Deliverable 1

Group name: liam-n-pham

# Introduction

This report will deliver the comprehensive information about the approaches for analysis the Sales and Customers dataset and provide the rationale behind all decisions that

# Dataset Overview

1. **Data Structure and Quality Assessment**

**Customer Data** (Customers\_v4.csv)

* **Records**: 1,001 customers
* **Attributes**: 9 variables including Customer ID, Demographics (Age, Gender), Geographic (City, Province, Region), and Membership information (Level, Tenure)
* **Key Fields**: Customer ID (unique identifier), Membership Level (Standard, Gold, Platinum), Customer Age (18-100 years), Tenure (years as customer)

**Sales Transaction Data** (Sales\_v4.csv):

* **Records**: 25,000 transactions
* **Date Range**: January 24, 2023 to August 20, 2025
* **Attributes**: 13 variables including Transaction ID, Customer ID, Product details (Category, Item), Financial metrics (Unit Price, Unit Cost, Discount Rate, Total Spent), and Operational data (Location, Payment Method, Date)

1. **Data Quality Issues Identified**

Duplicate Transaction IDs

* **Discovery**: 339 duplicate Transaction IDs affecting 673 rows
* **Pattern**: Transaction IDs were unique within each date but reused across different dates (daily counter reset pattern)
* **Evidence**: Same Transaction ID (e.g., *TXN\_101364*) appeared on different dates (2023-03-28 and 2024-09-11) with completely different customer, product, and amount data

Future Dates

* **Issue**: Transactions dated beyond October 24, 2025
* **Findings**: Small number of transactions in late August 2025
* **Decision**: Flagged but not remove

Referential Integrity

* **Issue**: Validation of Customer IDs between transactions and customer master table
* **Findings**: 100% of Customer IDs in Sales data matched Customers data

Value Validation

* **Issue**: Check for negative quantities, prices, or costs
* **Findings**: No invalid negative values found after cleaning

Missing Values

* **Issue**: Check for missing data in critical fields
* **Findings**: No missing values in Transaction ID, Customer ID, Date, Total Spent

1. **Data Cleaning Implementation & Decisions**

**A. Transaction ID Duplication Resolution**

We found 339 duplicate Transaction IDs (673 affected rows out of 25,000). Each duplicate Transaction ID referred to completely different transactions:

* Different Customer IDs
* Different transaction dates
* Different products, quantities, and amounts
* Different locations and payment methods

Finding: Transaction IDs were unique within each date but reused across dates (daily counter reset)

Resolution Strategy:

* Created composite primary key: Transaction\_ID\_Date = Transaction\_ID + "\_" + Date (YYYYMMDD)
* Added new columns:
  + Original\_Transaction\_ID: Preserved original Transaction ID for reference
  + Date\_Component: YYYYMMDD string extracted from Date field
  + Transaction\_ID\_Date: Composite key used as new unique identifier

Examples:

* Original: TXN\_101364 (appeared on 2023-03-28 and 2024-09-11)
* Resolved: TXN\_101364\_20230328 and TXN\_101364\_20240911

Rationale:

* Preserves original Transaction ID structure
* Ensure the uniqness by using the timestamp
* Eliminates duplication: 25,000 rows → 25,000 unique IDs (100% uniqueness)

**B. Date Validation**

There are transaction dates beyond today (October 24, 2025). We choose October 24, 2025 as a day for validating all the dates. We checked the transactions along with the customers data but all the records are valid so those records could be pre-orders

Resolution Strategy:

* Flagged but retained all future-dated transactions

Rationale:

* Small number of affected transactions
* Flagging allows filtering in analysis if needed
* Retaining preserves potential valid data

**C. Referential Integrity Validation**

We need to validate all Sales Customer IDs exist in Customer master table by using the formula

Result:

* All 25,000 transactions have valid Customer IDs
* No orphaned transactions (Customer ID in Sales but not in Customers)

**D. New columns**

Added new columns to support the statistical calculation and hypothesis

|  |  |  |
| --- | --- | --- |
| Type | Column name | Description |
| Transaction | Transaction Key | New unique key |
| Sale | Log\_Total\_Recalc | Logarit of total revenue |
|  | Gross Profit | Gross profit = price per unit – actual cost per unit |
|  | Pre\_Recalc | Actual cost of the product = quantity \* unit price |
|  | Gross\_Margin\_% | Percentage of gross margin = Gross after discount / total revenue |
|  | Gross\_Profit\_After\_Discount | Gross profit after discount = Total Gross – Cost per quantity |
|  | Discount\_Zero\_Flag | Base on discount rate, if discount rate = 0 then use “No discount”, otherwise “Discounted” |
| Date | Year | YYYY format |
|  | Month |  |
|  | Week |  |
|  |  |  |
|  |  |  |

1. **Clean the data & finalize the dataset**

The original dataset was split across 2 separate files (Sales and Customers). Therefore we would need to merge those dataset into a single dataset

Rationale for Merging:

* **Enable Customer Segmentation Analysis**: To analyze purchasing patterns by membership level (Platinum vs Standard), customer age groups, and tenure, we needed customer attributes joined with transaction data
* **Support Hypothesis Testing**: Several hypotheses require both transaction metrics and customer characteristics in a single analysis-ready dataset (credit card preference or membership by age)
* **Comprehensive EDA**: Exploratory analysis requires examining relationships between transactional behavior (what was purchased, when, how much) and customer characteristics (who purchased it, demographics)
* **Operational Efficiency**: A denormalized structure (Sales\_Cleaned with customer attributes) eliminates repeated and complex operations and helps to buid pivot table directly for analysis
* **Data Integrity**: Merge validated referential integrity - all Customer IDs in transactions successfully matched to customer master table

Merge execution:

* **Join Key**: Customer\_ID (common field between both datasets)
* **Join Type**: LEFT JOIN (Sales as base table, Customers as lookup table)

Original Data:

* Sales transactions: 25,000 rows
* Customers: 1,001 records
* Issues identified: 339 duplicate Transaction IDs

