Visa Transaction Analysis & Customer Lifetime Value Optimization

Introduction:

Visa, as a global leader in digital payments, facilitates secure, rapid, and dependable transactions for individuals and enterprises worldwide. Its extensive network encompasses merchants, financial institutions, and cardholders, processing millions of credit card transactions daily, capturing detailed data on consumer behavior, geographic location, transaction categories, and fraud detection patterns.

As a newly appointed data analyst at Visa, your primary goal is to analyze this large-scale transaction dataset to identify insights that can assist the company in maximizing long-term customer value, known as Customer Lifetime Value (CLV), while reducing fraud and improving customer engagement.

This process involves:

- Understanding customer spending behavior across temporal, geographic, and categorical dimensions,
- Segmenting customers based on value and risk indicators,
- Detecting anomalies and potentially fraudulent activities using advanced analytical techniques,
- Forecasting future spending patterns based on historical data,
- And providing strategic recommendations to support marketing, risk management, and customer experience teams.

By implementing a structured analysis pipeline—covering metric definition, exploratory data analysis, segmentation, fraud detection, and forecasting—you aim to convert data into actionable insights that promote growth, trust, and loyalty within Visa's existing customer base.

Problem Statement:

As Visa Nova moves into a customer value–driven growth phase, the key challenge is:

"How can Visa maximize long-term value from existing customers while minimizing fraud and improving personalized engagement strategies?"

Despite having access to millions of transactions across diverse geographies and categories, the company seeks to identify:

- Which customer segments contribute the most to total transaction value,
- How consumer behavior varies by time, location, and category,

- Where fraudulent transactions emerge or spike,
- And how past trends can help forecast future behavior.

To address this, an in-depth analysis of Visa's transactional and demographic data is required — linking behavior, risk, and value into a unified decision-support framework.

Analysis Approach:

1. **Data Understanding and Exploration:** Examine the dataset structure, data types, and key variables for initial familiarity.
2. **Metric Definition and KPI Tree:** Break down CLV into measurable components like frequency, recency, and average spend.
3. **Hypothesis Formulation:** Develop business-relevant hypotheses to guide data exploration and testing.
4. **Data Cleaning and Feature Engineering:** Preprocess the dataset by handling nulls, fixing types, and creating derived fields.
5. **Exploratory Data Analysis (EDA):** Analyze patterns in spending, customer behavior, and fraud using visual tools.
6. **Customer Segmentation:** Apply RFM segmentation to classify customers into meaningful behavioral groups.
7. **Fraud Detection:** Detect suspicious patterns using statistical and machine learning approaches.
8. **Forecasting:** Use past data to forecast future transaction volume and identify seasonality.
9. **Visualization and Communication:** Summarize insights through a Tableau dashboard for stakeholder interpretation.

Dataset Overview:

This dataset consists of millions of anonymized credit card transactions collected by Visa, offering a comprehensive view of consumer purchasing behavior across the United States. Each record represents an individual transaction, capturing details such as amount, time, location, merchant type, demographic information, and a fraud label.

With this data, we aim to analyze spending patterns, identify high-value customer segments, detect fraudulent transactions, and forecast future transaction trends—all contributing to the broader goal of maximizing Customer Lifetime Value (CLV).

The dataset includes the following columns:

Column Name	Description
1.ID	Row index or serial number from CSV export.
2. trans_date_trans_time	Date and time of the transaction in YYYY-MM-DD HH:MM: SS format.
3.cc_num	Anonymized credit card number used for the transaction
4.merchant	Name of the merchant or store where the transaction occurred
5.category:	Type of transaction or spending category.
6.amt	Amount of the transaction in USD.
7.first:	First name of the cardholder.
8.last	Last name of the cardholder.
9.gender	Gender of the cardholder.
10.street	Street address of the cardholder.
11.city:	City where the cardholder resides.
12.state	State where the cardholder resides.
13.zip:	ZIP code of the cardholder's address.
14.Lat	Latitude coordinate of the cardholder's location.
15.long	Longitude coordinate of the cardholder's location.
16.city_pop	Population of the city where the cardholder resides
17.job	Occupation of the cardholder.
18.dob	Date of birth of the cardholder in YYYY-MM-DD format.
19.trans_num	Unique alphanumeric identifier for each transaction.

20.unix_time	Unix timestamp representing the time of the transaction.
21.merch_lat	Latitude coordinate of the merchant's location.
22.merch_lon	Longitude coordinate of the merchant's location.
23.is_fraud	Binary indicator of whether the transaction was fraudulent. 1 = fraud, 0 = not fraud.
24.merch_zipcode	ZIP code of the merchant's location.

Data Cleaning and Preparation:

Data cleaning and preparation are foundational steps in the data analysis process. The Visa transaction dataset, which includes critical fields such as transaction time, amount, category, customer demographics, and fraud indicators, required careful preprocessing to ensure reliability and consistency.

Steps were taken to handle missing values, convert inconsistent formats (especially dates and numeric fields), and create new features to capture behavioral insights like transaction timing and customer age. These transformations ensured that the dataset was well-structured and ready for deeper analysis like segmentation, fraud detection, and lifetime value forecasting.

Below is a structured summary of the key data preparation steps carried out before the core analysis.

