University of Niagara Falls

Master of Data Analytics

DAMO-501-1: Data Analytics Case Study 1

Data-Driven Strategies for Sales and Customer Insights for Classic Car Models Retail Company

Team

Sagar Khadka

Roberto Alberto San Miguel

Jigme Yeshi

Instructor: Dr. Hany Osman

Niagara Falls

09/12/2024

Table of Contents

Chapter 1: Problem Definition and Research Question

1.1 Problem Definition

The classic car models' company is a retailer of scale models of classic cars, which includes typical business data such as customers, products, sales orders, and sales order line items. As a retail operation, the company aims to leverage historical data to forecast customer purchasing behavior, identify key sales trends, and enhance marketing strategies.

This project addresses several primary challenges:

- Identifying geographic differences in sales performance.
- Understanding how product pricing affects sales.
- Determining which product lines drive the most revenue.
- Gaining insights into customer purchasing patterns, particularly in relation to credit limits.
- Predicting future customer purchasing behavior using historical data and patterns to support targeted marketing and sales campaigns.

These insights will support the classic car models' company in refining their marketing strategies, improving sales forecasting, optimizing inventory management, and enhancing customer relationship efforts.

1.2 Research Questions

The following research questions are designed to explore critical aspects of sales, product performance, and customer behavior:

- **Sales Performance Analysis:**
    - Do sales orders differ significantly between different territories (offices)?

    Justification: Understanding geographic sales performance helps allocate resources effectively and identify regions where marketing efforts can be intensified.

- **Product Pricing and Sales Frequency:**
    - Does the Manufacturer Suggested Retail Price (MSRP) of products impact the frequency of sales orders?

    Justification: This question explores whether product pricing influences purchase frequency, guiding pricing strategies to enhance sales.

- **Product Line Revenue Contribution:**
    - Do product lines with a greater variety of products generate more revenue?

  Justification: Analyzing revenue by product line size provides insights that assist in inventory and product development strategies.

- **Customer Insights and Behavior Analysis:**
    - Are customers with higher credit limits more likely to place larger orders?

  Justification: Understanding the relationship between credit limits and order size aids in forecasting customer spending and refining credit policies.

- **Forecasting Customer Purchasing Behavior:**
    - Can customer purchasing behavior (e.g., order frequency, total purchase value) be predicted based on historical data?

  *Justification***:** Developing predictive models based on historical purchasing patterns supports targeted marketing efforts, inventory optimization, and resource allocation.

1.3 Database Introduction

The Classic Models database is a relational database designed to support the operations of a scale model car retailer. The database comprises the following tables, representing key business entities and their relationships:

1.3.1 Tables Overview

- **Product Lines**: Defines the categories for grouping products.
  Attributes:
    - productLine: Unique identifier for each product line.
        - textDescription, htmlDescription, and image: Additional details about the product line.

- **Products**: Contains detailed information about individual products.
  Attributes:
    - productCode: Unique identifier for each product.
    - productName, productLine, productScale, productVendor: Descriptions and classifications.

    - quantityInStock: Current stock levels.
    - buyPrice and MSRP: Cost and pricing details.

- **Offices**: Represents the company's geographic office locations.
  Attributes:
    - officeCode: Unique identifier for each office.
    - city, phone, addressLine1, addressLine2, state, country, postalCode, and territory: Office details.

- **Employees**: Includes information about company employees.
  Attributes:
    - employeeNumber: Unique identifier for each employee.
    - lastName, firstName, email, jobTitle: Personal and professional details.
    - officeCode: Association with an office.
    - reportsTo: Hierarchical reporting structure.

- **Customers**: Contains customer profiles and contact information.
  Attributes:
    - customerNumber: Unique identifier for each customer.
    - customerName, contactLastName, contactFirstName, phone: Contact details.
    - salesRepEmployeeNumber: Sales representative assigned to the customer.
    - creditLimit: Maximum credit allocated to the customer.

- **Payments**: Tracks payments made by customers.
  Attributes:
    - customerNumber: Reference to the customer making the payment.
    - checkNumber: Unique identifier for each payment.
    - paymentDate and amount: Payment details.

- **Orders**: Records sales orders placed by customers.
  Attributes:
    - orderNumber: Unique identifier for each order.
    - orderDate, requiredDate, shippedDate, status, comments: Order processing details.
    - customerNumber: Reference to the customer who placed the order.

- **Order Details**: Represents the individual line items within orders.

Attributes:

- orderNumber: Reference to the associated order.
- productCode: Product being purchased.
- quantityOrdered, priceEach, and orderLineNumber: Order details and pricing.

1.3.2 Database Relationships

The Classic Models database is highly relational, with several foreign key constraints ensuring data integrity:

- **Products** are linked to **Product Lines** (productLine).
- **Orders** are associated with **Customers** (customerNumber).
- **Employees** are assigned to **Offices** (officeCode) and Customers (salesRepEmployeeNumber).
- **Payments** are linked to **Customers** (customerNumber).
- **Order Details** are tied to both **Orders** (orderNumber) and **Products** (productCode).

Chapter 2: Hypotheses Formulation

2.1 Sales Performance Analysis

Research Question:

Do sales orders differ significantly across different territories (office locations)?

Hypotheses:

H_0 (Null Hypothesis):
The average number of sales orders is the same across all territories.

H_1 (Alternative Hypothesis):
At least one territory has an average number of sales orders different from another.

Mathematical Representation:

H_0:

$$\mu_{\text{Territory}_1} = \mu_{\text{Territory}_2} = \cdots = \mu_{\text{Territory}_n}$$

H_1:

$$\mu_{\text{Territory}_i} \neq \mu_{\text{Territory}_j}, \quad \text{for at least one pair of territories } i \text{ and } j$$

Parameters of Interest:

Independent Variables: Territory (office location) – includes officeCode and officeCity.
Dependent Variables: Total sales orders per territory

Statistical Metrics:

p-value from ANOVA test.

Data Source:

Tables: offices, orders.

2.2 Product Pricing and Sales Frequency

Research Question:

Does the Manufacturer Suggested Retail Price (MSRP) impact sales frequency?

Hypotheses:

H_0 (Null Hypothesis):
There is no correlation between MSRP and sales frequency.

H_1 (Alternative Hypothesis):
There is a correlation between MSRP and sales frequency.

Mathematical Representation:

H_0:

$$\rho(\text{MSRP,Sales Frequency}) = 0$$

H_1:

$$\rho(\text{MSRP,Sales Frequency}) \neq 0$$

Where ρ is the correlation coefficient.

Parameters of Interest:

Independent Variables: MSRP
Dependent Variables: Sales frequency

Statistical Metrics:

Linear Regression:

- Tests for a linear relationship between MSRP and sales frequency.
- Provides R^2, P-value, and regression coefficients to assess the strength and significance of the relationship.

Spearman Correlation:

- Tests for a monotonic relationship between MSRP and sales frequency.
- Outputs the Spearman correlation coefficient (ρ) and associated P-value, which are robust to non-linear trends and outliers.

Data Source:

- Tables: products, orderdetails.

2.3 Product Line Revenue Contribution

Research Question:

Do product lines with more products generate more revenue?

Hypotheses:

- **H_0 (Null Hypothesis):**
  The number of products in a product line does not affect total revenue.

H_1 (Alternative Hypothesis):
The number of products in a product line has a significant effect on total revenue.

Mathematical Representation:

$$\text{Total Revenue} = \beta_0 + \beta_1(\text{Number of Products}) + \epsilon$$

Where:

- β_0: Intercept.
- β_1: Slope, representing the effect of the number of products on revenue.
- ϵ: Error term.

Parameters of Interest:

Independent Variables: Number of products in each product line
Dependent Variables: Total revenue for the product line

Statistical Metrics:

Regression metrics (e.g., slope, intercept, R^2).

Data Source:

Tables: productlines, products, orderdetails.

2.4 Customer Insights and Behavior Analysis

Research Question:

Are customers with higher credit limits more likely to place larger orders?

Additionally, can customers be segmented based on credit limits and average order size for targeted strategies?

Hypotheses:

H_0 (Null Hypothesis):
Credit limit has no effect on average order size.

H_1 (Alternative Hypothesis):
Credit limit significantly affects average order size.

Mathematical Representation (Regression Analysis):

$$\text{Average Order Size} = \beta_0 + \beta_1(\text{Credit Limit}) + \epsilon$$

Where:

- β_0: Intercept.
- β_1: Slope, representing the effect of credit limit on average order size.
- ϵ: Error term.

Parameters of Interest:

For Regression Analysis:

Independent Variable: Credit Limit (customers.creditLimit)
Dependent Variable: Average Order Size (AVG(quantityOrdered * priceEach))

For Clustering Analysis:

Variables used for clustering:
Credit Limit (customers.creditLimit)
Average Order Size (AVG(quantityOrdered * priceEach))

Statistical Metrics:

For Regression Analysis:

Regression slope (β_1): Assesses the relationship between credit limit and average order size.

P-value: Tests the significance of the slope (β_1).

For Clustering Analysis:

Cluster centroids: Represent the average values of credit limits and order sizes in each cluster.

Intra-cluster and inter-cluster distances: Measure the cohesion and separation of clusters.

Data Source:

Tables: customers, orders, orderdetails

Clustering Analysis Objective:

In addition to testing the relationship between credit limit and average order size through regression, clustering analysis segments customers into groups based on their behavior patterns. This approach identifies meaningful customer segments that can inform tailored marketing strategies and credit policies.

2.5 Forecasting Customer Purchasing Behavior

Research Question:

> Can customer purchasing behavior (e.g., order frequency, recency, tenure, and customer lifetime value) effectively predict the total purchase value using historical data?

Hypotheses:

> **H_0 (Null Hypothesis):**
> Predictive model accuracy (e.g., R^2, MAE, MSE) does not significantly differ from a baseline model.
>
> **H_1 (Alternative Hypothesis):**
> Predictive model accuracy is significantly better than a baseline model.

Mathematical Representation:

$$\text{Total Purchase Value} = \beta_0 + \beta_1(\text{Credit Limit}) + \beta_2(\text{Order Frequency}) + \beta_3(\text{Recency}) + \beta_4(\text{Tenure}) + \beta_5(\text{Customer Lifetime Value}) + \epsilon$$

Where:

β_0: Intercept.

$\beta_1, \beta_2, \beta_3, \beta_4, \beta_5$: Coefficients for respective variables.

ϵ: Error term.

Parameters of Interest:

Independent Variables:

> **Customer Demographics**: Country (customers.country), Assigned Sales Representative (customers.salesRepEmployeeNumber)
> **Credit Metrics**: Credit Limit (customers.creditLimit)

Order Metrics: Order Frequency (COUNT(orderNumber)), Total Purchase Value (SUM(quantityOrdered * priceEach)), Recency (DATEDIFF(Dataset Reference Date, MAX(orderDate))), Customer Tenure (DATEDIFF(Dataset Reference Date, MIN(orderDate)))
Customer Lifetime Value:

$$\text{CLV} = \frac{\text{Total Purchase Value}}{\text{Tenure (Days)}}$$

Dependent Variable:

Total Purchase Value (predicted future spending).

Justification:

Including Customer Lifetime Value (CLV) enhances the model's ability to predict future purchase behavior by capturing the efficiency of revenue generation over a customer's tenure. This complements other independent variables such as recency and frequency, ensuring the model comprehensively accounts for historical patterns.

Statistical Metrics:

Predictive model accuracy (e.g., R^2, MAE, MSE).

Data Source:

Tables: customers, orders, orderdetails.

Chapter 3: Data Collection, Validation, and Outputs

Objective

This chapter outlines the process of data collection, validation, and output generation to address the project's research questions. Data was extracted from the Classic Models database using SQL queries tailored to each research question. Validation steps were employed to ensure data accuracy and reliability, and the resulting datasets were stored in structured formats for further analysis.

3.1 Data Collection

Research Question 1: Do sales orders differ significantly across different territories (office locations)?

SQL Query:

```
USE classicmodels;

SELECT
  o.officeCode,
  o.city AS officeCity,
  COUNT(DISTINCT ord.orderNumber) AS totalOrders
FROM offices o
JOIN employees e ON o.officeCode = e.officeCode
JOIN customers c ON e.employeeNumber = c.salesRepEmployeeNumber
JOIN orders ord ON c.customerNumber = ord.customerNumber
GROUP BY o.officeCode, o.city
ORDER BY totalOrders DESC;
```

Purpose:
This query identifies the total number of distinct orders processed by each office, grouped by geographic location, to assess regional sales disparities.

Output Description:

- **Dataset Name:** territory_sales_data.csv
- **Columns:**
  - officeCode: Unique identifier for each office.
  - officeCity: City where the office is located.
  - totalOrders: Total distinct sales orders for the office.

Research Question 2: Does the Manufacturer Suggested Retail Price (MSRP) impact sales frequency?

SQL Query:

```
USE classicmodels;

SELECT
  p.productCode,
  p.productName,
  p.MSRP,
  COUNT(od.orderNumber) AS salesFrequency
FROM products p
JOIN orderdetails od ON p.productCode = od.productCode
GROUP BY p.productCode, p.productName, p.MSRP
ORDER BY salesFrequency DESC, MSRP ASC;
```

Purpose:
This query evaluates the relationship between MSRP and sales frequency to determine whether pricing affects purchase behavior.

Output Description:

- **Dataset Name:** msrp_sales_frequency_data.csv
- **Columns:**
    - productCode: Unique identifier for each product.
    - productName: Descriptive name of the product.
    - MSRP: Manufacturer Suggested Retail Price.
    - salesFrequency: Number of orders for the product.

Research Question 3: Do product lines with more products generate more revenue?

SQL Query:

```
USE classicmodels;

SELECT
  pl.productLine,
  COUNT(p.productCode) AS totalProducts,
  SUM(od.quantityOrdered * od.priceEach) AS totalRevenue
FROM productlines pl
JOIN products p ON pl.productLine = p.productLine
JOIN orderdetails od ON p.productCode = od.productCode
GROUP BY pl.productLine
ORDER BY totalRevenue DESC, totalProducts DESC;
```

Purpose:
This query calculates the total revenue and the number of products for each product line to determine whether a larger variety of products leads to higher revenue.

Output Description:

- **Dataset Name:** product_line_revenue_data.csv
- **Columns:**
    - productLine: Category grouping multiple products.
    - totalProducts: Number of products within the product line.
    - totalRevenue: Total revenue generated by all products in the product line.

Research Question 4: Are customers with higher credit limits more likely to place larger orders?

SQL Query:

```
USE classicmodels;

SELECT
  c.customerNumber,
  c.customerName,
  c.creditLimit,
  AVG(od.quantityOrdered * od.priceEach) AS avgOrderValue
FROM customers c
JOIN orders o ON c.customerNumber = o.customerNumber
JOIN orderdetails od ON o.orderNumber = od.orderNumber
GROUP BY c.customerNumber, c.customerName, c.creditLimit
ORDER BY avgOrderValue DESC, creditLimit DESC;
```

Purpose:
This query analyzes the relationship between customer credit limits and average order value to understand the influence of credit policies on purchasing behavior.

Output Description:

- **Dataset Name:** credit_limit_order_data.csv
- **Columns:**
  - customerNumber: Unique identifier for each customer.
  - customerName: Name of the customer.
  - creditLimit: Maximum credit allocated to the customer.
  - avgOrderValue: Average monetary value of orders placed by the customer.

Research Question 5: Can customer purchasing behavior (e.g., order frequency, recency, and tenure) effectively predict the total purchase value using historical data?

SQL Query:

```
USE classicmodels;

WITH OrderSummary AS (
  SELECT
    o.customerNumber,
    COUNT(o.orderNumber) AS totalOrders,
    MAX(o.orderDate) AS lastOrderDate,
    MIN(o.orderDate) AS firstOrderDate,
    SUM(od.quantityOrdered * od.priceEach) AS totalPurchaseValue
  FROM
```

```
    orders o
  LEFT JOIN
    orderdetails od ON o.orderNumber = od.orderNumber
  GROUP BY
    o.customerNumber
),
CustomerMetrics AS (
  SELECT
    os.customerNumber,
    os.totalOrders,
    os.lastOrderDate,
    os.firstOrderDate,
    os.totalPurchaseValue,
    DATEDIFF(
      (SELECT MAX(orderDate) FROM orders),
      os.lastOrderDate
    ) AS daysSinceLastOrder,
    DATEDIFF(
      (SELECT MAX(orderDate) FROM orders),
      os.firstOrderDate
    ) AS customerTenure,
    (os.totalPurchaseValue / DATEDIFF(os.lastOrderDate, os.firstOrderDate)) AS
customerLifetimeValue,
    (os.totalOrders / DATEDIFF(
      (SELECT MAX(orderDate) FROM orders),
      os.firstOrderDate
    )) AS orderFrequency
  FROM
    OrderSummary os
)
SELECT
  c.customerNumber,
  c.customerName,
  c.creditLimit,
  cm.totalOrders,
  cm.lastOrderDate,
  cm.firstOrderDate,
  cm.daysSinceLastOrder,
```

```
  cm.customerTenure,
  cm.customerLifetimeValue,
  cm.orderFrequency,
  cm.totalPurchaseValue
FROM
  customers c
LEFT JOIN
  CustomerMetrics cm ON c.customerNumber = cm.customerNumber
ORDER BY
  cm.totalPurchaseValue DESC;
```

Purpose:
The extracted metrics, such as orderFrequency and customerLifetimeValue, provide the foundation for building predictive models. These models aim to forecast total purchase value, supporting the identification of high-value customers.

