Unveiling Profits: Leveraging Data Science for Predictive Profitability

By

Hajarat Titilope OLUFADE
Ajarah Omowunmi AMBALI
Nneka OKEKE
Bashirat
Amara

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion
- Appendix

Executive Summary

- **Objective**: Use data science to analyze and predict profitability for strategic insights.
- **Dataset**: Sales data with metrics on region, product type, revenue, costs, and profit.
- **Methods**:
  - Data preprocessing and feature engineering.
  - Regression modeling.
- **Key Models**: Linear, Random Forest, Gradient Boosting, and ridge regression optimized for accuracy.
- **Outcomes**:
  - High accuracy in profit predictions.
  - Identified key profit drivers: revenue and profit margin.
- **Impact**: Supports data-driven financial planning and growth strategy.

Introduction

- **Purpose**: Develop a data-driven solution to analyze and predict profitability.

- **Problem**: Identifying the factors that impact profit margins and forecasting future profits.

- **Approach**:
  - Utilize historical sales data to train machine learning models.
  - Apply regression analysis to identify and predict profit trends.

- **Goal**: Enable informed decision-making for sustainable business growth through accurate profit forecasting

METODOLOGY

Data Collection

- **Source**:
  - Kaggle dataset with historical sales transaction data.

- **Key Columns**:
  - Region, Item Type, Sales Channel, Units Sold, Unit Price, Unit Cost, Total Revenue, Total Cost, Profit Margin, Total Profit.

- **Tools Used**:
  - Data imported into a Pandas DataFrame for preparation and analysis.

Libraries and Modules Import

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from scipy import stats
import numpy as np
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_squared_error, mean_absolute_error, r2_score
from sklearn.ensemble import RandomForestRegressor
from sklearn.ensemble import GradientBoostingRegressor
from sklearn.linear_model import Ridge
```

Data Import and Preparation

- Loaded the dataset into a Pandas DataFrame as **wdf**.
- Created a copy as **df** to preserve the integrity of the original data in **wdf**.
- Verified successful data import and readiness for further analysis.
- Used one-hot encoding to transform categorical variables (Region, Item Type, Sales Channel) for analysis

```
wdf = pd.read_csv(r"C:\Users\HP\Desktop\Webfala Project\Advertising Data Set.csv")
```

```
df = wdf.copy()
```

```
df.shape
```

```
(10000, 14)
```

```
df.head(3)
```

	Region	Country	Item Type	Sales Channel	Order Priority	Order Date	Order ID	Ship Date	Units Sold	Unit Price	Unit Cost	Total Revenue	Total Cost	Total Profit
0	Sub-Saharan Africa	Chad	Office Supplies	Online	L	1/27/2011	292494523	2/12/2011	4484	651.21	524.96	2920025.64	2353920.64	566105.00
1	Europe	Latvia	Beverages	Online	C	12/28/2015	361825549	1/23/2016	1075	47.45	31.79	51008.75	34174.25	16834.50
2	Middle East and North Africa	Pakistan	Vegetables	Offline	C	1/13/2011	141515767	2/1/2011	6515	154.06	90.93	1003700.90	592408.95	411291.95

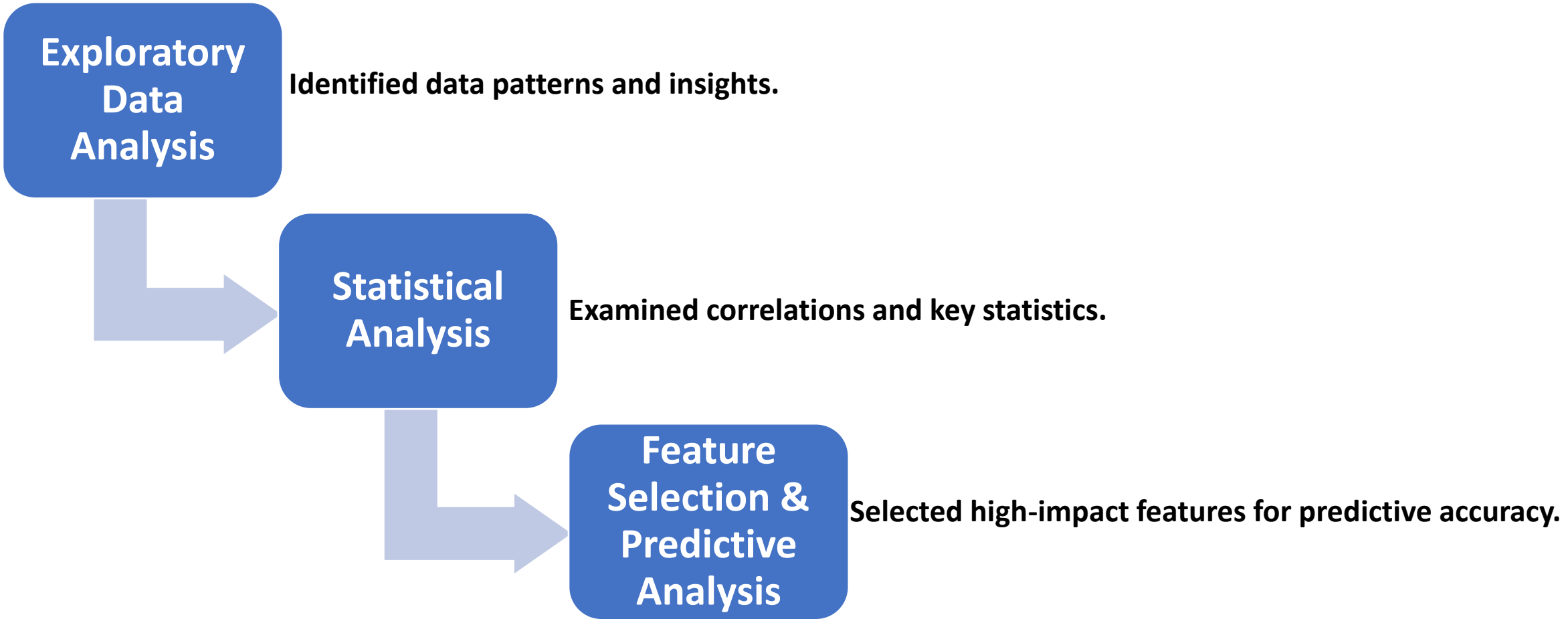
Missing Values Check

- Verified the dataset for missing values
- Observed that the dataset was clean with no missing entries

