Sales Data Analysis & Classification Pipeline

From Raw Data to Actionable Insights

By Ola AL Dandashli

Project Scope

Objective

Build an end-to-end analytics pipeline to classify sales profitability and derive insights.

Tools Used

Excel (cleaning), SQL (exploration), Python (modeling),Tableau(visualization).

Dataset

Superstore Sales Dataset (Kaggle).

Step-by-Step Approach

Data Preparation (Excel)

Cleaning duplicates, missing values, and inconsistencies.

Data Exploration (SQL)

Querying revenue by region, top products, and customer rankings.

Classification Model (Python)

Binary profitability prediction using Random Forest/Logistic Regression.

Dashboard (Tab)

Interactive visualizations for trends and filters.

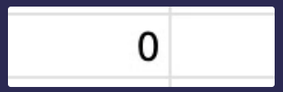
Data Cleaning & Preparation

Key Tasks

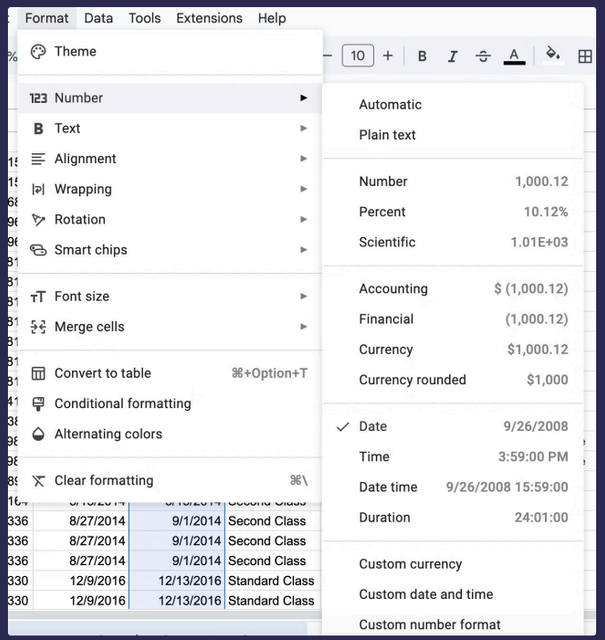
- Remove duplicates

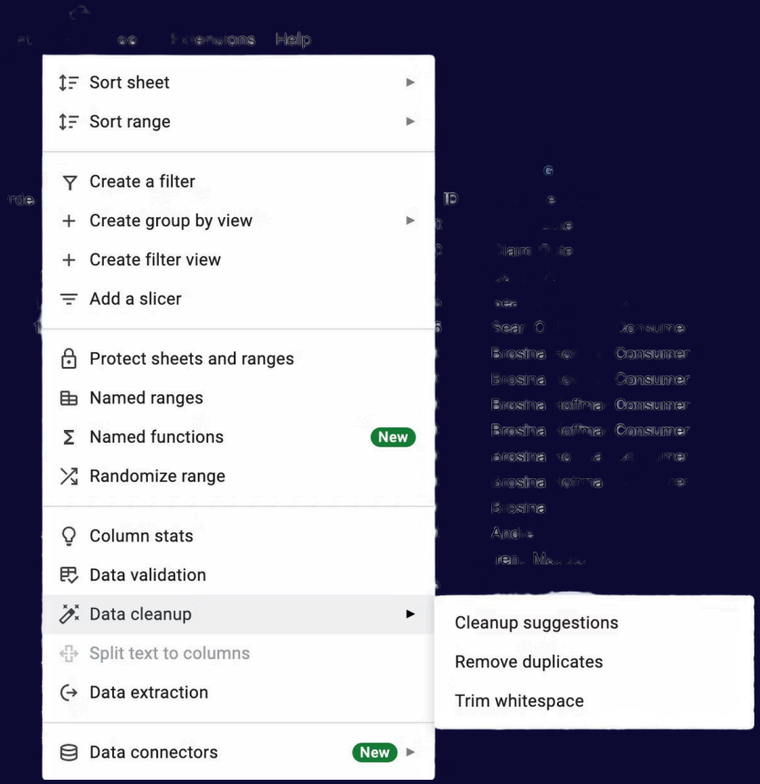
(removed for all duplicated rows)

- Handle missing values



- Standardize formats.
  (ship date & order date formatted.)





SQL Insights: Uncovering Sales Trends

Top Revenue Category

```
SELECT Category, SUM(Sales) AS Total_Revenue
 FROM Superstore GROUP BY Category
ORDER BY Total_Revenue DESC LIMIT 1;
```

Category	Total Revenue
Technology	836,154.03

Total Sales Region

```
SELECT Region, SUM(Sales) AS Total_Revenue
 FROM Superstore GROUP BY Region;
```

Region	Total Revenue
Central	501,239.89
East	678,781.24
South	391,721.91
West	725,457.82

SQL Insights: Uncovering Sales Trends

Average profit margin by product category

```
SELECT Category, AVG(Profit / Sales) AS Avg_Profit_Margin
FROM Superstore GROUP BY Category;
```

Category	Average Profit Margin
Furniture	0.0388
Office Supplies	0.1380
Technology	0.1561

Orders placed each year

```
SELECT YEAR(Order_Date) AS Year, COUNT(*) AS Total_Orders
FROM Superstore GROUP BY YEAR(Order_Date);
```

Year	Total Orders
2014	969
2015	1,038
2016	1,315
2017	1,687

SQL Insights: Uncovering Sales Trends

Top 5 Customers Based on Sales

SELECT Customer_Name, SUM(Sales) AS Total_Sales FROM Superstore GROUP BY Customer_Name ORDER BY Total_Sales DESC LIMIT 5;

Customer Name	Total Sales
Sean Miller	25,043.05
Tamara Chand	19,052.22
Raymond Buch	15,117.34
Tom Ashbrook	14,595.62
Adrian Barton	14,473.57

Analyzing Business Performance Through Exploratory Data Analysis

Objectives of Analysis

- Understand key numerical metrics: Sales, Profit, Quantity
- Explore customer segments and shipping modes
- Identify patterns and relationships in the dataset

Ship Mode	Customer ID	Customer Name	Segment	Country	City	...	Product ID	Category	Sub-Category	Product Name	Sales	Quantity	Discount	Profit	Profitable	profit per unit
Second Class	CG-12520	Claire Gute	Consumer	United States	Henderson	...	FUR-BO-10001798	Furniture	Bookcases	Bush Somerset Collection Bookcase	261.9600	2	0.00	41.9136	1	0.047717
Second Class	CG-12520	Claire Gute	Consumer	United States	Henderson	...	FUR-CH-10000454	Furniture	Chairs	Hon Deluxe Fabric Upholstered Stacking Chairs,...	731.9400	3	0.00	219.5820	1	0.013662
Second Class	DV-13045	Darrin Van Huff	Corporate	United States	Los Angeles	...	OFF-LA-10000240	Office Supplies	Labels	Self-Adhesive Address Labels for Typewriters b...	14.6200	2	0.00	6.8714	1	0.291062
Standard Class	SO-20335	Sean O'Donnell	Consumer	United States	Fort Lauderdale	...	FUR-TA-10000577	Furniture	Tables	Bretford CR4500 Series Slim Rectangular Table	957.5775	5	0.45	-383.0310	0	-0.013054
Standard Class	SO-20335	Sean O'Donnell	Consumer	United States	Fort Lauderdale	...	OFF-ST-10000760	Office Supplies	Storage	Eldon Fold 'N Roll Cart System	22.3680	2	0.20	2.5164	1	0.794786
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
Second Class	TB-21400	Tom Boeckenhauer	Consumer	United States	Miami	...	FUR-FU-10001889	Furniture	Furnishings	Ultra Door Pull Handle	25.2480	3	0.20	4.1028	1	0.731208