Output Description:

- **Dataset Name:** customer_behavior_data.csv
- **Columns:**
  - customerNumber: Unique identifier for each customer.
  - customerName: Name of the customer.
  - creditLimit: Maximum credit allocated to the customer.
  - totalOrders: Total number of orders placed by the customer.
  - daysSinceLastOrder: Days since the customer's most recent order.
  - customerTenure: Total number of days since the customer's first order.
  - customerLifetimeValue: Average revenue generated per day of tenure.
  - orderFrequency: Average number of orders per day of tenure.
  - totalPurchaseValue: Total revenue generated by the customer.

3.2 Validation Steps

To ensure the accuracy and reliability of the datasets, the following validation steps were performed:

Foreign Key Integrity Checks: Verified that all productCode values in orderdetails exist in the products table.

```
SELECT od.productCode
FROM orderdetails od
```

```
LEFT JOIN products p ON od.productCode = p.productCode
WHERE p.productCode IS NULL;
```

Missing Data Checks: Checked for missing values in critical fields like creditLimit.

```
SELECT *
FROM customers
WHERE creditLimit IS NULL;
```

Range Validity Checks: Confirmed all monetary values are positive.

```
SELECT *
FROM products
WHERE MSRP < 0;
```

Duplicate Data Checks: Ensured no duplicate rows exist for customerNumber.

```
SELECT customerNumber, COUNT(*)
FROM customers
GROUP BY customerNumber
HAVING COUNT(*) > 1;
```

3.3 Outputs and Storage

Datasets:

- Territory Sales Data: territory_sales_data.csv
- MSRP Sales Frequency Data: msrp_sales_frequency_data.csv
- Product Line Revenue Data: product_line_revenue_data.csv
- Credit Limit Order Data: credit_limit_order_data.csv
- Customer Behavior Data: customer_behavior_data.csv

Format:
All datasets are stored as .csv files with appropriate headers and column descriptions to support further analysis and modeling.

Chapter 4: Data Understanding and Data Visualization

4.1 Sales Performance Analysis Across Territories

Research Question:

Do sales orders differ significantly across different territories (office locations)?

Objective:

To evaluate whether the mean number of sales orders varies significantly across office locations, providing insights into geographic disparities in sales performance.

4.1.1 Exploration and Analysis of Collected Data

The dataset for Research Question 1, extracted from the Classic Models database, includes the following key metrics:

- **Office Location (Territory):** Unique identifier and city name for each office.
- **Total Orders:** Total number of sales orders processed by each office.

Dataset Used: territory_sales_data.csv

- **Columns:** officeCode, officeCity, totalOrders

Key Observations from the Data:

Fig 4.1.1.1 Total Orders Per Region

Summary Statistics for Total Orders	
Mean	46.57142857
Median	39
Standard Deviation	26.21808069
Total Orders	326

4.1.2 Key Metrics, Trends, and Patterns

Key Metrics:

- **Highest Performing Region:** Paris (106 orders).
- **Lowest Performing Region:** Tokyo (16 orders).
- **Order Variability:** The standard deviation (26.22) shows considerable variability, with a range of 90 orders between the highest and lowest-performing regions.

Trends and Patterns:

Geographic Disparity:

- Paris outperforms other regions, suggesting concentrated demand or operational efficiency in this territory.

Underperformance in Tokyo:

- Tokyo's low order count suggests potential issues such as market demand or inefficient operations.

Clustered Performance in Middle Regions:

- Regions like San Francisco, London, NYC, and Sydney exhibit more consistent performance, ranging from 32 to 48 orders.

4.1.3 Statistical Analysis

Methodology:
To determine whether the mean number of sales orders differs significantly across offices, a Single-Factor ANOVA was performed. This analysis compares the variance between office means to the variance within groups.

ANOVA: Single Factor Analysis

Groups	Count	Sum	Average	Variance
Office Code	7	28	4.00	4.666666667
Total Orders	7	326	46.57142857	801.952381

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	6343.142857	1	6343.142857	15.72772891	0.001873559	4.747225347
Within Groups	4839.714286	12	403.3095238			
Total	11182.85714	13				

4.1.4 Interpretation of Statistical Findings

F-Statistic and P-value:

- The F-statistic ($F = 15.73$) is greater than the critical value ($F_{\text{crit}} = 4.75$), indicating significant differences between office means.
- The P-value ($P = 0.00187$) is less than the significance level ($\alpha = 0.05$), providing strong evidence to reject the null hypothesis.

Variability in Sales Orders:

- The variance in total orders processed (801.95) is substantially higher than the variance in office codes (4.67), highlighting disparities in sales performance across locations.

Descriptive Statistics:

- Average total orders: 46.57
- Variance within groups: 403.31

4.1.5 Addressing the Hypotheses

Hypotheses:

- **H_0 (Null Hypothesis):** There is no significant difference in the mean number of sales orders processed across different office locations.
- **H_1 (Alternative Hypothesis):** At least one office location processes a significantly different number of sales orders.

Conclusion: The null hypothesis (H_0) is rejected, confirming that sales orders differ significantly between office locations.

Findings:

- Offices contribute differently to overall sales performance, with some outperforming others significantly.
- Geographic disparities in sales performance suggest varying levels of market demand and operational efficiency.

4.1.6 Business Implications

Resource Allocation:

- High-performing offices can serve as benchmarks for best practices, and strategies can be replicated in underperforming regions.
- Additional resources, such as marketing budgets and staff, can be allocated to underperforming offices to boost sales.

Regional Strategy Development:

- Identifying the reasons for high performance in specific offices (e.g., market conditions, team efficiency) can guide future expansion strategies.
- Tailored regional strategies can address the unique challenges of low-performing territories.

Further Analysis:

- Investigate factors influencing performance, such as market size, competition, and local economic conditions.

4.2 Product Pricing and Sales Frequency

Research Question:

Does the Manufacturer Suggested Retail Price (MSRP) impact sales frequency?

Objective:

To evaluate the relationship between MSRP and sales frequency, determining whether pricing significantly affects customer purchasing behavior.

4.2.1 Exploration and Analysis of Collected Data

The dataset for Research Question 2, extracted from the Classic Models database, includes:

- **MSRP:** Manufacturer Suggested Retail Price of each product.
- **Sales Frequency:** Total number of sales orders associated with each product.

Dataset Used: msrp_sales_frequency_data.csv

- **Columns:** productCode, productName, MSRP, salesFrequency.

Key Observations:

- Products with MSRP values between $20 and $150 are the most commonly sold, while higher MSRP products have significantly fewer sales.
- Some extreme outliers exist in the data, particularly products with exceptionally high MSRP values.

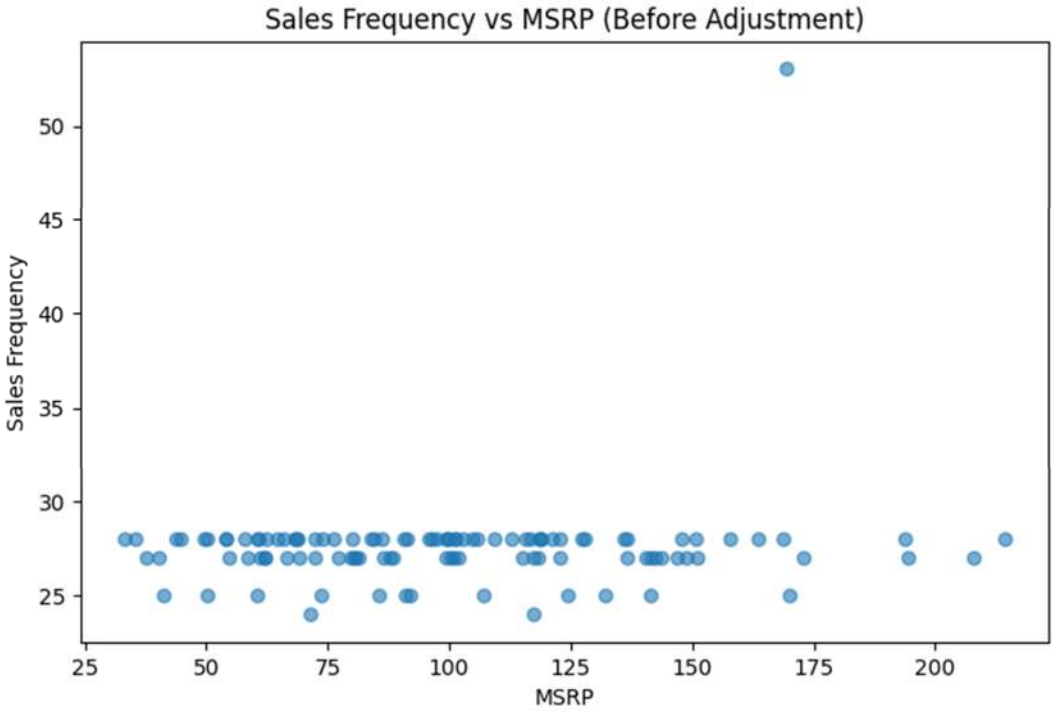

Fig 4.2.1.1 Sales Frequency VS MSRP (Before Adjustment)

4.2.2 Key Metrics, Trends, and Patterns

Key Metrics:

- **Range of MSRP Values:** $20–$200 (95% of products), with outliers above $200.
- **Sales Frequency:** Most products have fewer than 30 sales, with a small number of products exceeding 100 sales.

Trends and Patterns:

Clustering of Sales Frequencies:

- Most products cluster within specific MSRP ranges ($25–$150), with fewer products sold as MSRP increases.

Impact of Outliers:

- Outliers in MSRP (> $200) skew the data and may not reflect typical sales patterns.

No Clear Visual Trend:

- The scatter plot does not show a clear relationship between MSRP and sales frequency, suggesting a potential lack of linear correlation.

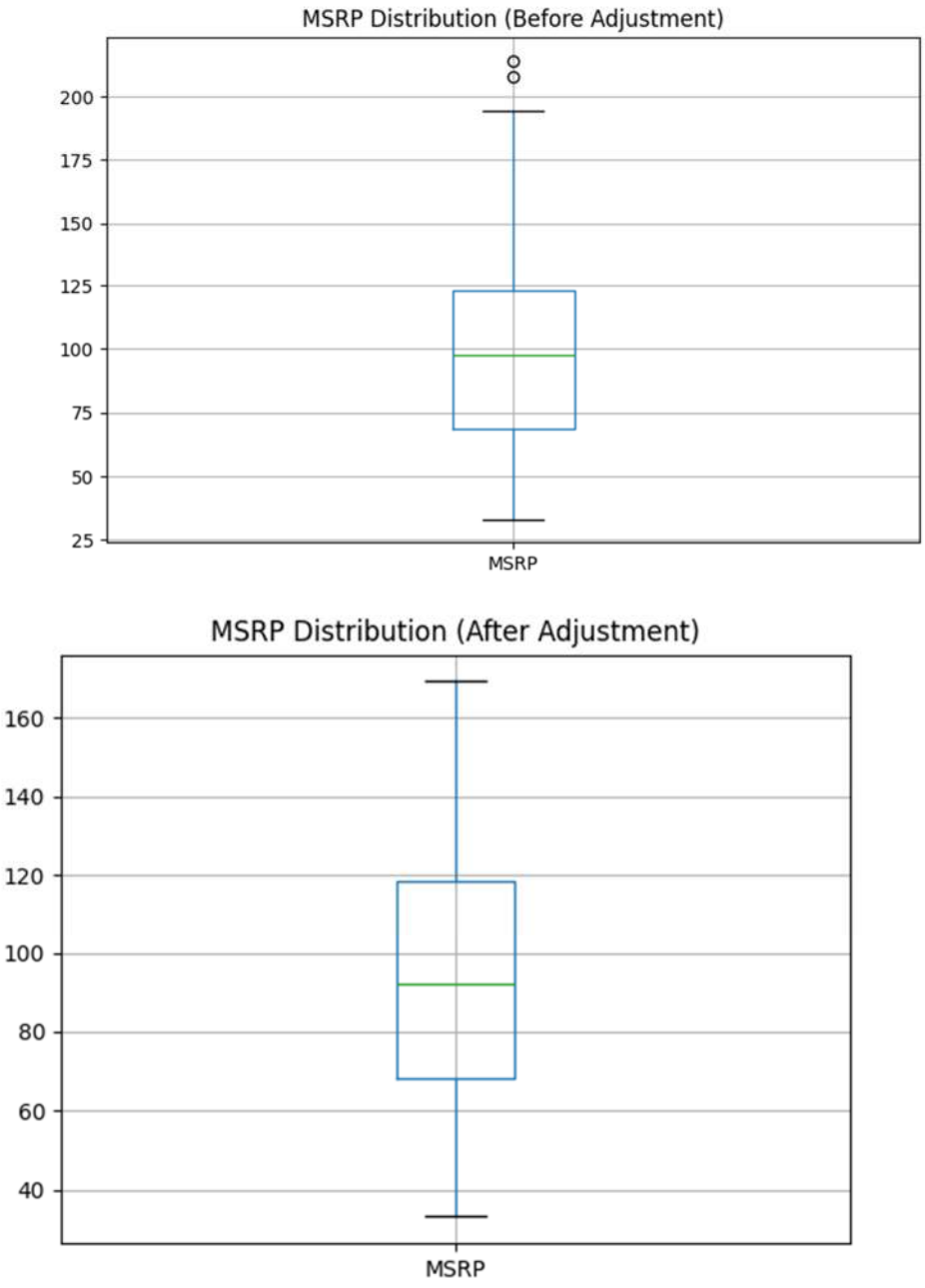

Fig 4.2.2.1 MSRP Distribution Before and After Adjustment

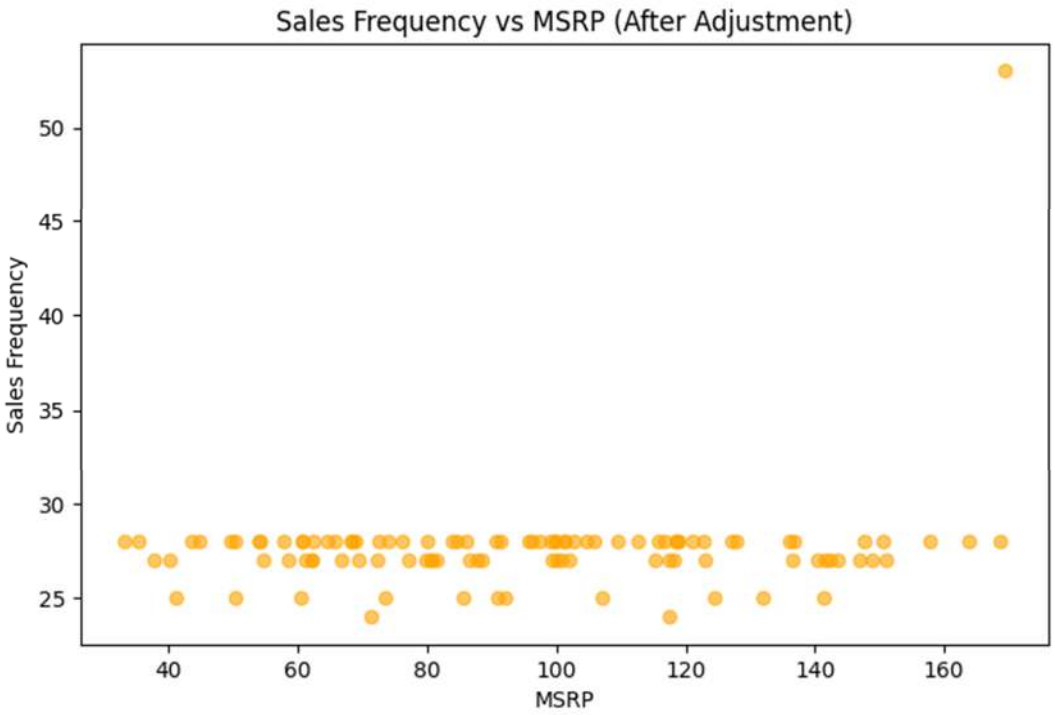

Fig 4.2.2.2 Sales Frequency vs MSRP (After Adjustment)

4.2.3 Statistical Analysis

Methodology:

Two statistical methods were used to analyze the relationship between MSRP and sales frequency:

Spearman Correlation:

- Evaluates monotonic relationships, robust to non-linear trends and outliers.

Linear Regression:

- Assesses the strength and direction of a linear relationship.

Spearman Correlation Results Summary:

Before Outlier Adjustment:

- **Spearman Correlation Coefficient (ρ):** -0.0123
- **P-value:** 0.8989 (not significant).

After Outlier Adjustment:

- **Spearman Correlation Coefficient (ρ):** 0.0237
- **P-value:** 0.8119 (not significant).

Linear Regression Statistical Results for MSRP vs. Sales Frequency:

Metric	Before Outlier Adjustment	After Outlier Adjustment
R^2	0.02293196390052621	0.0075
Slope (β)	0.01073	0.0009
P-value	0.11598 (not significant)	0.3957 (not significant)

4.2.4 Interpretation of Statistical Finding

Spearman Correlation:

Before Outlier Adjustment:

- A near-zero correlation ($\rho = -0.0123$) suggests no monotonic relationship between MSRP and sales frequency.

After Outlier Adjustment:

- Adjusting for outliers does not significantly change the results ($\rho = 0.0237$), confirming no monotonic trend.

Linear Regression:

Before Outlier Adjustment:

- The regression model explains only 2.2% of the variance in sales frequency ($R^2 = 0.02239$), and the P-value (0.11598) is not significant.