```
df.isnull().sum()
```

```
Region            0
Country           0
Item Type         0
Sales Channel     0
Order Priority    0
Order Date        0
Order ID          0
Ship Date         0
Units Sold        0
Unit Price        0
Unit Cost         0
Total Revenue     0
Total Cost        0
Total Profit      0
dtype: int64
```

Data Analysis Approach



Features and Target Variable Definition

- **Independent Variables:**
  - **Units Sold**: Quantity of items sold per order
  - **Unit Price**: Price per unit item
  - **Unit Cost**: Cost per unit item
  - **Total Revenue**: Total earnings from each order
  - **Total Cost**: Total expenses associated with each order
  - **Profit Margin**: Profit relative to revenue
  - **Order Priority**: Priority assigned to orders (e.g., high, medium, low, critical)
  - **Sales Channel**: Platform or method used to complete the sale
  - **Region**: Geographic region of the sale
  - **Item Type**: Category of the product sold (e.g., beverage, electronics)
- **Target Variable:**
  - **Total Profit**: Net profit per order (calculated by subtracting total cost from total revenue)

Data Splitting for Training and Testing

- **Purpose of Split:**
  To evaluate the model's performance on unseen data.

- **Split Ratio:**
  - 80% Training Set
  - 20% Testing Set

- **Hyperparameters Used:**
  - **test_size = 0.2**: Defines the split ratio
  - **random_state = 42**: Ensures reproducibility

- **Method:**
  - Implemented using **train_test_split** from **sklearn.model_selection**

Models Used in Training

- **Linear Regression**
  - Simple and interpretable model suitable for initial insights.
- **Random Forest Regressor**
  - Ensemble model leveraging multiple decision trees for higher accuracy.
- **Gradient Boosting Regressor**
  - Boosting technique that iteratively improves model predictions.
- **Ridge Regression**
  - Linear model with regularization to reduce overfitting.

Model Evaluation

- **Evaluation Metrics Used:**
  - **Mean Squared Error (MSE)**: Measures average squared difference between predicted and actual values, capturing overall accuracy.
  - **Mean Absolute Error (MAE)**: Measures the average absolute difference, indicating the model's precision.
