IST652 - Scripting for Data Analysis Final Project: Financial Transactional Analysis and Fraud Detection

Alex Grimm – [agrimm02@syr.edu](mailto:agrimm02@syr.edu)

ChihHao Yuan – [cyuan13@syr.edu](mailto:cyuan13@syr.edu)

Siddhant Kasture – [sikastur@syr.edu](mailto:sikastur@syr.edu)

1. Introduction

Credit cards are a very big part of people's daily lives. They are used to pay for a wide range of products and services such as electricity bills and groceries. People often use credit cards to purchase larger items. For example, computers and televisions are typically more expensive items that are often purchased using a credit card. This type of payment method allows individuals to charge purchases to their card and have a structured payment plan to pay for these purchases. While this method of payment is very convenient for customers, it is important to understand that they are regulated by financial institutions and banks. These institutions issue, regulate, and monitor credit card purchases. Our analysis aims to explore the relationships surrounding credit cards and their purchasing history.

1. Data source:

(Link:<https://www.kaggle.com/datasets/computingvictor/transactions-fraud-datasets/data?select=transactions_data.csv> )

The datasets used in this project are obtained from Kaggle, representing a synthetic yet realistic financial environment. They simulate transaction behaviors, cardholder activities, and expense-related data from a banking institution throughout the 2010s.

Dataset Components:

# 1. Transaction Data (transactions\_data.csv)

* Contains detailed records of individual transactions, including amount, merchant, timestamp, and transaction type through the 2010s
* Total 13,305,915 transaction records
* 12 Columns:

|  |  |  |
| --- | --- | --- |
| **Column Name** | **Description** | **Data Type** |
| id | Unique transaction ID | int64 |
| date | Timestamp of the transaction | object |
| client\_id | Customer identifier | int64 |
| card\_id | Linked card identifier | int64 |
| amount | Transaction amount (currency string) | object |
| use\_chip | Whether chip was used (Yes/No) | object |
| merchant\_id | ID of the merchant involved | int64 |
| merchant\_city | City of the merchant | object |
| merchant\_state | State of the merchant | object |
| zip | Merchant zip code | float64 |
| mcc | Merchant Category Code | int64 |
| errors | Transaction error flags, if any | object |

# 2. Card Information (cards\_dat.csv)

* Credit and debit card details
* Includes card limits, types, and activation dates
* Links to customer accounts via card\_id

|  |  |  |
| --- | --- | --- |
| **Column Name** | **Description** | **Data Type** |
| card\_id | Unique identifier for each card | int64 |
| client\_id | ID linking to the cardholder | int64 |
| card\_brand | Brand of the card (e.g., Visa, Amex) | object |
| card\_type | Type of the card (e.g., credit, debit) | object |
| card\_number | Numeric card number | int64 |
| expires | Card expiration date | object |
| cvv | Card verification value | int64 |
| has\_chip | Indicates if card has chip (Yes/No) | object |
| num\_cards\_issued | Number of cards issued to the client | int64 |
| credit\_limit | Credit limit assigned to the card | object |
| acct\_open\_date | Account opening date | object |
| year\_pin\_last\_changed | Year when PIN was last changed | int64 |
| card\_on\_dark\_web | Flag for dark web exposure (Yes/No) | object |

# 3. Fraud Labels (train\_fraud\_labels.json)

* Binary classification labels for transactions
* Indicates fraudulent vs. legitimate transactions

# 4. User Data (users\_data)

* Demographic information about customers

1. Data Processing:

# Combining Datasets:

Three files were merged:

transactions\_data.csv was joined with train\_fraud\_labels.json using transaction\_id. The resulting dataframe was then merged with cards\_data.csv on card\_id. Then we have a combined dataframe (merged\_df) that combines transaction details, fraud labels, and card information for further analysis. This required two merges. First we merged the transaction data with the fraud labels. We then merged the combined dataframe with the card data. This gave us the final dataset that we would use for our analysis.

# Cleansing and Formatting:

The amount and credit\_limit columns contained currency symbols (e.g., $), which were removed, and the values were converted to float. The date column was converted to datetime format for time-based analysis. Binary categorical variables use\_chip, has\_chip, target, and card\_on\_dark\_web were label encoded (Yes = 1, No = 0). Finally, we extracted the month and year from the date column for monthly and yearly trend analysis.

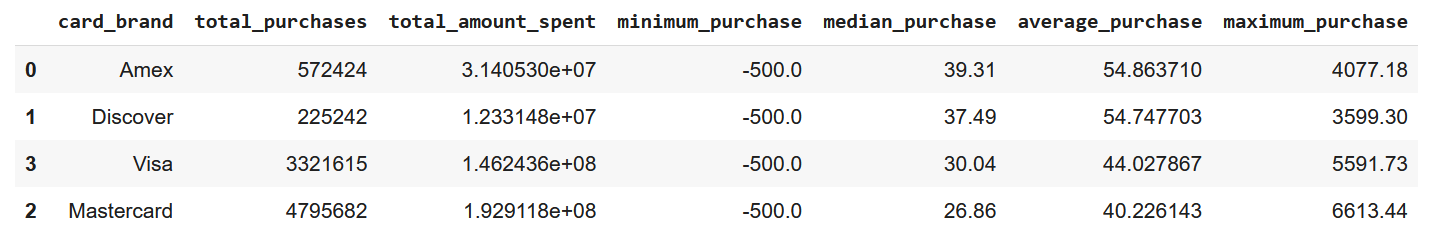
1. Describe your methods of analysis

# Research Question 1:

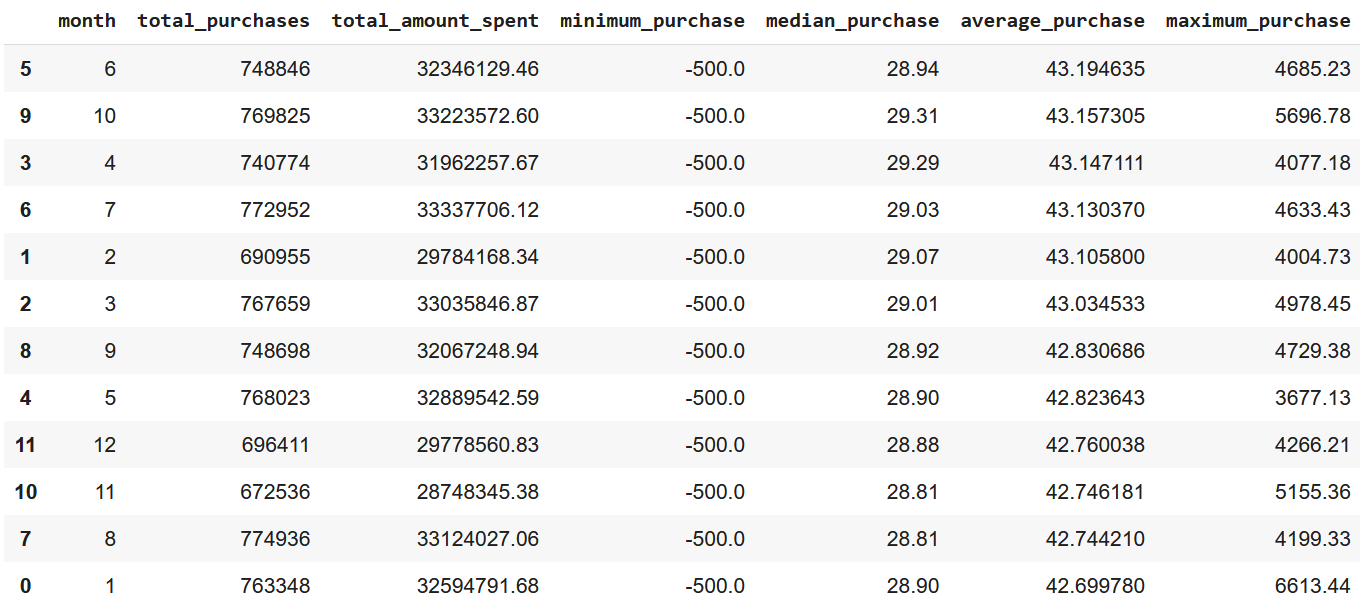
# Which card brands are used the most? Can we determine which card brands spend the most per purchase? Do these values change depending on the month of year?

