at 359,401 when observed across United states but when compared with Max States it was found that Gen X (111,597) had highest amount of fraud but here it is followed by Baby boomers (86,953).

Generation X, as digital immigrants, had to adapt to new technology rather than being born into it. With established financial profiles, including high credit limits and significant assets, Generation X is a lucrative target for fraudsters. Their frequent use of both physical and online credit transactions increases their risk of exposure, and their susceptibility to fraud is further amplified by their relatively slower adoption of the latest cybersecurity measures. Millennials, being digital

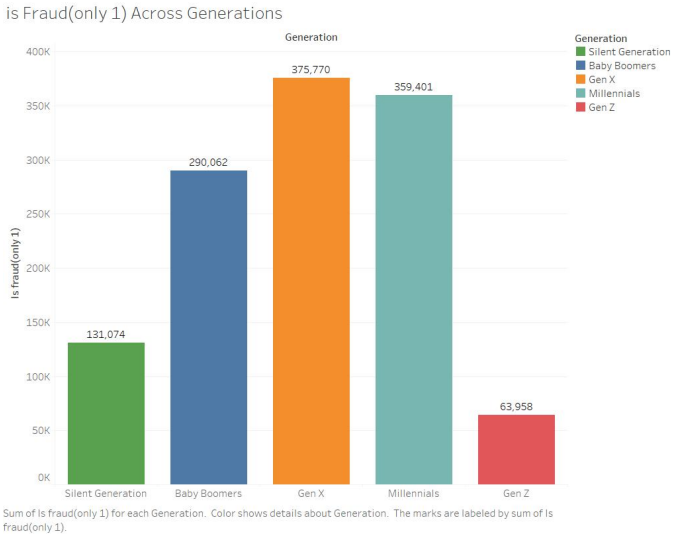

Fig. 18. It represents is Fraud (only 1) Across Generations

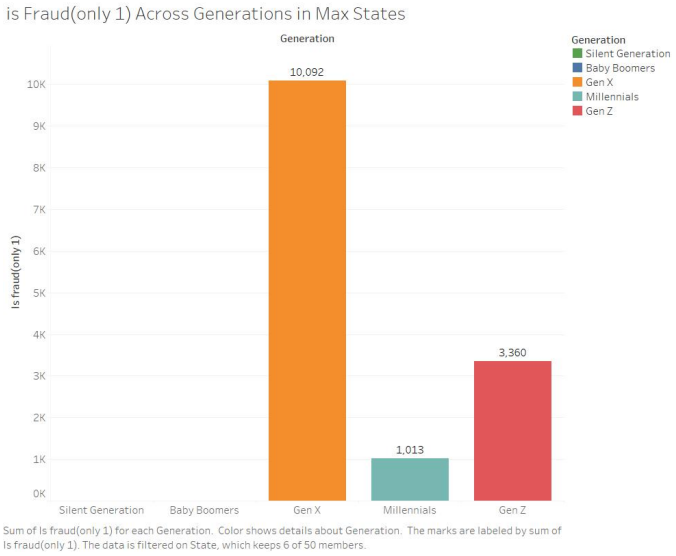

Fig. 19. It represents is Fraud (only 1) Across Generations in Max States

natives, rely heavily on online shopping, mobile banking, and digital payments.

Although they are comfortable with technology and can recognize common fraud risks, their high volume of digital transactions increases the likelihood of being targeted by cyber-attacks. Baby Boomers, as members of the late majority on the technology adoption curve, are slower to adopt the latest cybersecurity practices. Despite their lower overall digital engagement, their growing presence in online spaces exposes them to risks. Generation Z is highly familiar with technology and often early adopters of new digital tools. However, their comfort with technology can lead to complacency, making them vulnerable to sophisticated scams, particularly through social media and peer-to-peer payment platforms. The Silent Generation has the lowest number of fraud cases, largely due

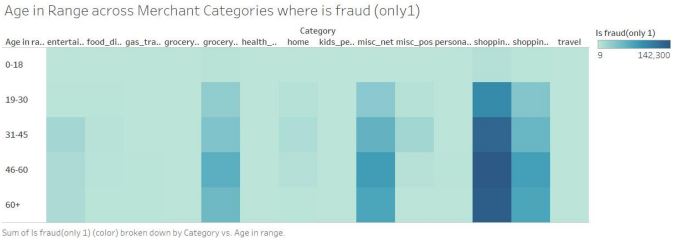

Fig. 20. Age in Range Across Merchant Categories where is fraud (only 1)

to their avoidance of extensive digital interaction, relying more on traditional banking methods like in-person banking and check writing.