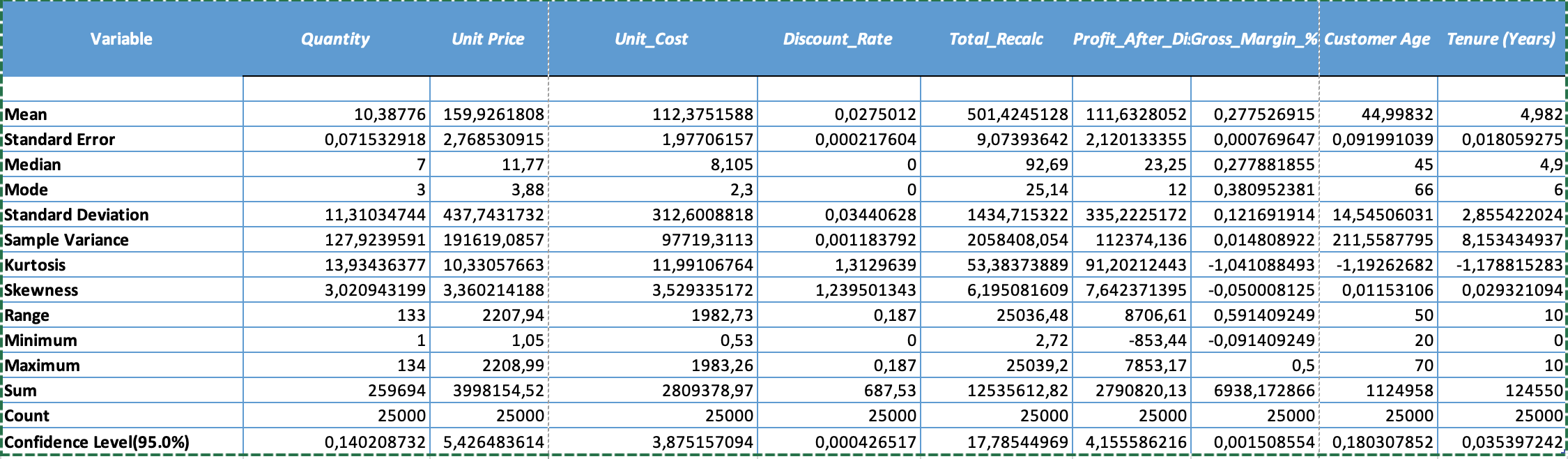
Cleaned Data (Sales\_Cleaned):

* A new sheet “Sales\_Cleaned” sheet was created
* **Records**: 25,000 valid transactions
* **Columns**: 31 variables (original 13 + 18 new features)
* **New Features Created**:
  + Composite Transaction Key (Transaction\_ID\_Date)
  + Date Components (Year, Month, Day, Quarter, Day\_of\_Week, Week\_of\_Year)
  + Profit Metrics (Gross\_Profit\_After\_Discount, Gross\_Margin\_%)
  + Calculated Fields (Total\_Recalc - verified total spending)
  + Customer Attributes (joined from customer table: Membership\_Level, Customer\_Age, Tenure\_Years, Region, Gender)
  + Data Quality Flags (for future dates, validation status)

# EXPLORATORY DATA ANALYSIS

With the cleaned data, we are proceed to the exploratory data analysis (descriptative statistic, visualization, correlation analysis then come up with the hypothesis)

1. **Descriptive Statistics Measures**



Rationale for using Mean, Median and Mode

* Mean: To calculate the total revenue, average transaction values and provide the understanding about overall business performance and use for hypothesis testing later
* Median: Represents the typical transaction unaffected by extreme outliers (e.g., very high-value purchases) and it is bettern than mean for skewed distributions (like in price)
* Mode: Identifies the most common purchase behavior (e.g., most frequent quantity ordered) that can helps understand a normal customer behavior and most popular product price tiers

Rational for using Standard Deviation, Range (Max – Min), Kurtosis and Skewness

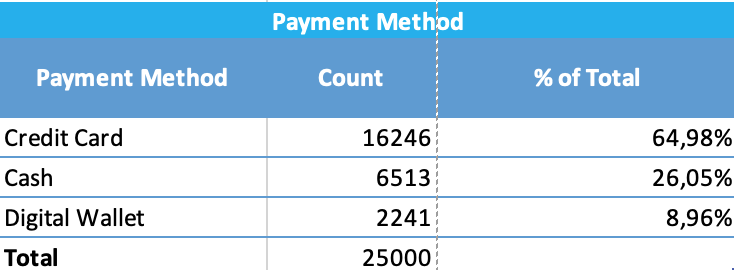
* Standard Deviation (SD): Quantify how spread out the data is around the mean, use for hypothesis testing (calculating test statistics) and understanding business risks
* Range (Max – Min): Identify extream values, use for data validation (e.g., negative profits might flag potential issues) and outlier detection
* Kurtosis: Identifies whether distribution has heavy tails (many outliers) or light tails (few outliers) which will be used to detect outliers and anomolies
* Skewness: Identifies whether data is symmetric, left-skewed (negative skewness) or right-skewed (positive skewness). For example, when we calculated the *Total\_Recalc* has skewness = 6.20 (highly right-skewed due to outlier high-value transactions)

1. **Pivot tables**

We need to calculate the pivot tables for summarize key features and visualize them using charts

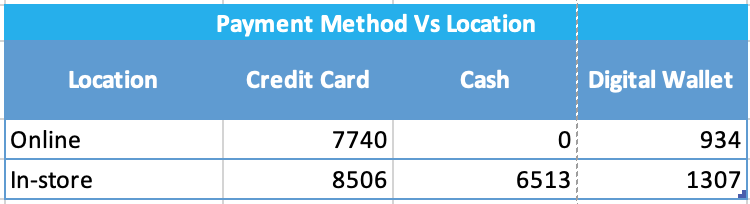
**Payment Method**:

We calculate the number of transactions that used credit card, cash and digital wallets and their proportion in our dataset.



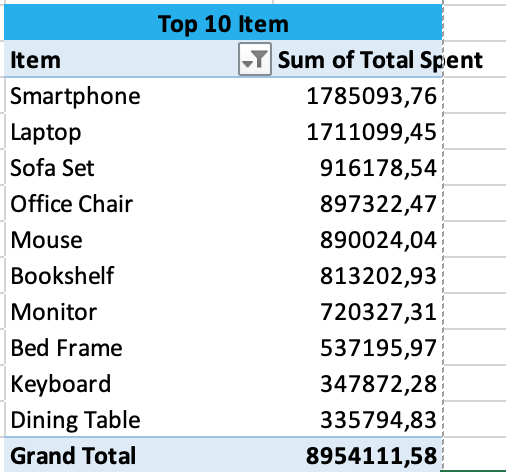
**Payment Method – Location**

The pivot table demonstrate the number of transaction that used in-store and online in different payment methods



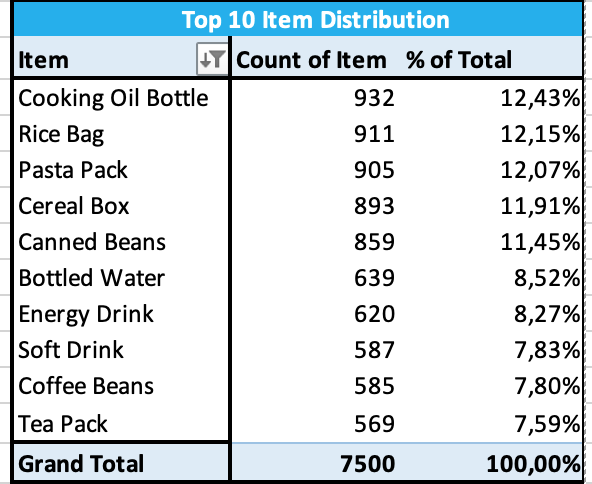
**Top 10 product’s revenue table**

The table demonstrates the top 10 of total revenue by each of category



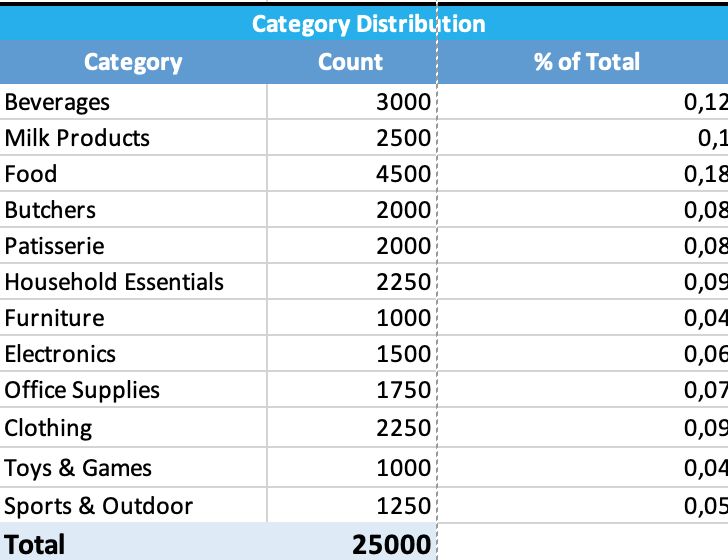
**Top 10 item distribution table**

This table used to visualize the top 10 items distribution by couting number of transactions