To begin the analysis, we created a subset of the data. We included the card type, card brand, date and amount variables. We then had to convert the amount to a numeric value by dropping the dollar sign and converting the value to a float type. We then converted date to a datetime object and extracted the month to create a new column.

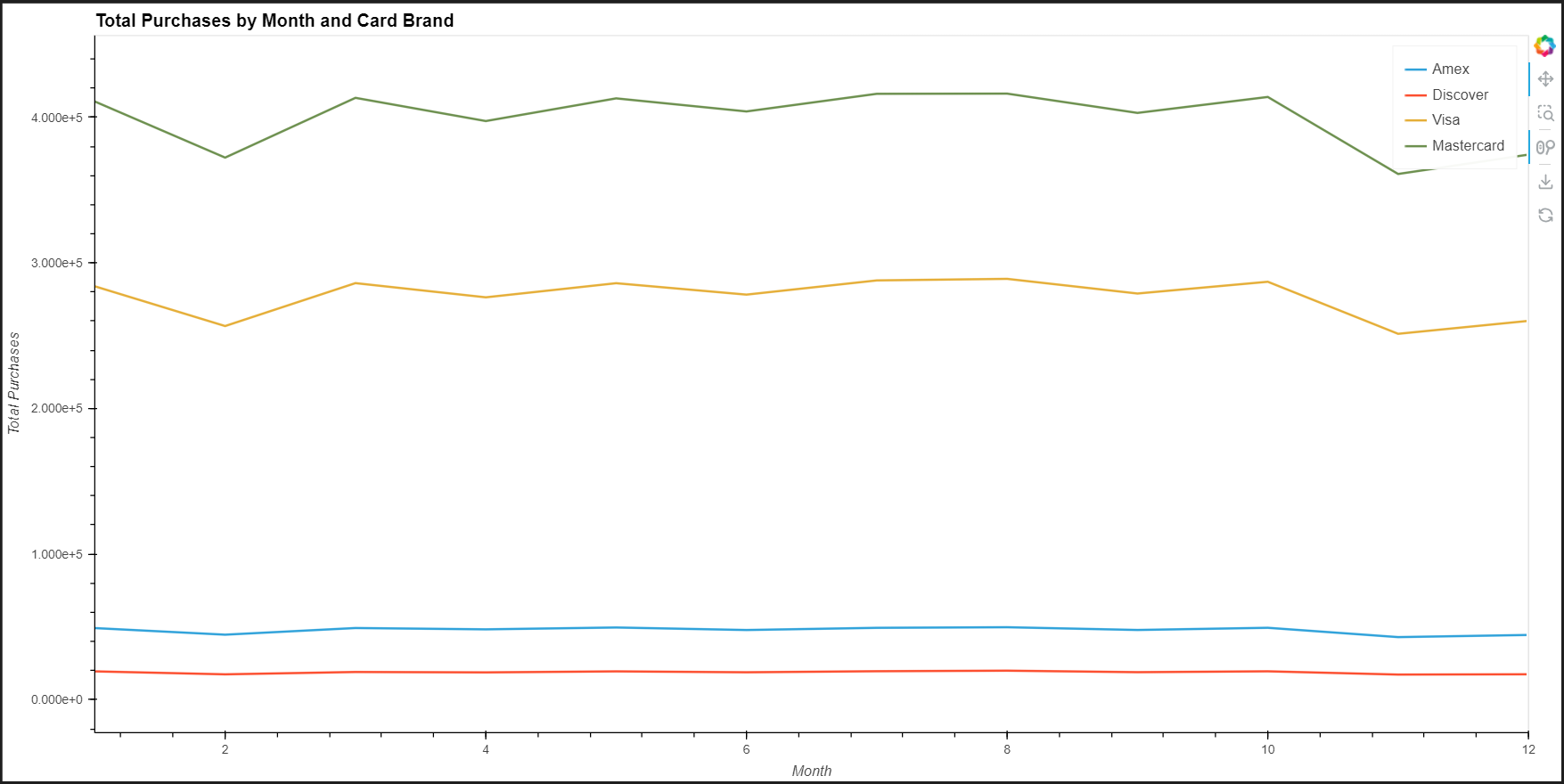
Now that the data is prepared we can move into the analysis. We first explore the transaction summary for the card brand. We group by card brand and create a new table that contains summary statistics regarding the transactions associated with each brand. To calculate these statistics, we use standard pandas functions like *.mean(), .median(),* and *.max()*.