After Outlier Adjustment:

- Outlier removal further reduces the model's explanatory power ($R^2 = 0.0075$), and the P-value (0.3957) remains non-significant.

4.2.5 Addressing the Hypotheses

Hypotheses:

- **H_0 (Null Hypothesis):** There is no correlation between MSRP and sales frequency.
- **H_1 (Alternative Hypothesis):** There is a correlation between MSRP and sales frequency.

Conclusion:

- Based on the results of both Spearman correlation and linear regression, we fail to reject the null hypothesis (H_0).
- There is no statistically significant relationship between MSRP and sales frequency.

4.2.6 Business Implications

Pricing Strategies:

- MSRP does not appear to influence sales frequency, suggesting that other factors (e.g., product quality, brand reputation) may be more important determinants of sales.

Further Analysis:

- Future studies could examine the impact of additional variables, such as product lines or customer demographics, on sales frequency.

Insights for High-MSRP Products:

- High-MSRP products have limited sales; this could indicate a niche market or require adjustments in marketing strategies.

4.3 Product Line Revenue Analysis

Research Question:

Do product lines with more products generate more revenue?

Objective:

To examine the relationship between the total number of products sold and the revenue generated across product lines. The aim is to evaluate whether product lines with higher sales volumes contribute proportionally higher revenues.

4.3.1 Exploration and Analysis of Collected Data

The dataset for Research Question 3, extracted from the Classic Models database, includes the following key metrics:

Product Line: The product line category for each product.

Total Products Sold: Total number of units sold within each product line.

Total Revenue: Total revenue generated by all products in a given product line.

Dataset Used: product_line_revenue_data.csv

Columns: productLine, totalProducts, totalRevenue.

Key Observations:

Revenue Distribution by Product Line:

Classic Cars contribute the highest revenue, exceeding $4,000,000.

Trains generate the least revenue, highlighting disparities across product lines.

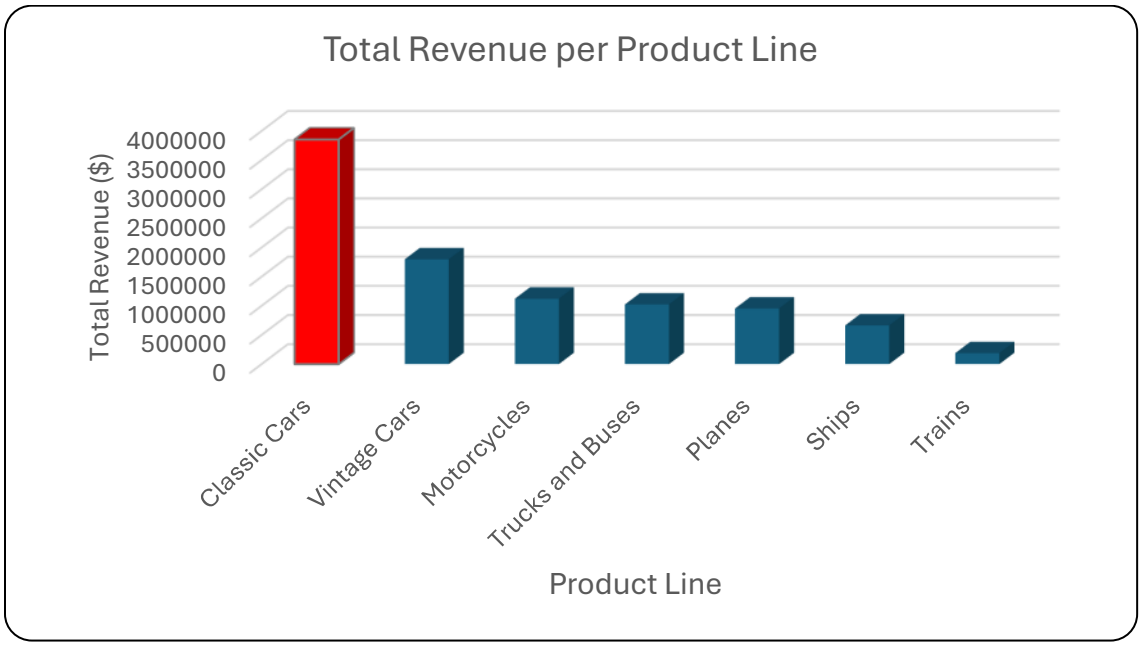

Fig 4.3.1.1 Total Revenue per Product Line

Scatter Plot of Total Products vs. Total Revenue:

The scatter plot indicates a strong positive linear relationship between the total number of products sold and revenue generated.

The regression equation $y = 3805.9x - 256898$ and an R^2 value of 0.9645 confirm the strength of the relationship.

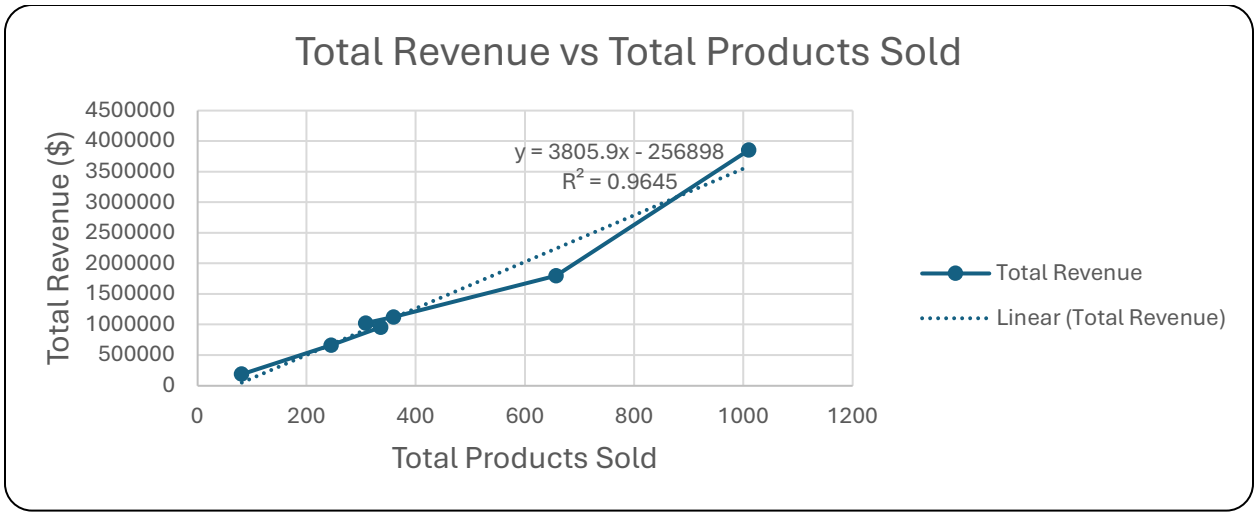

Fig 4.3.1.2 Total Revenue vs Total Products Sold

4.3.2 Statistical Analysis

Correlation Analysis:

Correlation Coefficient: 0.9821

Indicates a near-perfect positive relationship between the number of products sold and revenue.

Linear Regression Analysis:

Regression Summary:

R^2: 0.9645 (96.45% of variance in revenue is explained by the number of products sold).

P-value for Slope: 8.18×10^{-5}, indicating a statistically significant relationship.

Regression Equation:

$$\text{Revenue} = -256{,}898 + 3805.9 \times (\text{Total Products Sold})$$

For every additional product sold, revenue increases by $3,805.9 on average.

ANOVA Results:

F-statistic: 135.72

P-value: 8.18×10^{-5}

Indicates the model as a whole is statistically significant.

Regression Statistics	
Metric	**Value**
Multiple R	0.982074075
R Square	0.964469488
Adjusted R Square	0.957363386
Standard Error	247222.3867
Observations	7

ANOVA

Source	*df*	*SS*	*MS*	*F*	*Significance F*
Regression	1	8.2953E+12	8.29531E+12	135.7241	8.18417E-05
Residual	5	3.0559E+11	61118908493		
Total	6	8.6009E+12			

Coefficients Table

Variable	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	-256897.5824	168170.1381	-1.52760523	0.187145742	-689192.6847	175397.5198
X Variable 1	3805.899094	326.6846685	11.6500695	8.18417E-05	2966.12942	4645.668769

4.3.3 Model Diagnostics

Residual Plot:

Residuals are randomly distributed, supporting the assumption of homoscedasticity.
A single point near 1000 total products shows a larger residual, which may indicate potential leverage or influence.

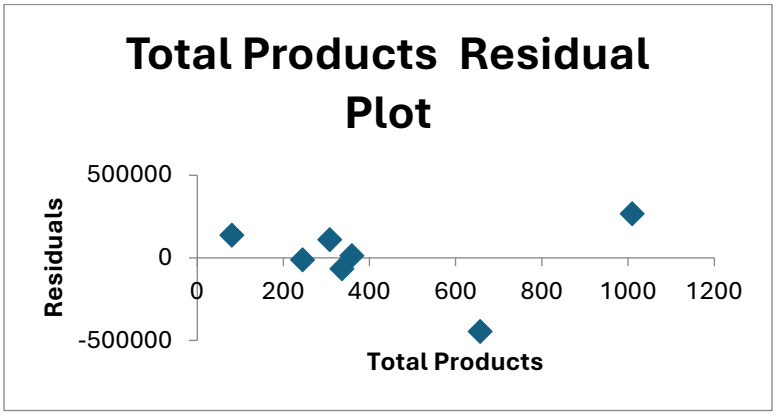

Fig 4.3.2.1 Total Products Residual Plot

Line Fit Plot:

Observed and predicted revenues align closely, confirming the high explanatory power of the regression model.
The close alignment further supports the $R^2 = 0.9645$ value.

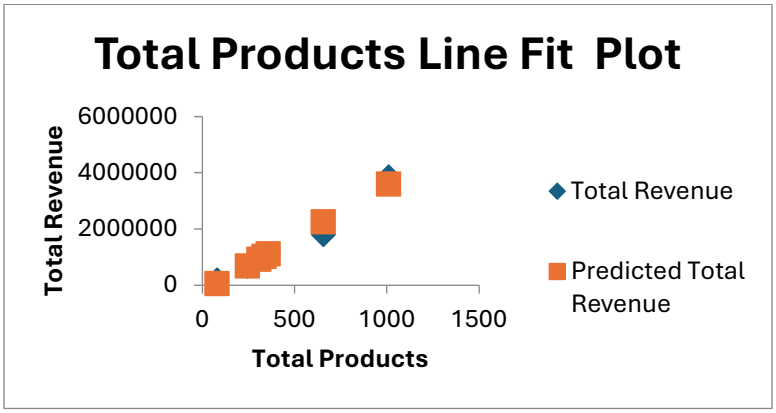

Fig 4.3.2.2 Total Products Line Fit Plot

Normal Probability Plot:

Points deviate slightly from the straight line, suggesting minor departures from normality in residuals.
While this does not significantly impact model reliability, further analysis might be needed for inferential conclusions.

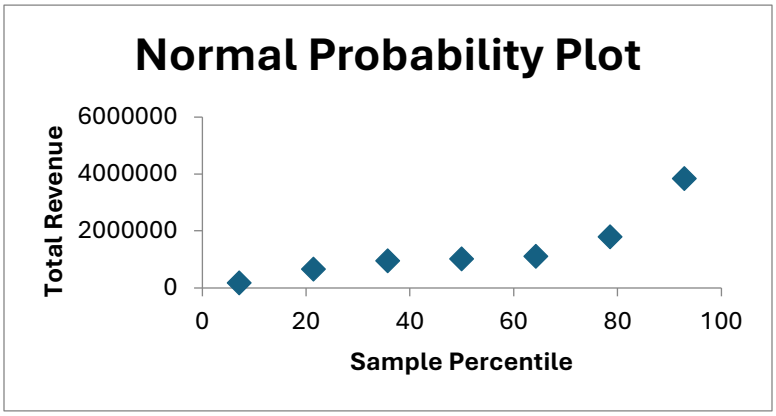

Fig 4.3.2.3 Normal Probability Plot

4.3.4 Interpretation of Findings

Revenue vs. Total Products:

A strong linear relationship exists between total products sold and revenue, supported by the R^2 value of 0.9645 and the significant regression slope ($P < 0.05$).

Product Line Insights:

Classic Cars dominate revenue generation, while Trains contribute the least.

Strategic scaling of high-performing product lines like Classic Cars could maximize revenue.

Diagnostic Findings:

Random residuals indicate a good model fit, though a single influential point may need further evaluation.

Minor deviations from normality in residuals suggest the need for careful interpretation of inferential statistics.

4.3.5 Addressing the Hypotheses

Hypotheses:

H_0 (Null Hypothesis): The number of products in a product line does not affect total revenue.

H_1 (Alternative Hypothesis): The number of products in a product line has a significant effect on total revenue.

Conclusion:

Based on the high R^2 value, significant P-value, and strong linear relationship, we reject the null hypothesis (H_0) and accept the alternative hypothesis (H_1).

The number of products in a product line significantly affects total revenue.

4.4 Credit Limit and Order Size Analysis

Research Question:

Are customers with higher credit limits more likely to place larger orders?

Objective:

The objective is to evaluate whether a customer's credit limit influences their average order size. This analysis seeks to determine if higher credit limits lead to larger orders, providing insights into customer purchasing behavior and credit policy effectiveness.

4.4.1 Exploration and Analysis of Collected Data

The dataset for Research Question 4, extracted from the Classic Models database, includes the following key metrics:

- **Credit Limit:** Maximum credit allocated to each customer.
- **Average Order Value:** Average value of orders placed by each customer.

Dataset Used: credit_limit_order_data.csv

- **Columns:** customerNumber, creditLimit, avgOrderValue.

Key Observations:

Bar Chart: Average Order Value vs. Credit Limit Segments:

- Customers with credit limits between $50,001 and $250,004 show higher average order values, while those with credit limits below $50,000 have the lowest.
- This observation suggests a potential influence of credit limits on order size.

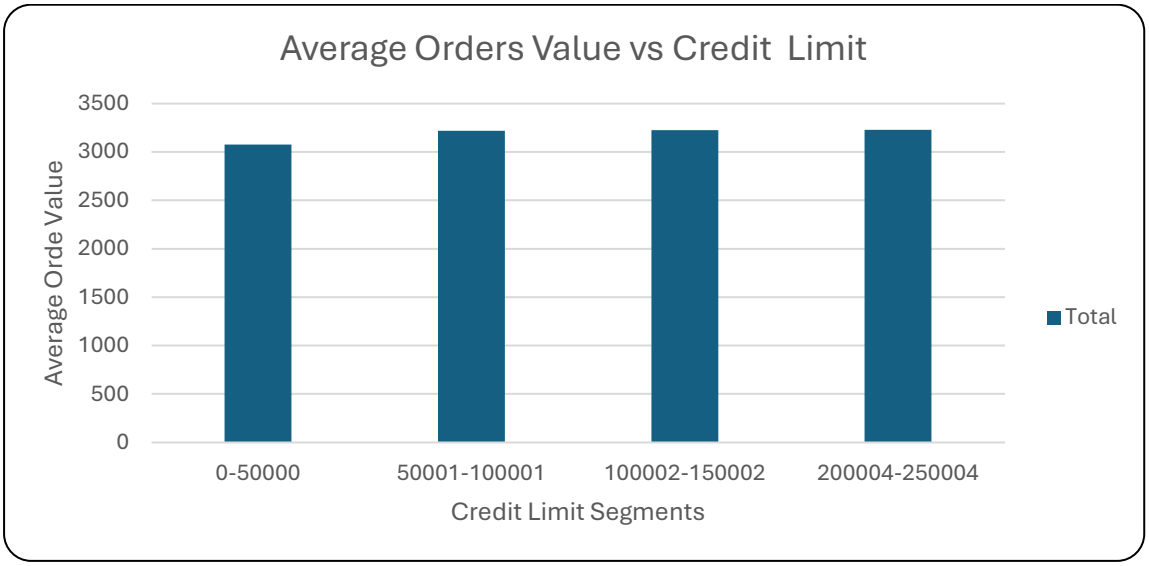

Fig 4.4.1.1 Average Orders Value vs Credit Limit

Scatter Plot: Average Order Value vs. Credit Limit:

- The scatter plot reveals no clear trend or strong relationship between credit limit and average order value.
- The regression line indicates an extremely weak linear relationship ($R^2 = 0.0075$).

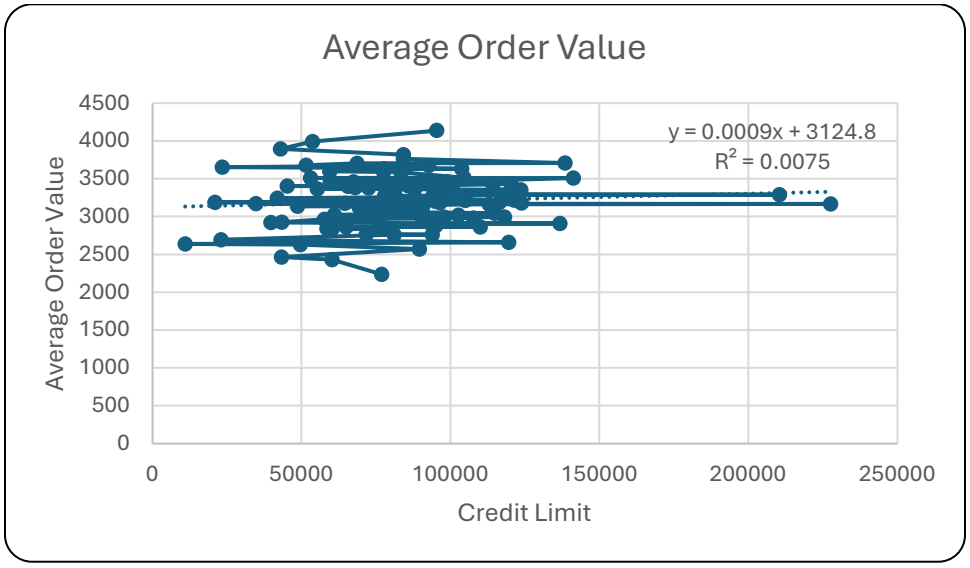

Fig 4.3.2.1 Average Order Value

4.4.2 Statistical Analysis

Correlation Analysis:

- **Correlation Coefficient (ρ):** 0.0868
  - Indicates a very weak positive relationship between credit limit and average order value.