  - **R-squared (R^2)**: Represents the proportion of variance explained by the model, indicating its goodness of fit.

Exploratory Data Analysis

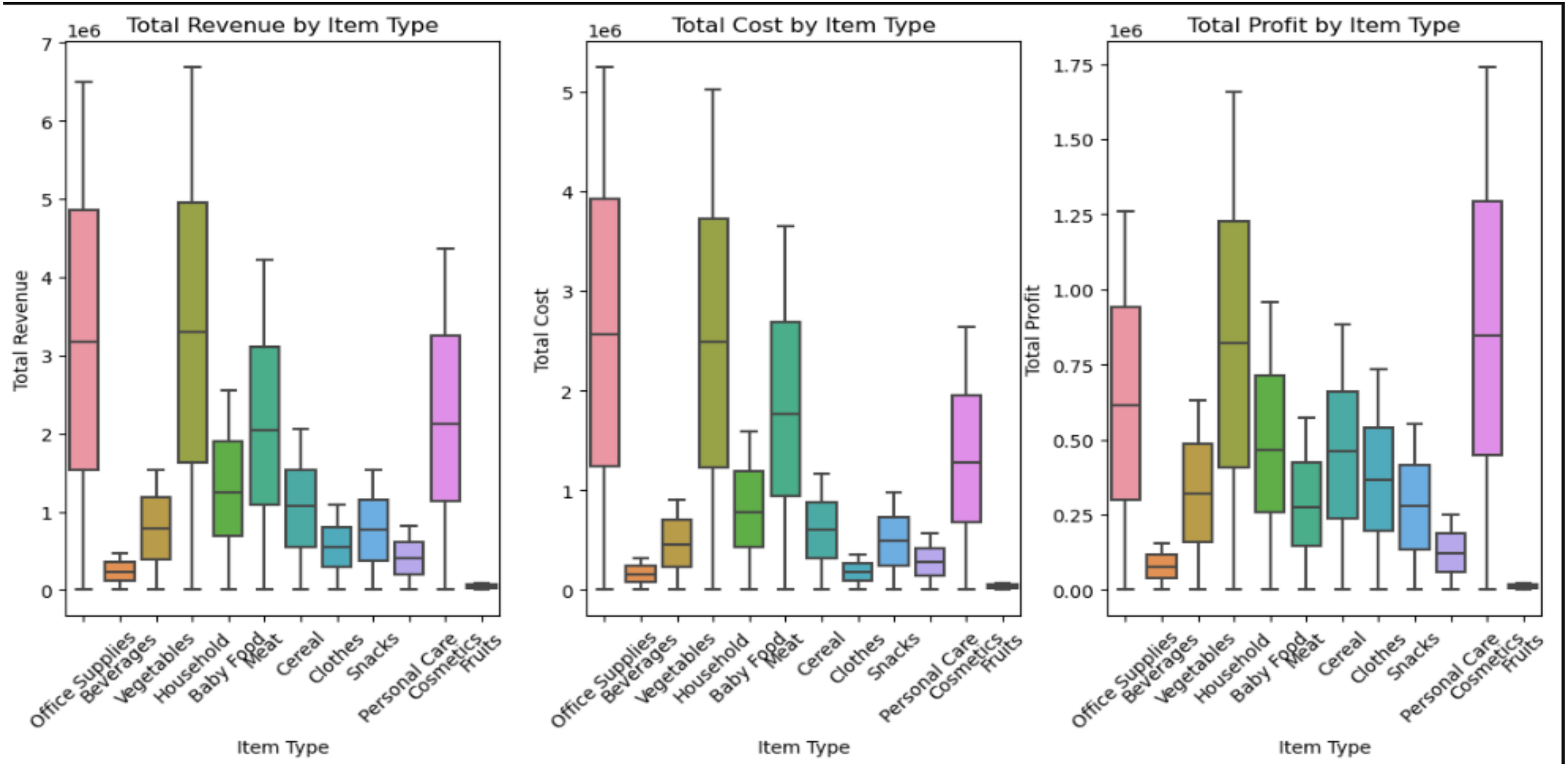
Descriptive Statistics Summary for Numerical Variables

	Units Sold	Unit Price	Unit Cost	Total Revenue	Total Cost	Total Profit	Profit Margin
Mean	5003	268.1	188.8	1333355.1	938265.8	395089.3	0.34
Standard Deviation	2874	217.9	176.4	1465026.2	1145914.1	377555.0	0.13
Minimum	2	9.33	6.9	167.9	124.6	43.4	0.14
25th Percentile	2531	109.3	56.7	288551.1	164785.5	98329.1	0.25
Median	4962	205.7	117.1	800051.2	481605.8	289099.0	0.36
75th Percentile	7472	437.2	364.7	1819143.4	1183821.5	566422.7	0.41
Maximum	10000	668.3	525.0	6680026.9	5241725.6	1738178.4	0.67

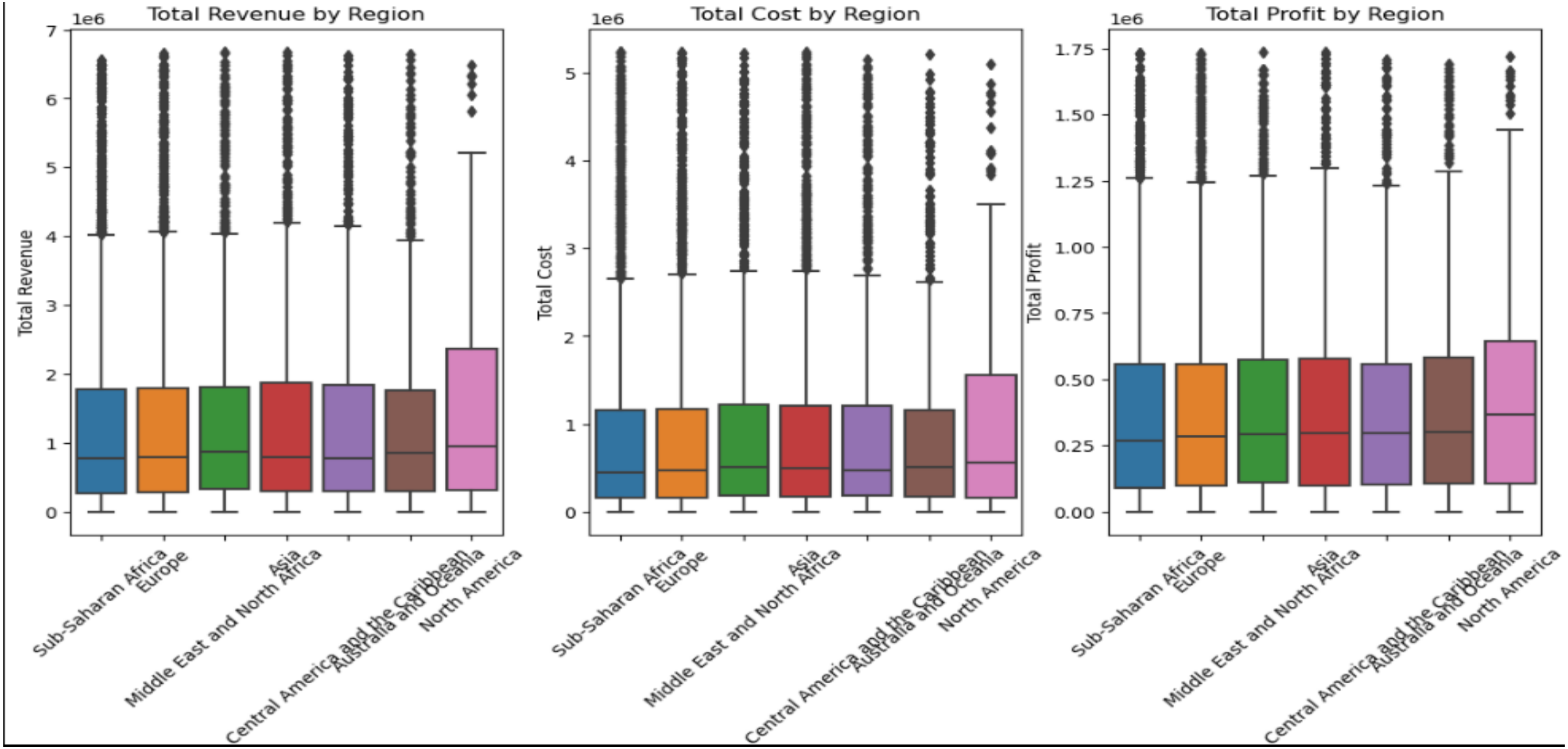
Unique Values Summary for Categorical Features

s/n	Region	Item Type	Sales Channel	Order Priority
1	Europe	Personal Care	Online	Critical
2	Sub-Saharan Africa	Household	Offline	High
3	Asia	Clothes		Medium
4	Middle East and North	Baby Food		Low
5	Africa	Office Supplies		
6	Central America and the Caribbean	Vegetables		
7	Australia and Oceania	Cosmetics		
8	North America	Cereal		