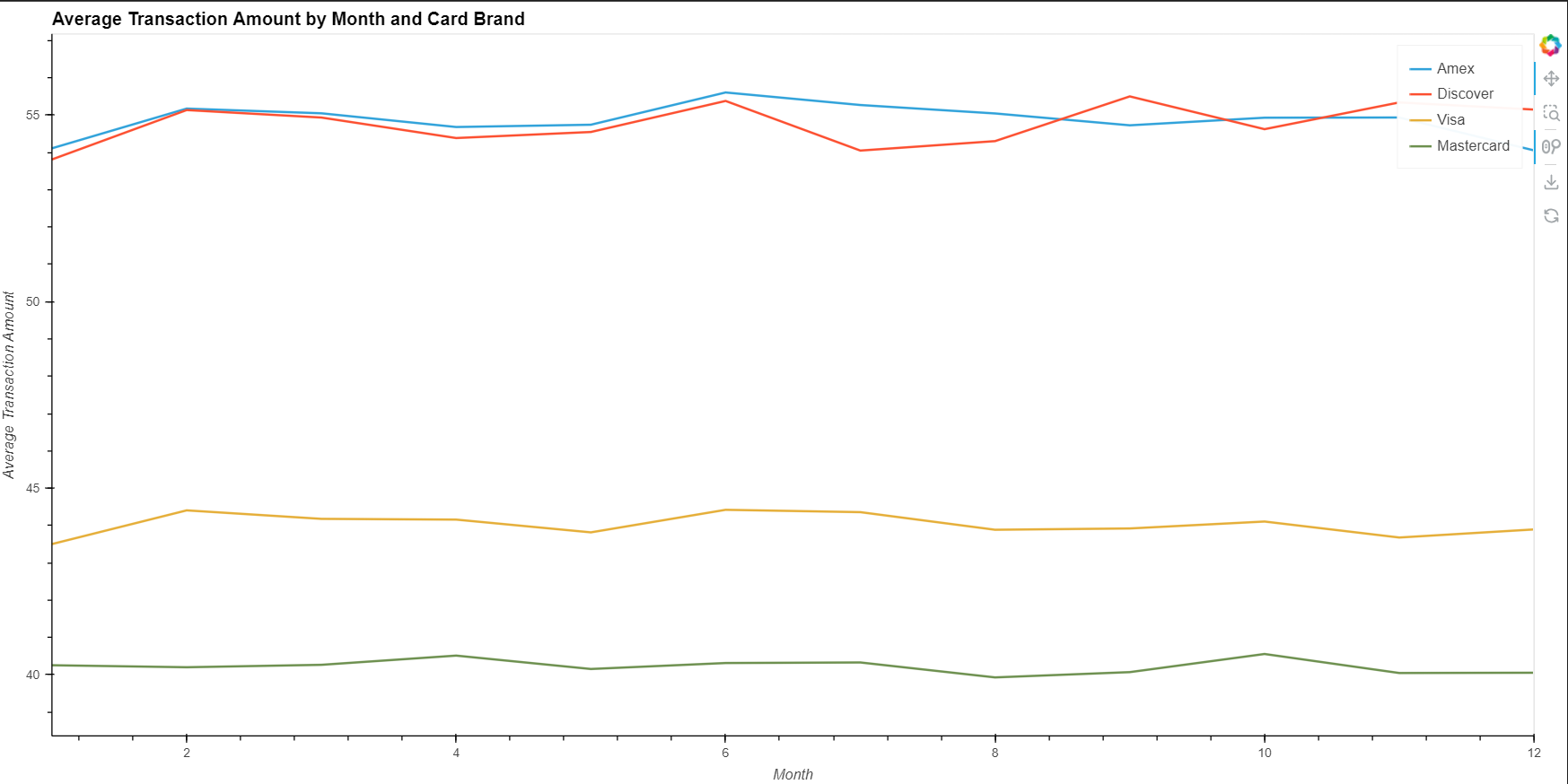


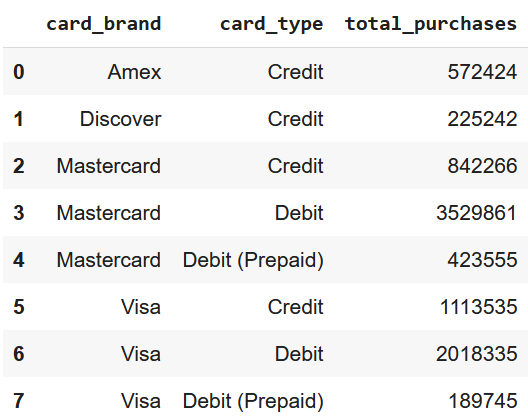
As you can see, American Express spends the most per transaction on average. We can also see that Visa Cards and Marstercards have significantly more transactions than the other card brands.

Next, we create a table that makes the same calculations above, but it is grouped by the month of the year instead of the card brand. We can see that on average the largest transactions occur in June and October. However it is worth noting that there is not a large variation in average transaction amount between all of the months. 

In the last table we create, we combine the first two tables. We group by the month of the year and card brand. We then calculate the summary statistics for every combination of month and card brand. The table is quite large and is hard to interpret, so we decided to create line plots to show the trends over the years.



The plot above shows the average number of purchases per card brand for every month of the year. Based on the plot, we can see that throughout the whole year, Visa and Mastercards have a noticeably larger number of purchases throughout the year than American Express and Discover cards

However, when we look at the next plot that shows the average transaction amount per month for each card brand, we see the opposite results. We can see that for almost every month, American Express cards and Discover Cards spend almost ten dollars on average more than Visa Cards and 15 dollars on average more than Mastercards. This caused us to wonder why this happens, so we created a table that shows the Card types that each Card Brand offers. 

We see that Mastercard and Visa provide Credit Cards, Debit Cards, and Prepaid Debit Cards. Since prepaid cards and Debit Cards can only be spent when there is money left on the card or in the account, I believe that people are less inclined to make more expensive purchases with these payment methods. However, it is worth noting that there are significantly more purchases made with these cards. This leads us to believe that people use them for more frequent, less expensive purchases. American Express and Discover cards only purchase restrictions are the credit limit associated with the account. We believe that this causes individuals to be more comfortable with spending larger amounts on purchases.

# Research Question 2: Analysis of the Relationship Between Credit Limit and Client Spending

# Do the clients who have higher credit limits tend to spend more?

## Data Used (Transaction Data, Card Information):

* client\_id: Identifies individual customers for grouping purposes.
* credit\_limit: The credit limit of each card.
* amount: The spending value of each transaction.
* card\_id: Used to link transactions to specific cards.
* date: Used to aggregate spending over time if needed.

## Analysis Process:

### Data Aggregation:

#### Total Spending of each client Calculation:

Using the merged dataset, transactions were grouped by client\_id, and the amount column was summed to obtain the total spending per client across the decade (2010s).

#### Average Credit Limit of each client Calculation:

Since many clients hold multiple credit cards, the data was also grouped by client\_id, and the average of the credit\_limit values

#### Combining Results:

The two resulting datasets were then combined using an inner join on the common key: client\_id.

#### Categorized Clients by Credit Limit:

I assigned clients to three credit limit groups — Low, Medium, and High.