L. Exhibit 7.1- Age in Range Across Merchant Categories where is fraud (only 1)

Interpretation- In the dataset, we grouped individuals of different ages into specific ranges by applying a formula in the Calculated Field in Tableau. We then created a visualisation displaying age ranges, shopping categories, and fraud occurrences where fraud = 1. A heatmap was selected to emphasise the categories in which fraud occurred and the corresponding age ranges, with the colour intensity representing the number of fraud cases. We used this type of visualisation as heat map seems ideal for displaying the relationship between multiple variables, such as age ranges, shopping categories, and fraud occurrences.

This analysis revealed that the ”Shopping net” category was particularly susceptible to fraud, especially among individuals aged 19 to 60. The analysis of fraud occurrences based on age ranges and shopping categories revealed that the ”Shopping net” category was particularly vulnerable to fraud, especially among individuals aged 19 to 60. The heatmap visualisation clearly highlighted that this category stood out due to the significant number of fraud cases it represented.

Several factors could explain why ”Shopping net” emerged as a hotspot for fraud. First, online shopping has experienced exponential growth in recent years, particularly among the 19-60 (19-30, 31-45, 46-60) age group. This demographic, encompassing young adults to those in their middle years, is highly active in online transactions, relying on e-commerce for a wide range of goods and services. The convenience of online shopping has led to an increase in transactions, and with more people using the internet for purchases, the risk of encountering fraud has also risen. Fraudsters often target online shoppers because the digital environment offers multiple avenues for attack, such as phishing scams, data breaches, and fraudulent websites.

Younger individuals, especially those in the 19-30 range, may not have fully developed digital literacy or the necessary caution required for secure online transactions. As a result, they might fall prey to deceptive practices such as fake online stores or phishing attempts. On the other hand, individuals in the older segment of this range (45-60) may be targeted because they often engage in larger financial transactions

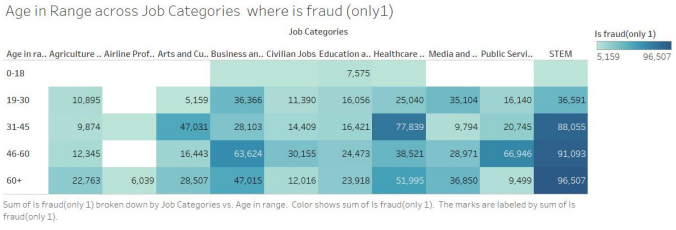

Fig. 21. Age in Range across Job Categories in States where is fraud (only1)

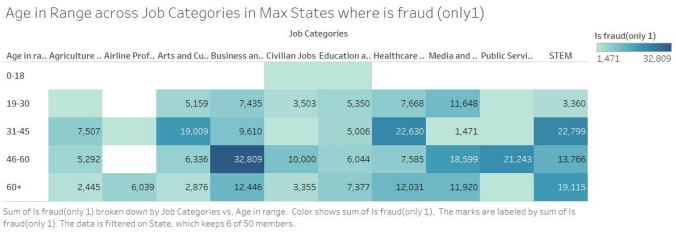

Fig. 22. Age in Range across Job Categories in States where is fraud (only1) in Max States

online, whether for investments, home-related purchases, or travel.

Moreover, the age group (31-45) and (45-60) tends to have a higher disposable income and spending power, making them attractive targets for fraudsters. This age group often makes frequent and high-value purchases online, whether for necessities, luxury items, or subscription services, increasing their exposure to potential fraud risks.

M. Exhibit 7.2- Age in Range across Job Categories in States where is fraud (only1)

Interpretation- We utilised a highlight table for visualisation, as it is effective for showing the intersection of two categorical variables (age ranges and job categories) and their associated fraud occurrences, with age ranges placed in the rows and job categories in the columns. The intensity of the colour represented the number of fraud occurrences (where is fraud = 1), while also displaying the actual count of fraud cases. The results revealed that the STEM category had the highest number of fraud occurrences in the 60+ age group, followed by the 45-60 and 31-65 age ranges. Other job categories with significant fraud occurrences included Healthcare and Medicine, Public Services, and Law.