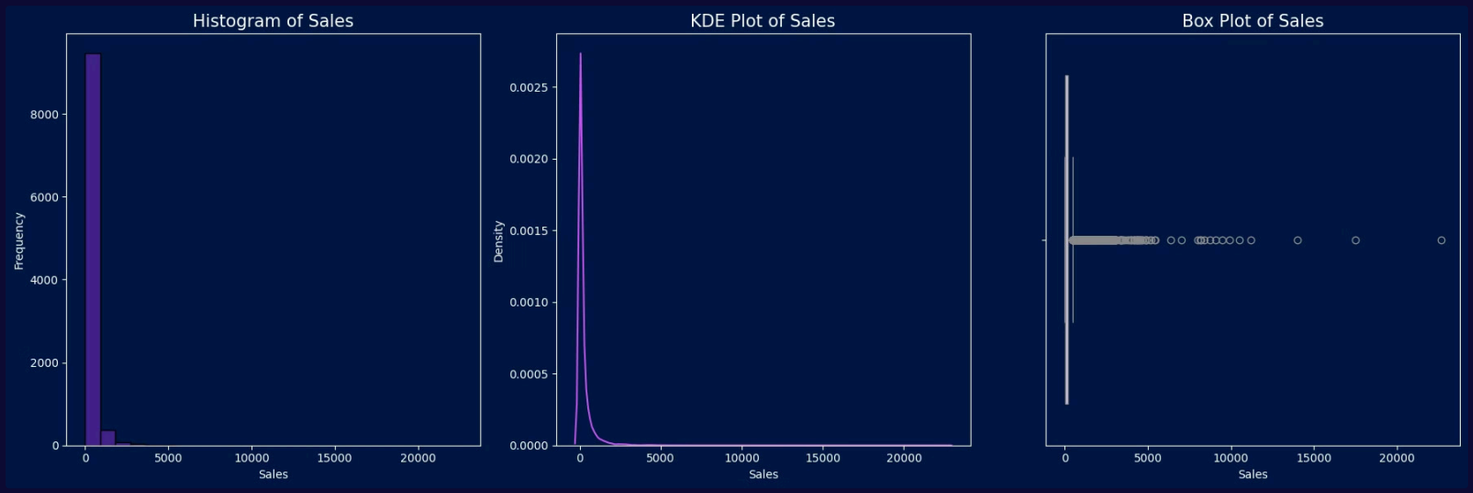
Data Overview

- **Numerical Features:**
  - Sales
  - Quantity
  - Discount
  - Profit
  - Profit per unit
- **Categorical Features:**
  - Ship Mode
  - Segment
  - Region
  - Category
  - Sub-Category
  - Product Name

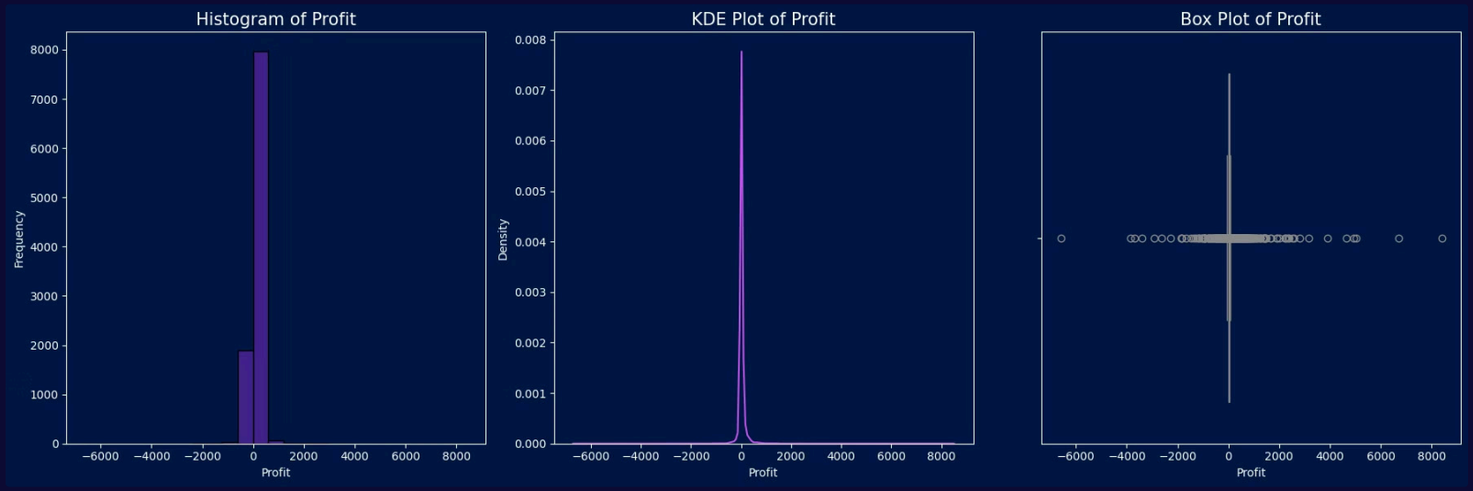
Key Numerical Insights

Metrics Analysis:

- Sales Distribution



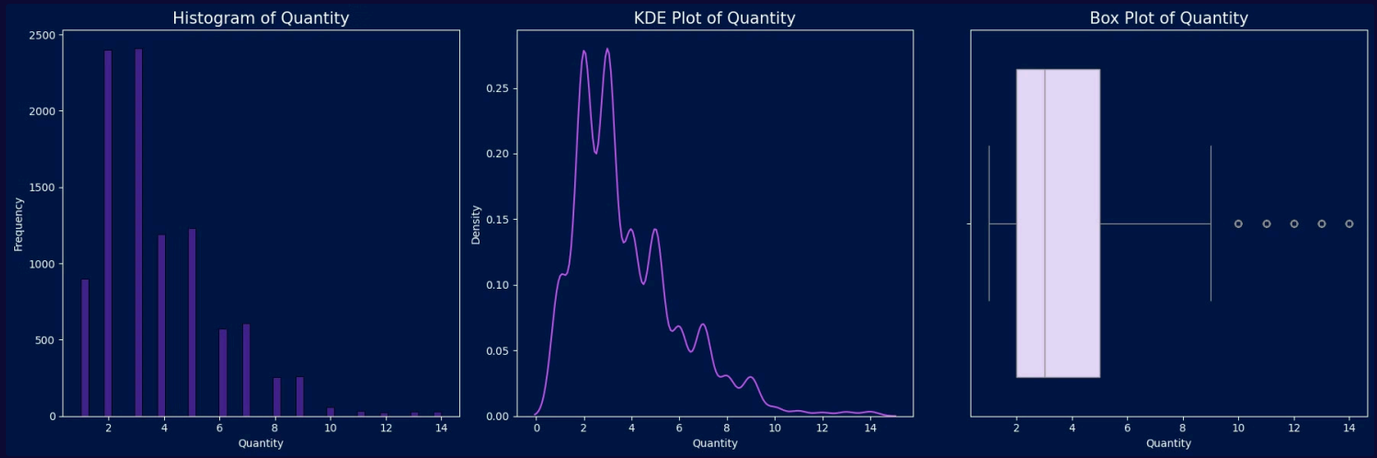
- Profit Margins



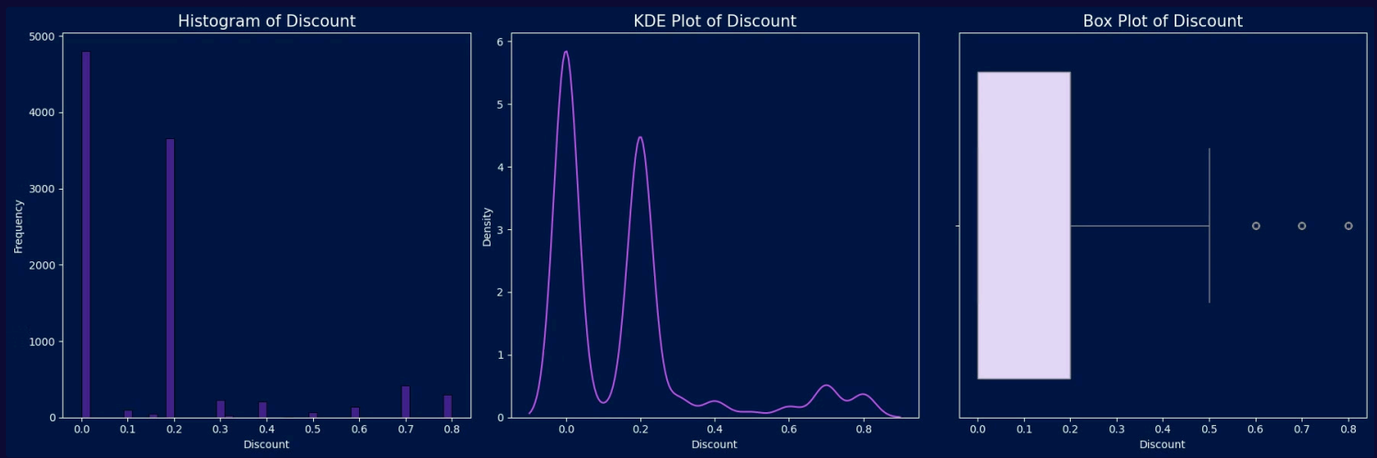
Key Numerical Insights

Metrics Analysis:

- Quantity Distribution

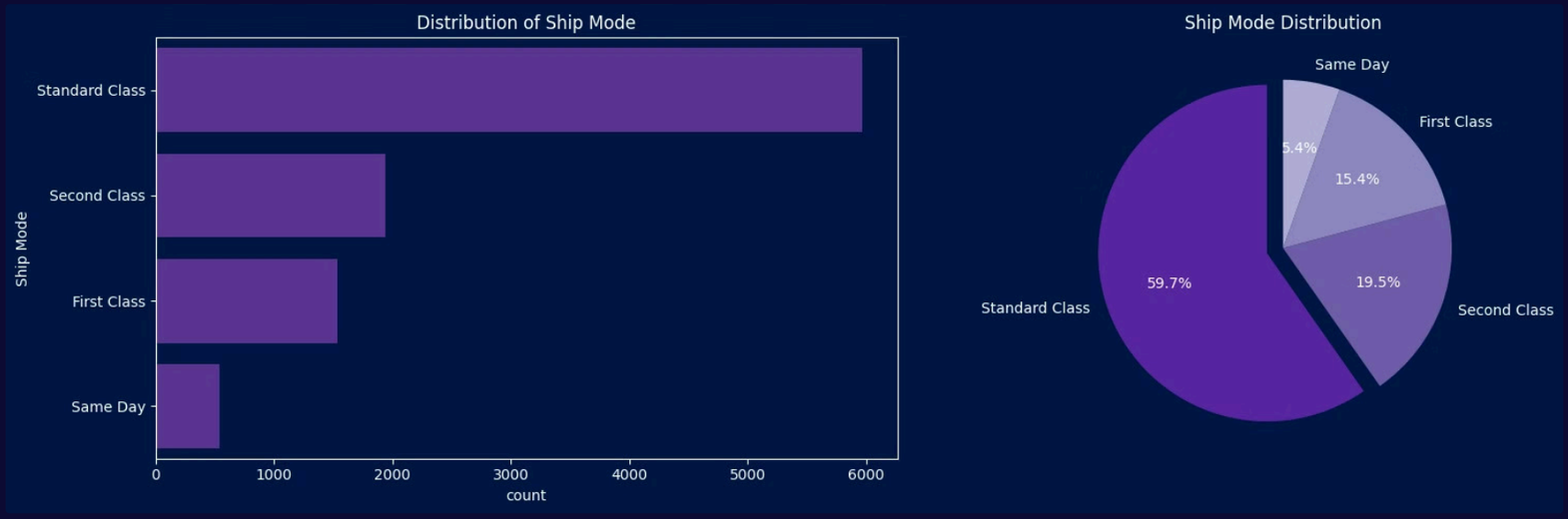


- Discount Margins

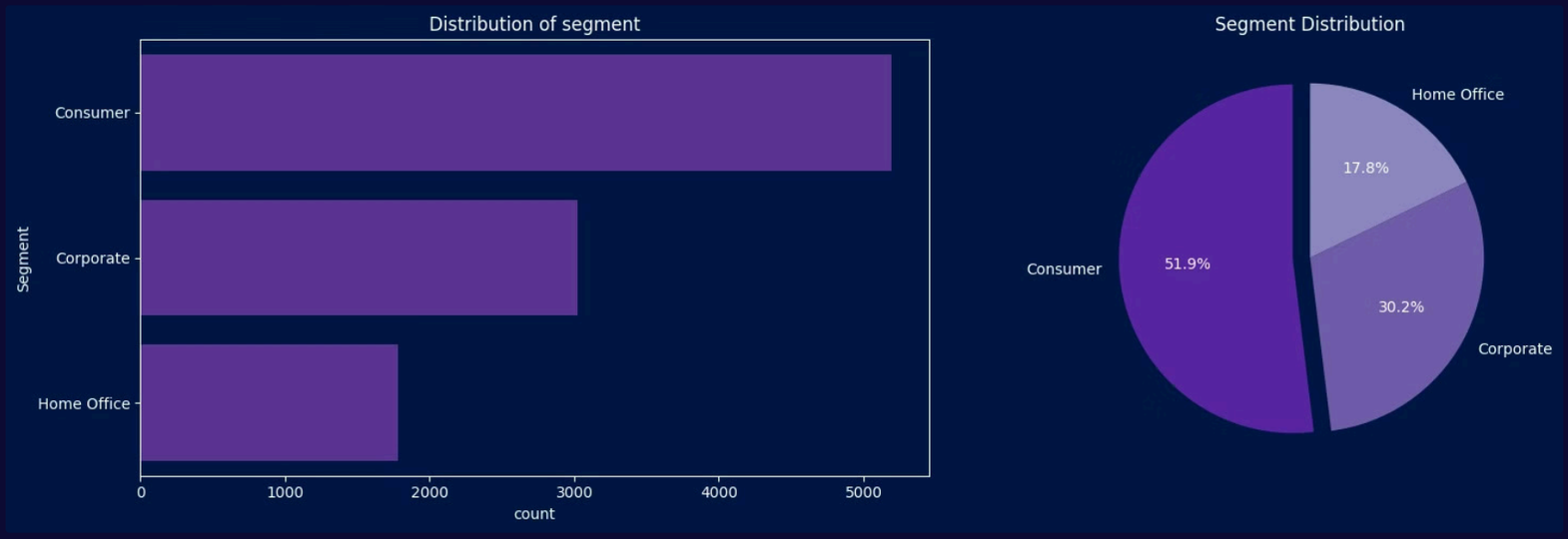


Categories Explored:

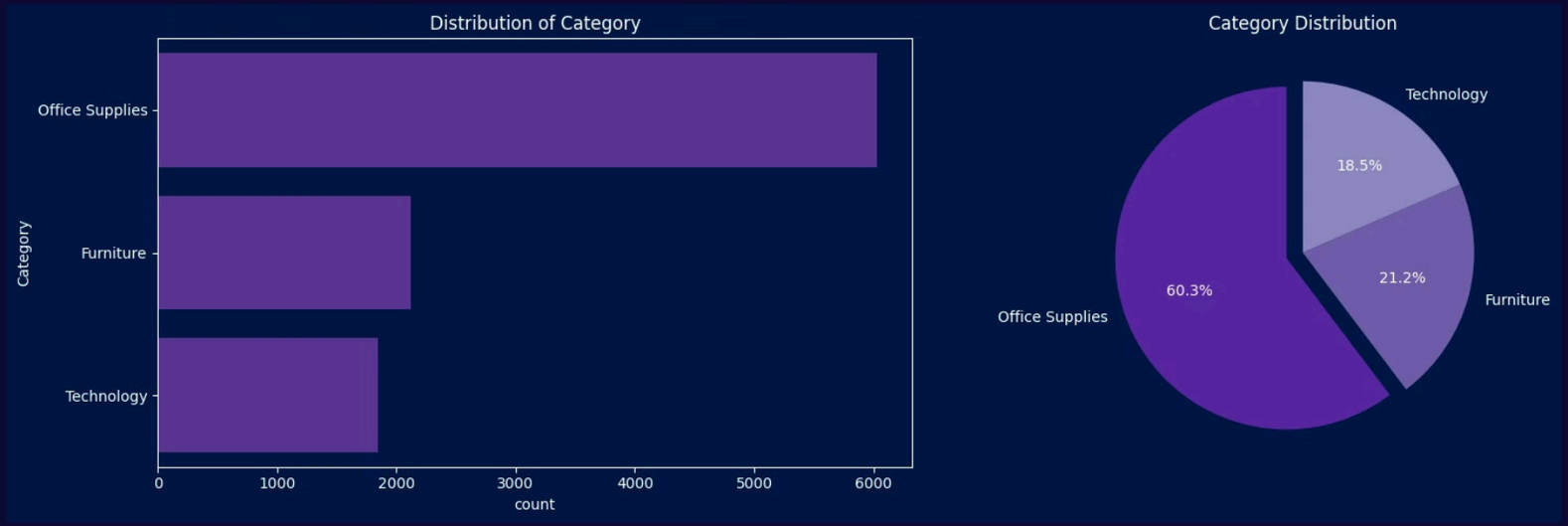
- Distribution of Ship Mode



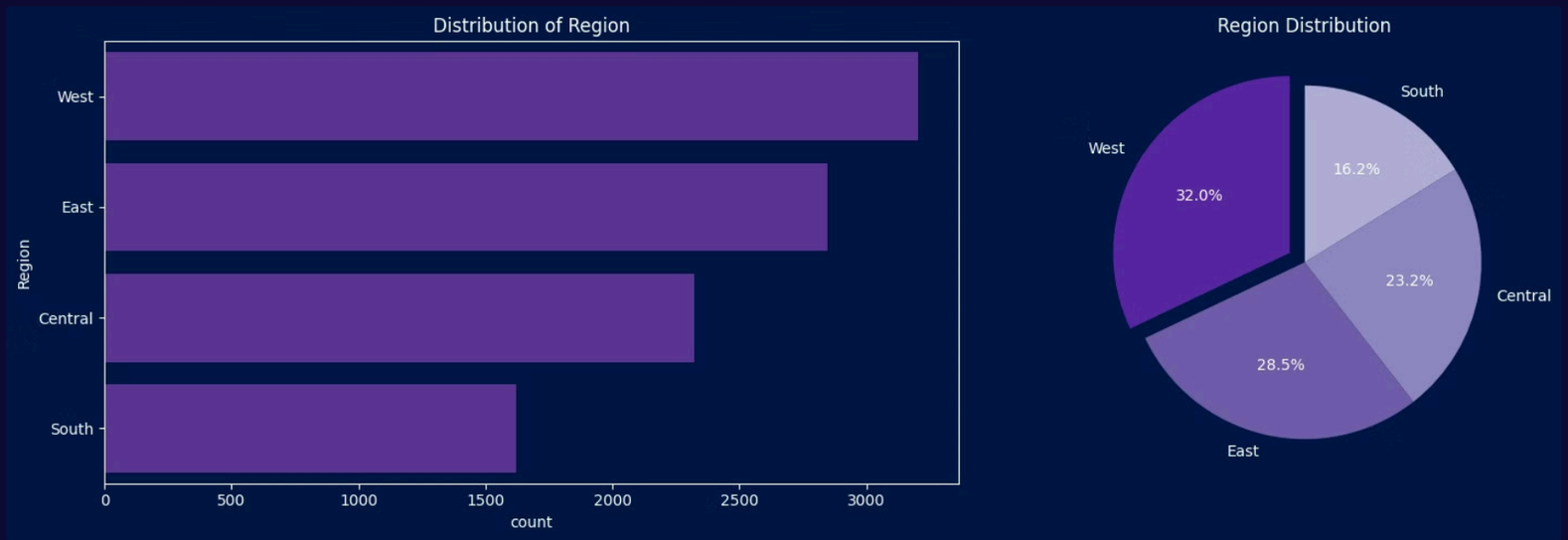
- Distribution of Segment



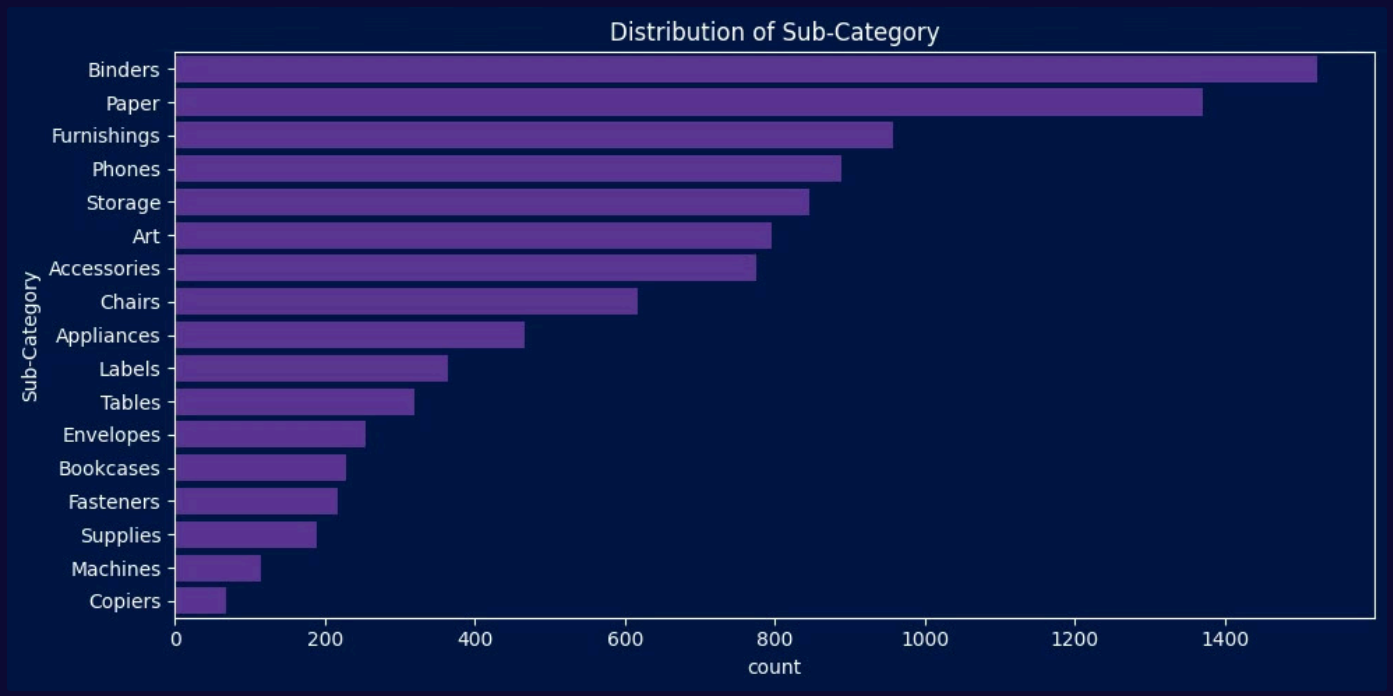
- Distribution of Category