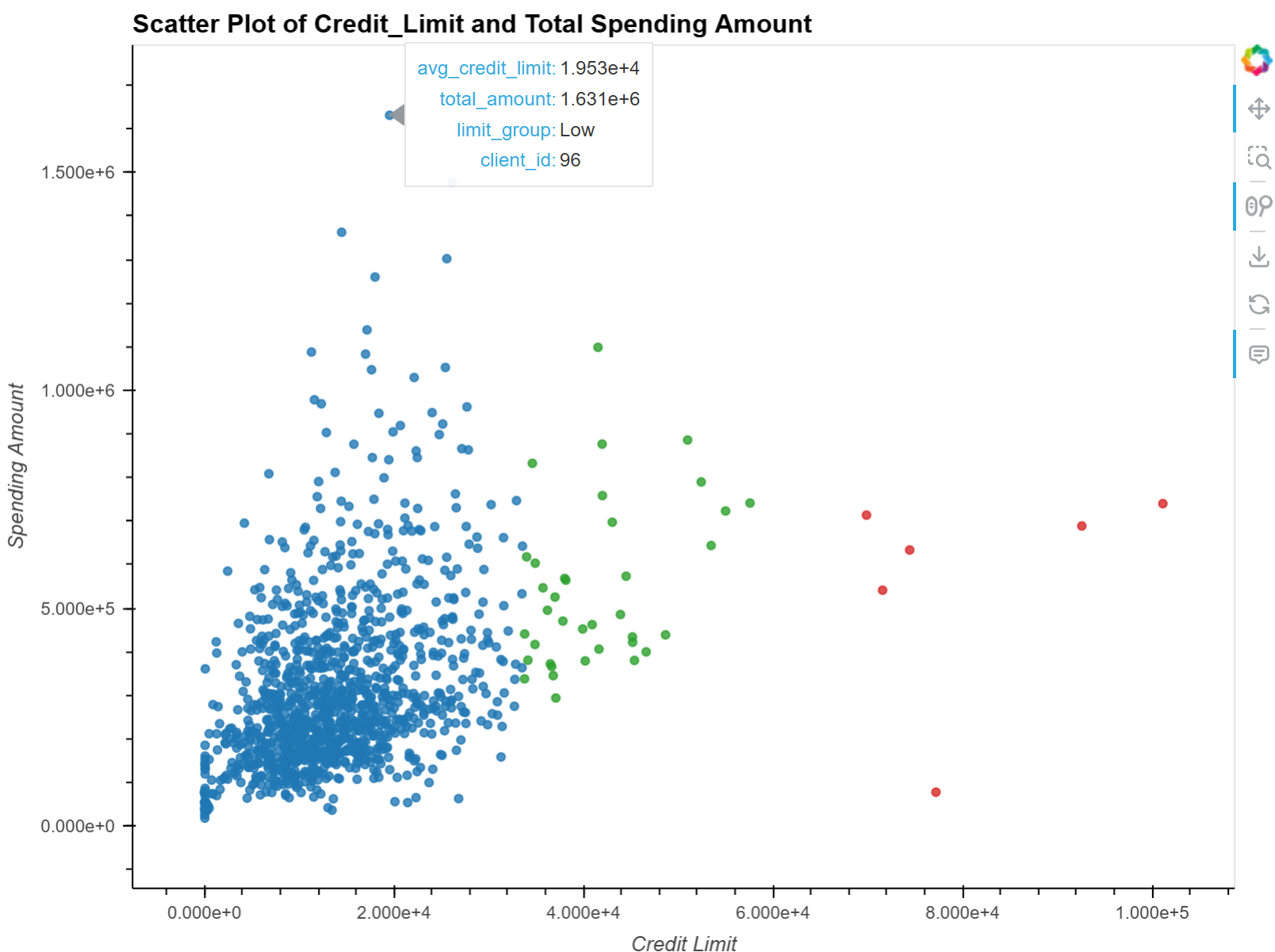
This grouping was done by dividing the avg\_credit\_limit into three equal-width ranges based on the numerical number itself, rather than using quartiles. Because, in real-world scenarios, the top-tier credit limit group typically contains fewer people, while most clients fall into the lower credit range.

This approach better reflects the actual distribution of credit limits and provides more realistic group segmentation for analysis.

### Analyze Spending Amount vs Credit Limit Relationship

To see the relationship between client spending and their credit limits, I used correlation analysis. The correlation coefficient between total spending amount and average credit limit is 0.4077, which indicates a moderate positive correlation. This means that as the credit limit increases, spending tends to increase as well.

To better illustrate this relationship, a scatter plot was used to visualize the distribution of clients across different credit limit and spending ranges. The plot reveals a general upward trend, with clusters of clients spending moderately within varying credit limit tiers.



Scatter Plot Analysis: Credit Limit vs. Total Spending

The scatter plot illustrates the relationship between each client's average credit limit and their total spending across the 2010s. The visual representation supports the earlier finding of a moderate positive correlation (r = 0.4077), suggesting that clients with higher credit limits tend to spend more.

Clients are color-coded based on their credit limit group:

* Blue: Low credit limit
* Green: Medium credit limit
* Red: High credit limit

As shown in the plot, the majority of low-limit clients (blue group) are mainly clustered in the lower-left region, indicating both lower credit and lower spending. In contrast, clients in the medium (green) and high (red) credit limit groups are generally positioned higher on the spending axis.

Although there is not a perfectly linear trend, the distribution of points broadly aligns along an upward diagonal direction, from the bottom-left (low limit, low spending) to the top-right (high limit, high spending). This pattern visually reinforces the correlation result and highlights the tendency for spending to increase with credit limits.

### Other Finding:

After creating the plot, we also noticed an outlier near the top of the spending axis, which means he/she spent the most in the decade among 1000+ clients but belonged to the low credit limit group (blue). This anomaly stands out from the general trend, as the client demonstrates an unusually high total spending despite having a low credit limit.

This caught our interest and led us to conduct a further investigation into the client’s payment and spending behaviors.

Putting ourselves in the shoes of a credit card company, this type of outlier could be highly significant. For instance:

* Does the client consistently pay on time, or are there signs of delayed or risky payment behavior, considering they are the top spender in the dataset?
* Or, could this be a valuable and loyal client who is worth additional promotions or a potential credit limit increase to strengthen their relationship with the company?