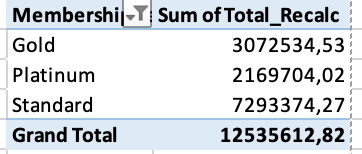
**Category distribution**

This table helps to summarize distribution of each categories and their proportion



**The total revenue by membership level**

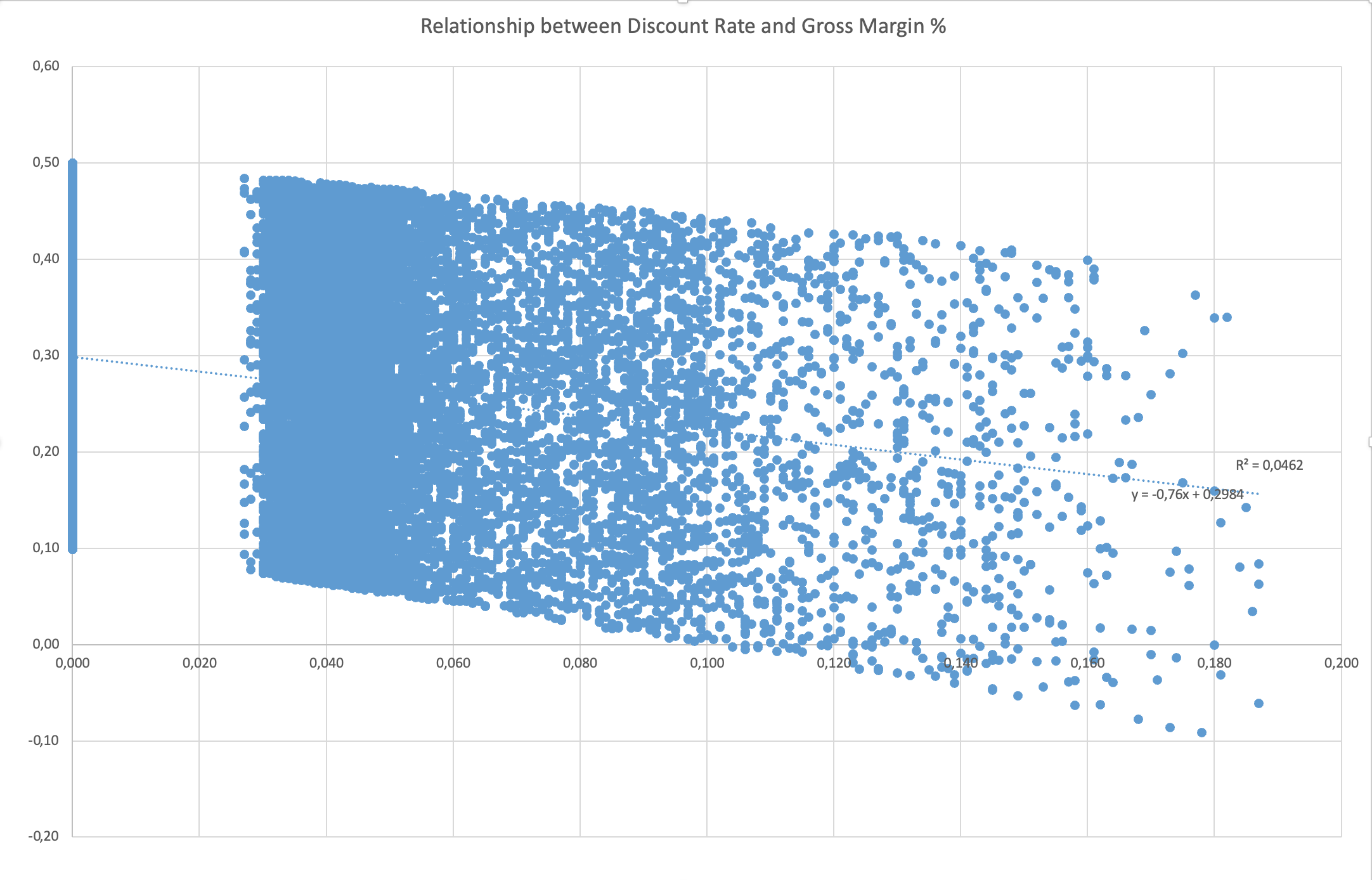
This table helps to demonstrate the total revenue by each of membership types and the relationship between membership levels and total recalc



1. **Visualization**

We have implemented 8 types of charts to demonstrate the relationship between data that we found (from “Charts” sheet)