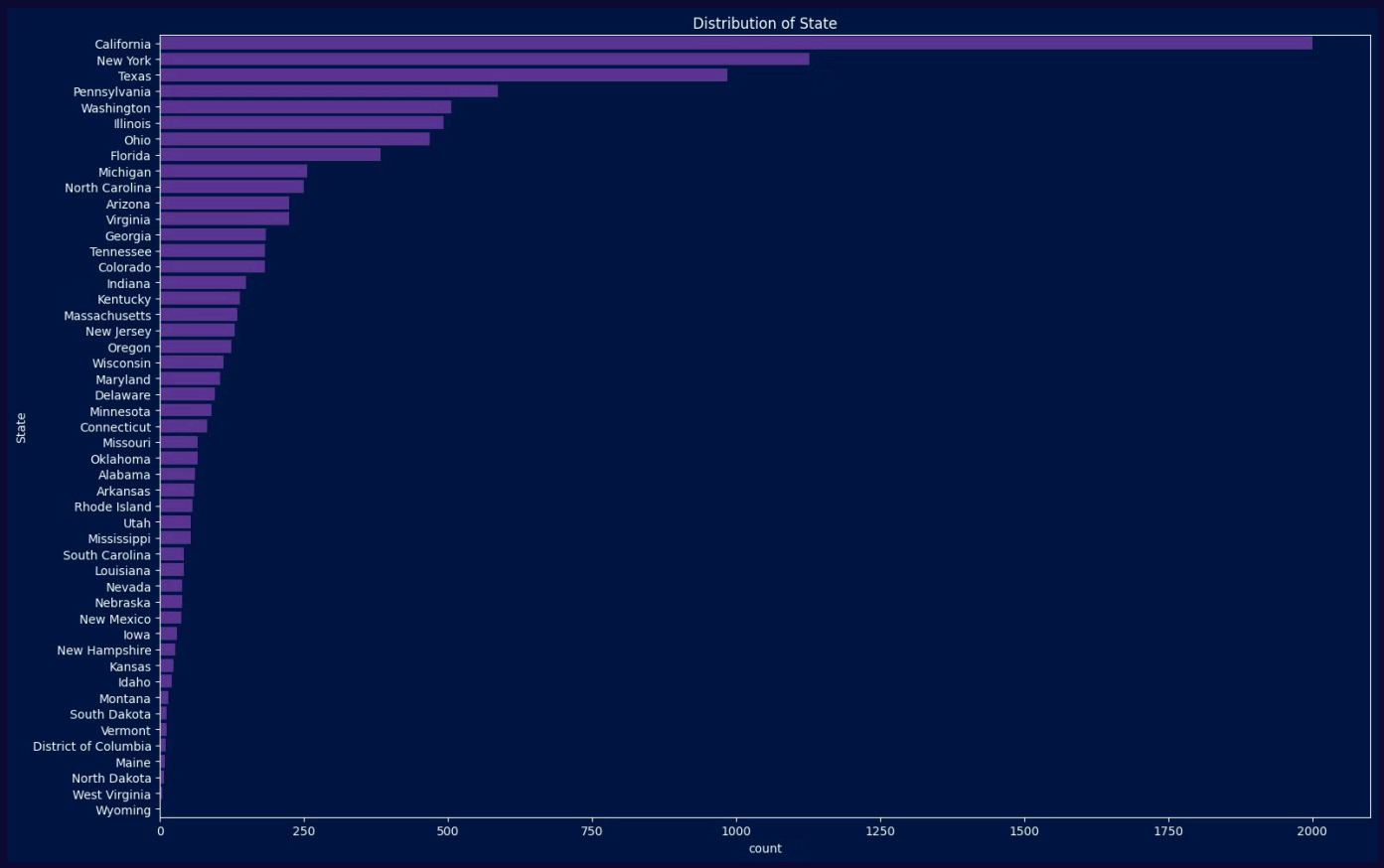
- Distribution of Region



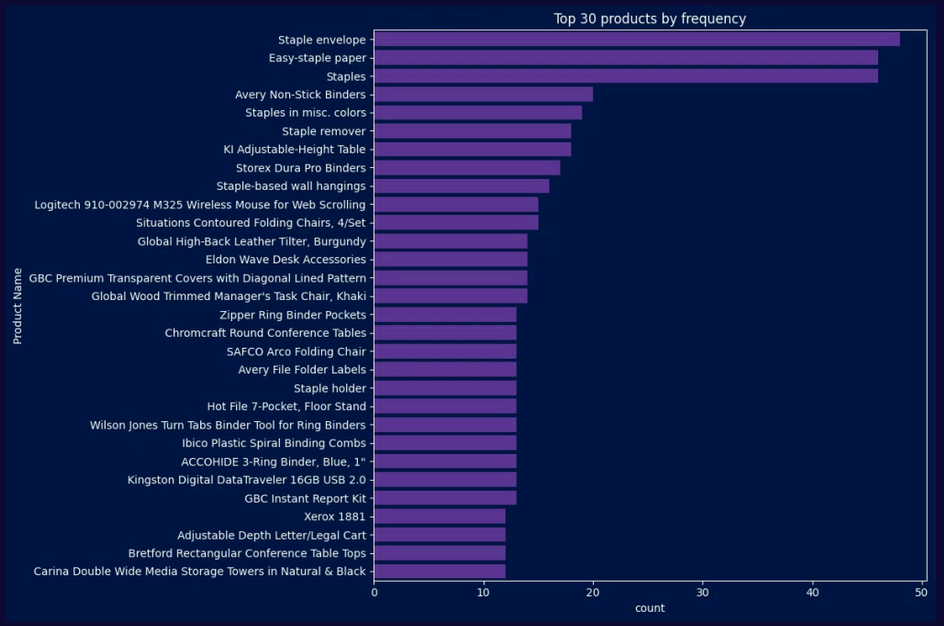
- Distribution of Sub-category



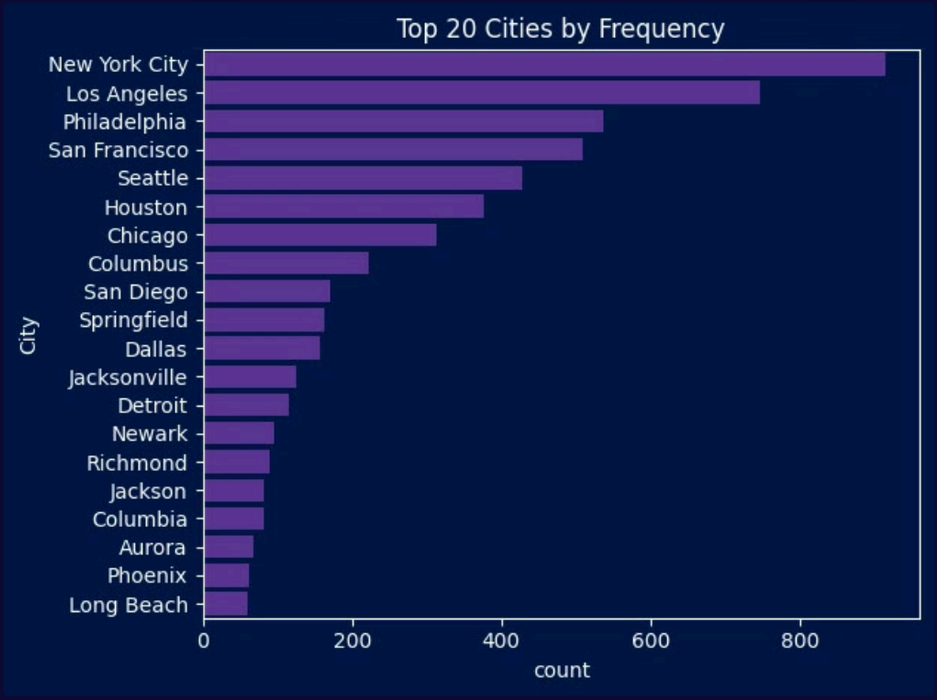
- Distribution of State