#### Background Check Based on User\_Data

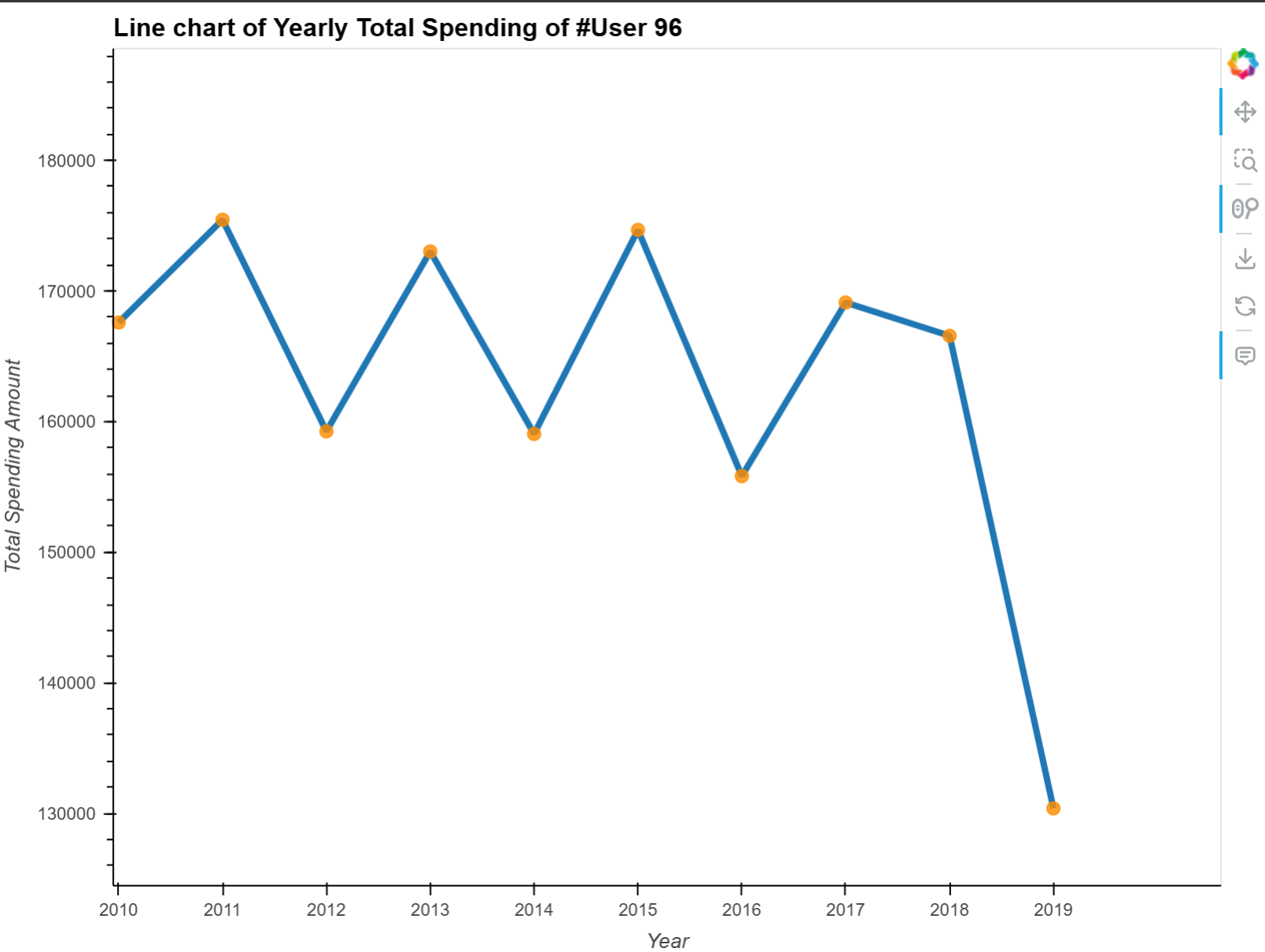
After identifying the outlier, we retrieved the corresponding client’s demographic and financial profile from the user dataset (user\_id = 96). Based on the data, we observed the following:

* She has a strong yearly income of $99,825, which indicates financial stability.
* Her total debt is relatively low, at $4,344, suggesting she is likely making timely payments.
* Her credit score is 685, which is considered a solid score within most credit scoring models.

These indicators point to a financially responsible and high-value client. As a result, we decided to further analyze her spending behavior to assess her potential for promotional offers.

#### Annual Spending Pattern Analysis (Client #96)

The results reveal that the client maintained a consistently high level of spending throughout the decade, with yearly totals generally ranging between $155,000 and $175,000. It was difficult to observe any clear trend or fluctuation from the raw numbers alone. To gain more straightforward insights, we created a line chart to visualize the client's yearly spending pattern. This graphical representation allowed for a clearer understanding of spending changes over time.



Upon visualizing the client's annual spending, we observed an interesting behaviors pattern:

She tends to spend more in odd-numbered years (e.g., 2011, 2013, 2015) and less in even-numbered years. Additionally, a steep drop in 2019 was noted, which we believe may be attributed to external events such as the COVID-19 pandemic.

This alternating pattern could reflect personal budgeting habits or life-cycle spending rhythms. If we were operating as a credit card company, this insight could inform targeted promotions, such as offering enhanced cashback or spending rewards during odd-numbered years when she is likely to be more active, since, compared to other cards, she might notice we have a better reward in her higher spending years. This strategy could help strengthen client loyalty and increase spending engagement.

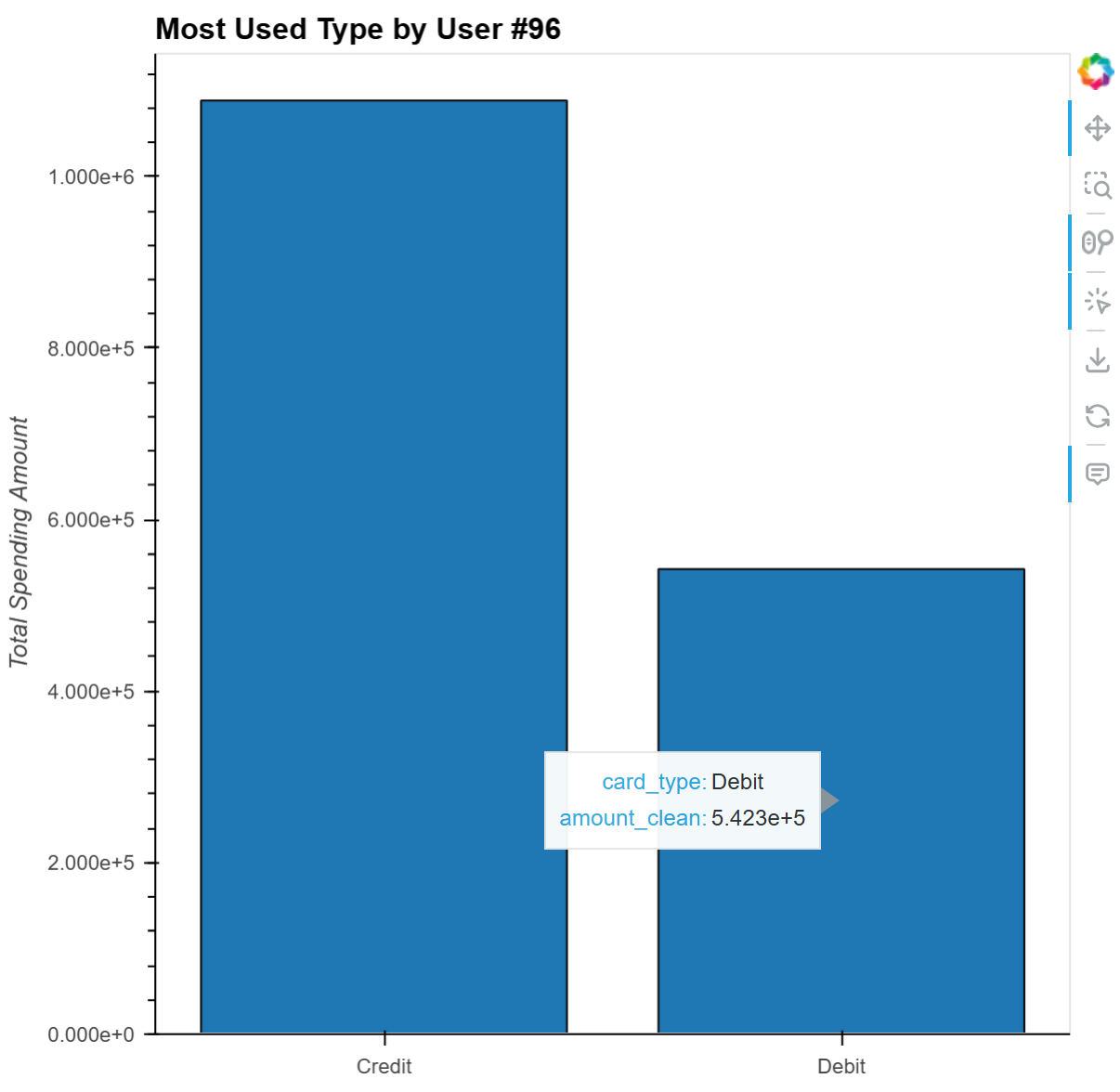
#### Card Type Spending Analysis (Client #96)

To further understand the client’s spending behavior, we analyzed her total spending by card type. Transactions associated with client\_id = 96 were grouped by the card\_type field, and the amount\_clean was summed for each group.