Next, we refined the visualisation by filtering the dataset to focus on the six states with the highest fraud occurrences (CA, PA, MI, TX, NY, OH). We examined whether the same job categories showed the highest fraud rates. The observations revealed a shift: the Business and Finance category now had the highest fraud occurrences in the 46-60 age range, while STEM and Healthcare and Medicine led in the 31-45 age range. Additionally, Public Services and Law saw higher fraud occurrences in the 46-60 age group.

The STEM category, particularly among individuals aged 60 and above, showed the highest fraud occurrences. This could be due to the fact that older professionals in STEM fields often have long-established careers with substantial incomes and savings. As they approach retirement, they may be less cautious with online transactions or more prone to phishing attacks, especially if they aren't as familiar with modern cybersecurity threats. Additionally, despite their technical backgrounds, they may not be as proactive in keeping up with evolving online security measures, making them susceptible to fraud.

The Business and Finance category had the highest fraud occurrences in the (46-60) age range. This demographic is typically at the peak of their financial and career stability, making them attractive targets for fraud. Individuals in this age group often handle larger financial transactions, manage multiple accounts, and engage in more complex financial activities, creating more opportunities for fraudsters to exploit vulnerabilities.

For individuals aged (31-45), STEM and Healthcare and Medicine job categories showed high fraud occurrences. This age group is likely to be highly active online and may engage in frequent online transactions, making them regular targets for cybercriminals. Healthcare professionals in particular have been targeted in recent years, with personal data often at risk due to the sensitive nature of their work and the large amount of personal information they handle. Additionally, professionals in STEM might be targeted due to the high-value nature of their work, despite their technical expertise.

The (46-60) age group also saw significant fraud occurrences in Public Services and Law. People in these professions are often targets for scams involving identity theft, as they may handle sensitive data or have access to government systems. Additionally, their high levels of trust within society and the nature of their work could make them less suspicious of fraudulent activities, especially if these scams appear to be official or government-related. As these individuals tend to be in trusted positions, fraudsters may also leverage this for more sophisticated schemes that prey on their professional roles.

V. Challenges and Limitations

During the course of this analysis, several challenges and limitations were encountered that could have influenced the findings. First, the dataset used primarily relied on transactional data, which may not fully capture all variables contributing to credit card fraud. The real-time variables, such as changes in economic conditions or technological advancements, were not accounted for, although they can significantly affect fraud trends.The geographical scope of the analysis also presents limitations. While the study covered specific states, fraud patterns likely vary across different regions and smaller jurisdictions, which were not fully explored. Another source of potential bias in this analysis arises from the categorization of jobs. The grouping of occupations was based on our interpretation and understanding, which may not align with industry-recognized classifications therefore certain nuances and distinctions between job roles might have been overlooked or oversimplified.

The analytical methods employed, such as visualisations and correlation analyses, while effective for identifying trends, do not account for deeper causal relationships, the use of advanced modelling techniques, such as machine learning-based fraud prediction models, could offer a more robust analysis. Another important limitation lies in the potential bias in the external sources used for validation. While U.S. websites were consulted to corroborate the analysis, online resources vary in reliability and may not provide up-to-date or region-specific data.

Together, these limitations highlight the complexity of analysing credit card fraud and emphasise the need for more comprehensive data, localised research, and advanced analytical methods in future studies.

VI. Conclusion

This analysis offers critical insights into the geographical, demographic, and transactional factors that shape credit card fraud patterns in the United States. The effective outlier detection in credit card fraud revealed that geographic and economic factors, rather than population size alone, drive fraud patterns. Gender, age, job categories, and specific transaction amounts also significantly influence fraud occurrences, with certain groups and transaction patterns showing higher vulnerability. Results of analysis shows that fraud is predominantly concentrated in economically significant eastern and midwestern states, such as New York, Pennsylvania, Michigan and Ohio, where high volumes of financial activity create increased exposure to fraudulent transactions. Notably, the analysis reveals no direct correlation between population size and fraud occurrences, suggesting that maybe economic dynamics, financial transaction volumes, and local consumer behaviours are more influential in shaping fraud risk. Demographic analysis highlights gender-based differences in fraud vulnerability, with men being more frequently affected in financially active states like New York and Ohio, while women show higher fraud involvement in states such as Texas and Pennsylvania. Furthermore, individuals aged 31 to 60, particularly those in high-income fields like STEM and Business and Finance, are disproportionately targeted, likely due to their financial engagement and larger transactional volumes.