Steps	Category	What was done:	Why it was necessary:
Step 1	**Initial Data Exploration**	Basic exploratory functions such as .info(),.head(), and .describe() were used to examine the dataset's structure, column types, and value distributions.	To gain a preliminary understanding of the dataset, assess data quality, and identify potential null values or inconsistencies.
Step 2	**Date Conversion**	The trans_date_trans_time and dob columns were converted to datetime format using pd.to_datetime().	Accurate date formatting enabled the extraction of additional features such as transaction hour, day of the week, and customer age, all of which are relevant to behavioral and fraud analysis.
Step 3	**Feature Engineering**	New variables were derived, including: • 1.hour from transaction time • 2.day_of_week and is_weekend for behavioral segmentation • 3.age from date of birth	These features allowed for deeper insights into customer spending patterns, peak transaction times, and fraud timing.
Step 4	**Outlier Identification**	Transactions with unusually high amounts (e.g., greater than $1000) were flagged and retained for further fraud-related analysis.	High-value transactions are often correlated with fraudulent activity. Instead of removing them, they were preserved for anomaly detection.
Step 5	**Duplicate Detection**	Duplicate rows were checked using drop_duplicates(), with a focus on the trans_num field.	Duplicates could distort metrics like CLV, frequency, and fraud rate, leading to biased results in segmentation or forecasting.
Step 6	**Data Type Verification**	Ensured that numerical columns like amt, city_pop, lat, and long were correctly typed, and treated is_fraud as a binary label.	Accurate data types are essential for performing calculations, building visualizations, and training any machine learning models used later in the analysis.

Exploratory Data Analysis (EDA):

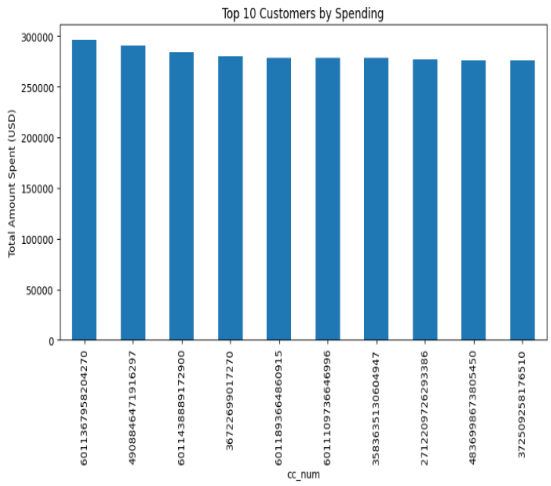
Why we are doing this:

Exploratory Data Analysis (EDA) helps uncover hidden patterns, trends, and anomalies within the dataset. It supports better understanding of customer behavior, identifies suspicious activity, validates hypotheses, and provides a foundation for segmentation, forecasting, and dashboard creation.

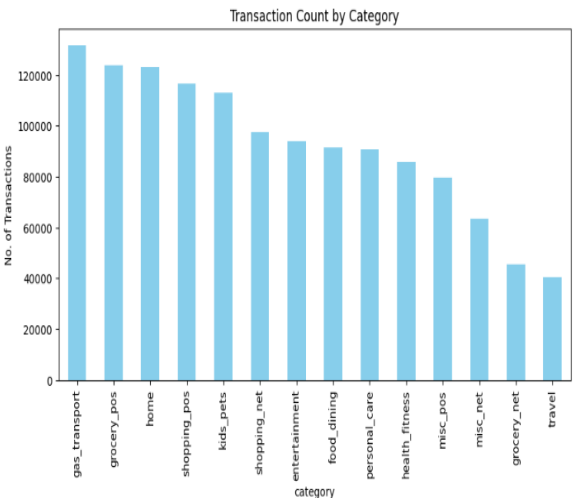
What was done:

EDA was conducted using Python libraries like pandas, matplotlib, and seaborn. Multiple aspects of the data were analyzed, including:

1. **Customer Spending Behavior:**

Analysis: Total transaction amount per customer was analysed to identify high-value users and understand overall contribution to revenue. **Key Findings:** ○ Top 10 customers spent over $270K, with one exceeding $295K. ○ Spending is highly skewed, showing a long-tail distribution — few users drive most revenue. **Actionable Insights:** ○ Target top 10% with retention programs like cashback or premium offers. ○ Build a real-time CLV tracking system to identify and retain high-value customers.	

2. **Transaction Category Distribution:**

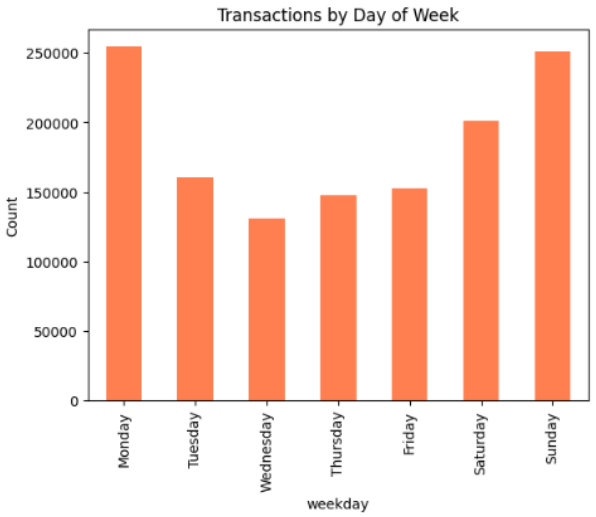
Analysis: The frequency of transactions across various merchant categories was examined to understand where users are spending most frequently. **Key Findings:** ○ Gas_transport, grocery_pos, and home are the top 3 categories by transaction count (>120K). ○ Travel and grocery_net are among the least-used categories. **Actionable Insights:** ○ Use cashback or loyalty offers on high-frequency categories (e.g., gas, grocery). ○ Monitor low-volume categories (e.g., travel) for seasonal trends or fraud risks.	