**Chart: "Relationship between Discount Rate and Gross Margin %"**

****

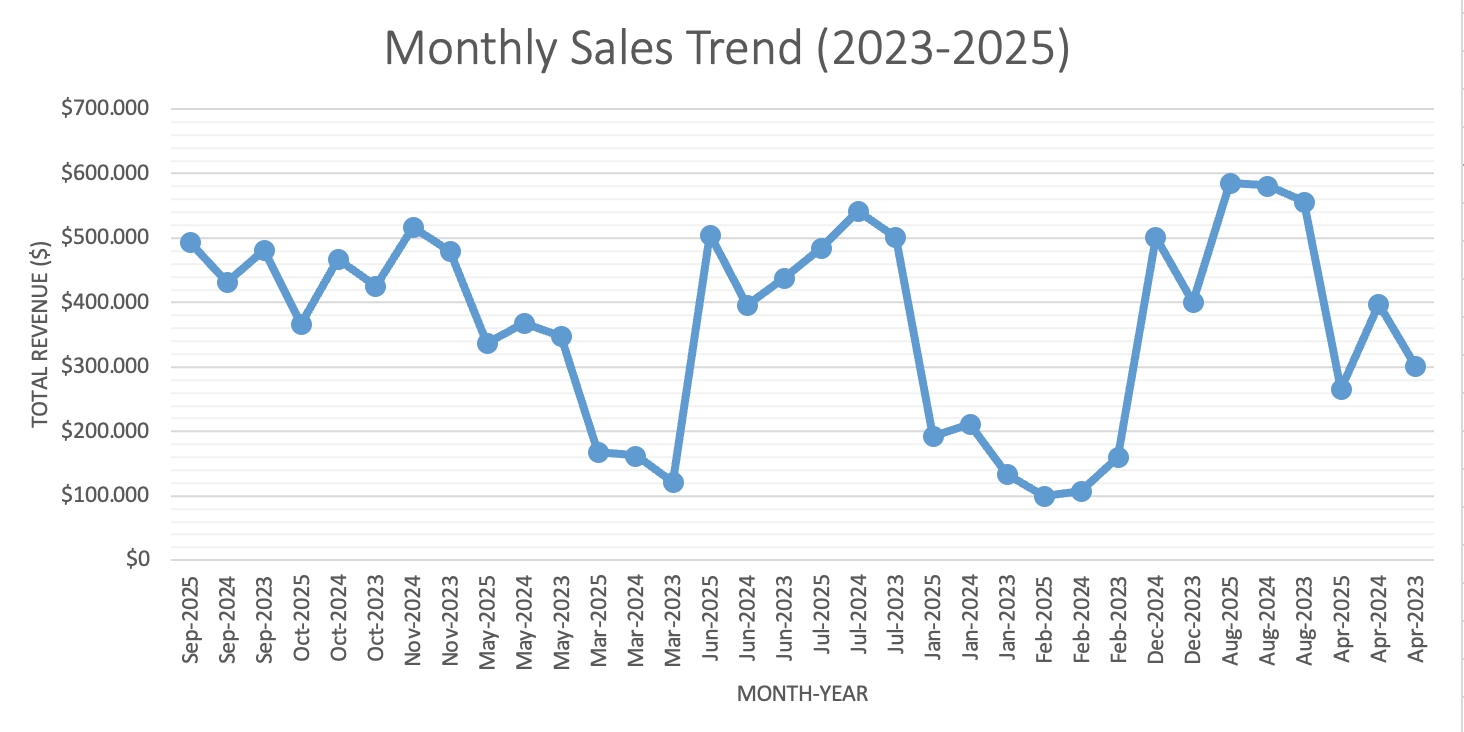
* **Overview**: This chart reveals the inverse relationship between discounting and profitability. Each dot represents a transaction, and the pattern shows how increasing discounts compress profit margins.
* **Rationale**: Scatter plots are ideal for visualizing relationships between two continuous numerical variables. By plotting each transaction as a point with discount rate on one axis and gross margin on the other, we can visually assess the strength and direction of their correlation.
* **Insight**: The negative correlation visible in the scatter pattern confirms that aggressive discounting erodes margins.

**Chart: "Monthly Sales Revenue Trend"**



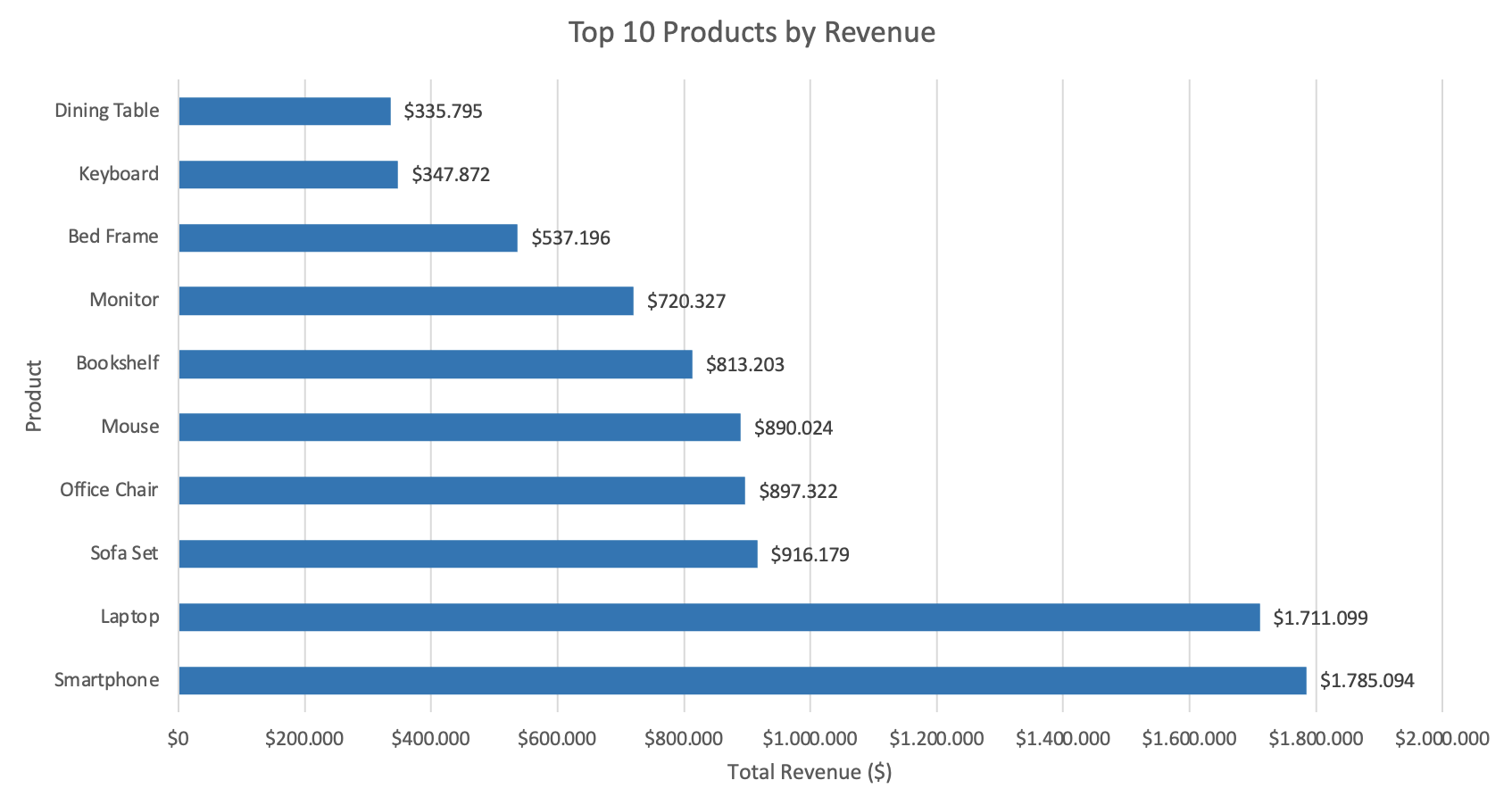
* **Overview**: This chart displays total revenue aggregated by month, allowing identification of seasonal patterns, growth trends, and potential cyclical behavior in sales
* **Rationale**: Line charts are the standard for time-series analysis, effectively showing how a metric changes over sequential time periods. The continuous line makes trends, seasonality, and anomalies immediately apparent.
* **Insight**: The line reveals temporal patterns in customer purchasing behavior