The data also identifies a specific fraud pattern within transaction ranges of 800–950, suggesting that fraudsters intentionally exploit this range to avoid detection by conventional security systems. Moreover, the "Shopping net" category emerges as a particularly high-risk area for fraud, reflecting the growing susceptibility of online shopping platforms to fraudulent activities, in nearly all age groups from 19 to 60.The results demonstrated a significant variation in fraud occurrences across the states of New York, Pennsylvania, Texas, and California, aligning with regional economic, demographic, and transaction volume factors. This supports the alternate hypothesis that certain states exhibit higher or lower proportions of fraud due to these regional factors. Thus, we reject the null hypothesis, confirming that there is indeed a meaningful difference in fraud distribution between the analysed states."

In summary, the findings suggest that understanding fraud involves not just examining transaction volumes and population statistics but also considering regional economic characteristics, demographic behaviours, and strategic fraud tactics. Financial institutions should enhance their fraud detection mechanisms by addressing these vulnerabilities. By doing so, they can more effectively mitigate the growing threat of credit card fraud.

VII. Acknowledgement

- **Ketki Bhatia and Niharika Suri (Dataset Preparation and Data Processing):** They collaboratively worked on cleaning, organising, and structuring the dataset to ensure consistency and accuracy. They handled missing values, corrected data inconsistencies, and pre-processed the dataset to make it suitable for analysis.
- **Akanksha (Calculated Fields in Tableau):** Akanksha developed and implemented calculated fields in Tableau to derive additional insights, such as generating age from date of birth and categorising transaction amounts. These fields provided essential metrics for the visualisations.
- **Akanksha and Ketki Bhatia (Visualisation Exhibit 1-5):** They created the initial set of visualisations (Exhibits 1-5), including bar charts, heatmaps. They visualised patterns such as fraud occurrence by state, gender-based fraud distribution, and transaction amount analysis.
- **Niharika Suri and Akanksha (Visualisation Exhibit 6-7):** They collaborated on Exhibits 6 and 7, which involved deeper visual analysis, focusing on specific states and gender-based fraud comparisons. They also employed specialised techniques such as state-level fraud mapping.
- **Niharika Suri (Research about notable characteristics and unique features of various U.S. states):** Niharika conducted research on the economic profiles, industries, and regional vulnerabilities of key U.S. states to provide context for the fraud trends observed. This research helped explain why certain states, like Texas and Pennsylvania, had high fraud occurrences.
- **Ketki Bhatia (Research about various generations and their characteristics):** Ketki performed research on generational behaviours, focusing on how different age groups engage with digital platforms and financial systems. This research was used to understand demographic factors in fraud vulnerability, particularly how older generations or younger tech-savvy individuals might be targeted differently by fraudsters.

References

[1] B. Cruz, "52 million Americans experienced credit card fraud last year," *Security.org*, Jul. 26, 2024. [Online]. Available: https://www.security.org/digital-safety/credit-card-fraud-report/: :text=60

[2] J. Egan, "Credit card fraud statistics," *Bankrate*, Jan. 12, 2023. [Online]. Available: https://www.bankrate.com/credit-cards/news/credit-card-fraud-statistics/.

[3] D. Carlin, "All 50 US states ranked by GDP [Report 2024]," *USA by Numbers*, Jan. 14, 2023. [Online]. Available: https://usabynumbers.com/states-ranked-by-gdp/.

[4] USAFacts, "How does gross domestic product differ by state?," *USAFacts*, Dec. 05, 2023. [Online]. Available: https://usafacts.org/articles/how-does-gdp-differ-by-state/.

[5] GlobalData, "Most populated states in the United States in 2021," [Online]. Available: https://www.globaldata.com/data-insights/macroeconomic/most-populated-states-in-the-us/.

[6] Gigafact, "Fact Brief: Is Nevada's economy heavily dependent on tourism?" Nov. 10, 2020. [Online]. Available: https://gigafact.org/fact-briefs/nevadas-economy-heavily-dependent-tourism.

[7] M. M. Miller and D. Lynch, "Alaska — History, flag, Maps, weather, cities, and Facts," *Encyclopedia Britannica*, Sep. 13, 2024. [Online]. Available: https://www.britannica.com/place/Alaska.