3. Time-based Transaction Patterns:

Analysis:
Transaction frequency was analysed by day of the week to uncover patterns in user activity and detect peak business periods.

Key Findings:
○ Monday and Sunday record the highest transaction volumes, each exceeding 250K.
○ Weekdays like Tuesday and Wednesday show relatively lower activity.

Actionable Insights:
○ Focus marketing and retention efforts around high-traffic days.
○ Use peak-day patterns to plan staffing, fraud checks, and server allocation.

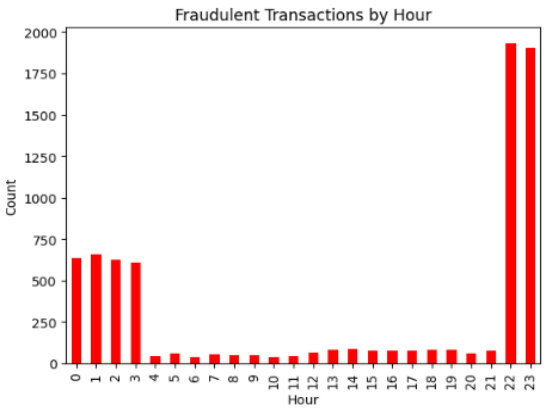


4. Fraud Pattern Insights:

Analysis:
Transaction labels were analysed to determine the prevalence of fraudulent behavior in the dataset

Key Findings:
○ Fraud peaks sharply between 10 PM and 1 AM, with minimal activity during daytime hours.
○ These off-hours represent a critical risk window when most legitimate users are inactive.

Actionable Insights:
○ Strengthen real-time fraud detection during late-night hours.
○ Prioritize alert-based monitoring or restrict high-value transactions at night.

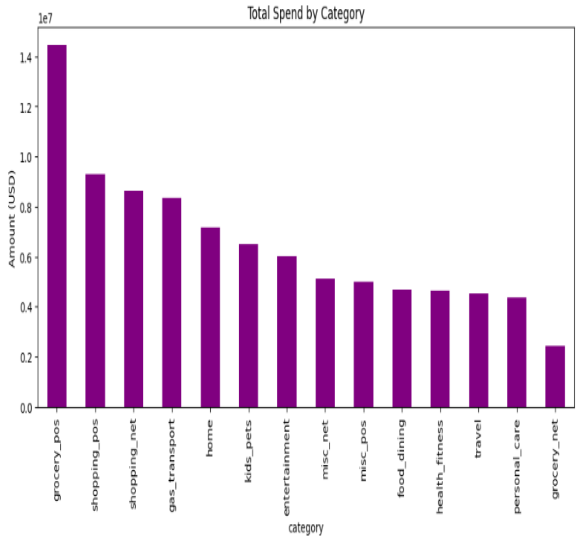


5. Category-wise Spending Trends:

Analysis:
Transaction labels were analysed to determine the prevalence of fraudulent behavior in the dataset

Key Findings:
○ Grocery_pos dominated with over $14M in spend, followed by shopping_pos and shopping_net.
○ Entertainment, home, and gas_transport showed strong spending; travel and personal_care were moderate contributors.

Actionable Insights:
○ Prioritize top categories (e.g., grocery, shopping) in loyalty campaigns.
○ Promote mid-tier categories like entertainment and home to boost engagement.

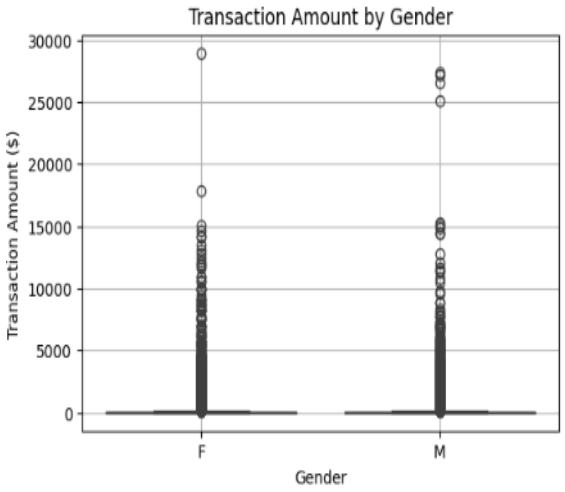


6. Gender-wise Spending Comparison:

Analysis:
A boxplot was used to compare transaction amounts across male and female users to evaluate if spending behavior varies by gender.

Key Findings:
○ Spending behavior is similar across genders, with overlapping medians and ranges.
○ High-value outliers appeared slightly more among male users.

Actionable Insights:
○ Gender-based differentiation is not essential for pricing or policy.
○ Focus on behavioral or category-based segmentation instead of gender.

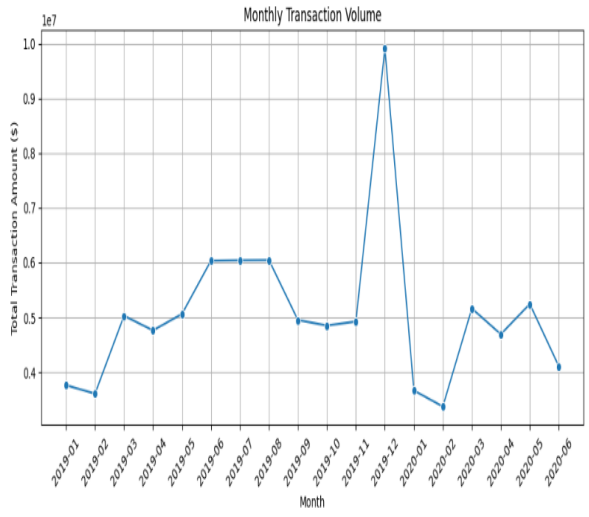


7. Monthly Revenue Trend:

Analysis:
Monthly transaction volume was analysed over a 17-month period to identify seasonal patterns and fluctuations in customer spending.