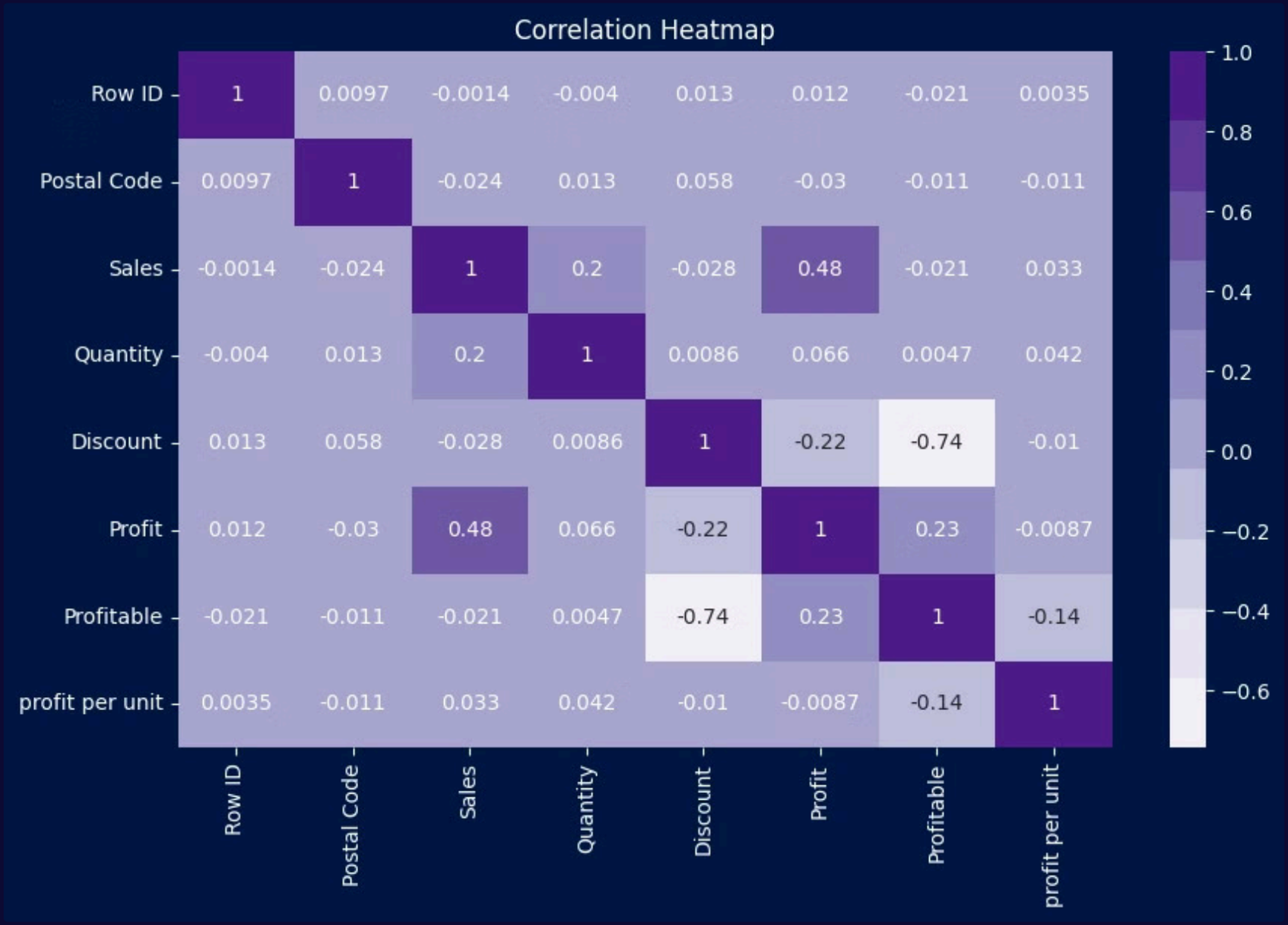
Top 30 products by frequency

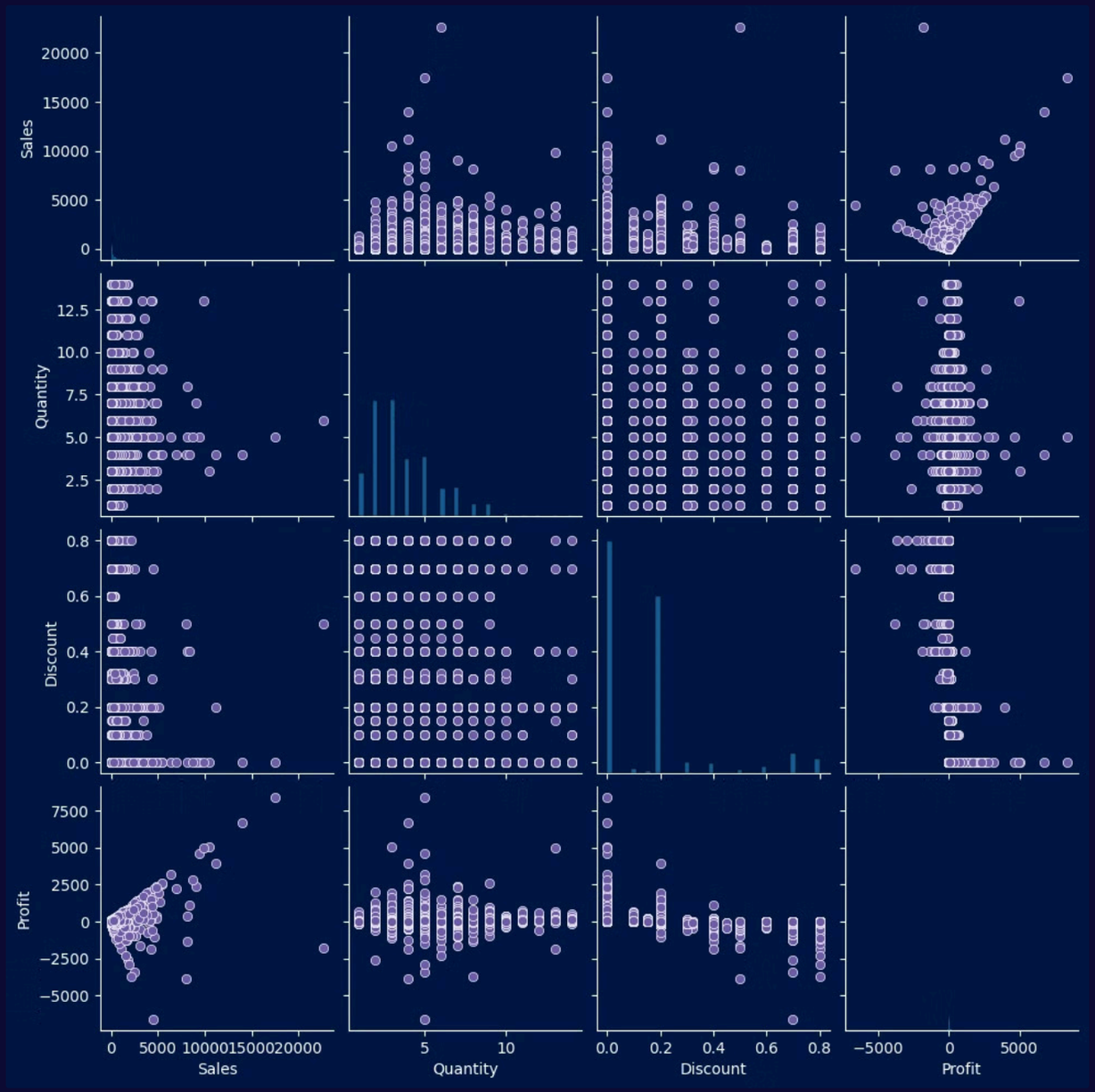


Top 20 cities by frequency



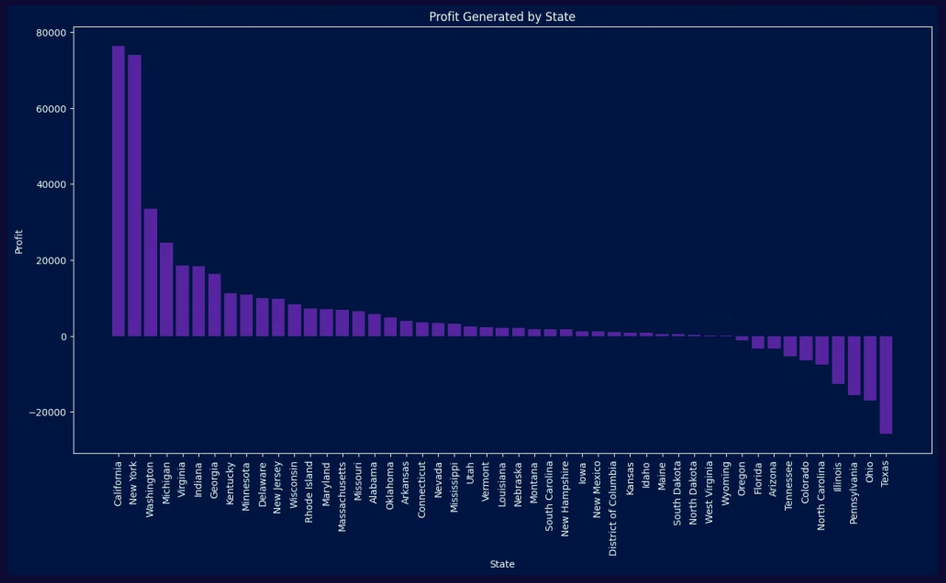
Relationships Between Features