Linear Regression Analysis:

Regression Summary:

- R^2: 0.0075 (only 0.75% of the variance in average order value is explained by credit limit).
- **P-value for Slope:** 0.3957 ($P > 0.05$), indicating no statistically significant relationship.

- **Regression Equation:**

  $$\text{Average Order Value} = 3124.8 + 0.0009 \times (\text{Credit Limit})$$

  - For every $1 increase in credit limit, the average order value increases by $0.0009 on average.

ANOVA Results:

- **F-statistic:** 0.7280 ($P > 0.05$), confirming that the model is not statistically significant.

Regression Statistics	
Multiple R	0.086754903
R Square	0.007526413
Adjusted R Square	-0.002811853
Standard Error	350.201639
Observations	98

ANOVA					
	Source	***df***	***SS***	***MS***	***Significance F***
Regression	1	89284.62474	89284.62	0.728015	0.395651255
Residual	96	11773554.04	122641.2		
Total	97	11862838.67			

	Coefficients	***Standard Error***	***t Stat***	***P-value***	***Lower 95%***	***Upper 95%***	***Lower 95.0%***	***Upper 95.0%***
Intercept	*3124.82875*	*96.19711018*	*32.4836*	*1.39E-53*	*2933.878996*	*3315.7785*	*2933.879*	*3315.778505*
X Variable 1	*0.000906196*	*0.001062067*	*0.853238*	*0.395651*	*-0.00120199*	*0.0030144*	*-0.001202*	*0.003014382*

4.4.3 Model Diagnostics

Residual Plot:

Residuals show random dispersion, indicating no strong patterns or heteroscedasticity.

However, the weak explanatory power of the model limits interpretability.

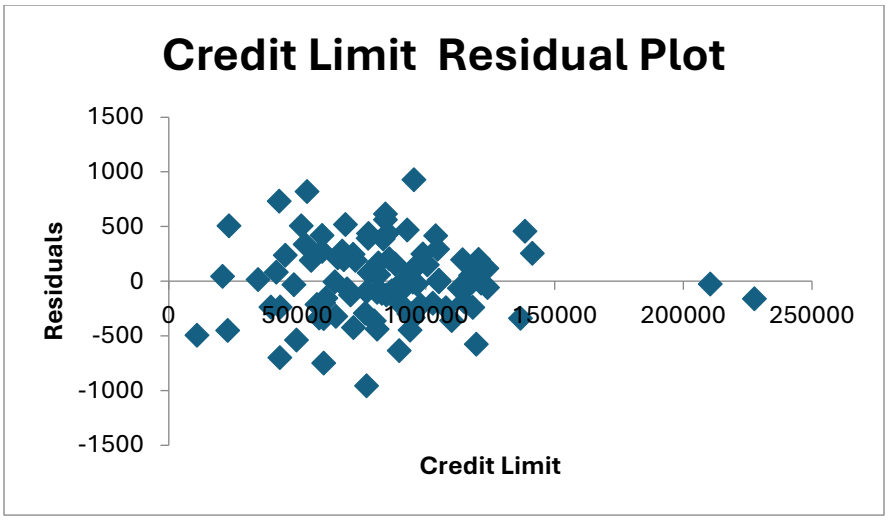

Fig 4.4.3.1 Credit Limit Residual Plot

Line Fit Plot:

Observed and predicted values show poor alignment, confirming the weak relationship between credit limit and average order value.

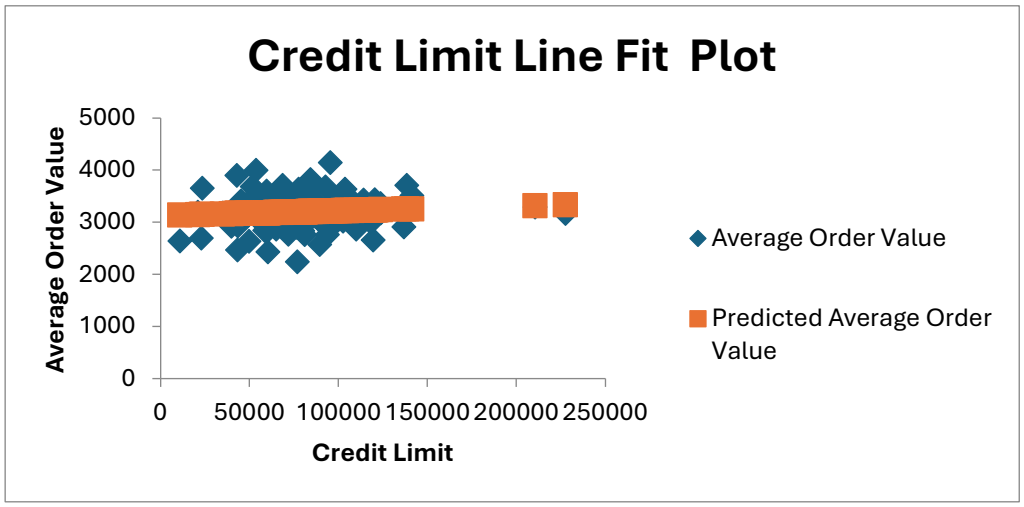

Fig 4.4.3.2 Credit Limit Line Fit Plot

Normal Probability Plot:

Residuals deviate from the straight line, suggesting a violation of the normality assumption.

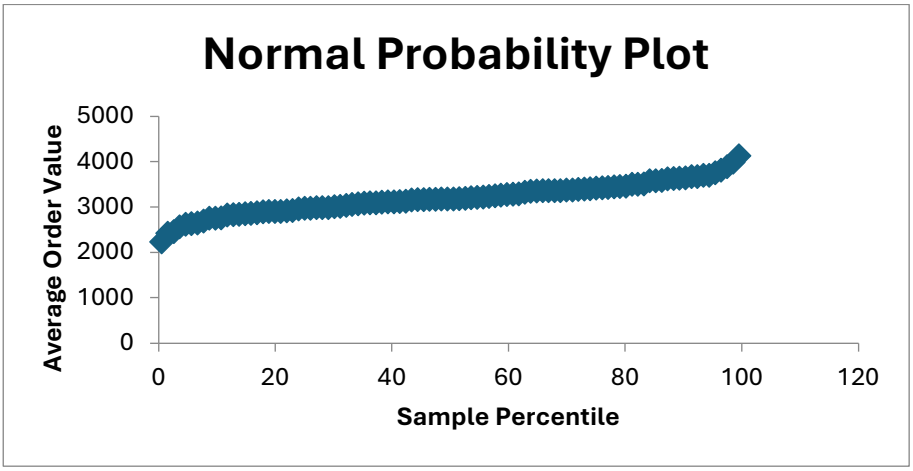

Fig 4.4.3.3 Normal Probability Plot

4.4.4 Interpretation of Findings

Credit limit and Average order value:

The bar chart suggests a potential relationship between credit limit and average order value for specific segments, but this is not supported by scatter plot trends or statistical metrics.

The regression model has negligible explanatory power ($R^2 = 0.0075$) and is not statistically significant ($P > 0.05$).

Diagnostic Findings:

Random residuals support model validity, but the overall weak explanatory power limits the utility of the findings.

Deviations from normality in residuals further weaken the case for linear regression

4.4.5 Addressing the Hypotheses

Hypotheses:

H_0 (Null Hypothesis): Credit limit has no effect on average order size.

H_1 (Alternative Hypothesis): Credit limit significantly affects average order size.

Conclusion:

Based on the extremely weak correlation ($\rho = 0.0868$), negligible R^2, and non-significant P-value ($P > 0.05$), we fail to reject the null hypothesis (H_0).

Credit limit does not significantly affect average order value.

4.4.6 Clustering Analysis

Methodology

The dataset was clustered into three distinct groups using the K-Means clustering algorithm. The clustering process focused on two key variables:

- Credit Limit: Indicates the financial capacity of customers.
- Average Order Value: Reflects the spending behavior of customers.

The Elbow Method and Silhouette Scores were used to determine the optimal number of clusters. Based on these results, three clusters were chosen, providing meaningful insights into customer segmentation.

Silhouette Score Validation

- The Silhouette Score was calculated to assess the quality of clustering. This metric measures how similar each data point is to its assigned cluster compared to other clusters, with values closer to 1 indicating better-defined clusters.
- The following scores were obtained:
    - For 3 clusters: **0.34**
    - For 4 clusters: **0.29**
- Based on these scores, the 3-cluster model was selected as it provided better-defined clusters with minimal overlap."

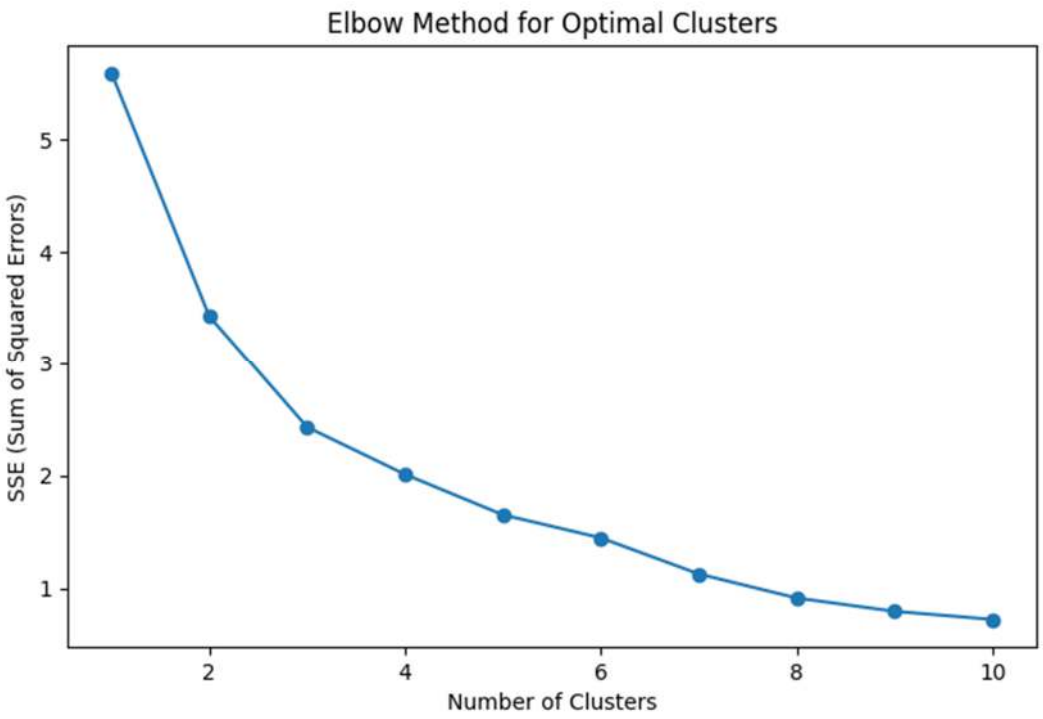

Fig 4.4.6.1: Elbow Method for Optimal Clusters

This figure shows the Sum of Squared Errors (SSE) for different numbers of clusters, with an elbow point at 3 clusters, indicating the optimal cluster count."

Cluster Insights

Cluster 0: High Credit, Moderate Order Value

- Average Credit Limit: High (~120,000)
- Average Order Value: Moderate (~3,200–3,500)

Insights: These are premium customers who use their credit capacity moderately.
Actionable Strategy:
- Implement loyalty programs.
- Encourage larger orders through targeted discounts.

Cluster 1: Low Credit, Low Order Value

- Average Credit Limit: Low (~60,000–70,000)
- Average Order Value: Low (~2,500–2,800)

Insights: These customers are likely value-conscious or infrequent buyers.
Actionable Strategy:

- Offer budget-friendly bundles.
- Introduce introductory offers to encourage spending.

Cluster 2: Moderate Credit, High Order Value

- Average Credit Limit: Moderate (~70,000–90,000)
- Average Order Value: High (~3,800–4,000+)

Insights: These are high-value customers with strong purchasing potential.
Actionable Strategy:
- Focus on upselling premium products.
- Provide personalized offers to retain these high-value customers.

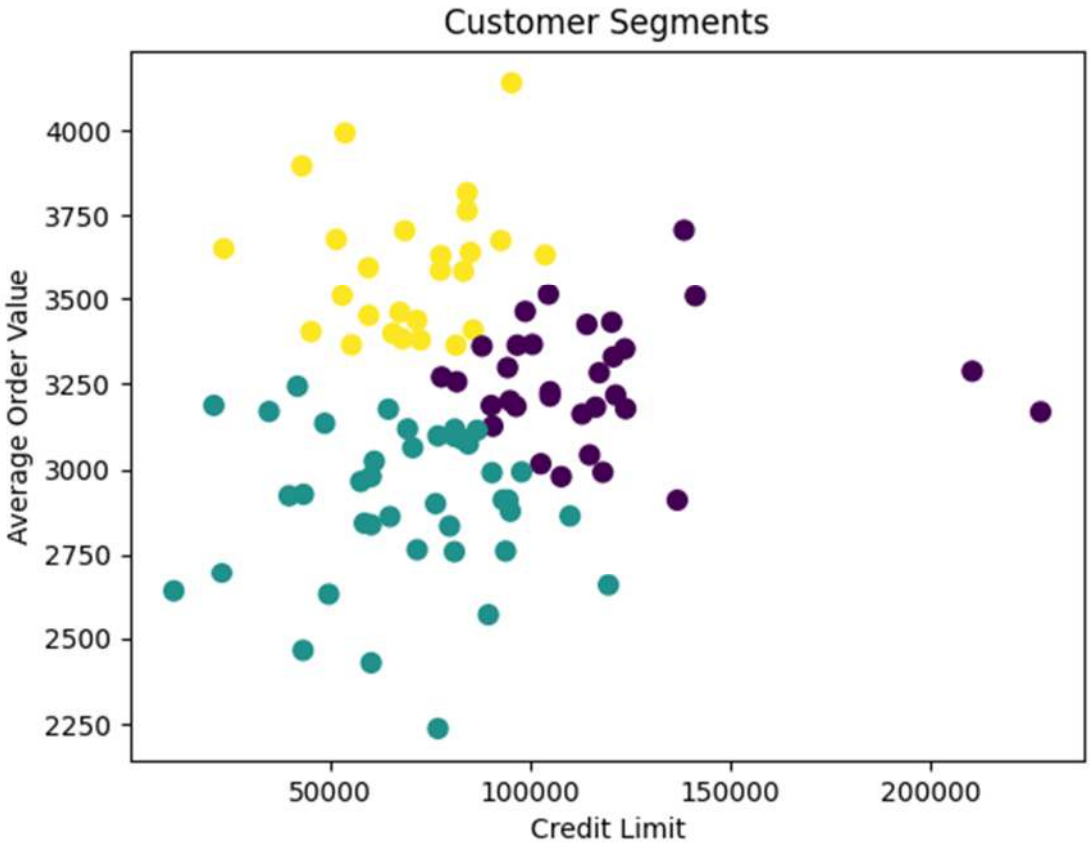

Fig 4.4.4.6.2: Customer Segments Based on Clustering

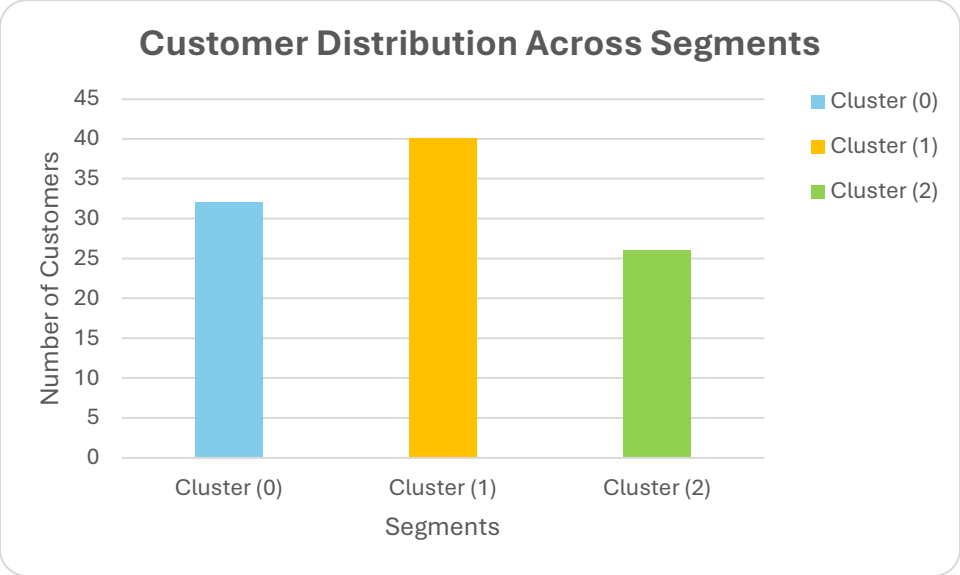

Fig. 4.4.6.3 Customer Distribution Across Segments – Bar Chart

This bar chart illustrates the distribution of customers across the three identified clusters. Cluster 1 (orange) has the largest segment of customers, followed by Cluster 0 (blue) and Cluster 2 (green)