**Chart: "Monthly Sales Trend (2023-2025)”**



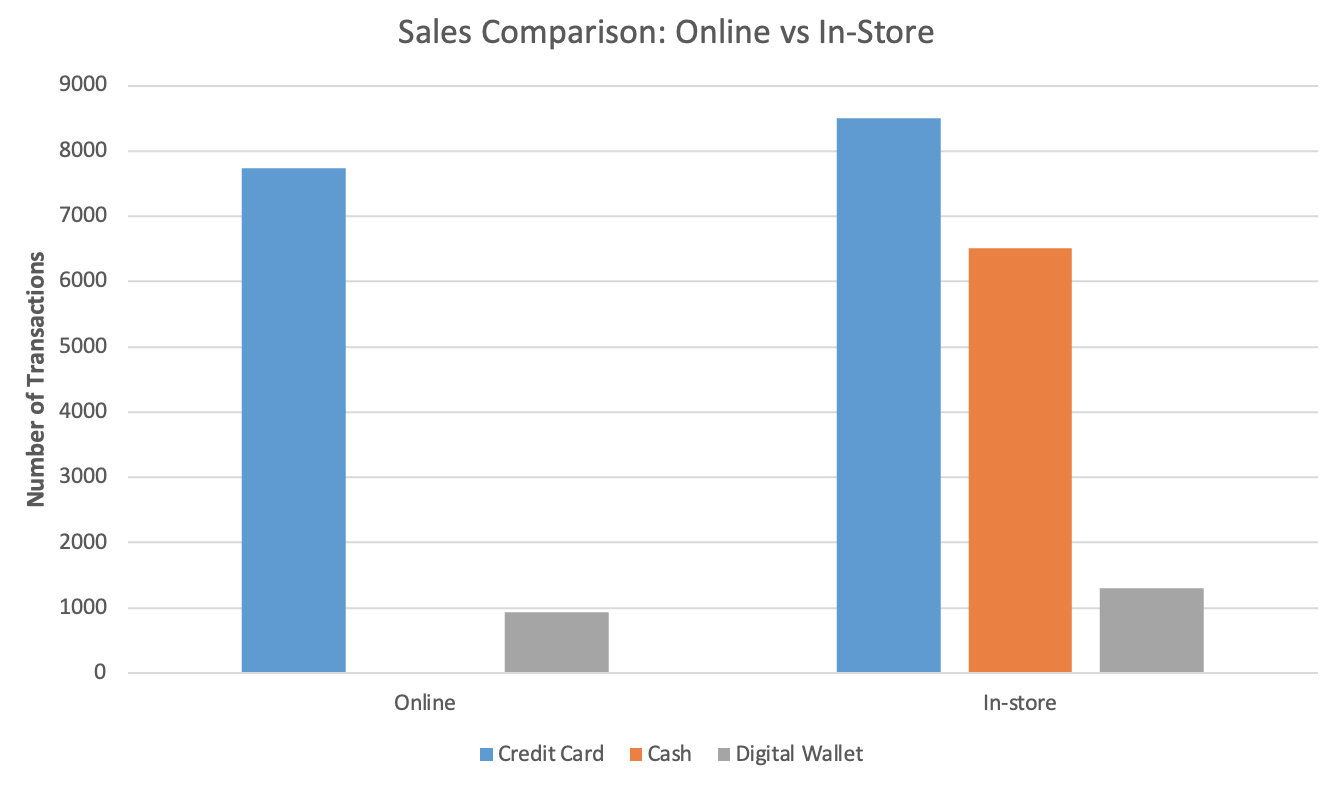
* **Overview**: This chart enables comparison of the same months across different years to identify consistent seasonal peaks or growth trajectories
* **Rationale**: Multi-year line charts allow year-over-year comparison while preserving the monthly granularity. This dual perspective shows both long-term growth and recurring seasonal patterns
* **Insight**: Helps determine if observed patterns (e.g., December peaks) are consistent across years or one-time events.

**Chart: "Top 10 Products by Revenue"**



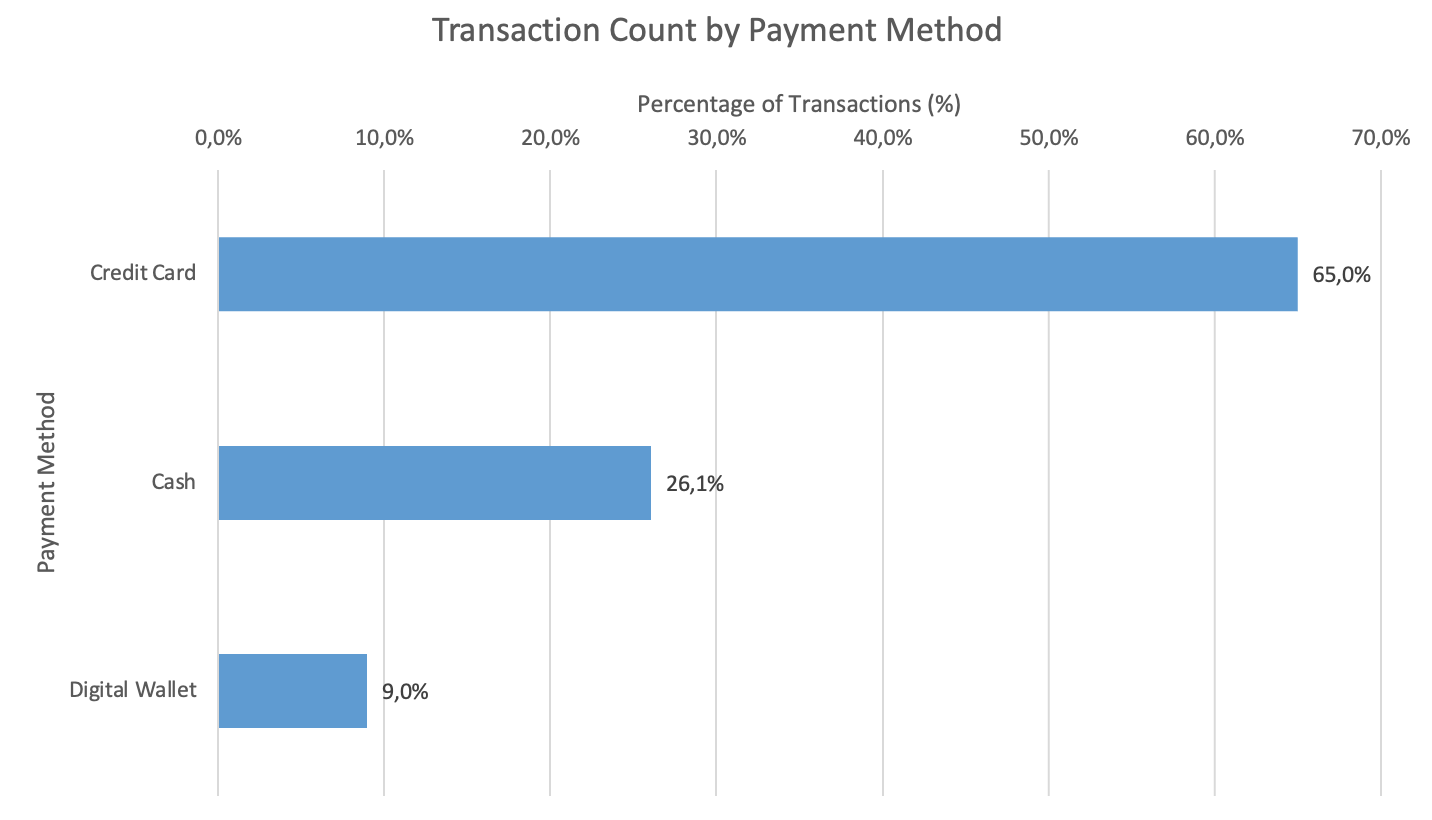
* **Overview**: This chart ranks the 10 highest revenue-generating products, using bar length to represent total sales revenue for each item.
* **Rationale**: Show the relationship between the items and the revenue, ranking items in descending order.
* **Insight**: Visual comparison immediately highlights that Smartphone and Laptop dominate revenue, generating $1.79M and $1.71M respectively, far exceeding other products.