9		Snacks		
10		Meat		
11		Fruits		
12		Beverages		

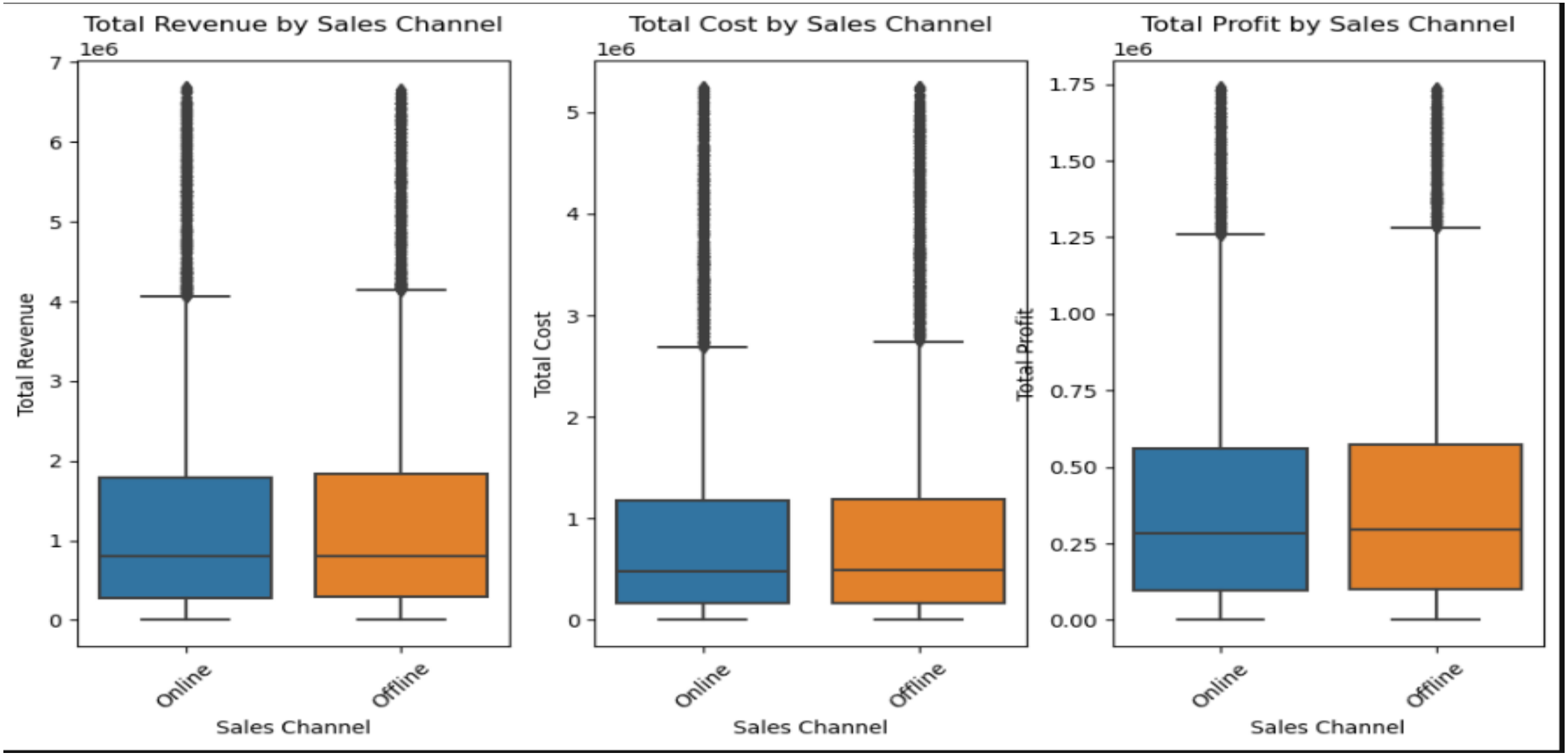
Comparative Analysis of Key Numerical Features by Item Type



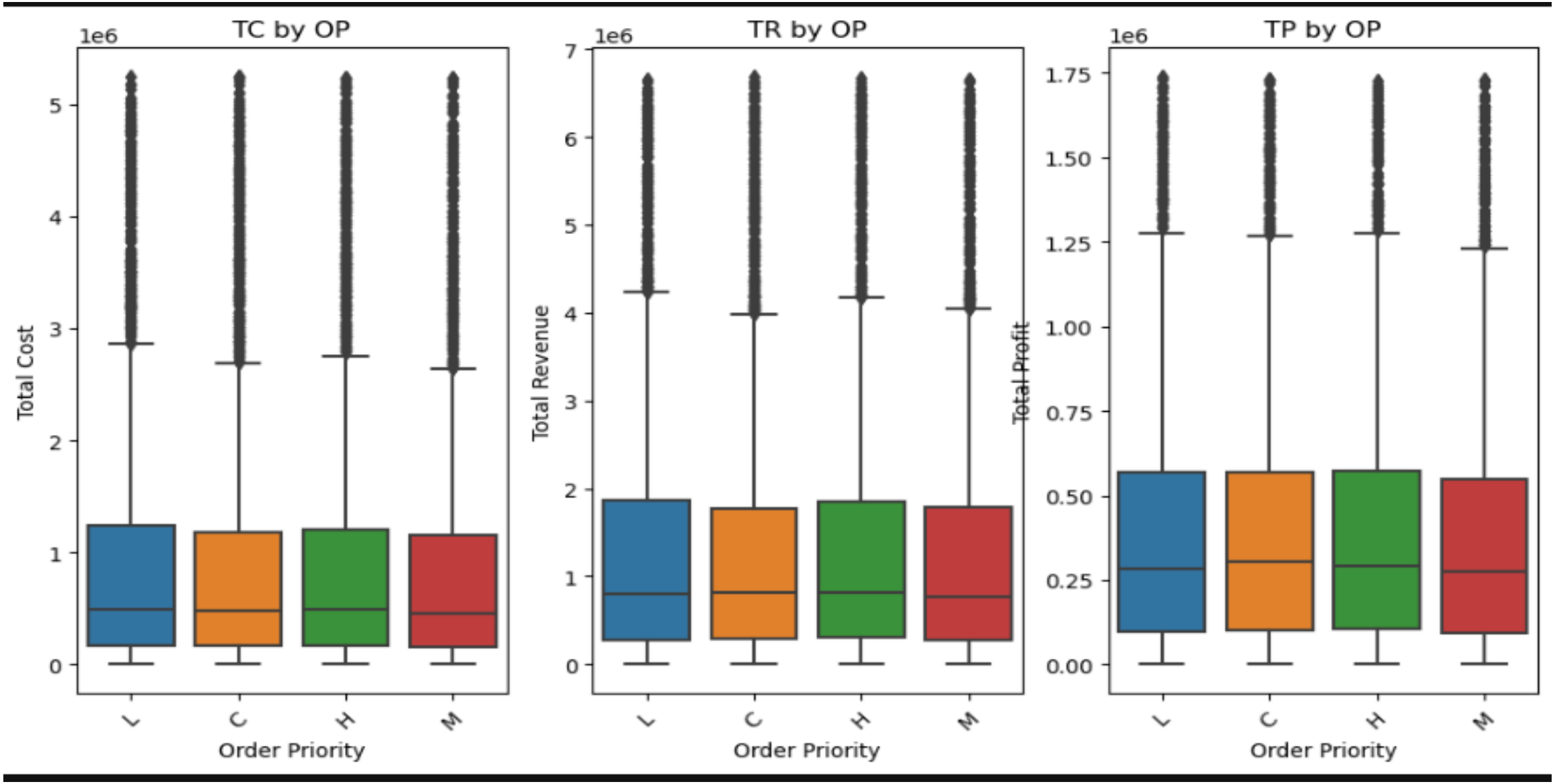
Comparative Analysis of Key Numerical Features by Region



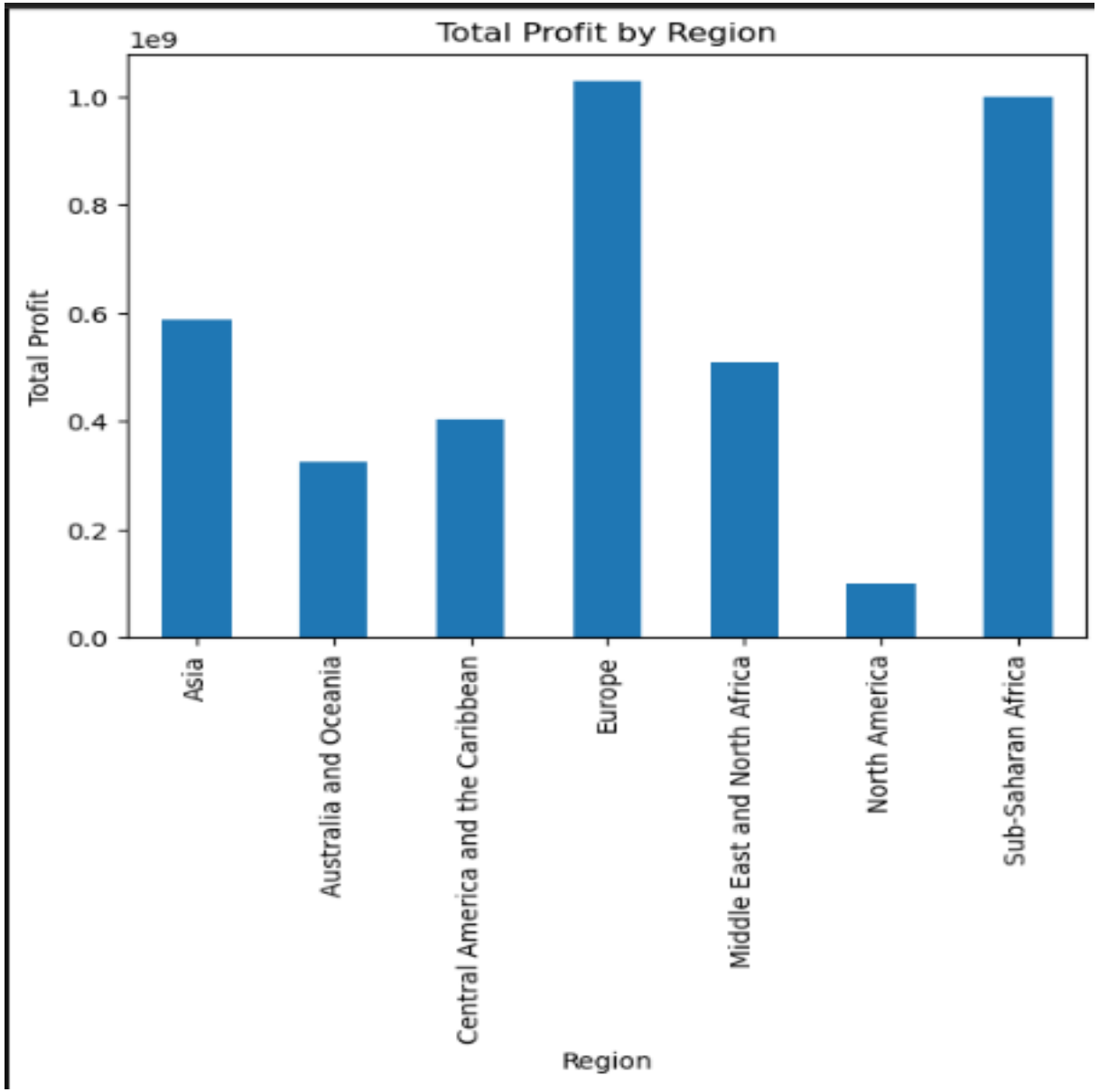
Comparative Analysis of Key Numerical Features by Sales Channel



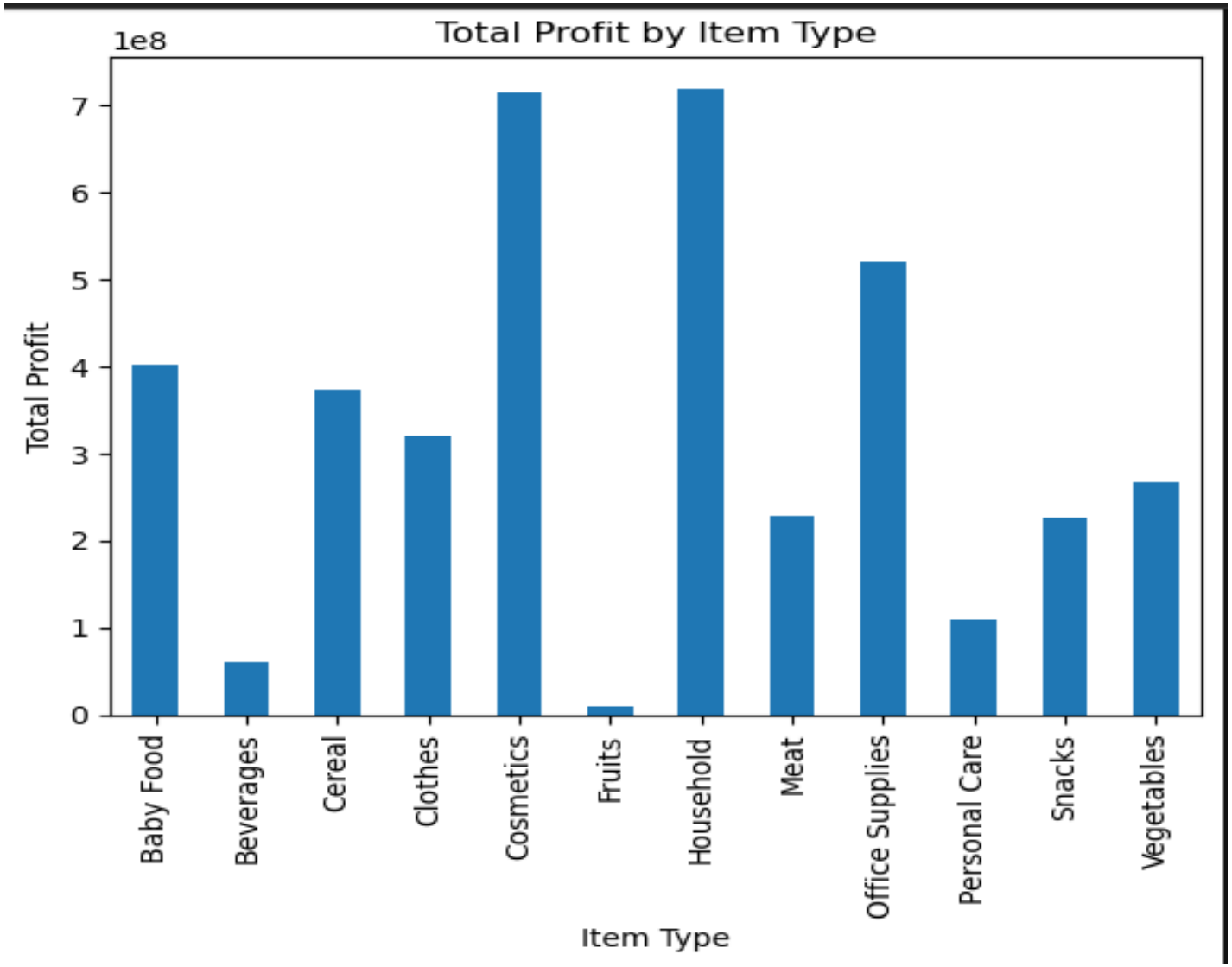
Comparative Analysis of Key Numerical Features by Order Priority



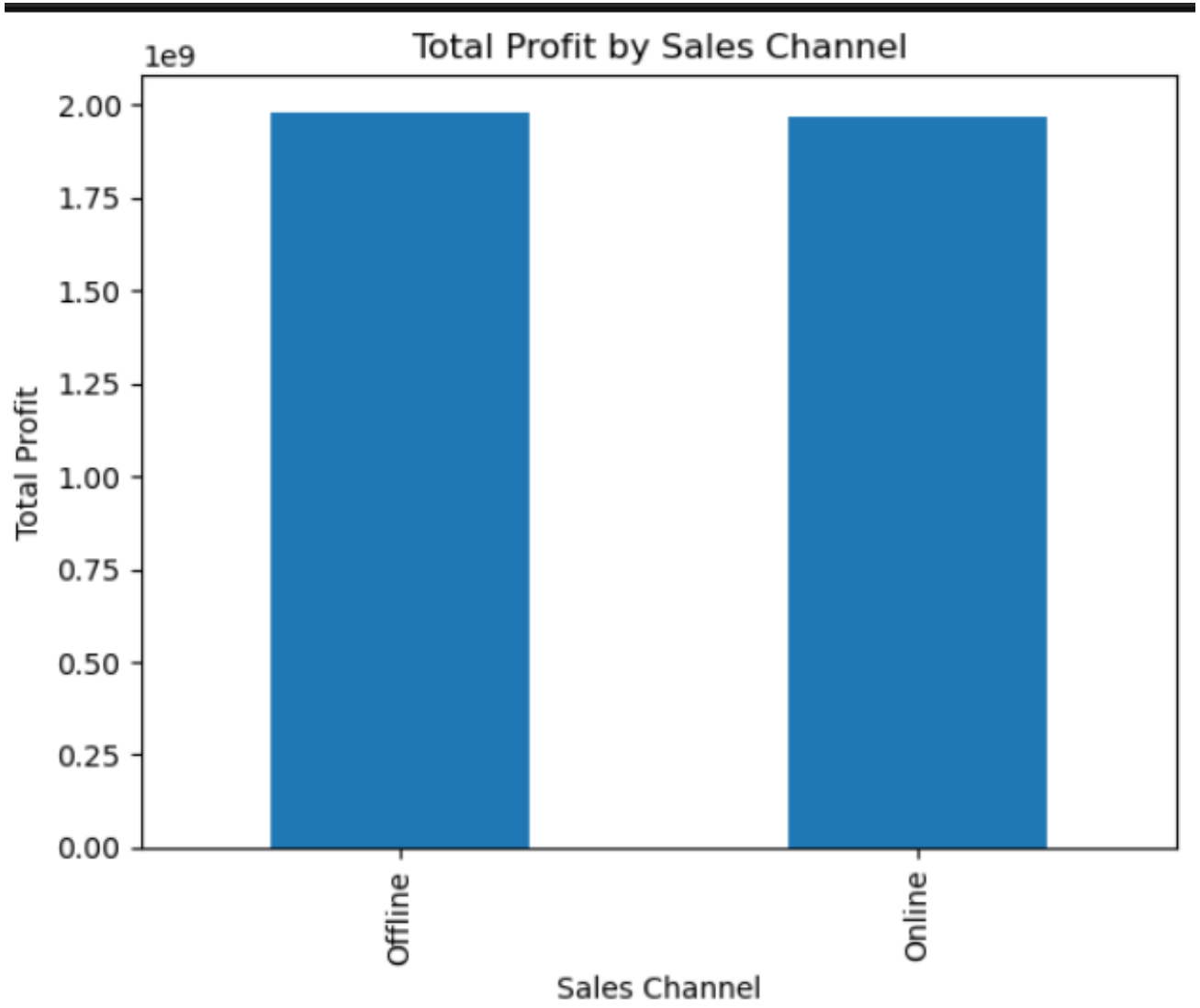
Profit Hotspots: Total Profit Analysis by Region



Profitability Snapshot: How Each Item Type Contributes to Total Profit

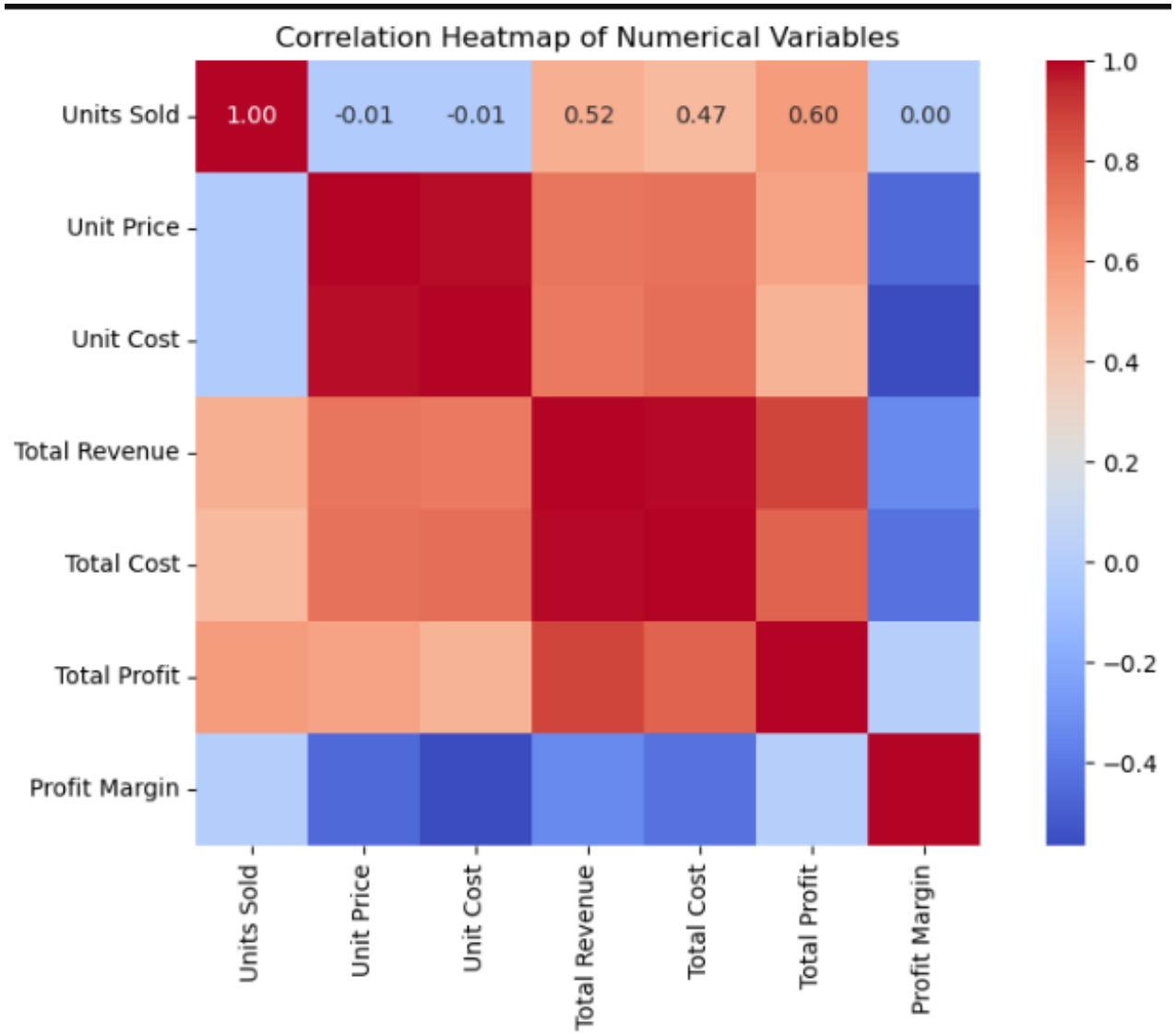


Profit Performance Across Sales Channels

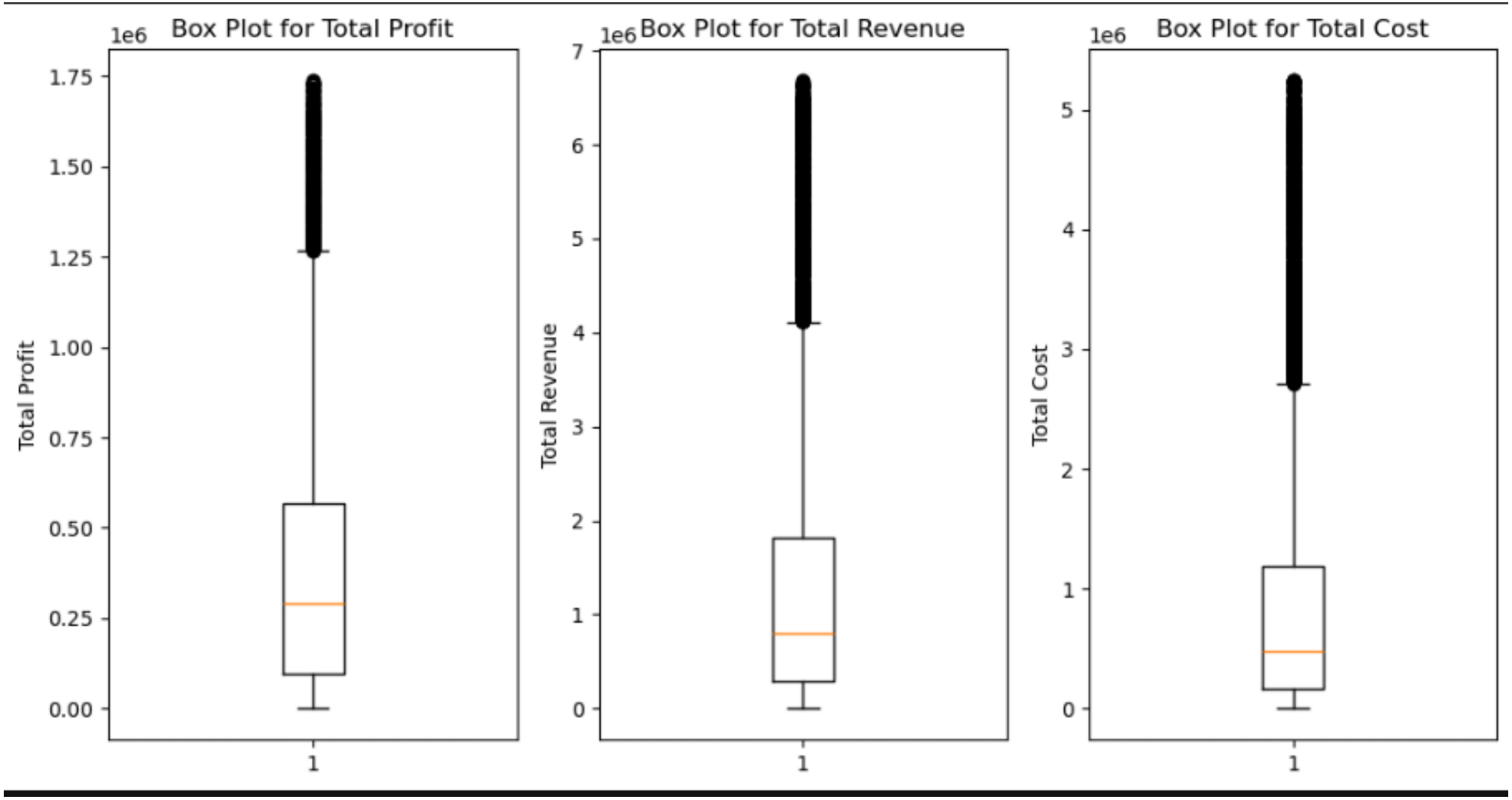


Profit Breakdown by Order Priority: Insights into High-Value Orders

Decoding Interactions: Correlation Heatmap of Key Numerical Variables



Z-Score Insights: Identifying Outliers in Numerical Data



Deriving the Profit Margin Metric

Profit margin as a derived feature

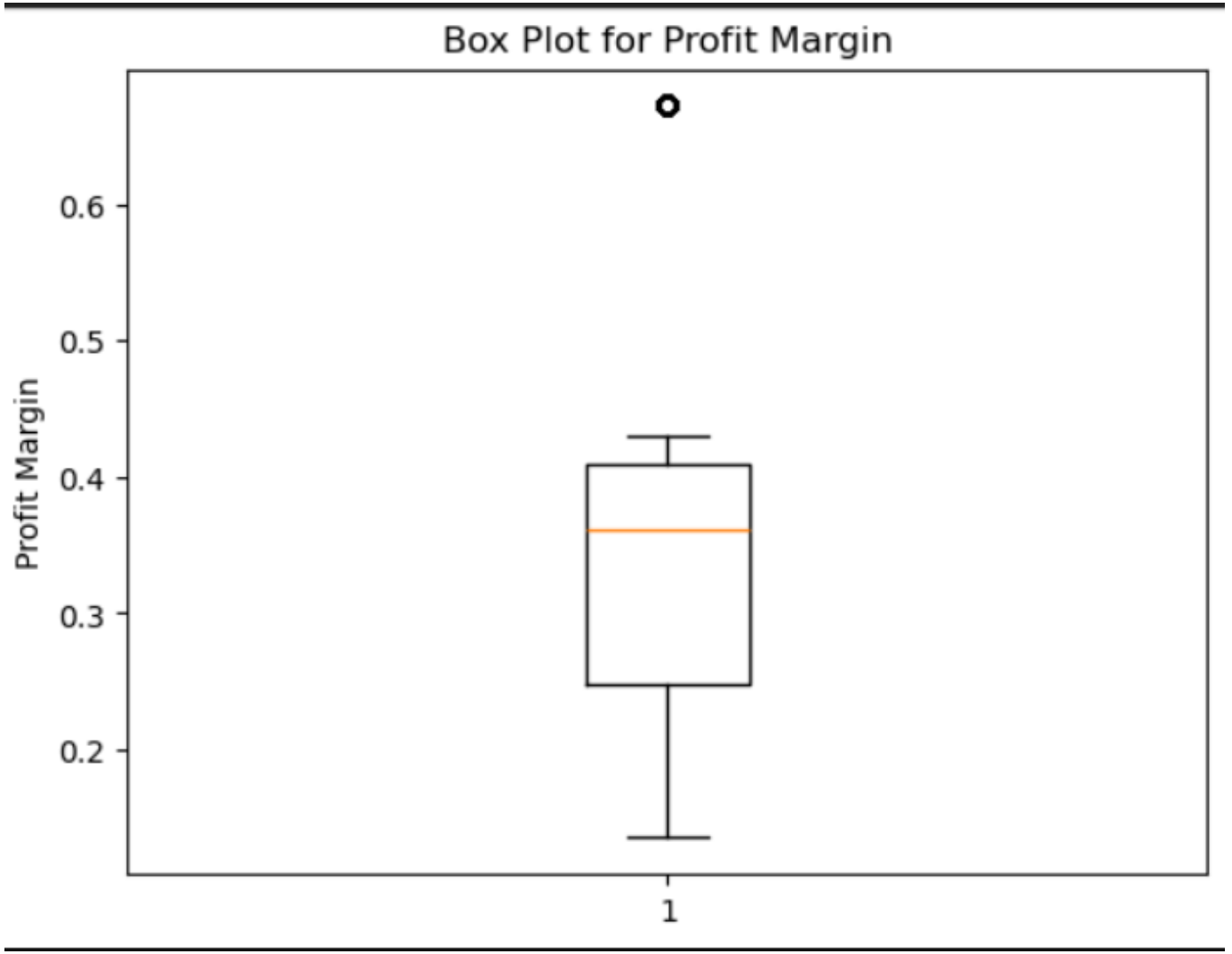
```
df = df[(df['Total Revenue'] != 0)]