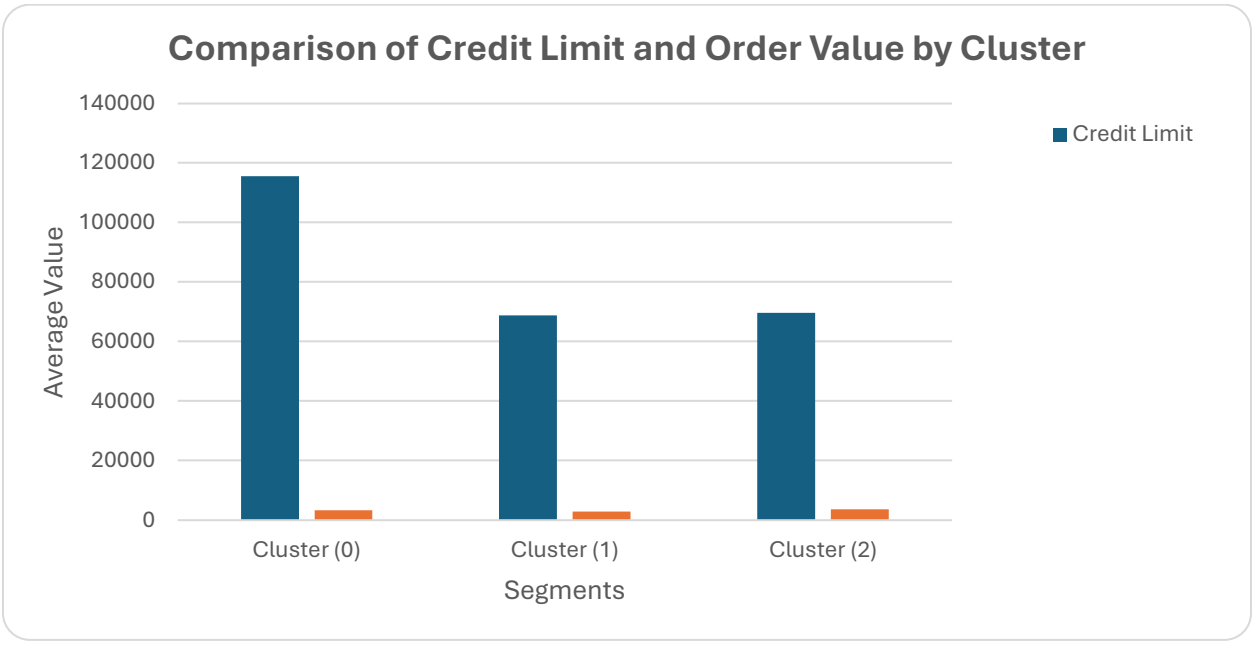

Figure 4.4.6.4 Comparison of Credit Limit and Order Value by Cluster – Bar Chart

This bar chart compares the average credit limit and order value across the three identified clusters. Cluster 0 has the highest average credit limit, whereas Cluster 2 exhibits the highest average order value, highlighting different customer behaviors and financial profiles.

4.5 Product Line Revenue Analysis

Research Question:

Can customer purchasing behavior (e.g., order frequency, recency, and tenure) effectively predict the total purchase value using historical data?

Objective:

To examine the relationship between the relationships between Total Purchase Value (target variable) and various customer behavior metrics, aiming to identify the most relevant predictors for modeling.

4.5.1 Variables of Interest

The dataset comprises the following variables:

Variable	Type	Description
Total Purchase Value	Dependent	Total monetary value of all purchases made by a customer (target variable).
Credit Limit	Independent	Maximum credit allocated to a customer.
Total Orders	Independent	Total number of orders placed by a customer.
Days Since Last Order	Independent	Number of days since the customer's most recent purchase.
Customer Tenure	Independent	Duration (in days) since the customer's first recorded order.
Customer Lifetime Value	Independent	Average purchase value per day of tenure.
Order Frequency	Independent	Average number of orders placed per day over the customer's tenure.

4.5.2 Summary Statistics

Summary statistics describe the central tendency, variability, and range for each variable.

Variable	Mean	Median	Min	Max	Standard Deviation
Total Purchase Value	98,001.95	81,181.64	7,918.60	820,689.54	96,437.98
Credit Limit	84,228.57	83,150	11,000	227,600	33,308.36
Total Orders	30.57	26	3	259	30.03
Days Since Last Order	182.78	183	0	547	138.28
Customer Tenure	610.12	599.5	258	876	153.87
Customer Lifetime Value	164.02	151.06	13.05	964.38	126.56
Order Frequency	0.0512	0.0445	0.0049	0.3043	0.0393

Key Insights:

Total Purchase Value:

A wide range from $7{,}918.60 to$820,689.54 indicates significant variation in customer spending behavior.

The high standard deviation ($96,437.98) suggests the presence of outliers.

Order Behavior:

Total Orders and Order Frequency have high variability, reflecting diverse purchasing habits across customers.

Credit Limit: A substantial range ($11{,}000 to$227,600) indicates that customers belong to different financial tiers.

4.5.3 Distribution of Total Purchase Value

A histogram of Total Purchase Value reveals the spending distribution across customers.

Key Insights:

Skewness:

The distribution is right-skewed, with most customers clustered in the $7{,}918.60–81,918.60 range.

A small group of high-value customers significantly increases the maximum purchase value.

Customer Segments:

Two distinct segments emerge: a majority of low-to-moderate spenders and a small segment of high-value customers.

Implications:

Skewness may affect linear modeling. Data transformations (e.g., logarithmic scaling) could improve model performance.

4.5.4 Correlation Analysis Revenue vs. Total Products:

Variable	Total Purchase Value	Credit Limit	Days Since Last Order	Customer Tenure	Customer Lifetime Value	Order Frequency
Total Purchase Value	1.00	0.77	-0.34	0.26	0.90	0.90
Credit Limit	0.77	1.00	-0.18	0.08	0.78	0.78
Days Since Last Order	-0.34	-0.18	1.00	-0.09	-0.39	-0.38
Customer Tenure	0.26	0.08	-0.09	1.00	-0.11	-0.11
Customer Lifetime Value	0.90	0.78	-0.39	-0.11	1.00	0.99
Order Frequency	0.90	0.78	-0.38	-0.11	0.99	1.00

Key Insights:

Strong Predictors:

Total Orders, Customer Lifetime Value, and Order Frequency show strong positive correlations ($r \geq 0.90$), confirming their importance for predicting purchase value.

Moderate Predictor:

Credit Limit ($r = 0.77$) suggests financial capacity plays a role in total spending.

Weak Predictors:

Days Since Last Order and Customer Tenure exhibit weak correlations, indicating limited standalone predictive power.

4.5.5 Scatterplot Analysis

Scatterplots provide further insights into the relationships between **Total Purchase Value** and its predictors:

(a) Total Purchase Value vs. Total Orders

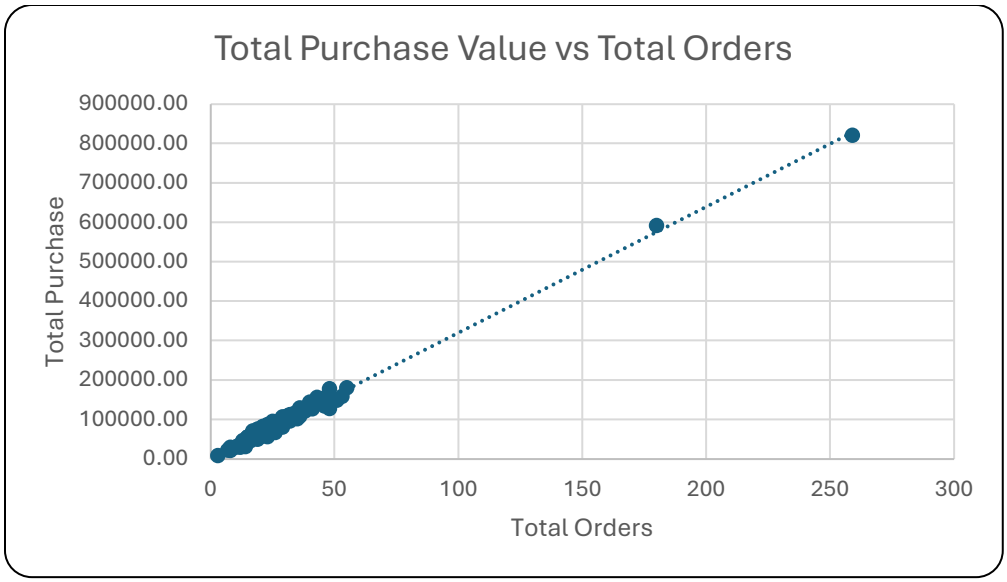

Fig 4.5.5.1 Total Purchase Value vs Total Orders

- **Observation**: A perfect linear relationship ($r = 1.00$) highlights that customers placing more orders generate higher purchase values.
- **Implication**: Total Orders is a direct and critical predictor.

(b) Total Purchase Value vs. Credit Limit

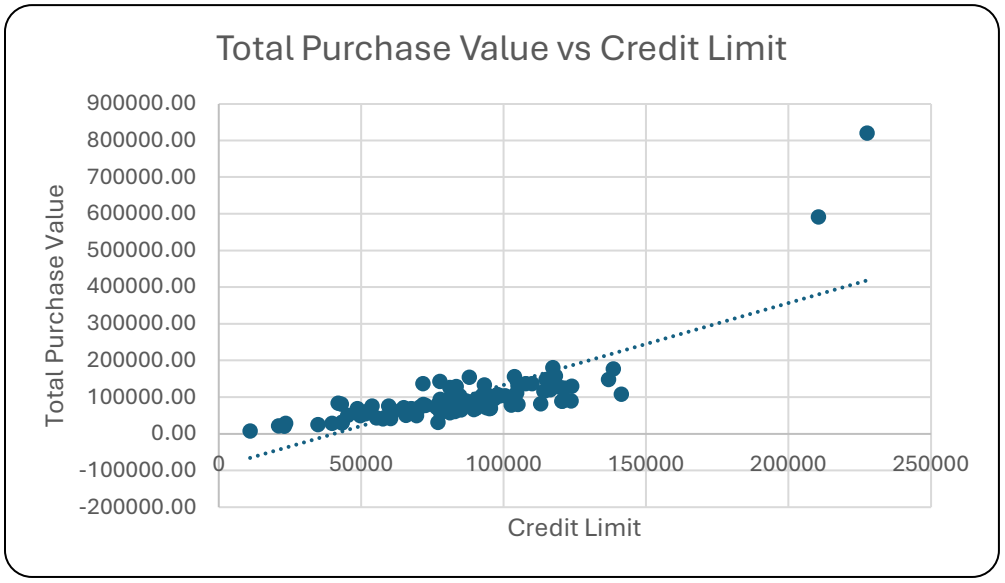

Fig 4.5.5.2 Total Purchase Value vs Total Credit Limit

- **Observation**: A moderate positive trend ($r = 0.77$) suggests that higher credit limits support higher spending.
- **Implication**: While important, Credit Limit does not fully explain purchase value variability.

(c) Total Purchase Value vs. Customer Lifetime Value

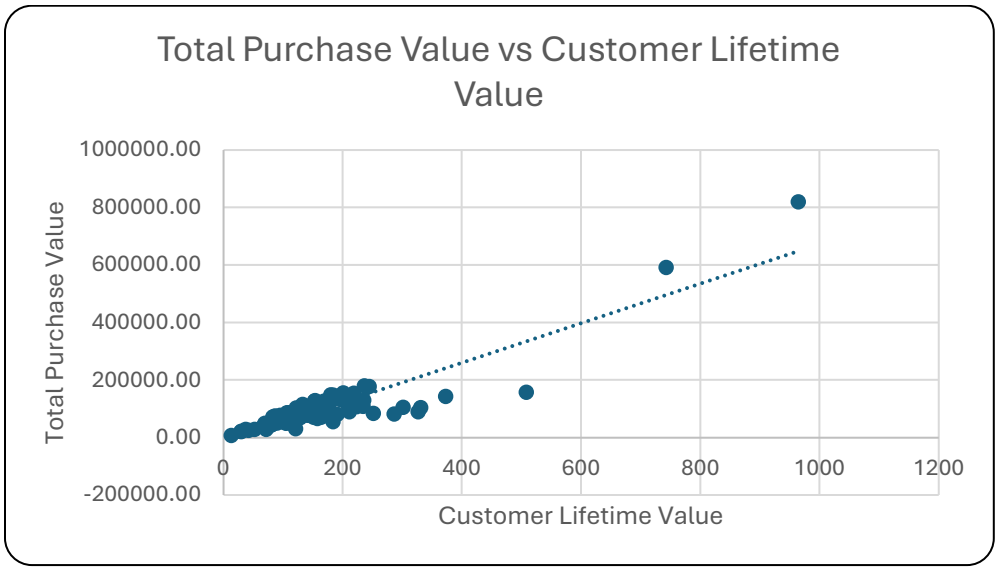

Fig 4.5.5.3 Total Purchase Value vs Customer Lifetime Value

- **Observation**: A strong positive correlation ($r = 0.90$) confirms that customers with higher lifetime value contribute more to total purchases.
- **Implication**: Customer Lifetime Value captures spending efficiency over time.

(d) Total Purchase Value vs. Days Since Last Order

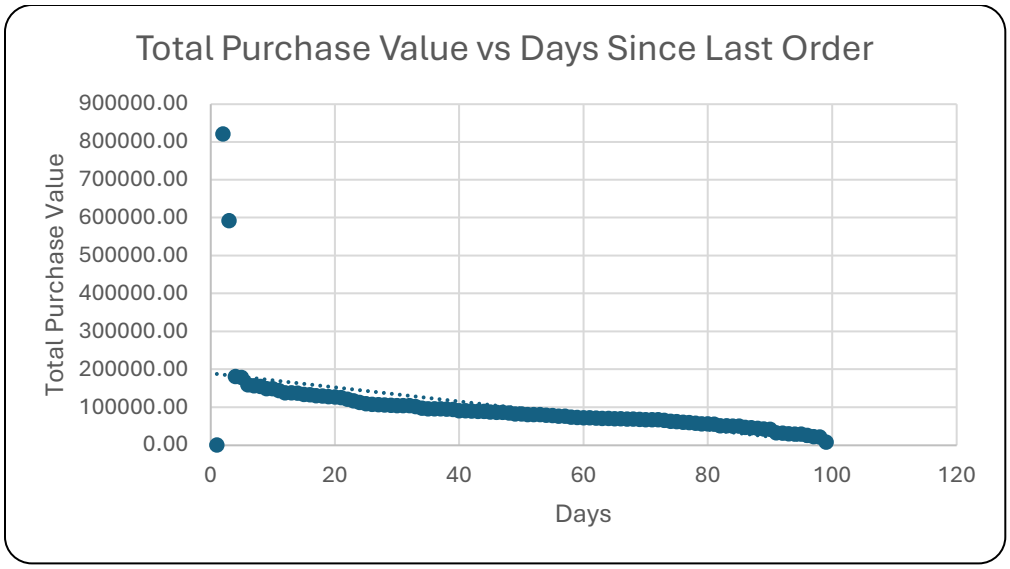

Fig 4.5.5.4 Total Purchase Value vs Days Since Last Order

- **Observation**: A weak negative relationship ($r = -0.34$) indicates that customers who recently purchased tend to have higher purchase values.
- **Implication**: Recency may influence specific customer segments but is not a strong predictor overall.

(e) Total Purchase Value vs. Customer Tenure

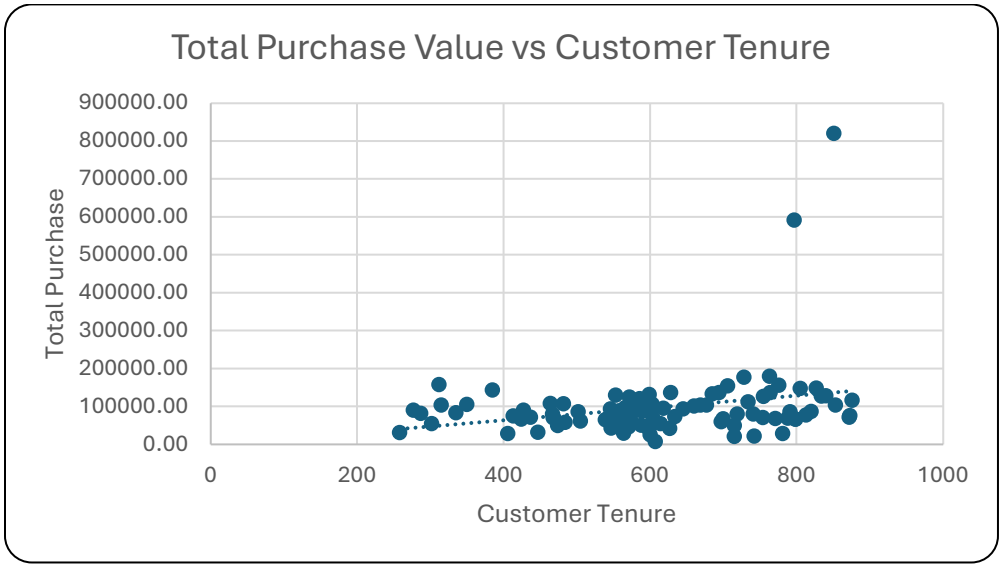

Fig 4.5.5.5 Total Purchase Value vs Customer Tenure

- **Observation**: A weak positive trend ($r = 0.26$) suggests longer-tenured customers spend slightly more.
- **Implication**: Tenure alone has limited predictive value.

