Exploratory Data Analysis Report: Example Key Point Interest

KeyStone Predictive

Produced by KeyStone Predictive
www.keystonepredictive.com

1 Introduction

This report shows exploratory data analysis for Example Business using simulated transaction data collected between January 1, 2023, and December 31, 2024. The dataset contains 1,000 transactions from 500 unique customers, with purchase amounts ranging from $5.02 to $74.84. We examined revenue, churn risk, customer behavior, and sales trends to identify areas for growth. The average transaction value is approximately $39.94. No missing data was detected in the dataset. Data confidentiality and security are addressed in client contracts. *A summary of the key variables used in this analysis is presented in Table 1 on the following page.*

Table 1: Variables Studied and Definitions

Variable	Definition
customer_id	Unique identifier for each customer in the dataset.
purchase_date	The date on which a transaction occurred.
amount_spent	Total dollar amount spent in a single transaction.
recency_days	Days since the customer's most recent purchase before cutoff.
frequency	Total number of purchases made by a customer before cutoff.
monetary_value	Total dollar amount spent by a customer across all transactions before cutoff.
days_since_first_purchase	Number of days since the customer's first purchase.
avg_days_between_tx	Average number of days between the customer's transactions.
tx_count_last_30_days	Count of transactions in the 30 days before the cutoff date.
had_tx_last_7_days	Indicator if the customer transacted in the last 7 days before cutoff.
total_days_active	Number of days between the customer's first and last purchase.
avg_tx_amount_last_30d	Average transaction amount in the last 30 days before cutoff.
is_weekend_dominant_buyer	Indicates if a customer makes more purchases on weekends.
avg_spend_per_tx	Average spend per transaction (monetary_value divided by frequency).
future_tx	Number of purchases made in the 30 days after the cutoff date.
churn_threshold	Dynamic threshold for inactivity, calculated as 1.5 times the average days between transactions.
churned_dynamic	Indicates if a customer's recency exceeds their dynamic churn threshold (1 = churned).

2 Sales Performance

2.1 Monthly Sales Trend

Figure 1: Monthly Sales Trend

This line chart shows the total sales processed each month from January 2023 to December 2024. By plotting month-over-month revenue, we can observe fluctuations in customer purchasing behavior and identify broader patterns across the two-year span. **Key takeaway:** Sales show consistent cyclicality with noticeable dips around March 2023, August 2023, and July 2024, and strong peaks in April 2023, October 2024, and a major spike in November 2024. The average monthly sales value across the period is approximately $1,665. These patterns suggest seasonal factors or effective promotional activity and can help guide decisions around staffing, inventory management, and targeted campaigns during high- and low-activity periods.

2.2 Quarterly Sales Trend

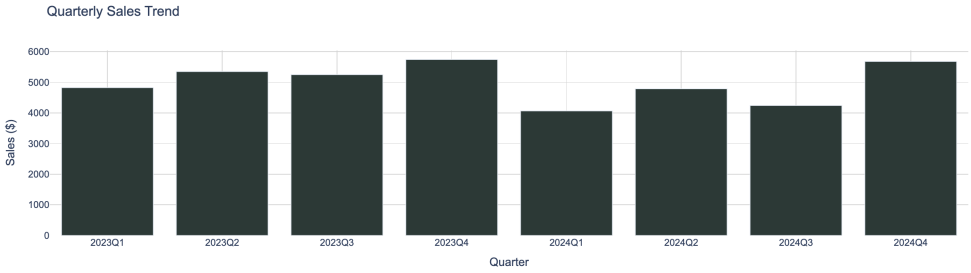

Figure 2: Quarterly Sales Trend

This bar chart summarizes total sales by quarter from Q1 2023 through Q4 2024. Aggregating monthly sales into quarterly periods helps smooth short-term fluctuations and highlight broader business performance trends. **Key takeaway:** Q4 2024 generated the highest sales of any quarter in the two-year span, followed closely by Q2 and Q4 of 2023. In contrast, Q1 2024 recorded the lowest sales. Notably, sales in Q4 2024 were approximately 41% higher than in Q1 2024. This sharp contrast underscores the importance of late-year quarters—particularly Q4—as key revenue drivers. Businesses may benefit from aligning inventory, staffing, and campaign efforts with these seasonal peaks.