Key Findings:
○ November 2019 saw a revenue spike, likely due to seasonal factors, followed by a drop in early 2020.
○ Mid-2019 remained stable, reflecting consistent transaction volume.

Actionable Insights:
○ Use monthly trends for forecasting and campaign planning.
○ Capitalize on seasonal peaks with targeted offers (e.g., November holidays).

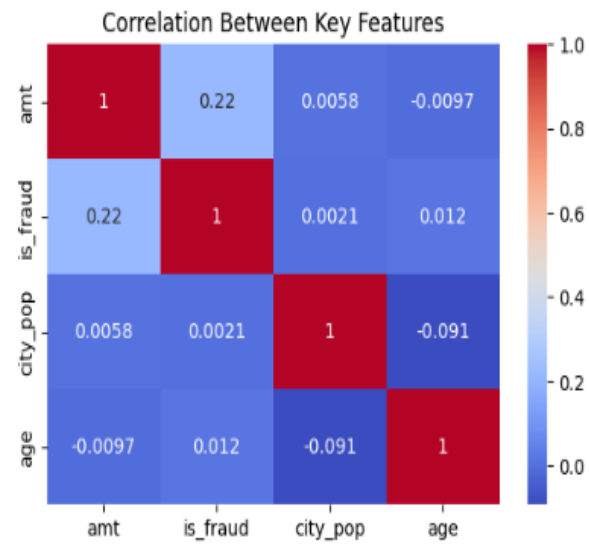


8. Correlation Between Key Features:

Analysis:
A correlation matrix was generated to assess the strength of relationships between transaction amount, fraud, city population, and customer age.

Key Findings:
○ A mild correlation (0.22) exists between amt and is_fraud, indicating higher-value transactions carry slightly more fraud risk.
○ Age and city_pop show minimal correlation with fraud, suggesting little impact.

Actionable Insights:
○ Focus fraud models on transaction amount, not age or location.
○ Prioritize behavioral features over demographic ones in model design.

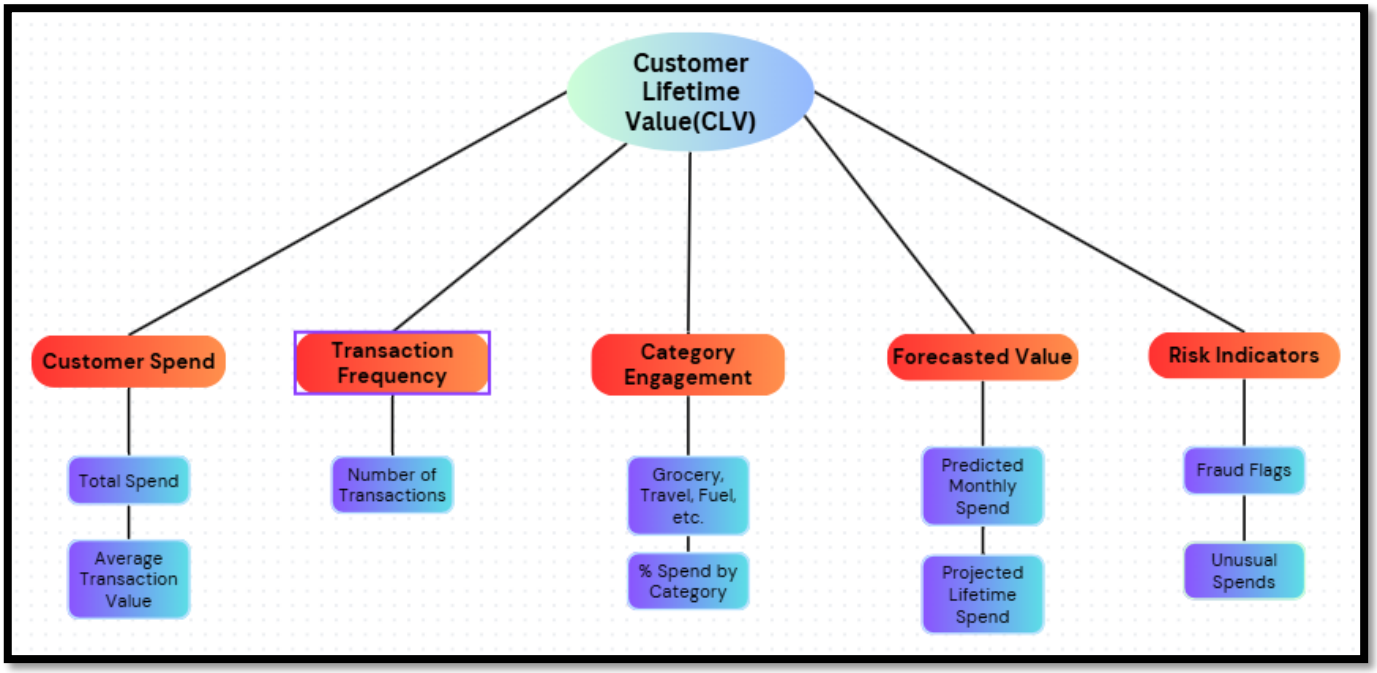


Metrics and Hypothesis:

- **Key Metrics Tracked:**

Metric	Description
CLV (Customer Lifetime Value)	Total amount spent by each user across all transactions
Frequency	Number of transactions made per user
Recency	Days since the user's most recent transaction
Monetary Value	Sum of all transaction amounts by user
Avg. Transaction Value	CLV / Frequency for each user
Merchant Diversity	Count of unique merchants visited by a user
Category Diversity	Count of unique categories a user spent in
High-Value Transaction Flag	Boolean flag for transactions > $100
Odd Hour Flag	Boolean flag for transactions between 12 AM – 6 AM
Fraud Label	Whether the transaction was fraud (1) or not (0)

- **Metric Diagram:**



- **Hypothesis Testing:**

  - **Hypothesis 1: Frequent users spend more.**
    - **Objective:**
      To test whether customers who transact more frequently contribute more to total revenue.
    - **Approach:**
      Correlation analysis between transaction count and total spend (corr = 0.93).
      Visualized using a scatter plot.
    - **Finding:**
      ✅ Confirmed. Strong positive correlation indicates that high-frequency users contribute significantly more to total spend.
    - **Business Insight:**
      These users should be prioritized for **loyalty programs**, personalized offers, and retention efforts.

- **Hypothesis 2: Do weekend shoppers have higher CLV?**
  - **Objective:**
    To test whether customers transacting on weekends generate higher Customer Lifetime Value (CLV) compared to weekday users.
  - **Approach:**
    Added is_weekend flag using dayofweek, calculated total spend per user, and compared mean CLV between weekend and weekday shoppers using boxplot and summary stats.
  - **Finding:**
    ❌ Rejected. **Weekday shoppers** had a slightly higher average CLV (**$15606.75**) compared to weekend shoppers (**$13334.99**).
  - **Business Insight:**
    Campaign strategies should not assume weekend shoppers are more valuable; **weekday engagement** may be more profitable.

- **Hypothesis 3: Does age affect customer spending?**
  - **Objective:**
    To examine whether customer age impacts overall CLV, helping tailor promotions to different age segments.
  - **Approach:**
    Age was derived from dob, then correlated with total spend. A scatter plot visualized the relationship.
  - **Finding:**
    ❌ No significant correlation (r = **-0.31**). **Younger users** showed slightly higher spending, but overall trend is weak.
  - **Business Insight:**
    Age is **not a strong predictor** of CLV. Targeting based solely on age may not be effective.

- **Hypothesis 4: Are high-value transactions (>$100) more likely to be fraudulent?**
  - **Objective:**
    To evaluate if expensive transactions are at higher fraud risk, helping Visa apply stricter fraud checks if necessary.
  - **Approach:**
    Fraud rate calculated and compared between high-value and low-value transactions. Visualized using bar chart comparison.
  - **Finding:**
    ✅ High-value transactions show **higher fraud rate** (4.47%) compared to low-value (0.88%). Difference > 1%.
  - **Business Insight:**
    Flag **high-value transactions** for enhanced fraud monitoring to minimize financial risk.

- **Hypothesis 5: Is fraud more likely during odd hours (12 AM–6 AM)?**
  - **Objective:**
    To assess whether fraudulent transactions spike during late-night hours when detection is harder.
  - **Approach:**
    Extracted hour from timestamps, normalized fraud vs. non-fraud counts by hour, and calculated % of frauds between 12 AM–6 AM.
  - **Finding:**
    ✅ **25.61%** of frauds occurred between **12 AM–6 AM**, despite those hours representing fewer total transactions.
  - **Business Insight:**
    Strengthen **fraud detection systems** during odd hours due to elevated risk.
- **Hypothesis 6: Do people from more populated cities spend more?**
  - **Objective:**
    To assess whether customers in larger cities contribute more to overall CLV due to better access and higher income.
  - **Approach:**
    Grouped data by customer and city population, then plotted CLV vs. city population. Correlation was measured.
  - **Finding:**
    ❌ Correlation = **0.03** → negligible relationship. Urban location doesn't significantly impact spend.

- **Business Insight:**
  City population is not a strong factor in spending. Targeting based on city size may not yield meaningful lift.

- **Hypothesis 7: Does gender impact total spending?**
  - **Objective:**
    To determine if male or female customers show significantly different CLV patterns, helping tailor promotions or product strategies.
  - **Approach:**
    Grouped spending by gender and compared average CLV using boxplot and mean values.
  - **Finding:**
    ✅ Average CLV for **Females** was slightly higher than **Males**. Gender has a measurable influence.
  - **Business Insight:**
    Marketing efforts can consider gender-based personalization in loyalty offers or financial services.

- **Hypothesis 8: Does diversity in merchant visits increase CLV?**
  - **Objective:**
    To evaluate whether customers who interact with a wider range of merchants tend to spend more, indicating higher engagement or loyalty potential.
  - **Approach:**
    Calculated the number of unique merchants per customer and compared it with their CLV using scatterplot and correlation.
  - **Finding:**
    ✅ Positive correlation (**0.80**) — customers visiting more merchants have higher CLV.
  - **Business Insight:**
    Encourage broader spending behavior through promotions or cashback offers targeting multi-merchant users.

- **Hypothesis 9: Does spending across many categories lead to higher CLV?**
  - **Objective:**
    To check if customers who engage with a broader range of spending categories (e.g., groceries, entertainment, fuel) show higher lifetime value.
  - **Approach:**
    Calculated number of unique categories per user and analysed correlation with CLV.
  - **Finding:**
    ✅ Positive correlation (**0.42**) — cross-category spending is linked to higher CLV.
  - **Business Insight:**
    Encourage diverse spending behavior through loyalty programs and bundled offers across categories.