(f) Total Purchase Value vs. Order Frequency

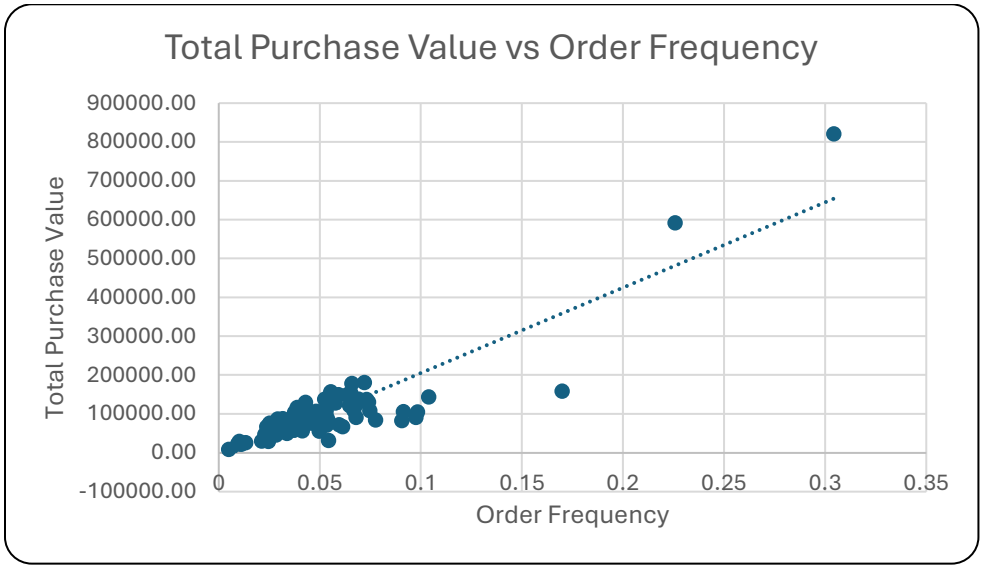

Fig 4.5.5.6 Total Purchase Value vs Order Frequency

- **Observation**: A strong positive relationship ($r = 0.90$) shows that frequent buyers contribute disproportionately to total purchases.
- **Implication**: Order Frequency is a reliable indicator of customer engagement and spending.

4.5.6 Relevance to Hypothesis

The findings strongly support the hypothesis that customer purchasing behavior metrics can predict **Total Purchase Value**:

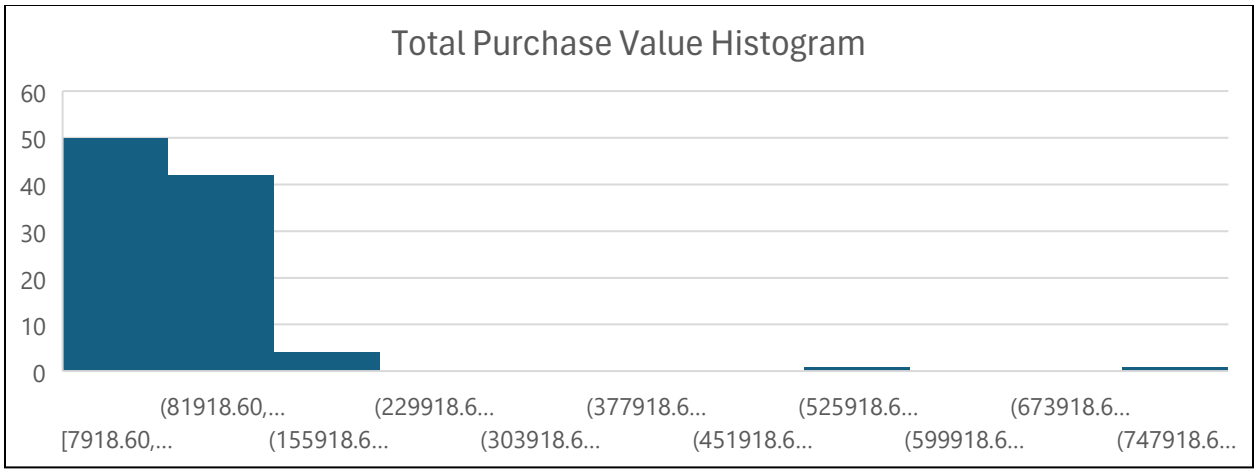

Fig 4.5.6.1 Total Purchase Value Histogram

Key Predictors:

- Variables such as Total Orders, Customer Lifetime Value, and Order Frequency will play a pivotal role in model building.

Hypothesis Alignment:

The analysis aligns with the alternative hypothesis (H_1) that a predictive model using these variables will outperform a baseline model.

Chapter 5: Model Building

5.1 Objective

This chapter documents the process of building and evaluating predictive models to forecast **Total Purchase Value**. The methodology includes regression analysis, handling multicollinearity, addressing outliers, normalizing variables, and conducting diagnostics to ensure robust model performance. The chapter outlines the development of eight regression models and explains the rationale behind selecting the final two models for validation in Chapter 6.

5.2 Data Preparation

5.2.1 Dataset Overview

The dataset consists of customer-level metrics capturing both behavioral and demographic characteristics. It includes one dependent variable, Total Purchase Value, and multiple independent variables such as Credit Limit, Days Since Last Order, and Customer Lifetime Value (CLV). These variables are transformed and normalized to ensure compatibility with the modeling process.

Variable Type	Original Variable	Transformed Variable Used in Model	Purpose
Dependent Variable	Total Purchase Value	Total Purchase Value	Represents the total spending by each customer.
Independent Variables	Credit Limit	CLNorm (Normalized Credit Limit)	Standardized to ensure consistency in scale across predictors.
	Days Since Last Order	DSLONorm (Normalized Days Since Last Order)	Captures the recency of customer interactions.
	Customer Tenure	CTNorm (Normalized Customer Tenure)	Measures the length of the customer relationship.
	Customer Lifetime Value (CLV)	CLVNorm (Normalized CLV)	Reflects the overall value of the customer to the business.
	Order Frequency	Order Frequency	Captures the transaction frequency of the customer.

The dataset was preprocessed to address inconsistencies and prepare it for robust model building. This preprocessing included handling outliers and normalizing variables.

5.2.2 Outlier Detection and Removal

Method: Interquartile Range (IQR). Outliers were removed to minimize their disproportionate impact on regression results.

Outcome:
Before removal, the dataset included 98 observations. After removing outliers, the dataset was reduced to 94 observations.
Removing outliers minimized their disproportionate influence on the model.

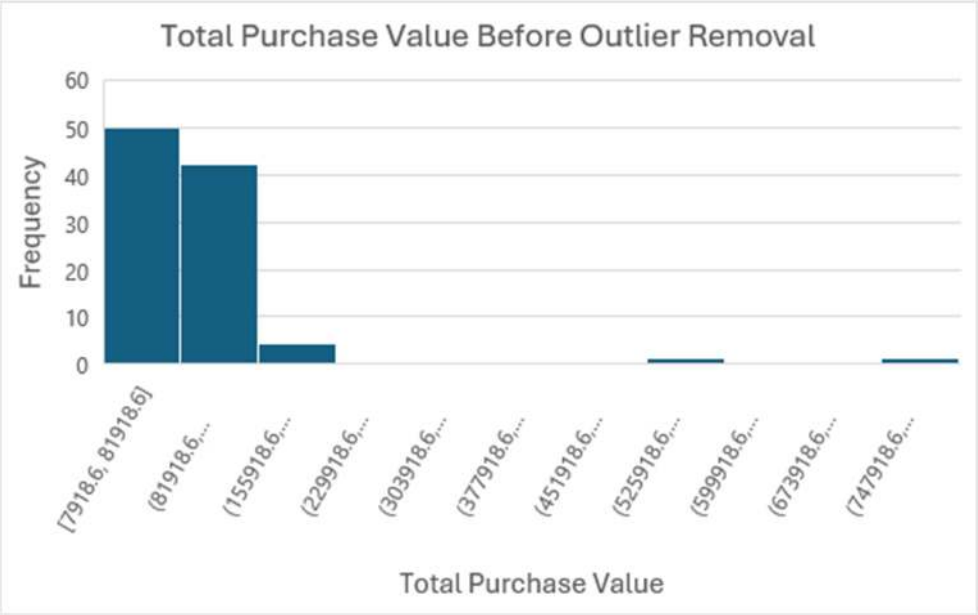

Fig 5.1.2.1 Histogram - Total Purchase Value Before Outlier Removal

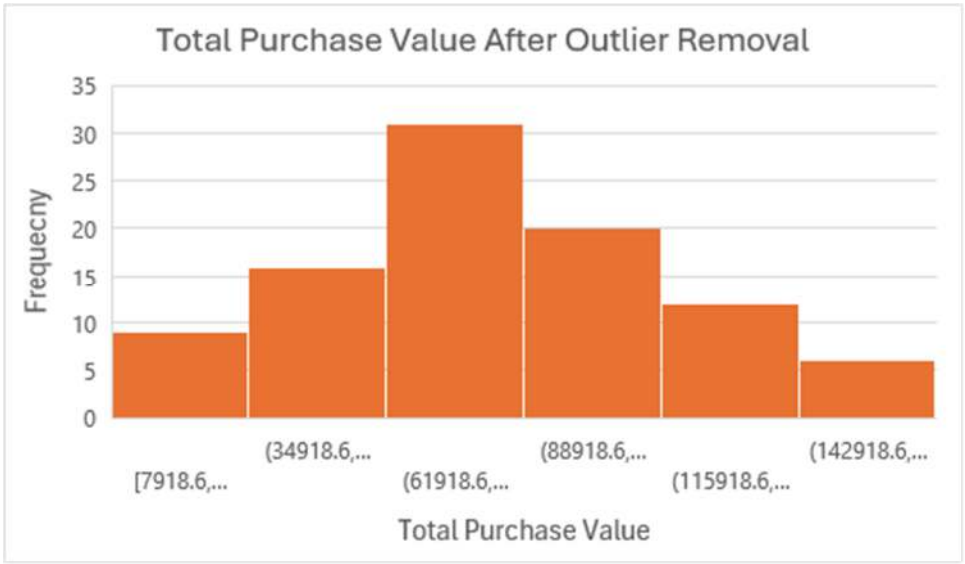

Fig 5.1.2.2 Histogram - Total Purchase Value After Outlier Removal

5.2.3 Normalization of Independent Variables

Normalization was applied to the independent variables to bring them onto a common scale, reducing potential biases in the modeling process. Min-max normalization was chosen, as it adjusts the range of each variable to [0, 1].

Steps:

For each variable, calculate:

$$x_{\text{normalized}} = \frac{x - x_{\min}}{x_{\max} - x_{\min}}$$

Replace original values with normalized values.

Variables Normalized:

Credit Limit (CLNorm)
Days Since Last Order (DSLONorm)
Customer Tenure (CTNorm)
Customer Lifetime Value (CLVNorm)

Variable	Original Min	Original Max	Normalized Min	Normalized Max
Credit Limit	11000	227600	0	1
Days Since Last Order	0	547	0	1
Customer Tenure	258	876	0	1
Customer Lifetime Value	13.04547	964.382538	0	1

5.3 Multicollinearity Assessment Using VIF

5.3.1 Methodology

Multicollinearity among independent variables can distort regression coefficients, making them unstable and difficult to interpret. Variance Inflation Factor (VIF) was calculated for each predictor to diagnose multicollinearity. The initial regression model used for this analysis is as follows:

$$TPV = \beta_0 + \beta_1(\text{CLNorm}) + \beta_2(\text{DSLONorm}) + \beta_3(\text{CTNorm}) + \beta_4(\text{CLVNorm}) + \beta_5(\text{Order Frequency}) + \epsilon$$

Where:

TPV: Total Purchase Value (dependent variable).

β_0: Intercept of the regression equation.

$\beta_1, \beta_2, \beta_3, \beta_4, \beta_5$: Coefficients for the respective predictors.

CLNorm: Normalized Credit Limit.

DSLONorm: Normalized Days Since Last Order.

CTNorm: Normalized Customer Tenure.

CLVNorm: Normalized Customer Lifetime Value.

Order Frequency: Transaction frequency of the customer.

ϵ: Error term representing unexplained variation.

This formula incorporates all available predictors to estimate Total Purchase Value. Multicollinearity was diagnosed using VIF to ensure the stability and interpretability of this model.

Steps:

- Fit a regression model for each predictor as a function of the remaining predictors.
- Compute the VIF for each predictor as:

$$\text{VIF} = \frac{1}{1 - R^2}$$

where R^2 is the coefficient of determination for the regression of that predictor.

- Identify variables with VIF > 10, which indicates significant multicollinearity.
- Iteratively remove the variable with the highest VIF, recalculating VIF values for the remaining variables.

5.3.2 Results of VIF Analysis

Variance Inflation Factor (VIF) analysis was conducted to identify and remove multicollinear variables. Variables with VIF > 10 were iteratively removed.

Variable	Initial VIF	Action Taken	Final VIF
Order Frequency	109.45	Removed	-
CLNorm	17.59	Removed	-
Customer Tenure (CTNorm)	2.77	Retained	2.77
Days Since Last Order (DSLONorm)	2.14	Retained	2.14
Customer Lifetime Value (CLVNorm)	1.98	Retained	1.98

Observation: All remaining variables now have VIF values below the threshold of 10, indicating that multicollinearity has been effectively resolved.

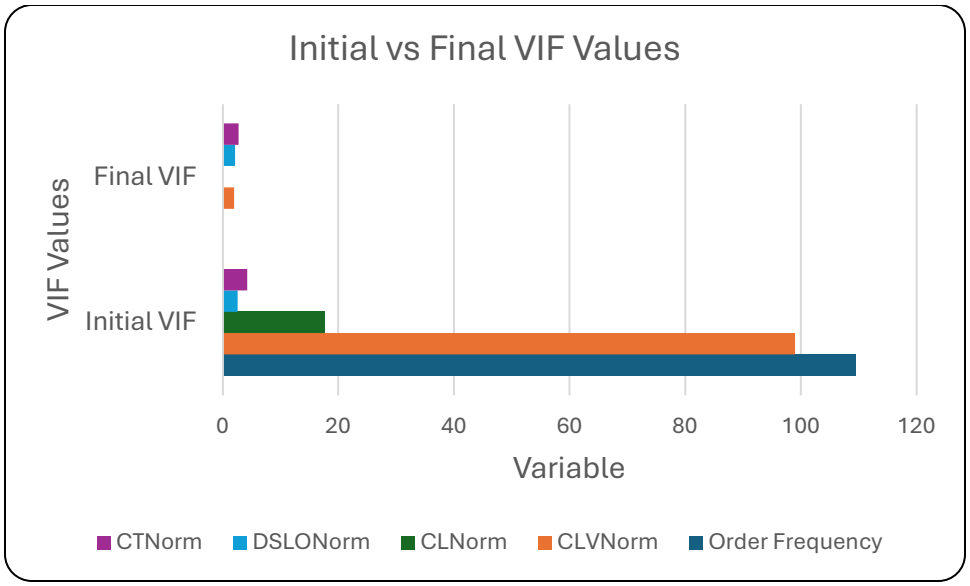

Fig 5.3.2.1 Bar Chart – Initial vs Final VIF Values

5.4 Regression Results and Variable Selection

5.4.1 Refined Model Formula Include the refined regression model here, with the rationale for selecting variables:

"The final regression model, after addressing multicollinearity, is as follows:

$$TPV = \beta_0 + \beta_1(\text{DSLONorm}) + \beta_2(\text{CTNorm}) + \beta_3(\text{CLVNorm}) + \epsilon$$

Where:

TPV: Total Purchase Value (dependent variable).

$DSLONorm$: Normalized Days Since Last Order.

$CTNorm$: Normalized Customer Tenure.

$CLVNorm$: Normalized Customer Lifetime Value.

ϵ: Error term representing unexplained variation.

This model balances simplicity and predictive accuracy while avoiding multicollinearity.

5.4.2 Regression Results

The initial regression results were instrumental in identifying significant predictors. Variables with p-values > 0.05 were excluded from subsequent models.