3 Customer Churn Analysis

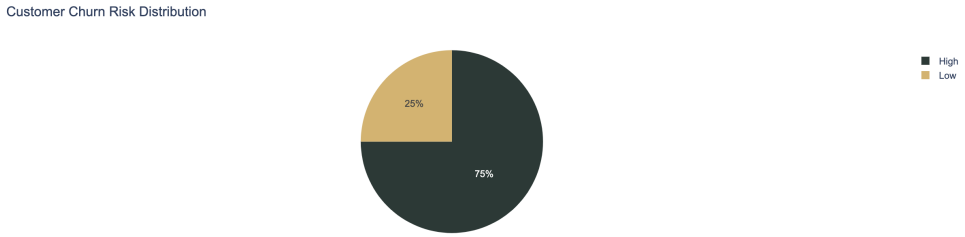

Figure 3: Customer Churn Risk Distribution

This pie chart segments customers into high or low churn risk based on a fixed 90-day inactivity cutoff. In this context, *churn* refers to customers who have not made a purchase within a defined period and are considered unlikely to return. Customers who have not transacted in the last 90 days are flagged as high risk. **Key takeaway:** Approximately 25% of customers fall into the high-risk category, signaling a substantial segment that may not return without intervention. While this fixed cutoff is helpful for exploratory analysis, a more adaptive churn definition based on each customer's purchase rhythm is introduced later during modeling to improve targeting accuracy.

4 Transaction Patterns

4.1 Transaction Amounts

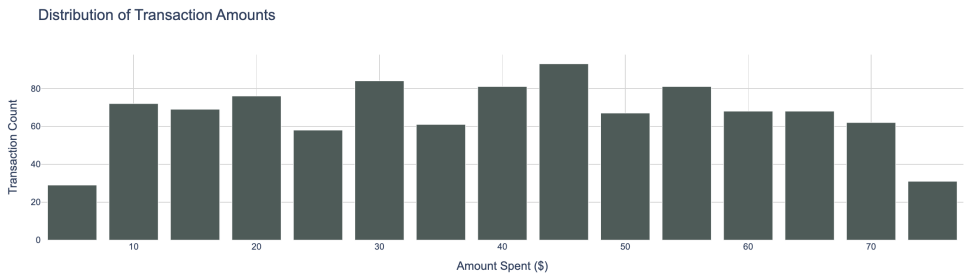

Figure 4: Distribution of Transaction Amounts

This histogram displays the distribution of transaction amounts ranging from \$5 to \$75. The data is grouped into equal-width bins, where each bar represents the number of transactions that fall within a specific spending range. This allows us to easily observe purchasing patterns across the full price spectrum. **Key takeaway:** Most purchases fall between \$10 and \$70, with noticeable peaks around \$10–\$20, \$30–\$35, and a sharp high point near \$45. This clustering suggests that customers are drawn to mid-tier price points, especially in the \$30–\$50 range. Strategically pricing popular items or bundles within these tiers may improve conversion rates and boost average order values.

4.2 Transactions Per Customer

Figure 5: Transactions Per Customer

This histogram shows how many times each customer made a purchase during the analysis period. The x-axis represents the number of transactions, and each bar—also called a bin—groups customers who share the same transaction count. **Key takeaway:** The majority of customers made only 1 or 2 purchases, with significantly fewer engaging more frequently. This long-tail distribution[1] highlights the importance of first- and second-time buyers. Businesses might consider onboarding campaigns or low-friction incentives to encourage repeat purchases from this segment and gradually shift more customers toward long-term retention.

[1] A long-tail distribution refers to a pattern where a large number of observations occur at low frequency (e.g., customers with 1–2 purchases), while a small number occur at high frequency (e.g., loyal or frequent buyers).