**Chart: “Sales Comparison: Online vs In-Store”**



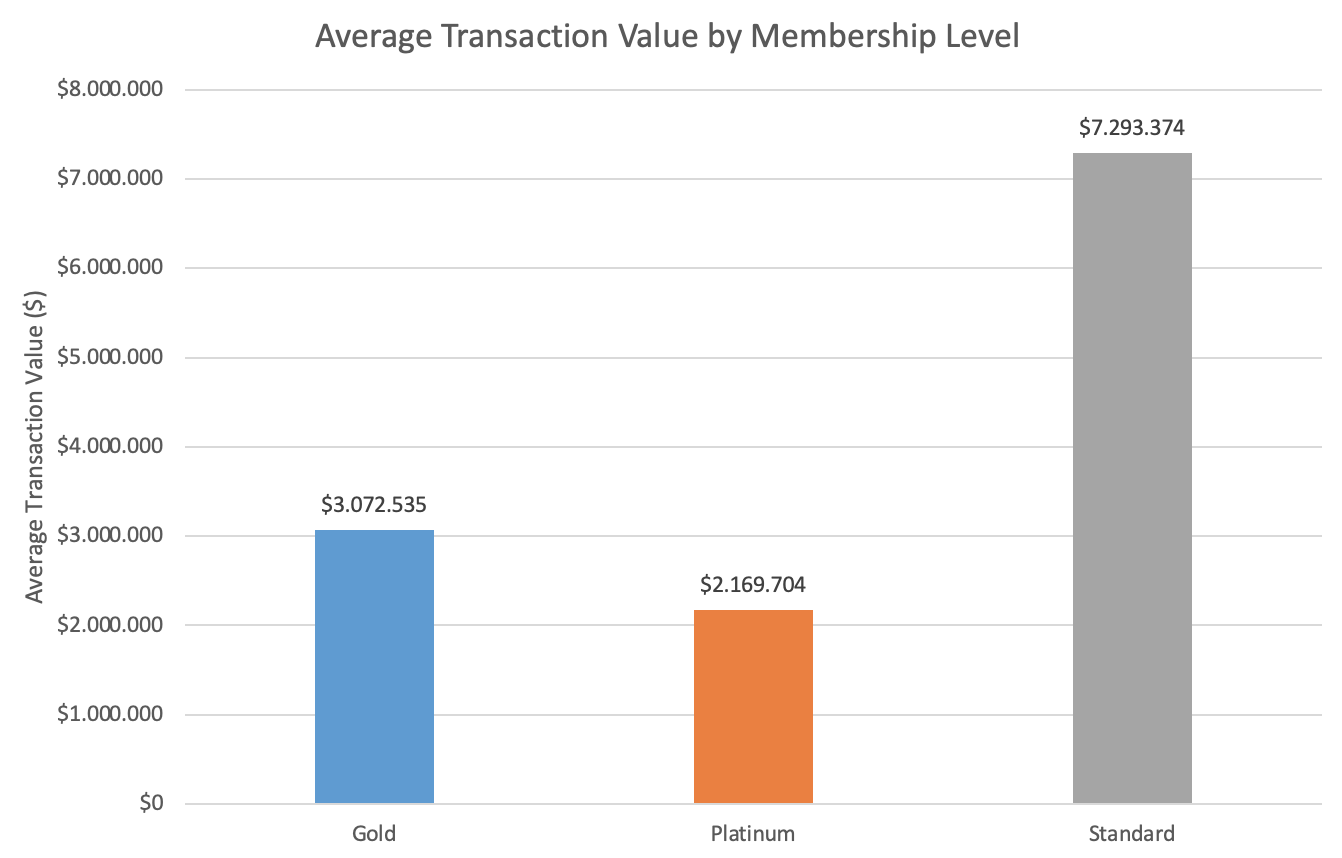
* **Overview**: This chart demonstrate the most common payment type in different locations
* **Rationale**: show the relationship between different type of payment methods with the location and their transactions, comparison between different type of locations
* **Insight**: Credit card was the popular choice by customers in both online and in-store’s locations

**Chart: “Transaction Count by Payment Method”**



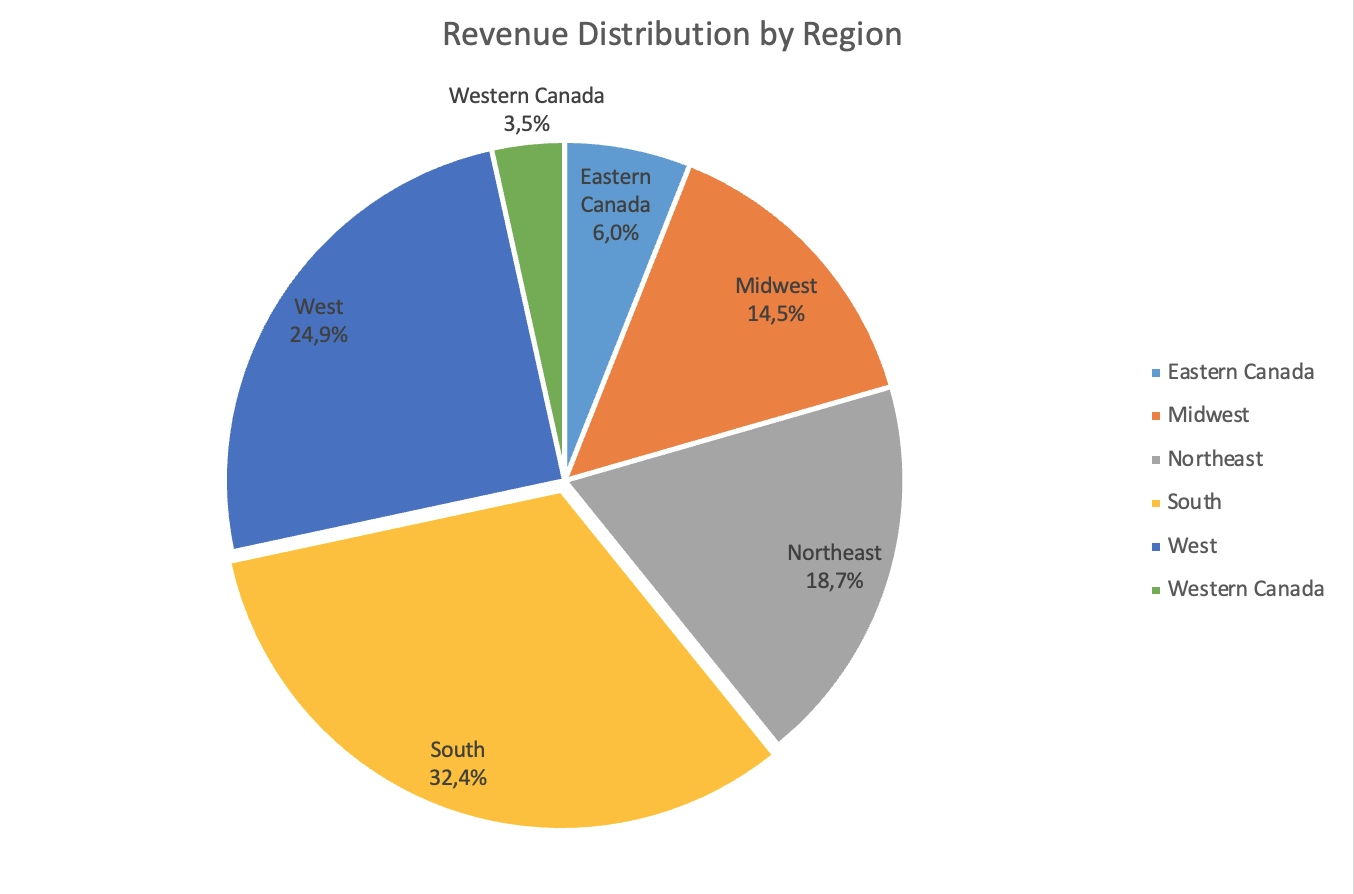
* **Overview**: This chart demonstrate the most usage payment method of the customers
* **Rationale**: This will help to identify the relationship between the total transactions with the payment method which will help for the hypothesis workflow later
* **Insight**: Credit card was the most common method among other payment types