# Calculate Profit Margin as a new feature
df['Profit Margin'] = df['Total Profit'] / df['Total Revenue']
```

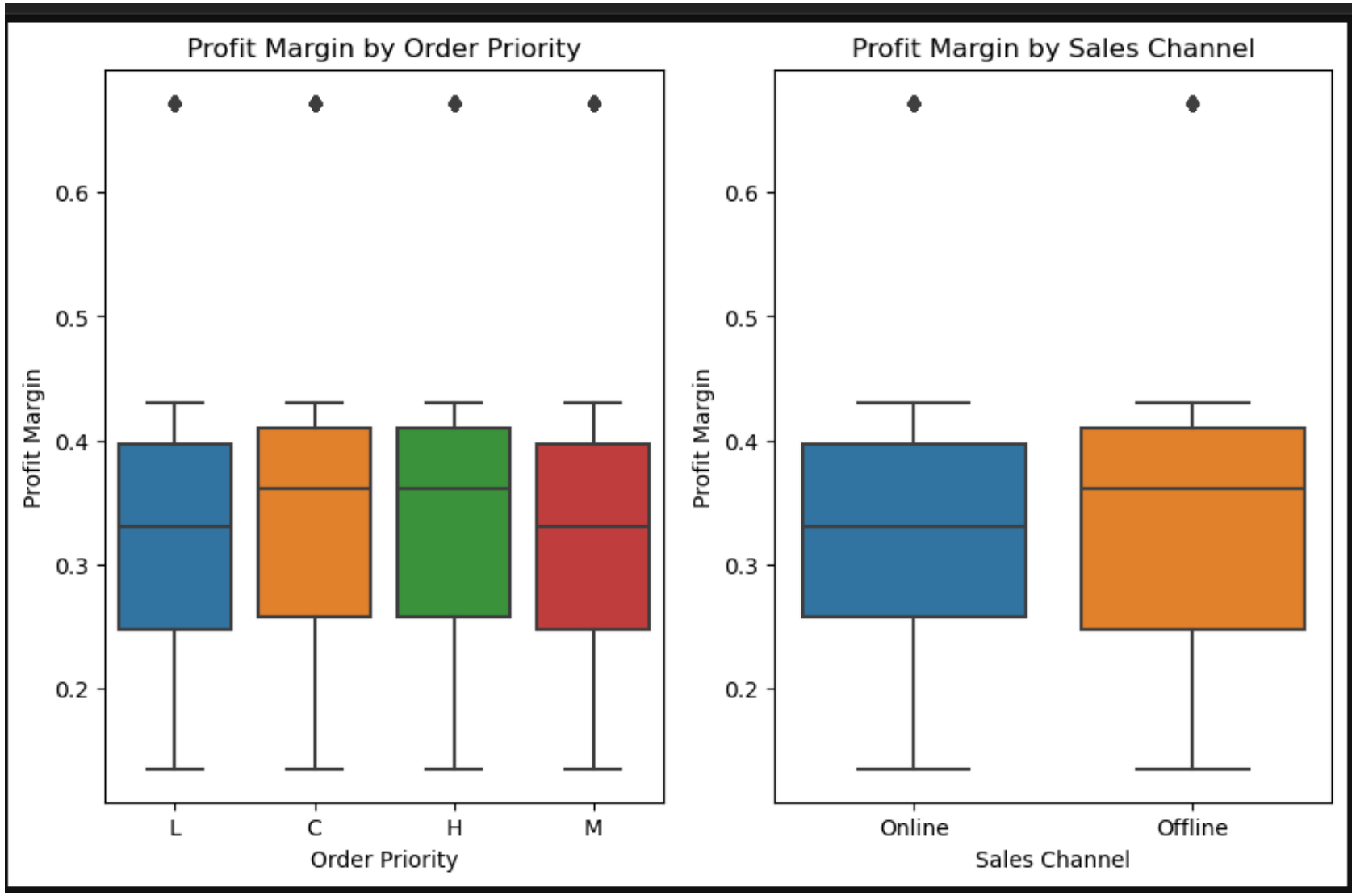
```
df['Profit Margin'].describe()
```