4.3 Daily Transaction Volume

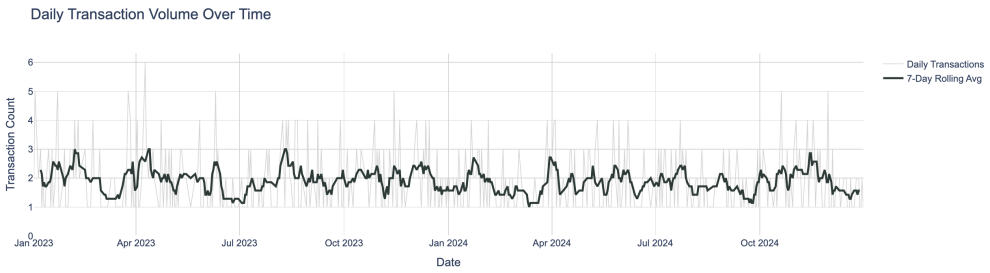

Figure 6: Daily Transaction Volume Over Time
Raw daily volume is shown in gray, and a 7-day rolling average is applied to highlight underlying trends in customer activity.

This line chart shows daily transaction counts from January 2023 to December 2024. Raw daily values (shown in light gray) often fluctuate due to natural variability in customer activity. To make underlying trends easier to spot, a 7-day rolling average (shown in bold) is overlaid to smooth out short-term noise. **Key takeaway:** Daily volume typically ranges between 1 and 4 sales per day, with occasional spikes reaching 5 or 6. While activity is generally consistent throughout the year, there are subtle lifts during late Q4—likely tied to holiday behavior—and slightly higher counts around spring. Businesses can use this smoothed trend to guide decisions on staffing, inventory, and timing for weekday promotions or flash deals.

5 RFM Correlation Insights

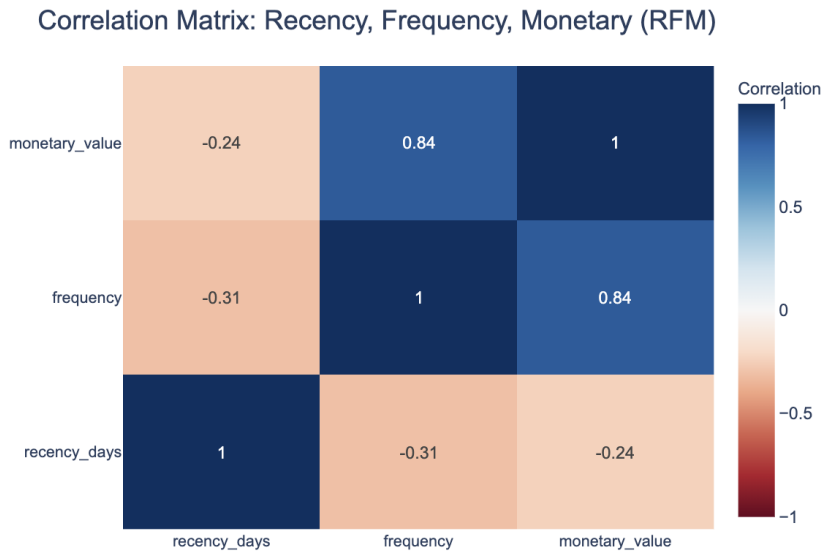

Figure 6: Correlation Matrix: Recency, Frequency, Monetary

This heatmap illustrates the strength of linear relationships between three key customer behavior metrics: recency (days since last purchase), frequency (total number of purchases), and monetary value (total amount spent). Each cell displays a Pearson correlation coefficient[2], which quantifies how closely two variables move together. The color gradient in the heatmap emphasizes these values: darker blue shades represent strong positive correlations, while darker red indicates strong negative correlations. Neutral or near-zero correlations appear in light tan.

Key takeaway: Frequency and monetary value are strongly positively correlated (0.84), meaning customers who purchase more often also tend to spend more. Recency is moderately negatively correlated with both frequency (–0.31) and monetary value (–0.24), implying that the longer it's been since someone last purchased, the less active and valuable they tend to be. These relationships can inform retention strategies—particularly by identifying lapsed customers with historically high activity who may still be worth targeting.

[2]Pearson correlation values range from –1 to 1. Values closer to 1 indicate a strong positive relationship (as one metric increases, so does the other), values closer to –1 indicate a strong negative relationship (as one metric increases, the other decreases), and values near 0 suggest little to no linear correlation.