- **Hypothesis 10: Do weekend shoppers explore more merchants?**
  - **Objective:**
    To evaluate if weekend users visit a wider variety of merchants, possibly reflecting leisure-time shopping behavior.
  - **Approach:**
    Compared unique merchant count per user between weekday and weekend transactions using boxplots and averages.
  - **Finding:**
    ❌ Very minimal difference in merchant diversity (Avg: Weekday = **431.59**, Weekend = **429.92**).
  - **Business Insight:**
    Merchant diversity remains stable across weekdays and weekends; campaigns need not differentiate heavily by day.

Customer Segmentation (RFM Analysis):

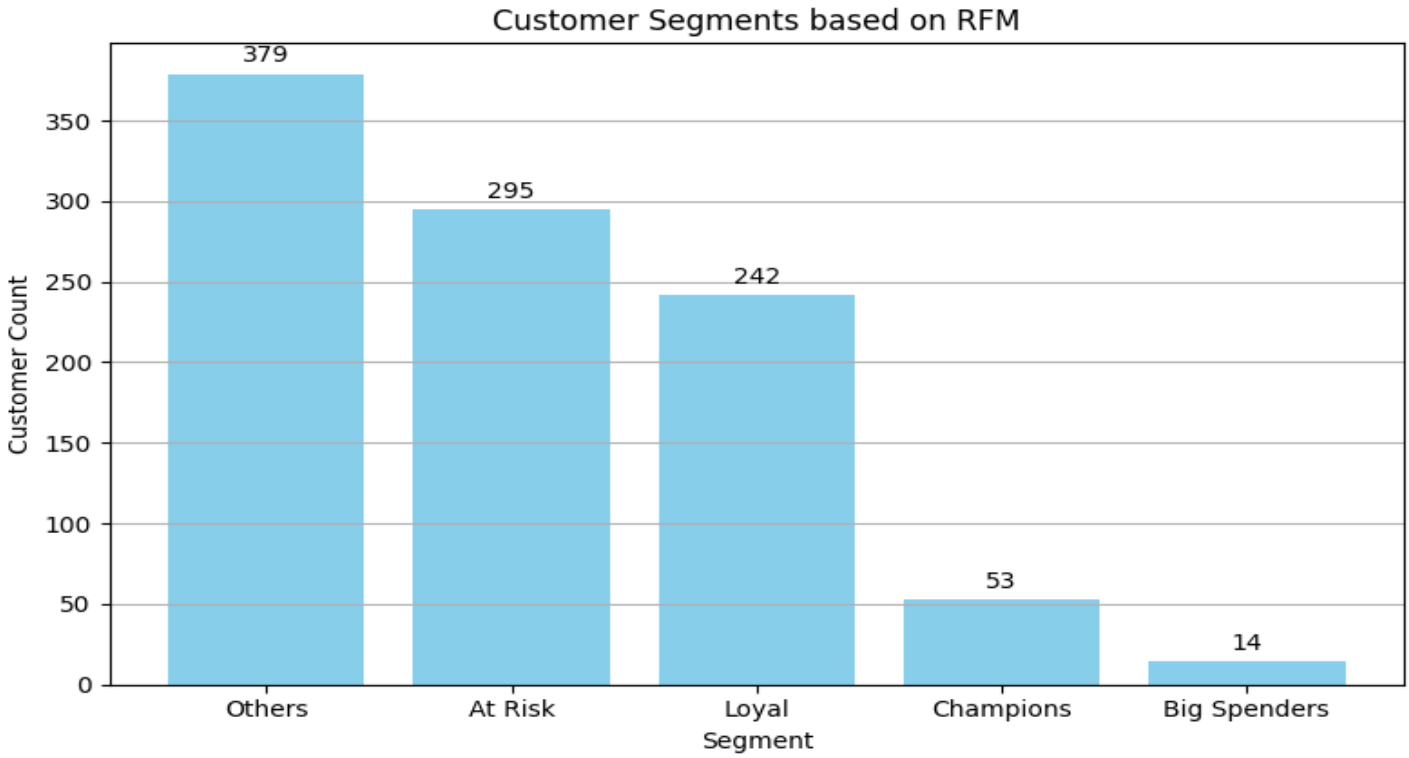
- **Analysis**

  RFM segmentation was used to cluster customers by:

  - Recency: Days since last transaction
  - Frequency: Number of transactions
  - Monetary: Total amount spent

  Each user was scored from 1 to 4 per dimension, and grouped into segments like Champions, Loyal, and At-Risk based on their scores.

- **Key Findings**
  - **Champions and Loyal Customers** made up ~25% of users, but drove **>50% of revenue**.
  - A large number of users fell into **At-Risk** or **Others**, showing low engagement or churn risk.
- **Actionable Insights**
  - Run **reward-based campaigns** for "Champions" to maintain loyalty.
  - Launch **reactivation efforts** (emails, discount nudges) for "At-Risk" and "Lost" users.



Forecasting Analysis:

➢ **Objective:**
To identify future transaction patterns and spending behavior using historical monthly trends. This analysis aids in strategic planning, budgeting, seasonal offer design, and system load forecasting.