**Chart: “Average transaction value by membership level”**



* **Overview**: This chart demonstrate transaction values by membership levels
* **Rationale**: This will help to identify the relationship between the average transaction by each of member types and from that can have a plan for further optimization, comparison between membership levels
* **Insight**: Standard member tends to have more transaction values than other member types

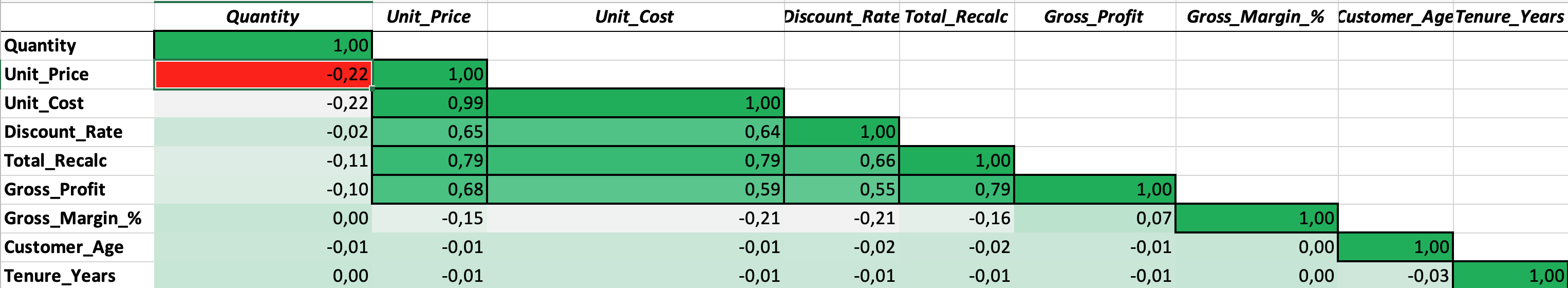
**Chart: “Revenue Distribution by Region”**



* **Overview**: This chart demonstrate the revenue by region
* **Rationale**: This will help to identify the relationship between the revenue distribution (%) by different regions in Canada, visualize the proportion
* **Insight**: South and West Canada are those regions that have the most distribution in term of revenue

1. **Correlation Analysis**

To quantifies the strength and direction of linear relationships between pairs of numerical variables, we created a correlation matrix to identify the strong and weak relationships



From the matrix, we found:

**Strong Positive Correlations (|r| > 0.7):**

* Unit\_Price ↔ Unit\_Cost (r = 0.99)
* **Interpretation**: Pricing is directly tied to product costs
* **Implication**: Cost structure determines pricing strategy; cost increases will require price adjustments
* Unit\_Price ↔ Total\_Recalc (r = 0.79)
  + **Interpretation**: Higher-priced items drive larger transaction totals
  + **Implication**: Revenue growth can be achieved by promoting higher-priced products
* Unit\_Cost ↔ Total\_Recalc (r = 0.79)
  + **Interpretation**: Items with higher costs contribute more to total transaction value
  + **Implication**: High-cost items are revenue drivers
* Total\_Recalc ↔ Gross\_Profit (r = 0.79)
  + **Interpretation**: Larger transactions generate higher absolute profit
  + **Implication**: Upselling and larger basket sizes directly increase profitability

**Moderate Relationships (0.5 ≤ |r| < 0.7):**

* Unit\_Price ↔ Discount\_Rate (r = 0.65)
  + **Interpretation**: Higher-priced items tend to have higher discount rates
  + **Implication**: Premium products are more frequently discounted
* Unit\_Cost ↔ Discount\_Rate (r = 0.64)
  + **Interpretation**: Higher-cost items are associated with higher discounts
  + **Implication**: Expensive items may require discounts to convert sales
* Unit\_Price ↔ Gross\_Profit (r = 0.68)
  + **Interpretation**: Higher unit prices lead to greater profit per transaction
  + **Implication**: Focus on premium products can improve profitability
* Discount\_Rate ↔ Total\_Recalc (r = 0.66)
  + **Interpretation**: Discounts are associated with higher transaction values
  + **Implication**: Discounts may be applied strategically to larger purchases
* Unit\_Cost ↔ Gross\_Profit (r = 0.59)
  + **Interpretation**: Higher-cost items generate higher profits (despite lower margins)
  + **Implication**: Selling expensive items (even with discounts) drives profit dollars
* Discount\_Rate ↔ Gross\_Profit (r = 0.55)
  + **Interpretation**: Discounts can coexist with profitability
  + **Implication**: Volume from discounts may offset margin compression

**Weak Correlations (0.1 ≤ |r| < 0.5):**

* Gross Margin % ↔ Gross Profit (r = 0.07)
  + **Interpretation**: Gross margin % doesn't vary much with profit

**Negative Correlations (r < 0):**

* Quantity ↔ Unit\_Price (r = -0.22)
  + **Interpretation**: Slight inverse relationship; bulk purchases at lower unit prices
  + **Implication**: Customers buying higher quantities tend to buy lower-priced items
* Quantity ↔ Unit\_Cost (r = -0.22)
  + **Interpretation**: low-cost items purchased in larger quantities
* Gross\_Margin\_% ↔ Discount\_Rate (r = -0.21)
  + **Interpretation**: Discounts reduce profit margin percentage
  + **Implication**: Balance needed between discount-driven volume and margin preservation
* Gross\_Margin\_% ↔ Unit\_Cost (r = -0.21)
  + **Interpretation**: Higher-cost items have slightly lower margin percentages
  + **Implication**: Expensive items may operate on thinner margins

1. **Detected patterns from visualizations and statistical**

**High Revenue Concentration in Electronics**

* Top 2 products (Smartphone: $1.79M, Laptop: $1.71M) generate 39% of top-10 revenue
* Evidence: Chart “Top 10 Products” and the pivot table “Top 10 Item”