[8] H. J. Critchfield, E. Clark, A. Augustyn, and G. L. McNamee, "Washington — State Capital, Map, History, cities, and Facts," *Encyclopedia Britannica*, Jul. 26, 1999. [Online]. Available: https://www.britannica.com/place/Washington-state.

[9] G. America, "Nevada ranked as having highest rate of financial fraud per capita," [Online]. Available: https://gamingamerica.com/news/9247/nevada-ranked-as-having-highest-rate-of-financial-fraud-per-capita.

[10] Oberlo, "Top US payment methods (2023–2027)," [Online]. Available: https://www.oberlo.com/statistics/top-us-payment-methods.

[11] J. Hallinen, "STEM — Description, Development, and Facts," *Encyclopedia Britannica*, Sep. 06, 2024. [Online]. Available: https://www.britannica.com/topic/STEM-education/STEM-educationref330962.

[12] R. A. Wooster, D. C. Reddick, and G. L. McNamee, "Texas — Map, population, History, and Facts," *Encyclopedia Britannica*, Sep. 14, 2024. [Online]. Available: https://www.britannica.com/place/Texas-state.

[13] C. L. Thompson and E. W. Miller, "Pennsylvania — Capital, Population, Map, Flag, Facts, and History," *Encyclopedia Britannica*, Sep. 14, 2024. [Online]. Available: https://www.britannica.com/place/Pennsylvania-state.

[14] N. Morgan and G. L. McNamee, "California — Flag, Facts, Maps, Capital, Cities, and Destinations," *Encyclopedia Britannica*, Sep. 14, 2024. [Online]. Available: https://www.britannica.com/place/California-state.

[15] P. J. Scudiere and A. K. Campbell, "New York — Capital, map, population, history, and facts," *Encyclopedia Britannica*, Sep. 13, 2024. [Online]. Available: https://www.britannica.com/place/New-York-state.

[16] F. R. Aumann, G. W. Knepper, and J. Wallenfeldt, "Ohio — History, capital, population, map, and Facts," *Encyclopedia Britannica*, Sep. 13, 2024. [Online]. Available: https://www.britannica.com/place/Ohio-state.

[17] R. J. Hathaway, S. Glazer, and R. J. Schaetzl, "Michigan — Capital, Map, Population, History, and Facts," *Encyclopedia Britannica*, Sep. 14, 2024. [Online]. Available: https://www.britannica.com/place/Michigan.

[18] A. Zelazko, "Millennial — Definition, characteristics, age range, and birth Years," *Encyclopedia Britannica*, Sep. 04, 2024. [Online]. Available: https://www.britannica.com/topic/millennialref356992.

[19] A. McKenna, "Generation X — Origin, Years, Characteristics, and Facts," *Encyclopedia Britannica*, Sep. 04, 2024. [Online]. Available: https://www.britannica.com/topic/Generation-Xref356273.

[20] P. Bump, "Baby boomer — Definition, Age Range, and Societal and Economic Impact," *Encyclopedia Britannica*, Sep. 13, 2024. [Online]. Available: https://www.britannica.com/topic/baby-boomers.

[21] A. Eldridge, "Gen Z — Years, Age Range, Meaning, and Characteristics," *Encyclopedia Britannica*, Sep. 13, 2024. [Online]. Available: https://www.britannica.com/topic/Generation-Z.

[22] J. Wallenfeldt, "Silent Generation — Years, characteristics, and name Meaning," *Encyclopedia Britannica*, Aug. 16, 2024. [Online]. Available: https://www.britannica.com/topic/Silent-Generation.

[23] The Editors of Encyclopaedia Britannica, "Eastern Seaboard — Map, Region, and Facts," *Encyclopedia Britannica*, Sep. 12, 2024. [Online]. Available: https://www.britannica.com/place/Eastern-Seaboard.

[24] The Editors of Encyclopaedia Britannica, "Midwest — History, States, Map, culture, and facts," *Encyclopedia Britannica*, Sep. 14, 2024. [Online]. Available: https://www.britannica.com/place/Midwest.

[25] S. Burga, "Why Gen Z is surprisingly susceptible to financial scams," *TIME*, Feb. 24, 2024. [Online]. Available: https://time.com/6802011/gen-z-financial-scams-fraud/.