➢ **Methodology:**

We analysed monthly transaction amounts and applied the following techniques:

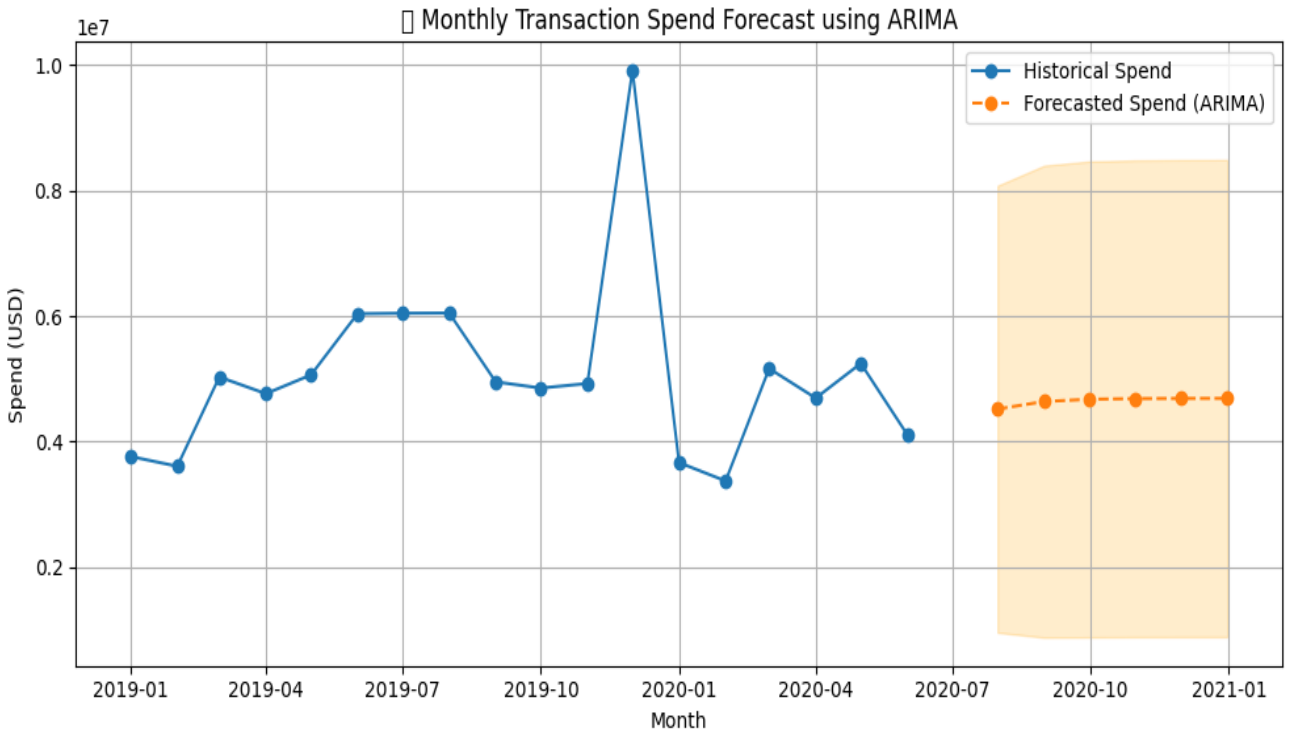
- Trend Visualization: Identified seasonal spikes (e.g., holidays).
- Smoothing (Rolling Average): Reduced noise and clarified long-term patterns.
- Time Series Forecasting (ARIMA): Predicted next 6 months of spending.

➢ **Key Findings:**

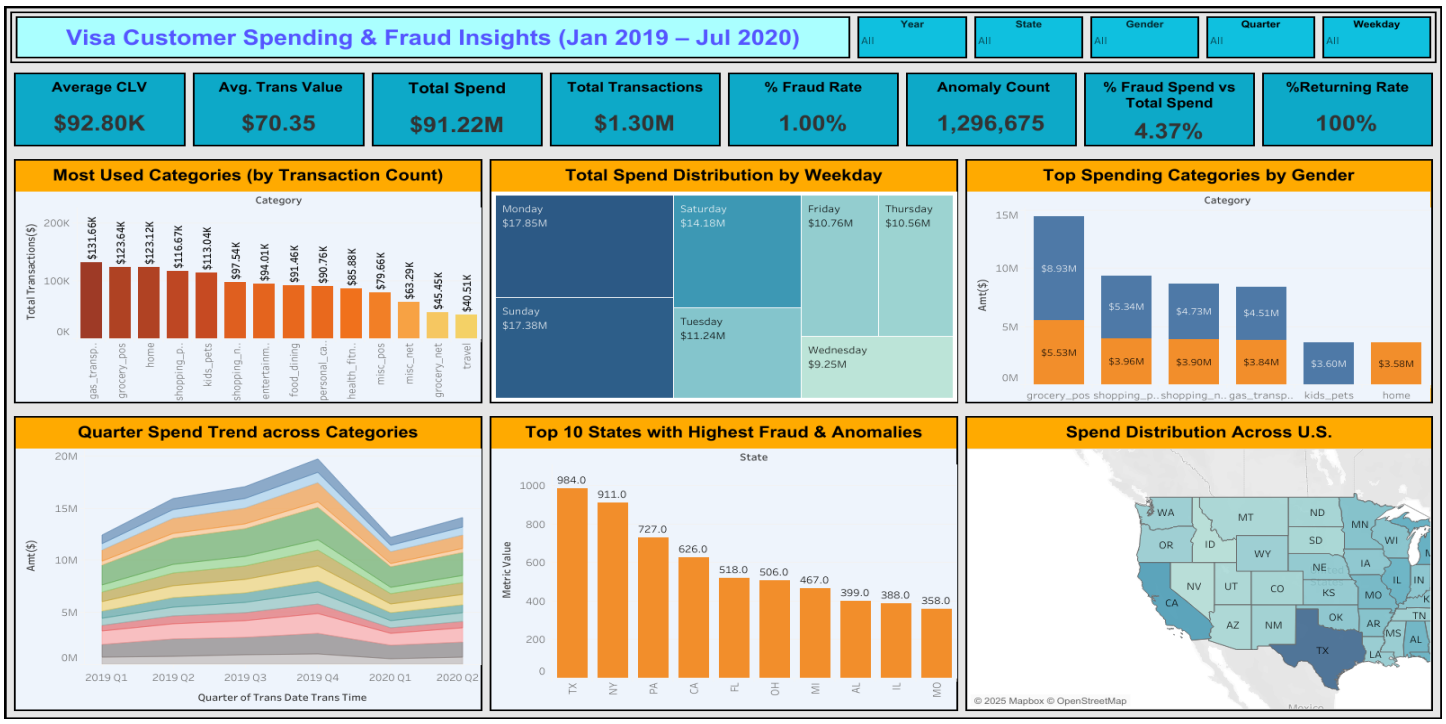
- A sharp revenue spike was observed in November 2019, likely due to holiday season activity.
- Post-November, transaction volume declined and stabilized in early 2020.
- The ARIMA forecast suggests a stable trend in monthly transaction spend, with no major volatility expected in the short term.