```
count    10000.000000
mean         0.344981
std          0.132653
min          0.135580
25%          0.247999
50%          0.361384
75%          0.409775
max          0.672035
Name: Profit Margin, dtype: float64
```

Box Plot Analysis of Engineered Profit Margin Feature



Comparative Analysis of Profit Margin by Order Priority and Sales Channel

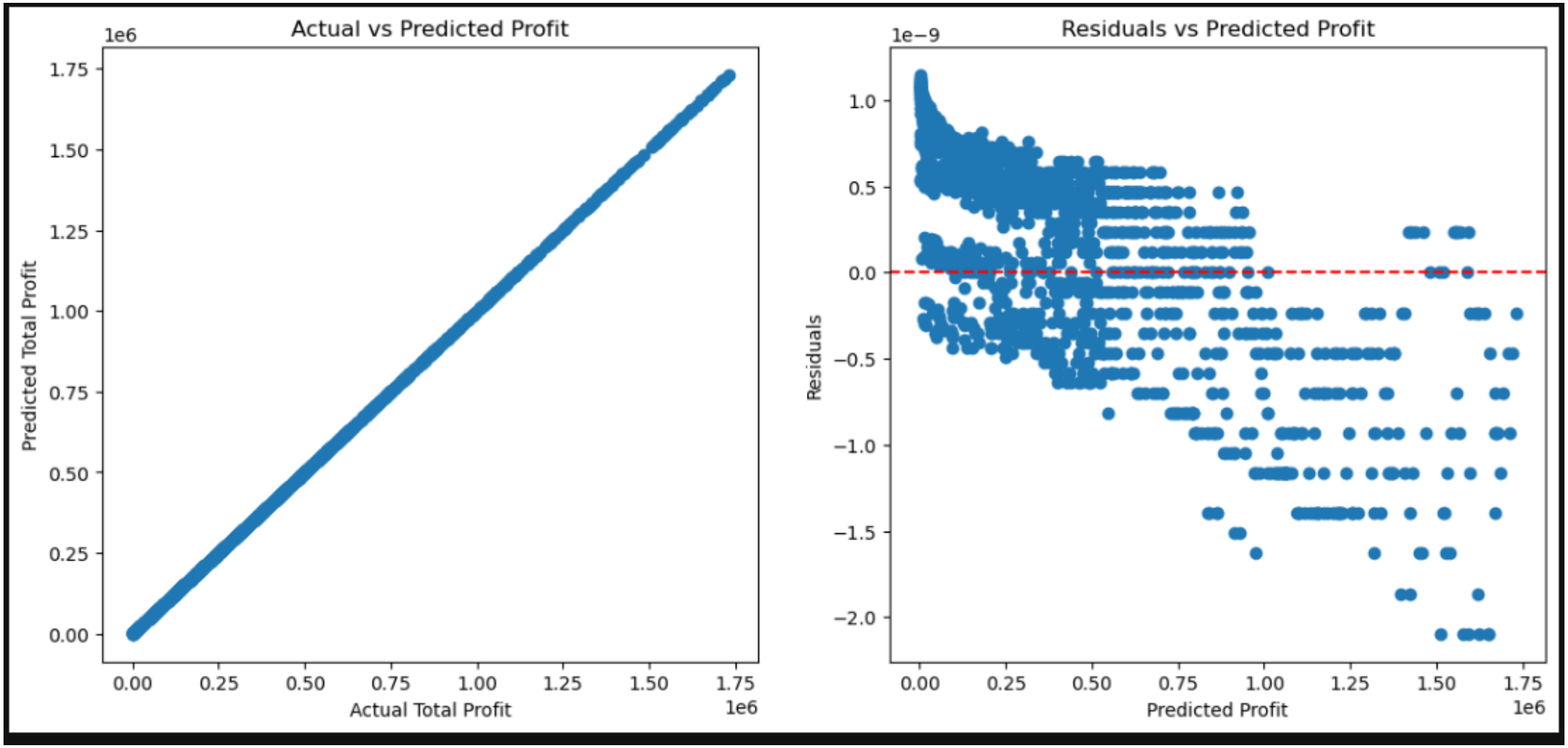


Predictive Profitability Analysis

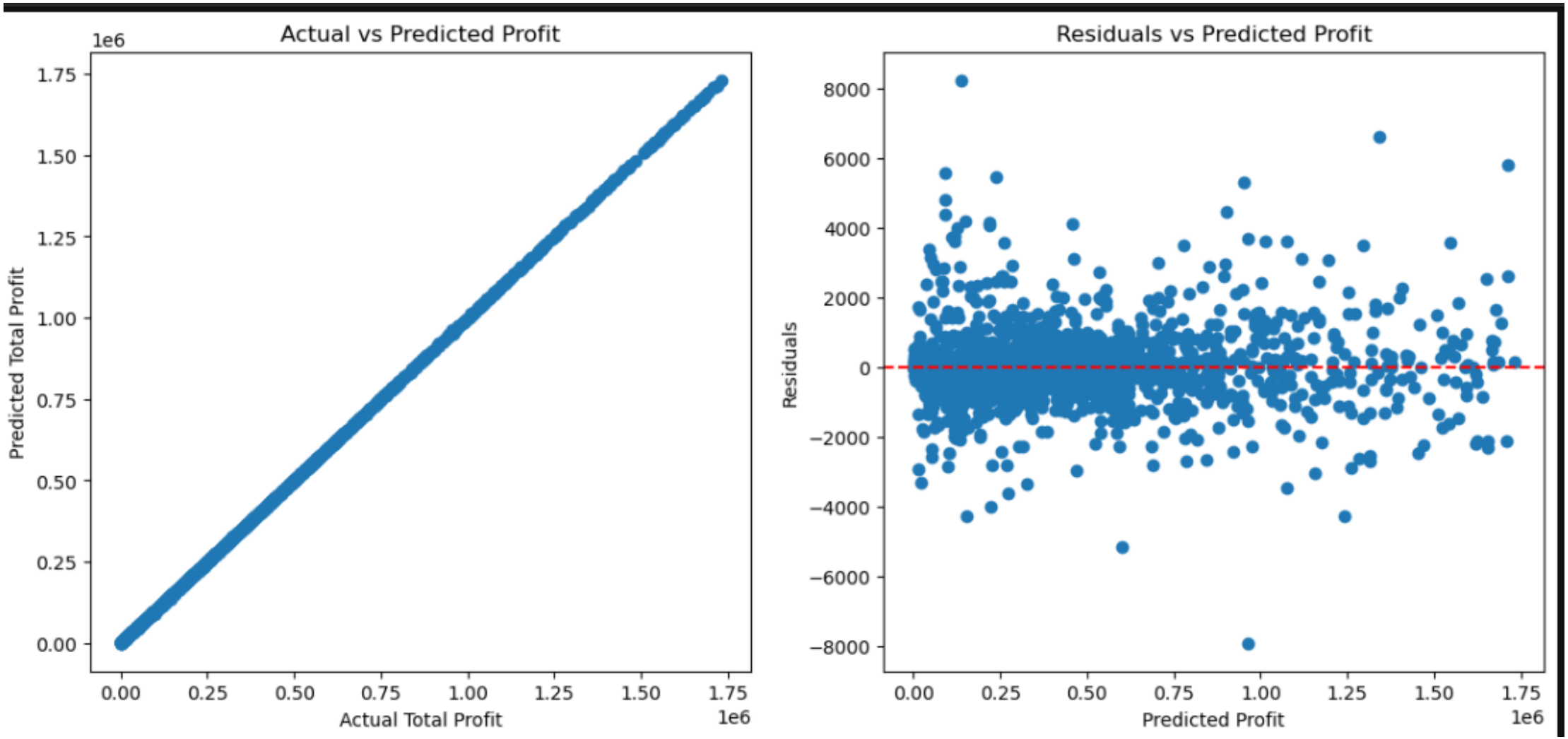
Model Performance Metrics Table

s/n	Model	Mean Square Error	Mean Absolute Error	R-squared
1	Linear Regression	4.13e-19	5.56e-10	1.000000
2	Random Forest Regression	9.94e+05	5.96e+02	0.999993
3	Gradient Boosting Regression	1.37e+08	8.48e+03	0.999070
4	Ridge Regression	7.68e-17	6.21e-09	1.000000

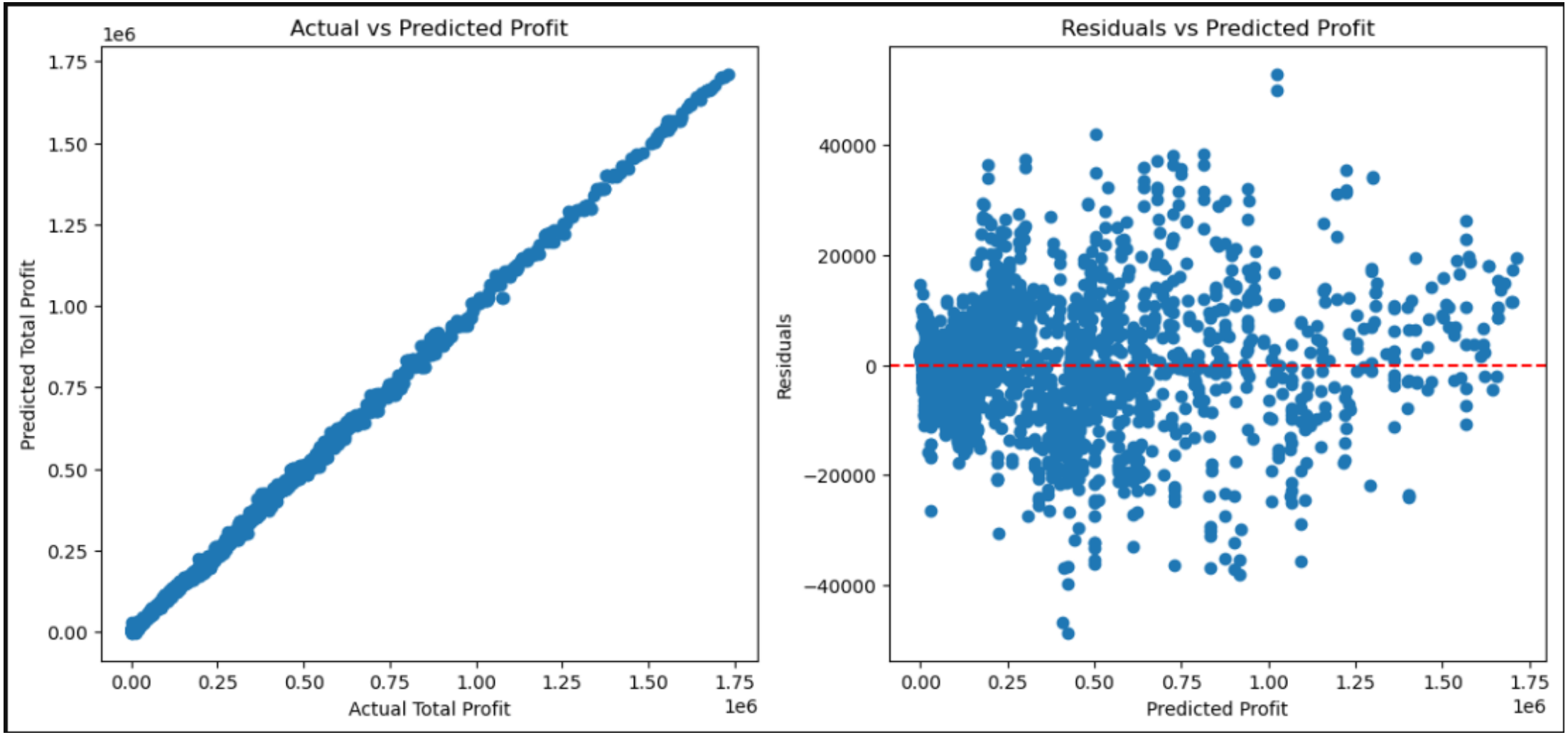
Performance Visualization of Linear Regression Model



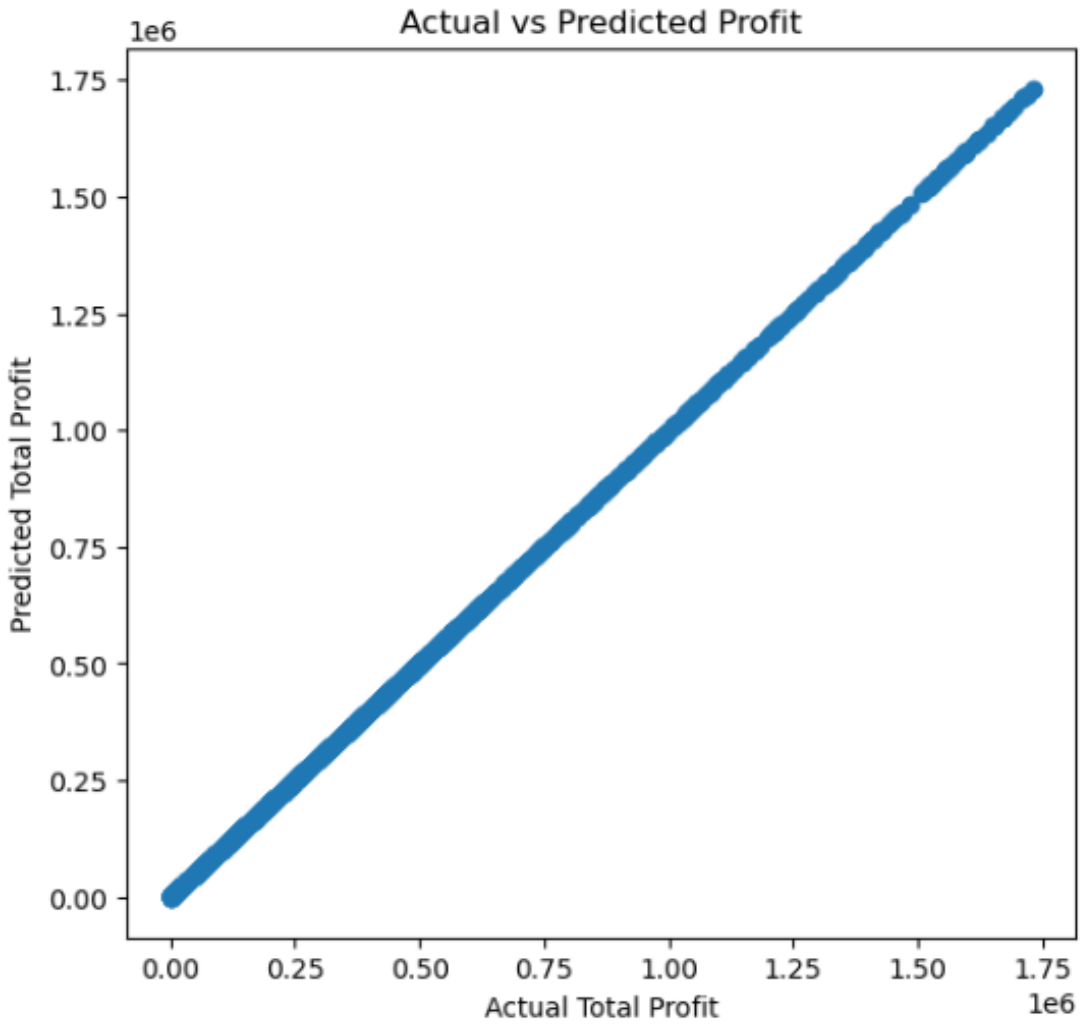
Performance Visualization of Random Forest Regressor Model

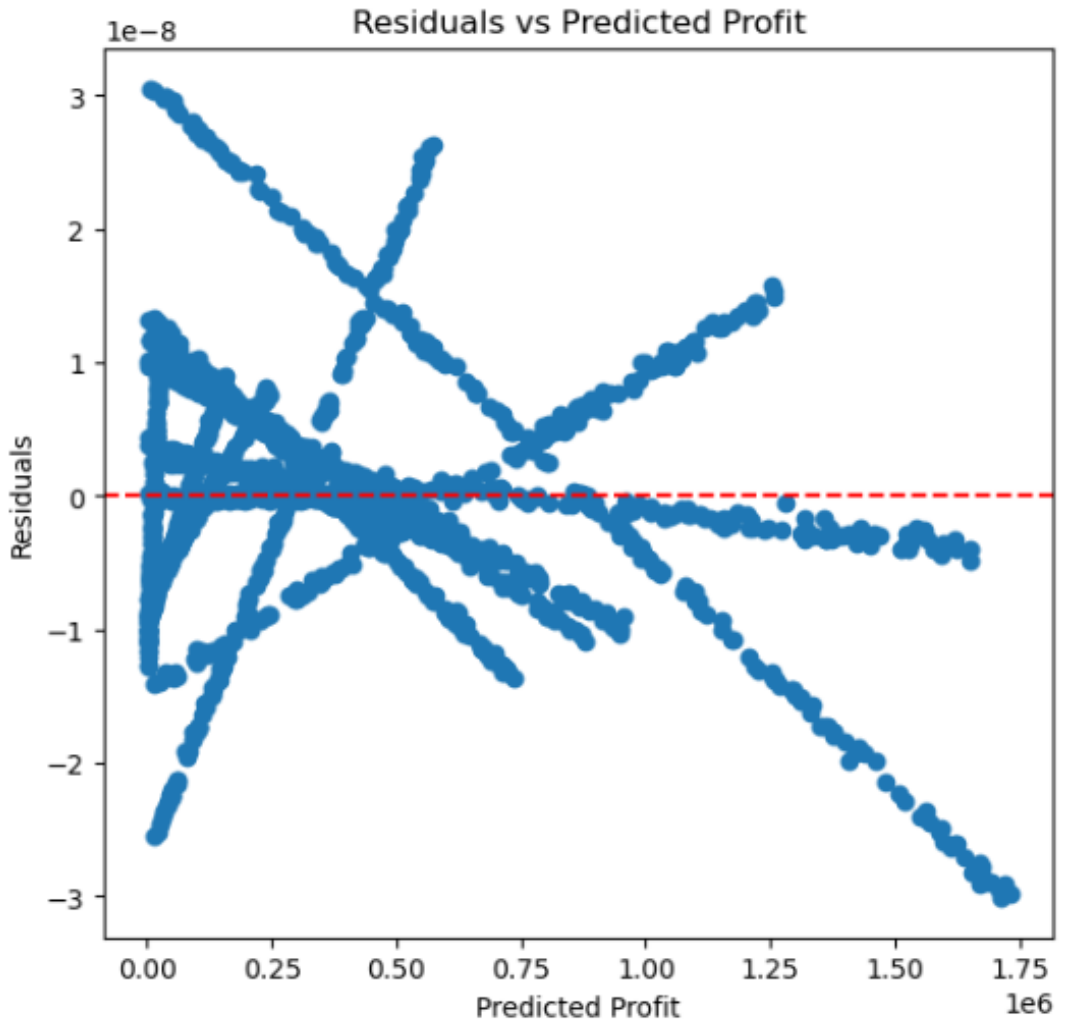


Performance Visualization of Gradient Boosting Regressor Model



Performance Visualization of Ridge Regression Model

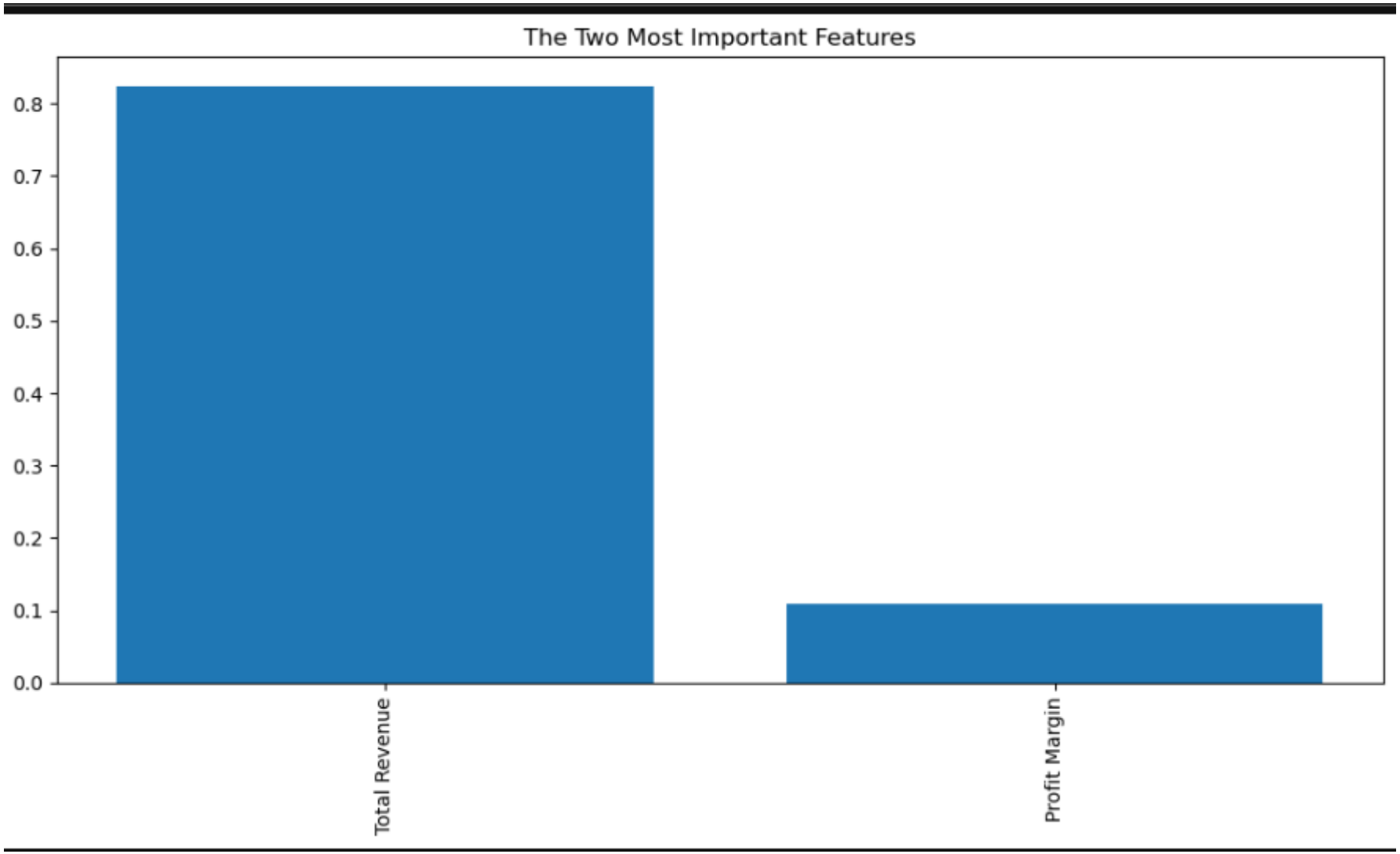




Feature Importance Rankings: Values in Order of Preference

Rank	Feature	Importance
1st	Total Revenue	8.2e-01
2nd	Profit Margin	1.1e-01
3rd	Units Sold	3.7e-02
4th	Total Cost	8.5e-03
5th	Unit Cost	3.7e-03
6th	Unit Price	1.3e-03
7th	Order Priority	2.3e-06
8th	Sales Channel	1.1e-06

Visual Representation of Key Features in Predictive Profitability Analysis



Conclusions

Key Insights from the Analysis

- **Regional Profitability**: Europe and Sub-Saharan Africa are profit hotspots.
- **Product Contributions**: Cosmetics and household items contribute significantly to total profit.
- **Sales Channel Impact:** No significant differences across sales channels.
- **Order Priority:** "Critical" orders yield the highest profits.