This breakdown allowed us to determine which type of card—credit, debit, or other—she relied on most throughout the decade.

Such insight is valuable from a business perspective, as it can inform which type of card promotions would be most appealing to this client.

To gain a clearer and more straightforward insight, I created a bar plot to visualize the client's total spending across different card types.



In this bar chart, we observe that while the client spent more using credit cards, she also maintained a significant amount of spending through debit cards.

From a business perspective, this presents an opportunity. If we were a credit card company, we could consider offering her a new credit card product or a targeted promotion, encouraging her to shift more of her spending from debit to credit.

This would not only provide her with greater rewards or benefits, but also increase card engagement and potential long-term loyalty.

# Research Question 3: Fraudulent transaction analysis using distributions, card-types & time-series data

**Section 1: Fraud Transactions and Distribution Across Lower Amounts**

**1. Objective**Understand how fraudulent transactions are distributed across different transaction amounts, with a particular focus on whether fraud is more prevalent at low or high monetary values. This insight can guide threshold-based monitoring and resource allocation for fraud detection.

**2. Data Preparation & Tools**

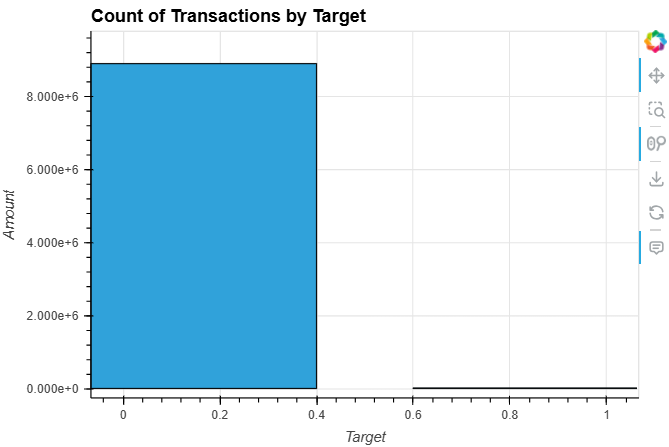
* **Dataset:** merged\_df, containing a binary target flag (0 = legitimate, 1 = fraudulent) and a cleaned transaction amount column, amount\_clean.
* **Visualization Libraries:**
  + HoloViews with the Bokeh backend for interactive bar charts.
  + Seaborn and Matplotlib for exploratory boxplots and summary statistics.

**3. Methodology**

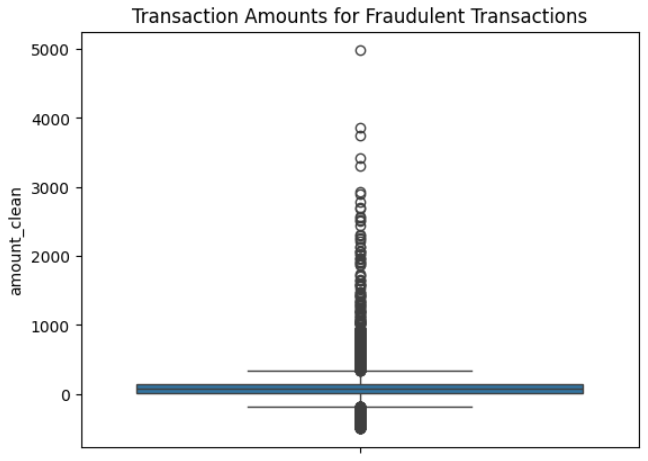
1. **Overall Class Imbalance:**
   * Count the number of legitimate vs. fraudulent transactions to gauge class imbalance.
   * Plot these counts using an interactive bar chart.
2. **Amount Distribution for Fraud Only:**
   * Filter merged\_df to retain only fraud cases (target == 1).
   * Use a Seaborn boxplot to visualize the spread of amount\_clean for these fraudulent transactions.
   * Compute and report descriptive statistics (count, mean, std, min, quartiles, max) on the fraud amounts.

**4. Results**

* **Class Imbalance**
  + The bar chart of transaction counts shows a stark imbalance: legitimate transactions vastly outnumber fraudulent ones. This “long tail” of fraud cases is typical in financial datasets and highlights the need for specialized techniques (e.g., resampling, anomaly detection) when modeling



* **Fraud Amount Distribution**
  + The boxplot of amount\_clean for fraud cases reveals that the bulk of fraudulent transactions cluster at the lower end of the monetary scale.



* + Descriptive statistics (e.g., median, interquartile range) confirm that most fraud amounts fall below a modest threshold, with only a few high-dollar outliers.

|  |  |
| --- | --- |
|  | **amount\_clean** |
| **count** | 13332.000000 |
| **mean** | 110.234682 |
| **std** | 213.736207 |
| **min** | -500.000000 |
| **25%** | 17.835000 |
| **50%** | 69.975000 |
| **75%** | 148.492500 |
| **max** | 4978.450000 |

**5. Key Observations**

* **Low-Value Targeting:** Fraudsters appear to favor smaller transactions, likely to minimize detection risk and blend in with routine customer activity.
* **Wide Range of Fraud Amounts:** While low amounts dominate, occasional larger frauds occur—these outliers may represent sophisticated schemes or automated high-value fraud attempts.
* **Implications for Detection:**
  + **Threshold Rules:** Consider implementing dynamic, lower-threshold alerts to catch high volumes of small-value fraud.
  + **Outlier Monitoring:** Supplement threshold rules with anomaly detection models to surface the less frequent high-value fraud cases.

**Section 2: Debit vs. Credit Cards—Fraud Analysis**

### 1. Objective

Compare the incidence of fraudulent transactions across card types—specifically Debit (including Prepaid) versus Credit cards—to pinpoint which card products are most at risk and tailor detection strategies accordingly.

### 2. Data & Tools

* **Dataset:** merged\_df with fields card\_type (e.g. “Debit”, “Debit (Prepaid)”, “Credit”) and binary target flag (0 = legitimate, 1 = fraud).
* **Libraries:**
  + Pandas for filtering and counting.
  + HoloViews + Bokeh for an interactive bar chart of fraud counts by card type.