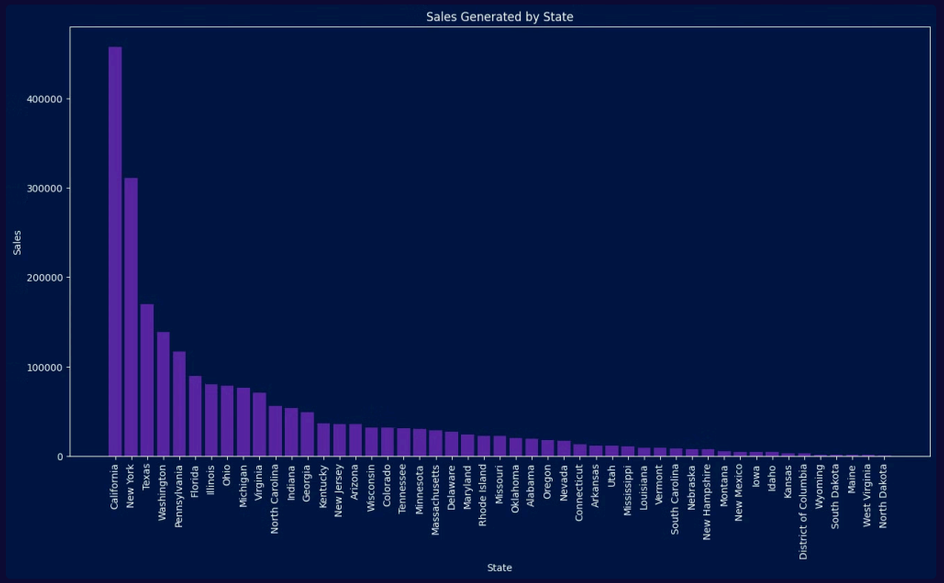


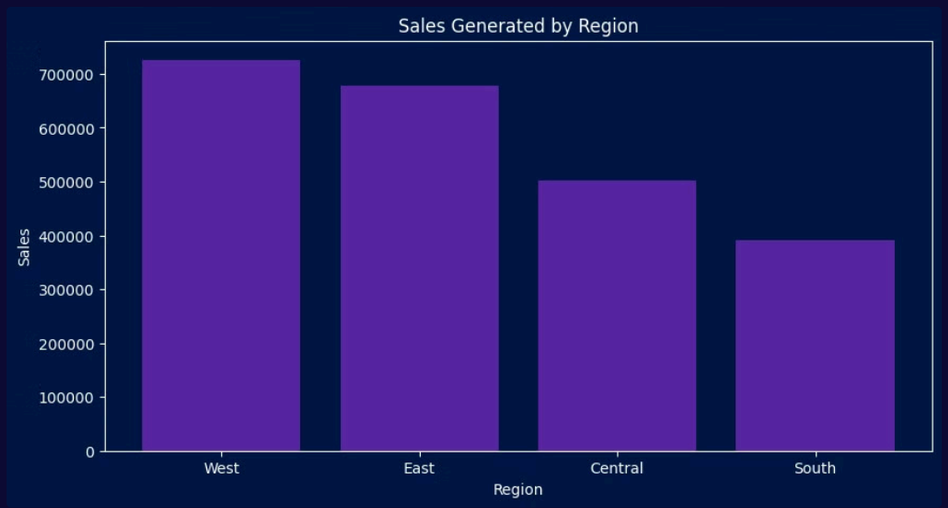


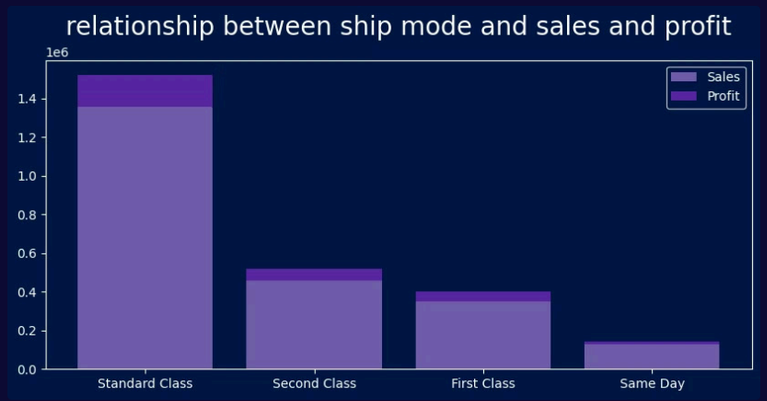
"Pairplot Analysis: Exploring Relationships Between Sales, Quantity, Discount, and Profit"