- **Correlation Insights:** Strong positive correlations found among unit price, unit cost, and total revenue. Medium negative correlations with profit margin.
- **Outliers:** Z-score analysis revealed outliers in total revenue, total cost, and total profit.

Model Performance Summary

- **High Predictive Accuracy**
  - Linear and Ridge Regression models achieved $R^2 = 1.0$, indicating perfect fit.
  - Random Forest and Gradient Boosting also displayed strong predictive power with R^2 near 1.
- **Model Selection Insight**
  - Linear & Ridge Regression are preferred for **simplicity** and **interpretability**.
  - Random Forest & Gradient Boosting offer **robustness** for capturing complex patterns.
- **Top Predictive Features**
  - **Total Revenue** and **Profit Margin** are the most impactful, emphasizing the role of revenue generation and cost efficiency in driving profitability.

Recommendations

- **Focus on Profitable Regions and Products**
  - Prioritize Europe, Sub-Saharan Africa, cosmetics, and household items for targeted marketing and operations.
- **Enhance High-Priority Orders**
  - Develop strategies to optimize delivery and reduce processing time for critical orders.
- **Optimize Sales Channels**
  - Investigate channel-specific customer behavior to identify potential profitability improvements.
- **Monitor Cost and Price Dynamics**
  - Regular analysis of pricing and cost management to maintain favorable profit margins.
- **Address Revenue and Cost Outliers**
  - Investigate outliers to identify operational inefficiencies or demand fluctuations for improved profit consistency.

Project Learnings

- Gained experience in **feature engineering** and **predictive modeling**.
- Enhanced understanding of **profitability drivers** in a real-world context.
- Developed skills in **model evaluation**, **data-driven recommendations**, and **insight extraction** for business applications.

Questions and Answers

Thank You for your Attention