### 3. Methodology

1. **Filter by Card Type**

**Extract Prepaid cards:  
  
prepaid\_df = merged\_df[merged\_df['card\_type'].str.contains('Prepaid')]**

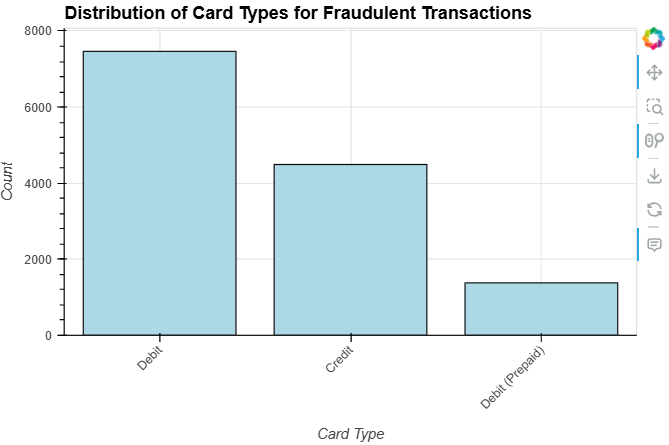
**Extract all Debit cards (including prepaid):  
  
debit\_df = merged\_df[merged\_df['card\_type'].str.contains('Debit')]**

1. **Count Fraud Cases**

**Count frauds in each subset:  
  
prepaid\_fraud\_count = prepaid\_df[prepaid\_df['target'] == 1].shape[0]**

**debit\_fraud\_count = debit\_df[debit\_df['target'] == 1].shape[0]**

1. **Visualize**
   * **Use HoloViews to plot the counts of fraudulent transactions by each card type (Debit, Credit, Debit (Prepaid)).**

****

### 4. Results

* **Raw Counts:**
  + Prepaid cards: 1,378 fraudulent transactions
  + Debit cards (total): 8,842 fraudulent transactions
  + Credit cards: (inferred from chart) ≈ 4,500 fraudulent transactions
* **Interactive Bar Chart:** A bar chart (Figure 2) clearly shows Debit cards account for the majority of fraud cases, followed by Credit cards, with Prepaid cards at the lowest level.

### 5. Key Observations

* **Debit Cards Are the Primary Target:** Debit cards (including non-prepaid) exhibit the highest absolute number of fraud cases—over twice as many as Credit cards.
* **Moderate Credit-Card Fraud:** Credit‐card fraud is non-negligible but significantly lower than debit. Fraud detection thresholds and monitoring rules should differ between these products.
* **Lower Incidence on Prepaid:** Prepaid cards see the fewest fraud incidents, suggesting either lower usage or better built-in risk controls.

**Section 3: Fraud Analysis Trend Using Time-Series Data**

### 1. Objective

Track and visualize the evolution of fraudulent transactions over time to uncover temporal patterns, spikes, and possible seasonality. These insights will inform proactive fraud-prevention strategies aligned with high-risk periods.

### 2. Data & Tools

* **Dataset:** merged\_df with a date column (converted to datetime), and binary target flag (1 = fraud).
* **Libraries:**
  + **Pandas for date manipulation and aggregation.**
  + **HoloViews + Bokeh for dynamic, interactive time-series plots.**
  + **Panel for embedding a year-selector slider.**

### 3. Methodology

1. **Preprocessing:**
   * **Convert merged\_df['date'] to datetime.**

**Filter to fraud cases:  
  
fraud\_df = merged\_df[merged\_df['target'] == 1].copy()**

1. **Aggregation:**

**Group by calendar date and count fraud occurrences:  
  
fraud\_trend = (**

**fraud\_df**

**.groupby(fraud\_df['date'].dt.date)**

**.size()**

**.reset\_index(name='count')**

**)**

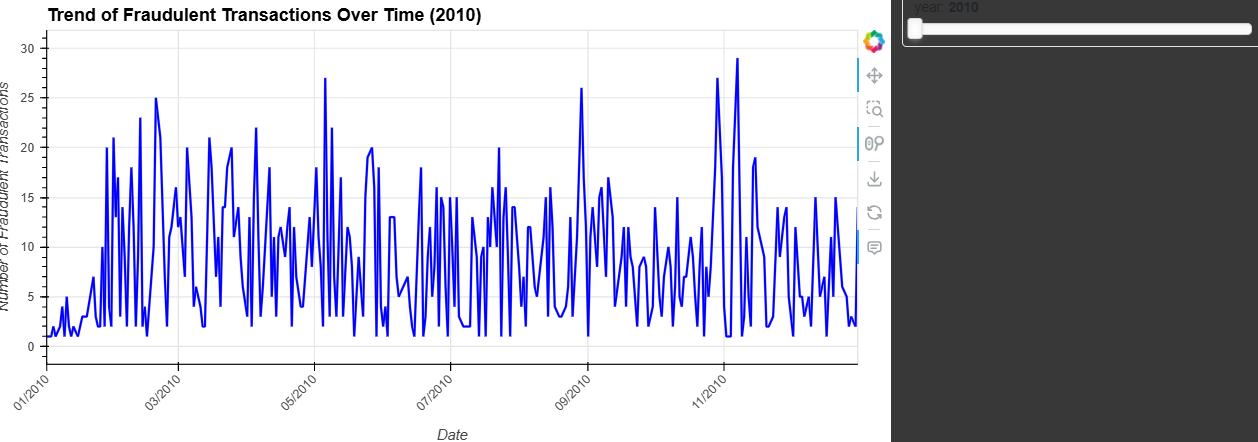
**fraud\_trend['date'] = pd.to\_datetime(fraud\_trend['date'])**

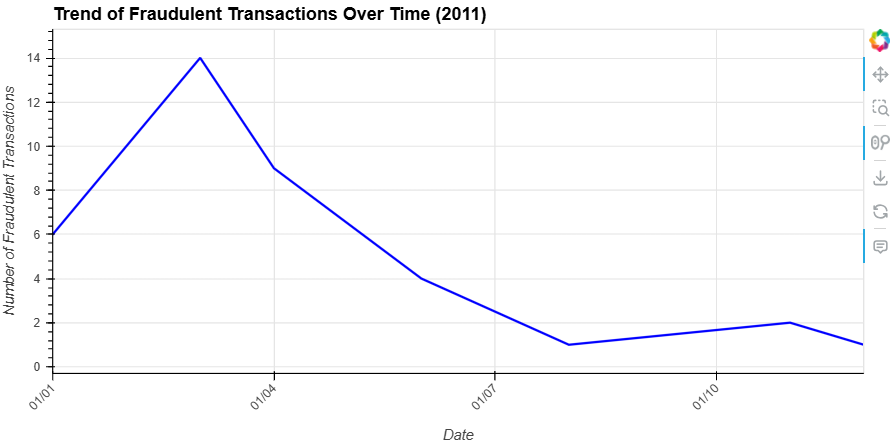
**fraud\_trend['year'] = fraud\_trend['date'].dt.year**

1. **Dynamic Visualization:**
   * Build a HoloViews DynamicMap that accepts a year parameter to filter fraud\_trend and render an interactive line plot (hv.Curve) of daily fraud counts for that year.
   * Expose the list of available years via a Panel slider for user-driven exploration.

**4. Results**

**Intermittent Spikes:**

* + 2010 exhibits pronounced volatility, with sharp daily peaks—indicative of isolated fraud campaigns or detection-system lapses.  
    