6 Summary and Next Steps

Summary Insights

- **Total Revenue:** $39,944.99 — shows total income generated over the two-year analysis period.
- **Number of Transactions:** 1,000 — reflects overall business activity and customer engagement.
- **Number of Unique Customers:** 500 — indicates the size of the active customer base.
- **Average Transaction Value:** $39.94 — helps identify typical spending behavior per visit.
- **Churn Risk Distribution:** 375 high risk, 125 low risk — reveals retention gaps and re-engagement opportunities across the customer base.
- **Top Purchase Range:** $30–$50 — represents the most common spending tier, with a noticeable spike near $45.
- **Customer Frequency:** Most customers made 1–2 purchases — highlighting a need to improve repeat engagement.
- **Transaction Timing:** Volume increases modestly in Q4 and spring — suggesting seasonal behavior patterns useful for campaign timing.
- **RFM Correlation:** Frequency and monetary value are highly correlated (0.84) — showing that loyal buyers also tend to be higher spenders.
- **Inactive High-Value Customers:** Some customers with historically high frequency and spend now show low recency — ideal targets for win-back campaigns.

Key Observations

- Sales show recurring peaks in April, October, and November.
  Helps businesses time promotions and inventory builds around seasonal demand.
- Q4 consistently outperforms other quarters.
  Supports targeting holiday and year-end periods with stronger marketing and staffing.
- Most purchases fall between $10 and $70, with a spike near $45.
  Suggests ideal price points for bundling, discounts, or featured product placement.

- The majority of customers only buy once or twice.
  Indicates a need for retention strategies targeting first-time buyers.
- Daily sales average 1–4 transactions, with subtle lifts in Q4 and spring.
  Useful for planning staffing and testing weekday promotions based on smoothed demand.
- About 25% of customers are at high churn risk.
  Re-engagement campaigns can prioritize those with high past spend or frequency but declining activity.

Possible Improvements

- Incorporate product-level and category data.
  Supports deeper analysis of purchase behavior and bundling opportunities.
- Integrate customer demographic or behavioral segments.
  Enables more personalized targeting, cohort analysis, and value-based retention strategies.
- Track campaign exposure and promotional timing.
  Allows for ROI measurement and optimization of marketing effectiveness.
- Expand beyond transactional data to include browsing or engagement metrics.
  Improves customer profiling and early churn prediction.
- Add weekly or hourly transaction granularity.
  Supports staffing forecasts, flash sale planning, and demand smoothing.
- Store historical snapshots of churn thresholds and segment status.
  Enables backtesting retention strategies and tracking changes over time.

Potential Machine Learning Applications

- **Customer Churn Prediction** (e.g., XGBoost, Logistic Regression, Survival Analysis)
  Identifies at-risk customers using dynamic thresholds, recency patterns, and 30-day future activity to inform timely retention strategies.
- **Sales Forecasting** (e.g., Prophet, ARIMA, LSTM)
  Predicts future sales volume at daily or weekly resolution to support inventory planning, budget forecasting, and staffing decisions.
- **Customer Segmentation** (e.g., K-Means, Hierarchical Clustering, DBSCAN)
  Groups customers by RFM profiles, churn risk, or behavioral similarity for tailored offers, VIP programs, and lifecycle targeting.

- **Behavioral Pattern Modeling** (e.g., Time Series Clustering, HMMs)
  Uncovers seasonal rhythms, weekend behavior, or high-value dormant segments to guide re-engagement tactics.
- **Anomaly Detection** (e.g., Isolation Forest, Autoencoders)
  Flags unexpected spikes or drop-offs in transactions that may signal campaign impact, fraud, or system issues.
- **Recommendation Systems** (e.g., Item-Based Collaborative Filtering, Matrix Factorization)
  Boosts order value by suggesting products based on past purchase behavior and price sensitivity ranges.

Interactive Dashboards (Direct Links)

The following sample dashboards are hosted on Dropbox and will open directly in your browser. This setup is for demonstration only and is not used for client-facing deliverables or data hosting.

- Monthly Sales Trend
- Quarterly Sales Trend
- Transaction Amount Distribution
- Transactions Per Customer
- Daily Transaction Volume
- RFM Correlation Matrix
- Customer Churn

These interactive dashboards are best viewed in Chrome, Firefox, or Edge.