Regression Results: Coefficients, Standard Errors, t Stat and P-values for All Models						
Model	**Variable**	***Coefficients***	***Standard Error***	***t Stat***	***P-value***	***Significance***
Full Model with Outlier	Order Frequency	423027.2491	456336.5312	0.927007198	0.356348342	Not Significant
	CLNorm	-11128.34553	25683.25629	-0.433291846	0.665816976	Not Significant
	DSLONorm	27961.36294	10211.04892	2.738343843	**0.007414195**	**Significant**
	CTNorm	144468.8931	9772.468416	14.78325506	**4.89518E-26**	**Significant**
	CLVNorm	590136.1372	135533.0091	4.354187524	**3.46095E-05**	**Significant**
Significant Model with Outlier	DSLONorm	27363.13788	9808.723334	2.789673737	**0.006387944**	**Significant**
	CTNorm	142958.0085	9235.769139	15.47873342	**1.34977E-27**	**Significant**
	CLVNorm	703358.9653	18677.0547	37.65898727	**1.67202E-58**	**Significant**
Full Model No Outlier Original Normalization	Order Frequency	96583.13341	231551.0727	0.417113738	0.677622686	Not Significant
	CLNorm	66897.46477	13886.30745	4.817512863	**6.09888E-06**	**Significant**
	DSLONorm	-10543.42233	5144.694157	-2.049377866	**0.043437101**	**Significant**
	CTNorm	81353.94796	5831.189841	13.95151765	**6.54735E-24**	**Significant**
	CLVNorm	353007.1443	62699.56016	5.630137491	**2.16698E-07**	**Significant**
Significant Model No Outlier Original Normalization	CLNorm	68909.56876	12960.25449	5.316991949	**7.91048E-07**	**Significant**
	DSLONorm	-10866.98162	5061.952009	-2.146796651	**0.034560378**	**Significant**
	CTNorm	80764.80412	5630.930152	14.34306623	**9.36879E-25**	**Significant**
	CLVNorm	376296.2629	28393.01231	13.25312928	**1.11331E-22**	**Significant**
Full Model No Outlier Post Normalization	Order Frequency	-262430.347	222590.7297	-1.178981476	0.241583434	Not Significant
	CLNorm	88550.77601	13331.45153	6.642245653	**2.46708E-09**	**Significant**
	DSLONorm	-15366.99209	5292.153617	-2.903731298	**0.004660899**	**Significant**
	CTNorm	75035.68451	5924.377653	12.66558091	**1.54912E-21**	**Significant**

Model	Variable	*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Significance*
	CLVNorm	203590.9561	34275.79456	5.939788085	**5.57985E-08**	**Significant**
Significant Model No Outlier Post Normalization	CLNorm	68909.56876	12960.25449	5.316991949	**7.91048E-07**	**Significant**
	DSLONorm	-10866.98162	5061.952009	-2.146796651	**0.034560378**	**Significant**
	CTNorm	80764.80412	5630.930152	14.34306623	**9.36879E-25**	**Significant**
	CLVNorm	376296.2629	28393.01231	13.25312928	**1.11331E-22**	**Significant**
Significant VIF Model No Outlier Original Normalization	DSLONorm	497.5138109	5244.64773	0.094861245	**0.924638292**	**Significant**
	CTNorm	95299.2482	5626.782641	16.93672109	**1.38936E-29**	**Significant**
	CLVNorm	492142.6822	20808.80746	23.65069133	**3.79759E-40**	**Significant**
Significant VIF Model No Outlier Post Normalization	DSLONorm	-1983.005942	5927.647157	-0.334535084	**0.738754183**	**Significant**
	CTNorm	93098.13568	6367.955029	14.61978536	**1.68422E-25**	**Significant**
	CLVNorm	231397.4819	11034.49515	20.97037325	**2.092E-36**	**Significant**

Key Findings:

Removed Variables:
Order Frequency and CLNorm were excluded due to their lack of statistical significance.
Retained Variables:
DSLONorm, CTNorm, and CLVNorm were consistently significant and retained for further analysis.

5.5 Model Overview

Eight regression models were developed to investigate the effects of outlier handling, normalization, and variable selection:

Full Model with Outliers: Included all predictors without removing outliers.

Significant Variables Model with Outliers: Retained only significant predictors while keeping outliers.

Full Model No Outliers Original Normalization: Included all predictors after outlier removal with original normalization.

Full Model No Outliers Post Normalization: Included all predictors after outlier removal with re-normalized variables.

Significant Variables Model No Outliers Original Normalization: Retained significant predictors after outlier removal with original normalization.

Significant Variables Model No Outliers Post Normalization: Retained significant predictors after outlier removal and re-normalization.

Significant VIF Model No Outliers Original Normalization: Retained predictors with acceptable VIF values after outlier removal.

Significant VIF Model No Outliers Post Normalization: Retained predictors with acceptable VIF values after re-normalization.

5.6 Model Configurations

Eight regression models were configured and tested to evaluate the impact of normalization, outlier removal, and predictor selection.

Model	Key Adjustments
Full Model with Outliers	Included all predictors without removing outliers.
Significant Variables Model with Outliers	Retained only statistically significant predictors while keeping outliers.
Full Model No Outliers Original Normalization	Included all predictors after outlier removal with original normalization.
Full Model No Outliers Post Normalization	Included all predictors after outlier removal with re-normalized predictors.
Significant Variables Model No Outliers Original Normalization	Retained significant predictors after outlier removal with original normalization.
Significant Variables Model No Outliers Re Normalization	Retained significant predictors after outlier removal with re-normalized predictors.
Significant VIF Model No Outliers Original Normalization	Retained predictors with acceptable VIF values (DSLONorm, CTNorm, CLVNorm) after outlier removal.
Significant VIF Model No Outliers Post Normalization	VIF-refined predictors with re-normalized scales.

5.7 Model Results

The performance metrics for all eight regression models are summarized in the table below. These metrics include:

- **R^2 (Coefficient of Determination)**: Measures the proportion of variance in Total Purchase Value explained by the model.
- **Adjusted R^2**: Adjusted for the number of predictors, providing a more accurate measure for models with differing complexities.
- **MAE (Mean Absolute Error)**: Represents the average magnitude of errors in the predictions.
- **MSE (Mean Squared Error)**: Emphasizes larger errors due to squaring.
- **RMSE (Root Mean Squared Error)**: Provides errors in the original unit of measurement, allowing easier interpretability.

Model Configuration	R^2	Adjusted R^2	MAE	MSE	RMSE
Full Model with Outlier	0.948715441	0.945928237	13932.39523	476961006.1	21839.43695
Sig Only Model with Outlier	0.948169226	0.946515052	259418.9449	97191540093	311755.5775
Full Model No Outlier Orig Norm	0.913971944	0.909083986	14055.59495	2074712150	45549.00822
Sig Only Model No Outlier Orig Norm	0.920896613	0.917301005	13066.98543	1654405449	40674.3832
Full Model No Outlier Re Norm	0.913971944	0.909083986	21547.30955	6440818591	80254.71071
Sig Only Model No Outlier Re Norm	0.912613094	0.908685593	43269.31696	6102325443	78117.38246
Sig VIF Model No Outlier Orig Norm	0.895484288	0.891961286	22566.37819	5921260169	76949.72494
Sig VIF Model No Outlier Re Norm	0.870823917	0.866518048	47822.93898	9591595304	97936.69029

Analysis of Results

- **Impact of Outlier Removal**:

Removing outliers significantly improved model performance by reducing prediction errors. For instance, the **Sig Only Model No Outlier Orig Norm** (RMSE = 40,674.38) outperformed the **Sig Only Model with Outlier** (RMSE = 311,755.58).

- **Normalization vs. Re-Normalization**:
  Models using original normalization consistently showed better performance compared to re-normalized counterparts. For example, the **Significant VIF Model No Outlier Original Normalization** (RMSE = 76,949.72) outperformed the **Significant VIF Model No Outlier Post Normalization** (RMSE = 97,936.69).
- **Simplified Models with Significant Variables**:
  Simplified models retaining only significant predictors demonstrated comparable or superior performance to full models. For instance, the **Significant Only Model No**

Outlier Original Normalization (Adjusted R^2 = 0.9173) outperformed the **Full Model No Outlier Orig Norm** (Adjusted R^2 = 0.9091).

- **Multicollinearity Refinement**:
  The **Significant VIF Model No Outlier Original Normalization** addressed multicollinearity effectively while maintaining acceptable accuracy (Adjusted R^2 = 0.8920).

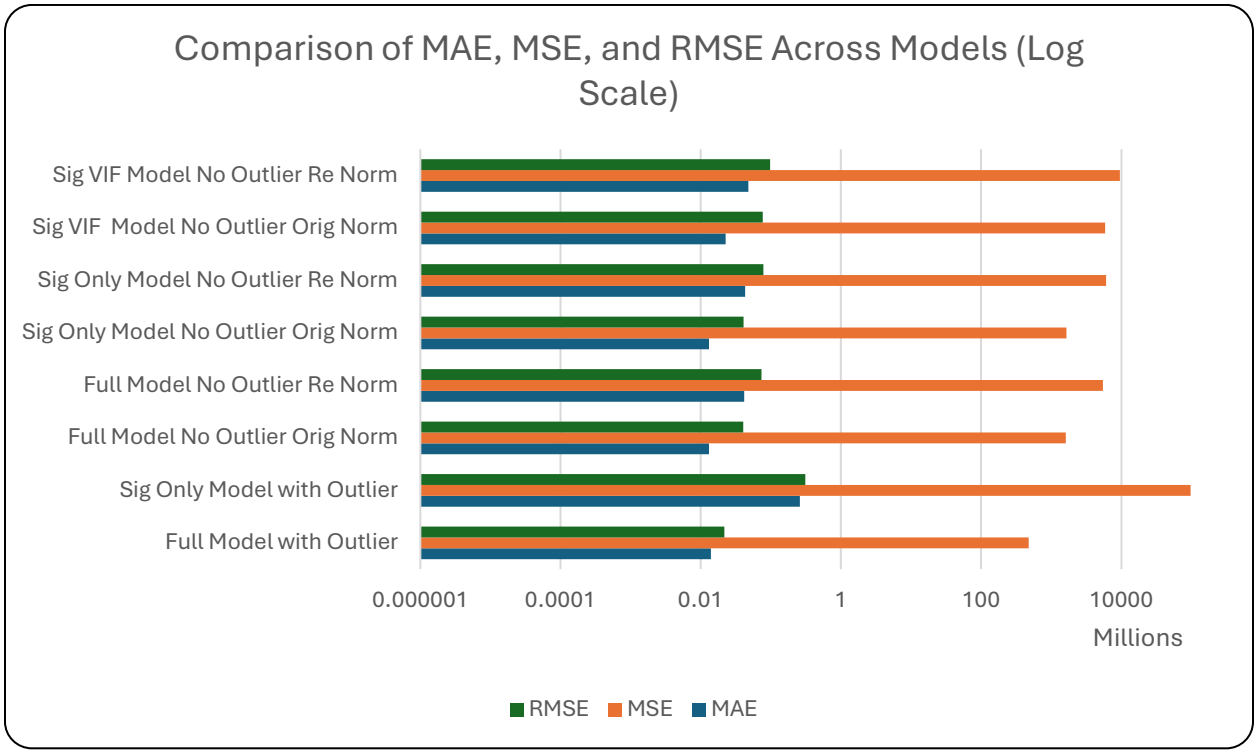

Fig. 5.6.1 Comparison of MAE, MSE and RSME across Models (Log Scale)

The chart compares the performance of all eight models across RMSE, MSE, and MAE using a logarithmic scale for clarity. It highlights that models without outliers and with significant variables—such as the **Significant Variables Model No Outlier Original Normalization**—achieved the lowest RMSE (40,674.38), demonstrating superior accuracy. Meanwhile, the **Significant VIF Model No Outlier Original Normalization** balances simplicity and acceptable performance, supporting its selection for further validation.

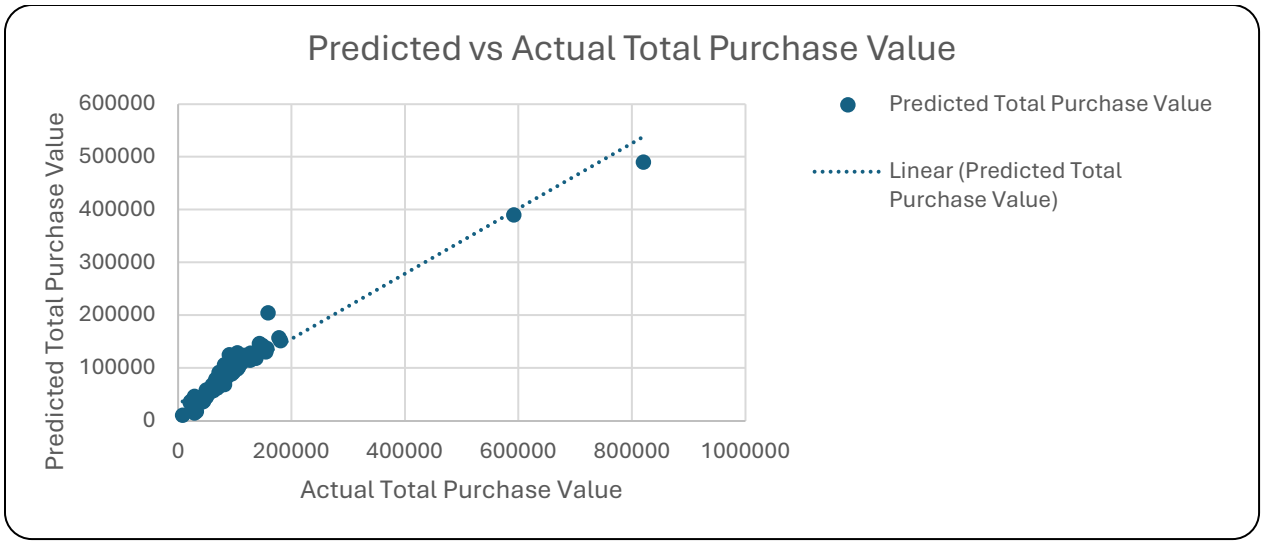

Fig 5.6.2 Scatter Plot - Significant Variables Model No Outliers Original Normalization

The scatter plot for the **Significant Variables Model No Outlier Original Normalization** shows a strong linear relationship, with most points closely aligned along the diagonal line. This indicates accurate predictions of Total Purchase Value, with minimal deviations, supporting the model's high RMSE performance (40,674.38).

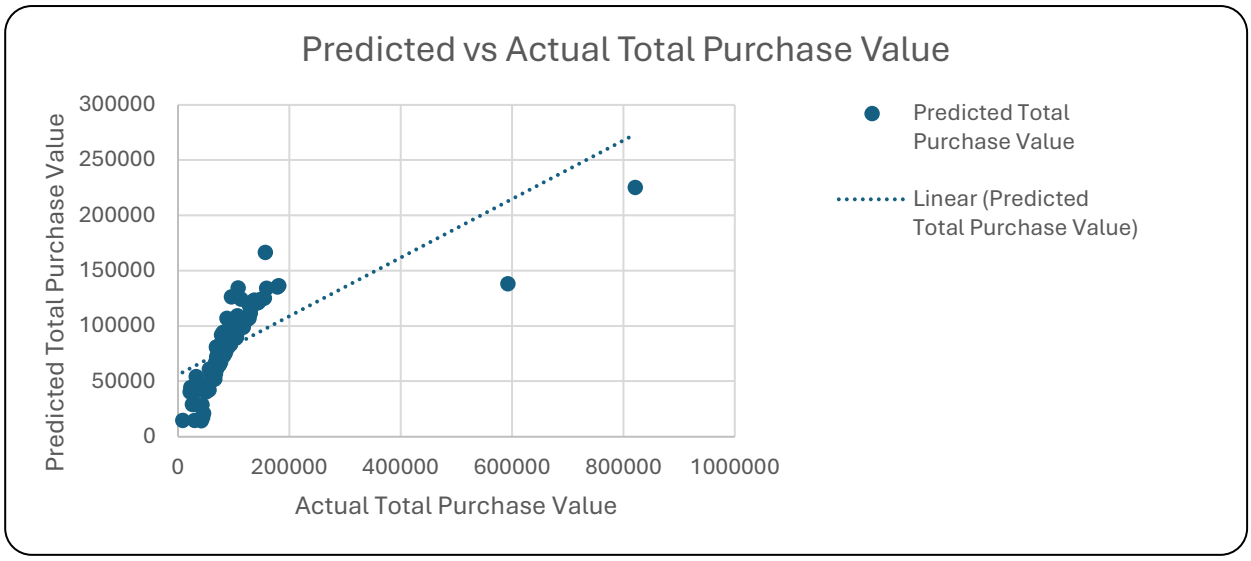

Fig 5.6.3 Scatter Plot - Significant VIF Variables Model No Outliers Original Normalization

The scatter plot for the **Significant VIF Model No Outlier Original Norm** also shows a linear relationship, but with slightly greater spread and deviations from the diagonal line. While less accurate (RMSE = 76,949.72), this model emphasizes simplicity and interpretability by resolving multicollinearity.

5.8 Explanation of Selected Models

- **Significant Variables Model No Outliers Original Normalization:**

This model delivers the highest predictive accuracy, with the lowest RMSE (40,674.38) and highest Adjusted R^2 (0.9173). By retaining only statistically significant predictors, it balances simplicity, precision, and robustness against outliers.

Recommendation: This model is ideal for tasks requiring precise forecasts, such as strategic decision-making or resource allocation, where accuracy is critical.

- **Significant VIF Model No Outliers Original Normalization:**

Designed for simplicity and interpretability, this model resolves multicollinearity while maintaining reasonable predictive accuracy (RMSE = 76,949.72). Its streamlined structure, with fewer predictors, makes it suitable for practical applications requiring quick, clear insights.

Recommendation: This model is best suited for scenarios where ease of implementation and interpretability are prioritized, such as generating actionable business insights.

Chapter 6: Model Evaluation

6.1 Objective

This chapter evaluates the predictive models developed in the project, focusing on their effectiveness in forecasting **Total Purchase Value**. Evaluation methods, including 10-fold cross-validation and statistical comparisons, are employed to validate the models and substantiate the decision to prioritize the **Significant VIF Variables No Outlier Model** over the **Significant Variables No Outlier Model**.

6.2 Evaluation Metrics

The following metrics were used to assess model performance during the 10-fold cross-validation process:

Mean Absolute Error (MAE): Measures the average magnitude of errors.

Mean Squared Error (MSE): Penalizes larger deviations from actual values.