* 2011’s activity peaks early (around March) and then tapers off, suggesting possible seasonal drivers or enhanced mid-year controls.

****

**Seasonality Hints:**

* Early-year (Q1) surges appear consistently across multiple years (e.g., 2010 & 2011).
* Late-year holiday periods (Q4), while less pronounced in 2011, showed activity spikes in other years (as you noted for 2013–2014, 2016–2018).

Program Description

We first begin by loading in the transaction, fraud, and card dataframes. The transaction and card datasets are structured csv files, so we utilized pandas read\_csv to load them in. The fraud dataset was a semi-structured JSON file, so we had to load the json file in and convert it to a pandas dataframe. We then had to convert the id column to a float, so it would match the ID column in the transaction data.

We then used the pandas merge function to merge the transaction and fraud datasets together using an inner join on the transaction ID columns. We then merged the merged data frame together with the card data on card ID. We then dropped the repeated columns.

Moving on to the first research question, we first create a new dataframe with a subset of variables including card brand, card type, date, and amount. We then remove the $ from the amount column and convert the column to a float. Date was then converted to datetime and we created a new column with the month extracted from date. We then create three tables. Each table has calculated summary statistics based on the grouping. The first table is grouped by card brand, the second table is grouped by the month, the final table is grouped by month. We then created separate dataframes with the card brands filtered from the last table. Two line plots were then created using holoviews. The first shows lines for each card brand representing total purchases per month. The second shows lines for each card brand for the average transaction amount for each month. Finally, we create a table that shows the card types for each brand and the total number of transactions for each.

In research question 2, the first step was to ensure that the formats of all the columns we needed for analysis were correct. We used ["amount," "credit\_limit," and "date"]. The "amount" and "date" columns were already cleaned and formatted correctly. Therefore, we cleaned the "credit\_limit" column by removing the $ sign and converting it to a float. We also extracted the year from the "date" column for further analysis. Next, we created client\_spend\_df, which grouped the dataset by client\_ID and summed the amount. We also created client\_credit\_limit\_df, which grouped the dataset by client\_ID and calculated the average credit\_limit. Then, we merged these two datasets to calculate the correlation between spending and credit limit and to gain clearer insights. Additionally, we categorized each client into three credit limit groups — "Low," "Medium," and "High" — based on their average\_credit\_limit values directly, rather than using quartiles, to better reflect a real-world situation. After calculating the correlation, we used a scatter plot from Holoviews to visualize the relationship more clearly.

In the plot, we identified an outlier and decided to conduct a deeper investigation. We loaded the user\_dataset and focused specifically on user 96, the outlier. We extracted user 96's information from merged\_df into a new dataset, grouped it by date\_year, and summed the amount to see her yearly spending habits. We then used a line chart (Holoviews) to visualize the trend more clearly. Finally, we grouped user 96’s data by cardtype and summed the amount spent on each card type, using a bar plot to visualize which cards she used most frequently.

In the third part of our analysis, we focused on fraud labels that were derived from the semi-structured dataset provided. Bringing together the insights from all three analyses—amount distributions, card-type breakdown, and temporal trends—paints a cohesive picture of where and when fraud risk is most concentrated, and how detection strategies can be deployed most effectively. The majority of fraudulent transactions occurred at relatively small dollar amounts, suggesting that fraudsters often “fly under the radar” by keeping individual losses minimal. However, occasional large-ticket frauds were still observed, highlighting the need for complementary outlier-detection methods alongside standard threshold rules. When analyzing card types, we found that debit card**s** accounted for the majority of fraud incidents—more than double the count seen with credit cards. Credit cards exhibited moderate fraud volume but tended to involve higher average amounts per incident. Prepaid cards, meanwhile, showed the least number of fraud cases. These findings emphasize the importance of tailoring fraud detection techniques based on card type behavior. In terms of temporal patterns, the analysis across the years 2010 to 2020 revealed that fraud activity tends to spike during early-year (Q1) and year-end (Q4/holiday) periods, reflecting clear seasonal trends. Some years, notably 2010, displayed sharp daily peaks, indicating short-lived fraud campaigns or potential gaps in detection during those periods.

The analysis of fraudulent transactions across amount distributions, card types, and temporal trends offers a comprehensive understanding of fraud patterns. Fraudsters primarily target low-value transactions to evade detection, although occasional high-value frauds highlight the need for dynamic thresholding and outlier detection. Debit cards emerge as the most frequent medium for fraudulent activity, while credit card fraud, although less common, tends to involve larger sums per incident. Seasonally, fraud peaks during the early and late parts of the year, suggesting heightened vigilance is needed during these periods. Overall, these insights underline the necessity for adaptive, card-specific, and seasonally-aware fraud detection strategies to effectively minimize financial risk and loss.

Program Output

***purchase\_by\_brand\_df:*** groups data by card brand and calculates total transactions and summary statistics

***purchase\_by\_month\_df:*** groups data by month and calculates total transactions and summary statistics

***purchase\_by\_month\_brand\_df:*** groups data by month and card brand and calculates total transactions and summary statistics

***Card\_brand\_type\_df:*** groups data by card brand and card type and calculates total transactions

***client\_spend\_df:*** groups data by client\_id and sum the spending amount

***client\_credit\_limit\_df:*** groups data by client\_id and calculate the average credit\_limit

***cor\_spend\_credit\_limit\_df:*** merge ***client\_spend\_df*** and ***client\_credit\_limit\_df*** and assign client to different credit\_limit groups based on avg\_credit\_limit.

***user96\_yearly\_spending\_df:*** Extract #user96 information to a new dataset and group by "date\_year" and sum the "amount\_clean".

***user96\_cardtype\_df:*** Extract #user96 information and group by "card\_type" and sum the "amount\_clean".

Conclusions From Results

Based on all of our results from the analysis in research question one,, we see that Mastercards are used the most, while American Express cards have a slightly higher average transaction amount than the other card brands. When we look at these trends month by month, the patterns stay relatively consistent, so we can say that these trends do not change when we look at them by the time of year.

In research question 2, we found that credit limits have a moderate positive correlation with spending amount, as reflected by a correlation coefficient of 0.4077. This suggests that as the credit limit increases, spending tends to increase as well. To better understand this relationship, we created a scatter plot to visualize the distribution between clients' average credit limits and their total spending throughout the 2010s.

Focusing on the next iteration of fraud detection on small-ticket debit anomalies during seasonal peaks, backed by a lightweight anomaly-detection layer for larger-value outliers, will likely yield the biggest immediate lift in both catch-rate and false-positive reduction. This targeted approach balances efficient resource use with broad coverage of fraud tactics

# Group Member Contribution:

**Alex Grimm**

* loading and merging the datasets.
* Conducted the analysis related to Research Question 1, which examined card brand usage and monthly spending trends.

**ChihHao Luca Yuan**

* Responsible for Research Question 2, which investigated the relationship between credit limits and spending behavior, including the outlier client analysis.

**Siddhant Kasture**

* Conducted the analysis for Research Question 3, which explored fraud transaction patterns based on transaction amount, card type, and time.