[26] M. Michaels, "All 50 states ranked for identity theft and credit card fraud, from most at risk to the least," *Business Insider India*, Feb. 15, 2018. [Online]. Available: https://www.businessinsider.in/all-50-states-ranked-for-identity-theft-and-credit-card-fraud-from-most-at-risk-to-the-least/articleshow/62937494.cms.

[27] N. Campisi, "The 10 most scammed states in America," *Forbes Advisor*, Aug. 04, 2023. [Online]. Available: https://www.forbes.com/advisor/personal-finance/most-scammed-states/.

[28] B. Cruz, "52 million Americans experienced credit card fraud last year," *Security.org*, Jul. 26, 2024. [Online]. Available: https://www.security.org/digital-safety/credit-card-fraud-report/: :text=60

[29] J. Egan, "Credit card fraud statistics," *Bankrate*, Jan. 12, 2023. [Online]. Available: https://www.bankrate.com/credit-cards/news/credit-card-fraud-statistics/.

[30] MSN, "Ohio ranks among states most affected by internet crimes according to FBI," *MSN*, Sep. 14, 2024. [Online]. Available: https://www.msn.com/en-us/news/us/ohio-ranks-among-states-most-affected-by-internet-crimes-according-to-fbi/ar-AA1qoOBP.

[31] D. Scofield and T. Carloss, "Ohio listed no. 7 for number of victims of cybercrime country-wide in 2021," *News 5 Cleveland WEWS*, Apr. 12, 2022. [Online]. Available: https://www.news5cleveland.com/news/local-news/ohio-listed-no-7-for-number-of-victims-of-cybercrime-country-wide-in-2021.

[32] S. Taber, "Internet Crime Report - Michigan SBDC," *Michigan SBDC*, Mar. 15, 2023. [Online]. Available: https://michigansbdc.org/cybersecurity/internet-crime-report/.

[33] M. S. Magda, J. B. B. Trussell, and S. K. Stevens, "Philadelphia — History, map, population, and Facts," *Encyclopedia Britannica*, Sep. 13, 2024. [Online]. Available: https://www.britannica.com/place/Philadelphia.

[34] P. R. Duis and C. Schallhorn, "Chicago — History, population, map, and facts," *Encyclopedia Britannica*, Sep. 12, 2024. [Online]. Available: https://www.britannica.com/place/Chicago.

[35] J. S. Lemons, "Rhode Island — Map, population, history, beaches, and Facts," *Encyclopedia Britannica*, Sep. 15, 2024. [Online]. Available: https://www.britannica.com/place/Rhode-Island-state.

[36] B. A. Martin and G. L. McNamee, "Idaho — History, Economy, People, and Facts," *Encyclopedia Britannica*, Sep. 15, 2024. [Online]. Available: https://www.britannica.com/place/Idaho.

[37] C. Lamothe, "5 reasons women are more likely to be targeted for financial theft scams," *GOBankingRates*, Sep. 14, 2024. [Online]. Available: https://www.gobankingrates.com/money/financial-planning/reasons-women-are-more-likely-to-be-targeted-for-financial-theft-scams/.

[38] G. Lankevich, "New York City — Layout, map, economy, culture, facts, and History," *Encyclopedia Britannica*, Sep. 15, 2024. [Online]. Available: https://www.britannica.com/place/New-York-City.

[39] The Editors of Encyclopaedia Britannica, "The South — Definition, States, Map, and History," *Encyclopedia Britannica*, Sep. 15, 2024. [Online]. Available: https://www.britannica.com/place/the-South-region.

[40] G. L. McNamee, "Las Vegas — History, Layout, Population, Map, Economy, and Facts," *Encyclopedia Britannica*, Sep. 15, 2024. [Online]. Available: https://www.britannica.com/place/Las-Vegas-Nevada.

[41] D. Delić, "Are women at more risk of online scams? The latest statistics in 2022," *ProPrivacy.com*, Jul. 06, 2022. [Online]. Available: https://proprivacy.com/blog/women-and-online-scams-latest-statistics-2022.

[42] C. Rodriguez, "Credit Card Ownership Statistics and Facts – By Income, Credit Score, Education Level and More [2024 Data Study]," *UpgradedPoints.com*, Sep. 13, 2024. [Online]. Available: https://upgradedpoints.com/credit-cards/credit-card-ownership-statistics/.