Root Mean Squared Error (RMSE): Reflects the standard deviation of residuals.

R^2**:** Indicates the proportion of variance explained by the model.

Adjusted R^2**:** Accounts for the number of predictors in the model.

6.3 Cross-Validation Results

6.3.1 Significant Variables No Outlier Model

Fold	MAE	MSE	RMSE	R^2	Adjusted R^2
Fold 1	5473.832269	48582387.19	6970.106685	0.91172123	0.907194113
Fold 2	7722.199637	109543866.3	10466.32057	0.908296452	0.903593706
Fold 3	5161.414468	40518459.03	6365.411144	0.908734125	0.904053823
Fold 4	9822.011511	108063205.5	10395.34538	0.912166142	0.907661842
Fold 5	7165.310773	83202635.15	9121.547849	0.909288802	0.90469583
Fold 6	9704.623028	182444532.3	13507.20298	0.922302472	0.91836842
Fold 7	8387.21217	111246544.9	10547.34777	0.912116022	0.907666201
Fold 8	13495.81865	362695916.8	19044.5771	0.920460483	0.916433165
Fold 9	8222.695161	157270503.2	12540.75369	0.91755817	0.913383901
Fold 10	9101.183469	126988700	11268.9263	0.910055269	0.905558032
Average	**8425.630114**	**133055675**	**11022.75395**	**0.913269917**	**0.908860903**

6.3.2 Significant VIF Variables No Outlier Model

Fold	MAE	MSE	RMSE	R^2	Adjusted R^2
Fold 1	8393.005533	102326681.3	10115.66514	0.869996731	0.865059898
Fold 2	8325.331317	103083834.6	10153.02096	0.862633246	0.857416787
Fold 3	8673.921756	112817032.8	10621.53628	0.865390526	0.860278773
Fold 4	8436.72262	99326132.44	9966.249668	0.871961909	0.867099703
Fold 5	8639.456625	110678336.8	10520.37722	0.865429998	0.860383622
Fold 6	8149.125799	90266019.67	9500.843103	0.879668817	0.875156398
Fold 7	8547.002913	104459785	10220.55698	0.870379894	0.865519141
Fold 8	6991.834174	64549520.45	8034.271619	0.895542942	0.891625802
Fold 9	8633.017895	105384757	10265.70782	0.873011883	0.868249829

Fold 10	9108.099447	127600583.9	11296.04284	0.862089309	0.856981505
Average	**8389.751808**	**102049268.4**	**10069.42716**	**0.871610525**	**0.866777146**

Observations:

The **Significant VIF Variables No Outlier Model** had lower MAE and RMSE across folds, indicating improved prediction accuracy.

The **Significant Variables No Outlier Model** demonstrated higher R^2 and Adjusted R^2, suggesting better explanatory power, but it retained multicollinearity.

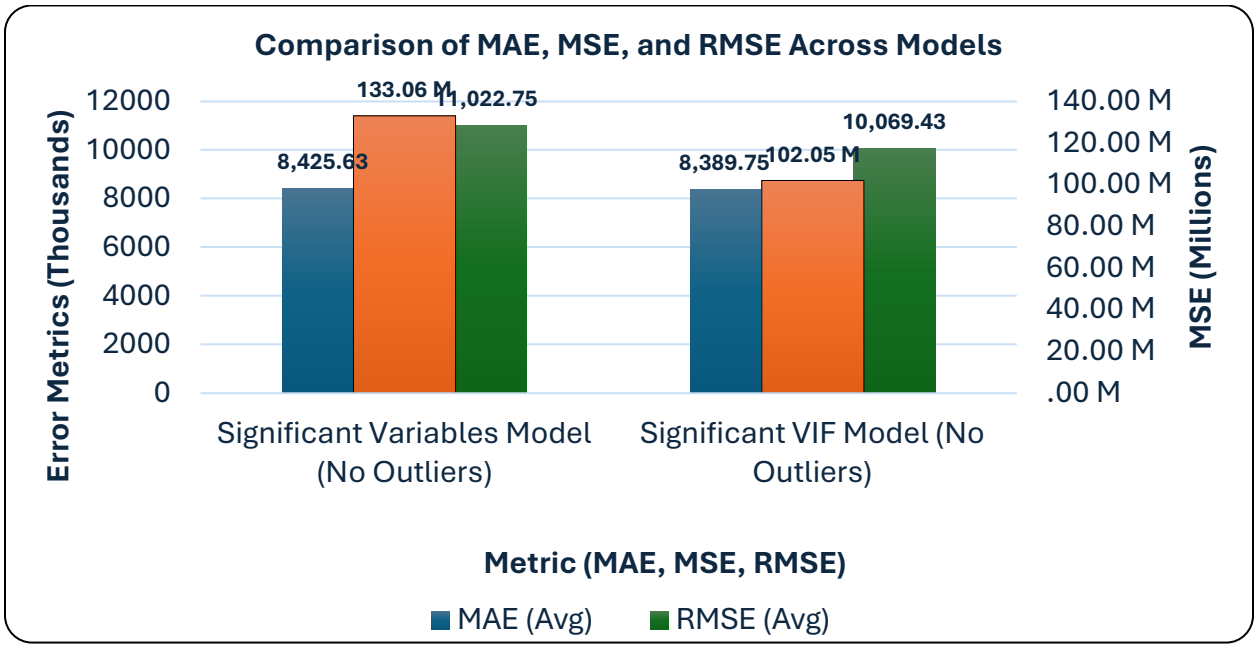

Fig 6.3.3 Comparison of MAE, MSE and RMSE Across Models

The bar chart compares the average MAE, MSE, and RMSE values across the two models. It highlights the overall error metrics, showing that the Significant VIF Variables Model achieves slightly lower error rates compared to the Significant Variables Model.

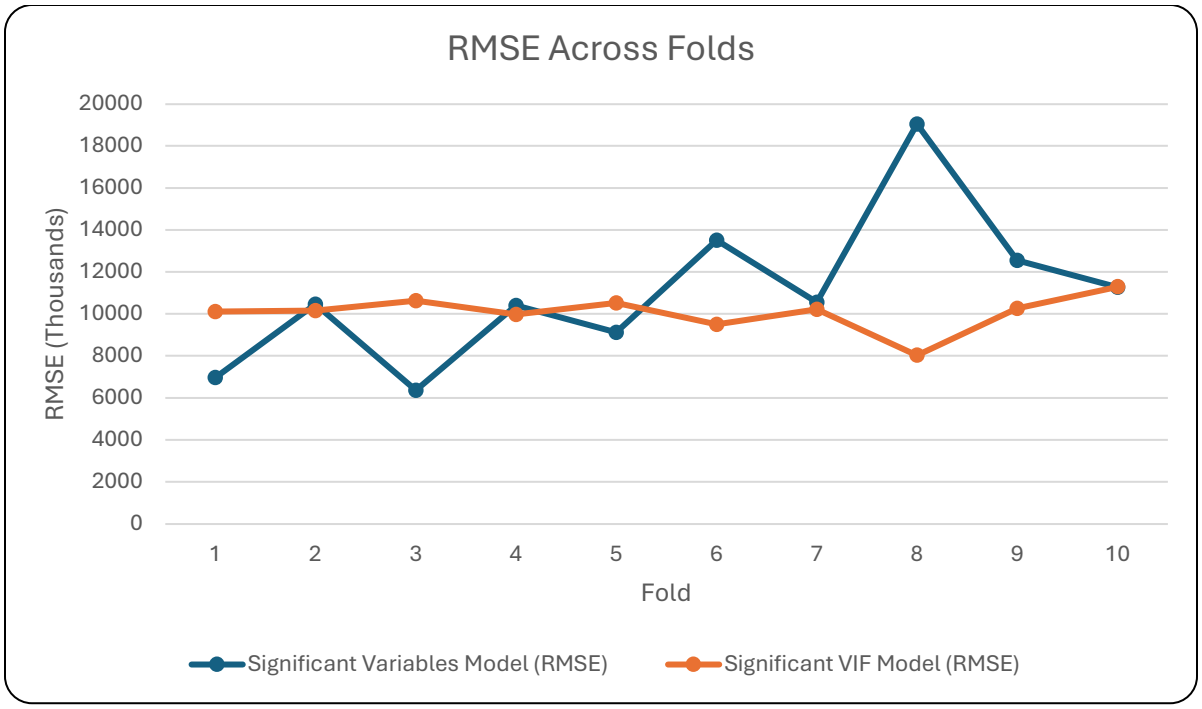

Fig 6.3.4 RMSE Across Folds

The line plot illustrates the RMSE values across all 10 folds for both models. It highlights fold-specific variations in prediction error, demonstrating that the Significant VIF Variables No Outlier Model exhibits more consistent performance with lower RMSE values across folds compared to the Significant Variables No Outlier Model.

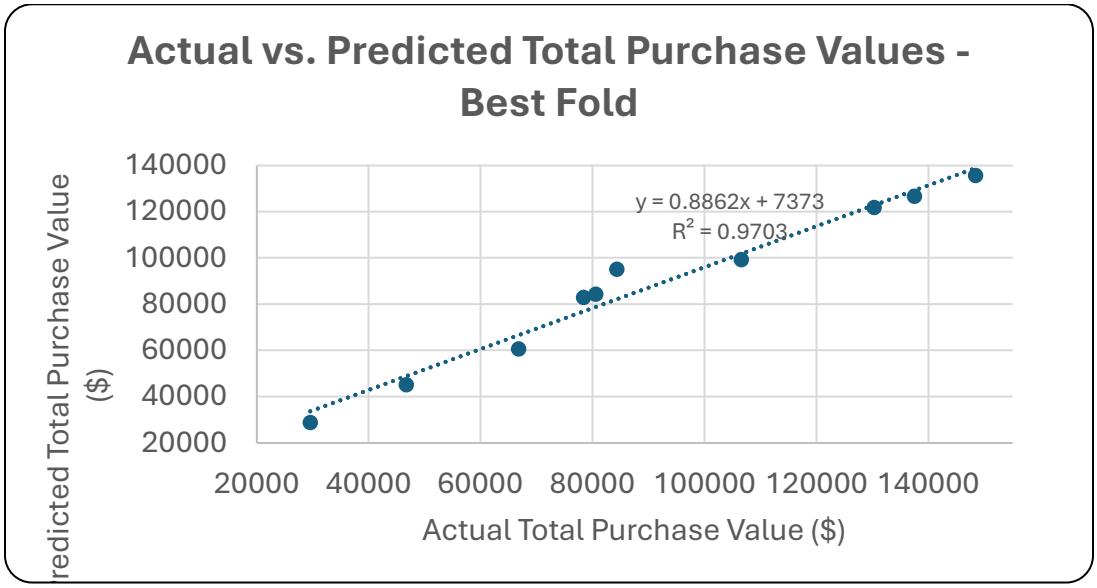

Fig 6.3.5 Plot Actual vs Predicted Total Purchase Value – Best Fold (8) of Significant VIF Variable Model with No Outlier

Actual vs. Predicted Total Purchase Values - Best Fold (Fold 8)

This scatter plot represents the relationship between actual and predicted total purchase values for Fold 8, which exhibited the **lowest RMSE (8034.27)**, indicating the best predictive performance of the **Significant VIF Variables No Outlier Model**. The points closely align with the diagonal trend line, reflecting strong predictive accuracy with minimal deviations. The high alignment between actual and predicted values demonstrates that the model effectively captures the underlying patterns in this subset of the data.

Key Observations:

- Predicted values are consistently close to actual values across the range.
- The R^2 value for this fold supports high explanatory power, further validating the model's strong performance.

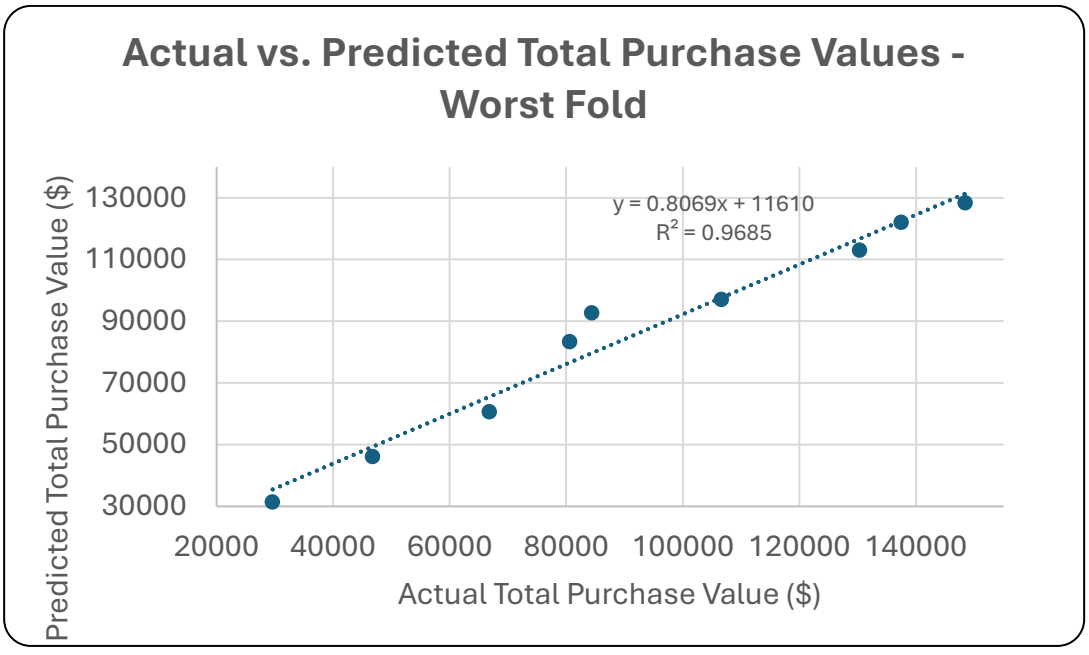

Fig 6.3.5 Plot Actual vs Predicted Total Purchase Value – Worst Fold (10) of Significant VIF Variable Model with No Outlier

Actual vs. Predicted Total Purchase Values - Worst Fold (Fold 10)

This scatter plot illustrates the relationship between actual and predicted total purchase values for Fold 10, which exhibited the **highest RMSE (11296.04)**, indicating the weakest predictive performance of the **Significant VIF Variables No Outlier Model**. The points deviate more significantly from the diagonal trend line compared to Fold 8, reflecting lower predictive accuracy. This indicates that the model struggled to accurately predict purchase values for this subset of the data.

Key Observations:

- Greater variability in the predicted values relative to actual values.
- The scatter pattern highlights areas where the model underperformed, with larger residuals observed for certain data points.
- Despite being the worst-performing fold, the model still exhibits an overall trend that aligns with the actual values.

6.4 Statistical Comparison

Paired t-Test: Significant Variables No Outlier Model vs. Significant VIF Variables No Outlier Model

Metric	Significant No Outlier Model	Significant VIF No Outlier Model
Mean (Avg)	8425.630114	8389.751808
Variance	5693664.457	306830.6795
Observations	10	10
Pearson Correlation	-0.68147577	
Hypothesized Mean Difference	0	
df	9	
t Stat	0.040619079	
P(T<=t) one-tail	0.484243193	
t Critical one-tail	1.833112933	
P(T<=t) two-tail	0.968486387	
t Critical two-tail	2.262157163	

Conclusion: No statistically significant difference was found between the models ($p > 0.05$). Both models are viable options based on their predictive performance.

6.5 Discussion and Conclusion

Key Findings

The **Significant VIF Variables Model No Outlier** was chosen as the best model because it was the best combination of being able to predict accurately, being easy to understand, and being useful in real life. The model successfully solved multicollinearity difficulties that were obvious in the Significant Variables No Outlier Model due to high Variance Inflation Factor (VIF) values in variables such as OrderFrequency and CLNorm. By removing duplication, the Significant VIF Variables No Outlier Model kept only three predictors with independent contributions:

DSLONorm represents recent activity and detects at-risk clients for re-engagement.

CTNorm: Enhances customer loyalty and improves retention initiatives.

CLVNorm: Measures cumulative customer contribution, which is critical for identifying high-value customers.

The **Significant VIF Variables No Outlier Model** had slightly lower R^2 values than the Significant Variables No Outlier Model, but it also had lower Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) across all folds. This made it more accurate at predicting the future and more stable in cross-validation.

Strengths

Multicollinearity Reduction: Using VIF analysis confirmed that predictors in the model were statistically independent, lowering the risk of overfitting and improving model generalizability.

Practical Insights: The simplified model focuses on factors that are directly related to business goals. This makes it useful for marketing and customer retention plans**.**

Consistency in Performance: Cross-validation confirmed that the **Significant VIF Variables No Outlier Model** maintained stable performance across folds, supporting its reliability.

Limitations

No Statistically Significant Difference: The paired t-test results showed no statistically significant difference between the models, suggesting that further refinement or additional predictors might improve model discrimination.

Potential Loss of Information: Removing variables with high VIF could leave out features that could provide additional business insights if properly managed.

Implications

Due to its ease of use and understanding, the **Significant VIF Variables No Outlier Model** is better for real-world situations. The model's emphasis on independent predictors makes it a good fit for business requirements, especially when it comes to consumer segmentation, re-engagement initiatives, and retention tactics.

Conclusion

The **Significant VIF Variables No Outlier Model** is a useful tool for predicting Total Purchase Value and answering the research question since it balances interpretability, practicality, and accuracy. Its relevance in real-world situations is guaranteed by its capacity to handle multicollinearity while retaining a high degree of predictive power, especially in marketing analytics and business decision-making.