**Discount Rate Impact on Profit Margins**

* Negative correlation (r = -0.21) between discounts and margins
* Evidence: Chart Discount Rate and Gross Margin %" & Correlation matrix

**Temporal Revenue Trends**

* Time-series patterns visible across 2023-2025
* Evidence: Charts Monthly Sales Trend
* Enables seasonal analysis and year-over-year growth tracking

**Payment Method Dominance by Channel**

* Credit Card: 65%, Cash: 26% (in-store only), Digital Wallet: 9%
* Evidence: Chart “Transaction Count by Payment Method”, “Average transaction value by membership level” and the pivot table for Payment method
* In-store transactions: 65% of total (16,326 of 25,000)

**Product Category Diversity**

* 12 categories with balanced distribution; Food leads at 18%
* Evidence: Category distribution table
* No single category exceeds 20% (reduces concentration risk)

**Strategic Discounting**

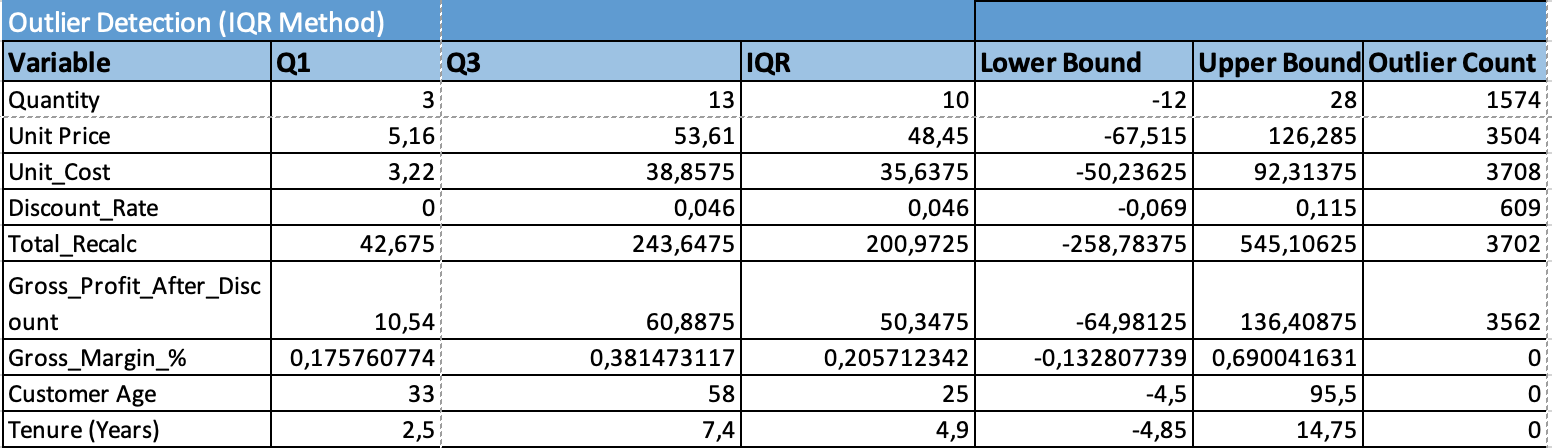
* Low average discount rate (2.75%) with median of 0% (Pattern “Discount Rate Impact on Profit Margins”)
* Discounts may be channel-specific or payment-method-specific (Pattern “Payment Method Dominance by Channel”)
* Most transactions (especially in-store cash purchases of staples) have no discount
* Selective discounting on big-ticket items maintains overall margin health

**Portfolio Balance**

* Electronics concentration (Pattern “High Revenue Concentration in Electronics”) creates revenue concentration risk
* Category diversity (Pattern “Product Category Diversity”) provides some hedge but not enough
* Food + Beverages (30% of transactions) generate steady traffic but lower revenue
* Electronics (6% of transactions per Pattern 6) drive 39% of top-10 revenue per Pattern “High Revenue Concentration in Electronics”

1. **Outlier detection and analysis**

From the statistical result, we found unusual values that are valid data points but extreme compared to the typical distribution. We applied few approaches to see the distribution of that data and evaluate them

IQR (Interquartile Range) Method

**Total\_Recalc**: 3702 outliers

* Threshold: Transactions > $545
* Interpretation: Very high-value transactions, likely from:
  + Corporate/B2B bulk buyers
  + Platinum membership customers
  + Large multi-item orders
  + Premium product purchases
* Business Insight: These outliers represent important high-value customer segments that should be analyzed separately and catered to with premium services
* Decision: Keep all outliers - they are valid and represent key revenue drivers

**Quantity**: 1574 outliers

* + Threshold: Orders > 28 units (when typical orders are 2-5 units)
  + Interpretation: Unusually large orders, possibly:
* Bulk purchases for events, offices, or resale
* Stock-up behavior during promotions
* Business customers or resellers
  + Business Insight: Opportunity to create bulk purchase programs or B2B pricing tiers
  + Decision: Keep all outliers - represent important transaction type

**Gross\_Profit\_After\_Discount**: 3562 outliers

* Threshold: Profit > $136
* Interpretation: High-profit transactions correlating with high revenue outliers
* Business Insight: These transactions have disproportionate impact on profitability; focus on replicating conditions that generate them
* Decision: Keep all outliers - critical to profit analysis

**Discount\_Rate**: 609 outliers

* Threshold: Discount > 11.5%
* Interpretation: Deep discount promotions, possibly:
* Clearance sales
* Special promotional events
* Loyalty rewards or membership benefits
* Price-match guarantees
* Business Insight: Assess ROI of deep discounts
* Decision: Keep all outliers - analyze effectiveness separately

From all of that, we decided to keep all the outliers in the analysis because:

* **Validity**: All outliers represent real, valid business transactions - not data entry errors or measurement errors
* **Business Value**: High-value outliers (revenue, profit) represent key customer segments that drive disproportionate value
* **Pattern Insights**: Outliers reveal important patterns (bulk buying, premium customers, promotional effectiveness)
* **Statistical Honesty**: Removing outliers would artificially narrow distributions and hide real business variability
* **Segmentation Opportunity**: Outliers may warrant separate analysis as distinct customer or transaction segments

1. **Hypotheses for Statistical Analysis**

Based on the patterns identified in earlier steps and outliers, we formulate the hypothesis for future statistical analysis

****

# PowerBI

# Project structure

Excel file: CP610\_D2\_Liam\_n\_Pham.xlsx

|  |  |
| --- | --- |
| Sheet name | Description |
| Insights | Hypothesis testing |
| Correlation | Correlation matrix and key findings about dataset |
| Charts | All charts, visualizations for the relationship between items |
| Customer checks | Data validation for customer datasets |
|  |  |
|  |  |
|  |  |
|  |  |

# Conclusion

We processed the 2 dataset Sales and Customers. From those data we have evaluate, identify the outliers and clean the data, prepare the new dataset for expoloratory data analysis. We used different approaches from statistical measures to visualization to find out the relationship and patterns between the categories within the dataset. In addition, we also identified unsual data (outliers) from the dataset and formulate hypothesis for future analysis