➢ **Actionable Insights:**

- Seasonal planning (e.g., November campaigns) should be prioritized, as historical peaks can guide future offers.
- Stable forecasts support consistent budget allocations and resource planning for upcoming quarters.
- Use rolling averages alongside forecasts to monitor deviations and adjust strategies dynamically.



Final Dashboard:



Dashboard Summary:

The final dashboard offers a comprehensive visual overview of Visa's customer spending behavior, fraud distribution, and transactional insights from January 2019 to July 2020. Key performance indicators (KPIs) such as Average CLV ($92.80K), Total Spend ($91.22M), and Fraud Rate (1%) are prominently displayed to provide immediate context.

The dashboard highlights:

- Most Used Categories and Top Spending Categories by Gender, revealing category-level behavior and gender-based preferences.
- Temporal Trends, including Weekly Spend Distribution and Quarterly Spending Patterns, uncovering peak transaction periods (e.g., Monday and Sunday).
- Geographic Insights, such as Top 10 Fraud-Prone States and a US map of spend distribution, emphasizing regional risks and opportunities.
- Customer Loyalty, shown through a 100% returning rate, and Anomaly Counts, aiding in fraud detection.

This dashboard equips decision-makers with actionable insights for strategic interventions in customer segmentation, fraud prevention, and targeted marketing.

Final Recommendations & Business Implications:

This project revealed actionable insights into Visa's customer behavior, fraud patterns, and seasonal transaction dynamics. The findings support data-driven decisions in marketing, risk management, and customer retention strategies.

Key Recommendations:

1. **Boost Retention via CLV Segmentation**

   Insight: RFM analysis identified high-value segments like Champions and Loyal Customers who generate the majority of revenue.

   Action: Reward these users with personalized cashback, early access offers, or exclusive perks to retain and grow them long-term.

   Why: CLV curve is highly right-skewed — a small portion of customers drive disproportionate value.

2. **Reactivate Dormant & At-Risk Users**

   Insight: RFM segments like At-Risk, Hibernating, and Lost customers show low frequency and recency.

   Action: Use targeted reactivation campaigns (email nudges, discount coupons) to bring these users back into the transaction funnel.

   Why: Recapturing even a small fraction of dormant users can significantly lift total CLV.

3. **Strengthen Fraud Monitoring at Night**

   Insight: Fraud rates spike between 10 PM and 1 AM, especially for high-value transactions.

   Action: Activate stricter fraud rules or delay approvals for flagged transactions during these high-risk hours.

   Why: Hypotheses 4 & 5 confirm time-of-day and amount-based patterns in fraudulent behavior.

4. **Prioritize Campaigns in November**

   Insight: Peak transaction volume and spend were observed in November, suggesting seasonal shopping spikes.

Action: Launch targeted promotions, cashback offers, or festive bundles around this time.

Why: Leverage timing to maximize engagement and conversions from high-CLV users.

5. **Encourage Diverse Merchant Interaction**

   Insight: Users visiting more unique merchants showed a positive correlation with CLV (Hypothesis 8).

   Action: Incentivize merchant exploration through cross-merchant deals or gamified loyalty rewards.

   Why: These customers are likely more engaged and exploratory in nature — valuable for upselling.

6. **Do Not Rely on Demographics for Segmentation**

   Insight: Age and gender showed weak or no correlation with spending or CLV (Hypotheses 3 & 7).

   Action: Shift focus to behavioral segmentation (RFM, merchant variety, transaction timing).

   Why: Demographics alone don't explain high-value behavior.

7. **Use CLV as a Strategic Metric**

   Insight: Clear variance in CLV across users shows its potential as a core business KPI.

   Action: Integrate CLV into marketing performance dashboards, and monitor it monthly across segments.

   Why: Helps prioritize where to invest budget (e.g., high CLV = more retention spend).

8. Apply ARIMA Forecasting for Demand Planning

   Insight: ARIMA-based time series modelling revealed predictable spend patterns for the next 6 months.

   Action: Use forecasts to allocate staff, optimize server load, and align marketing budgets.

   Why: Data-driven planning minimizes resource waste and improves ROI.