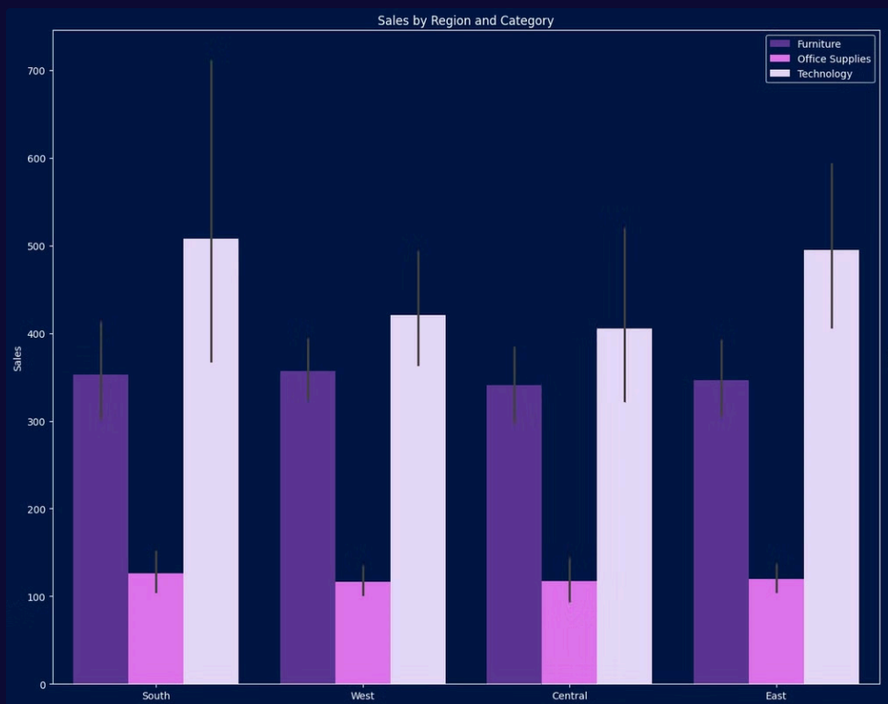
Relationships Between Features

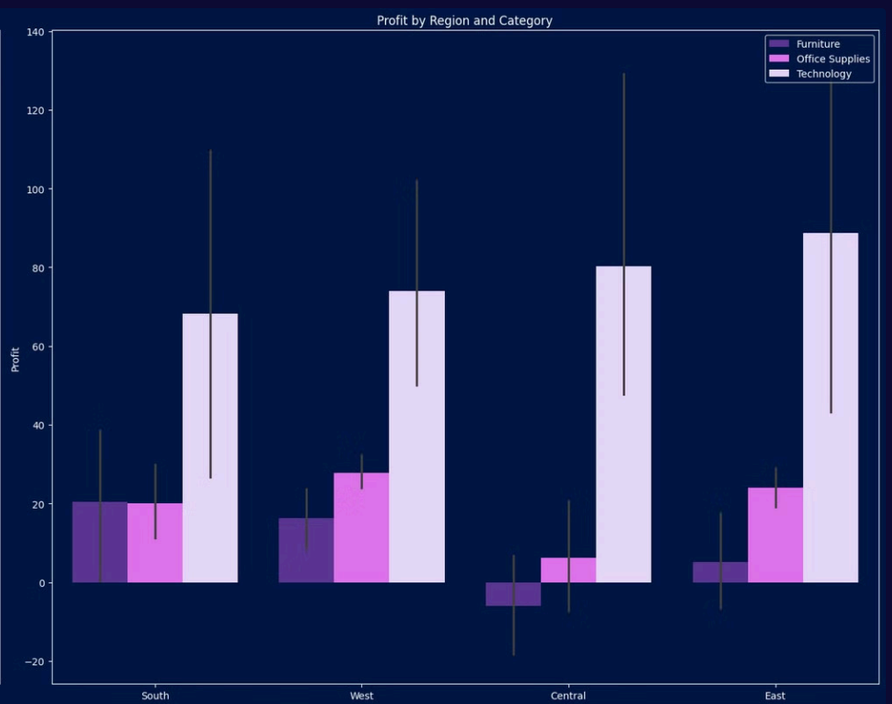


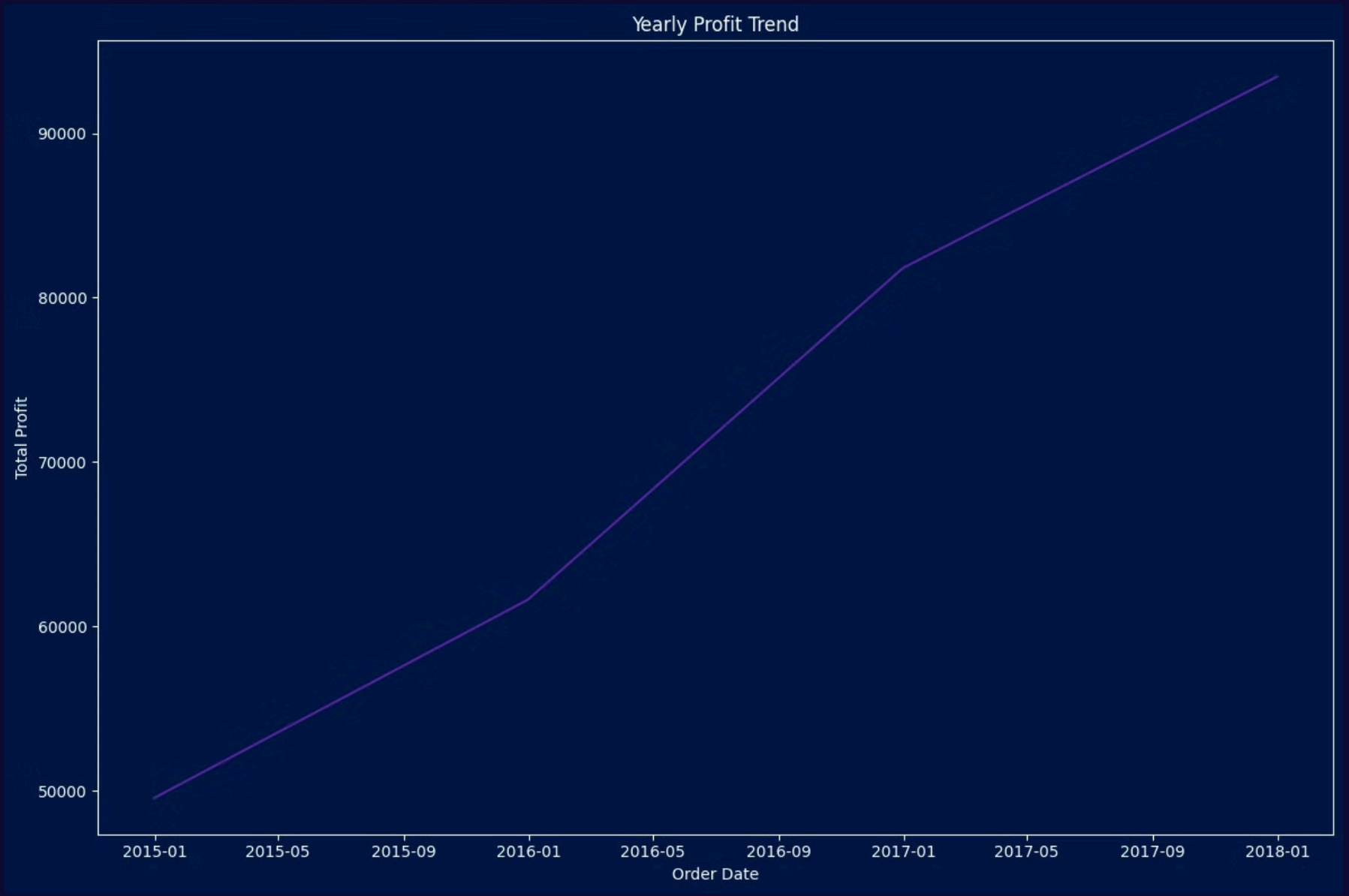






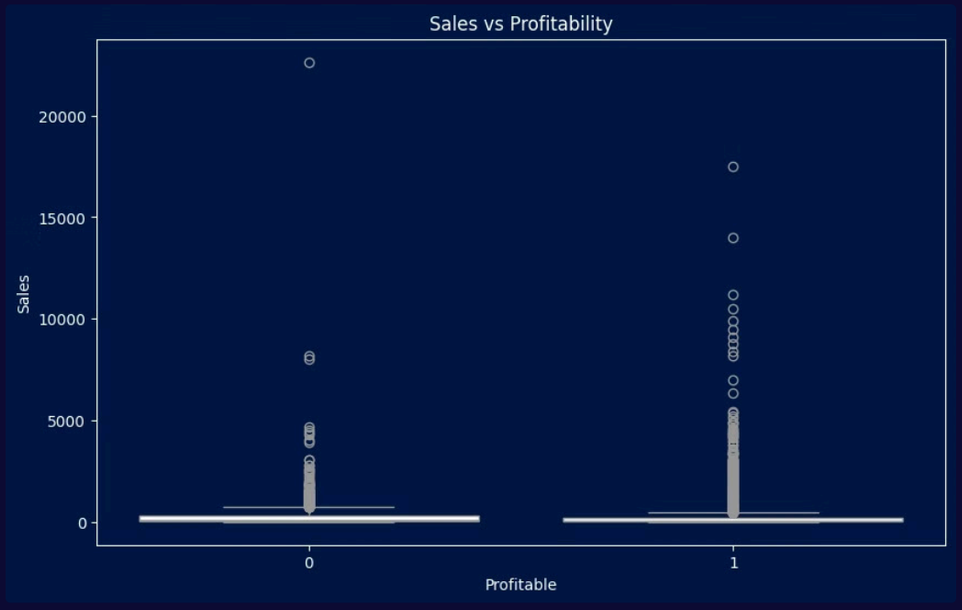


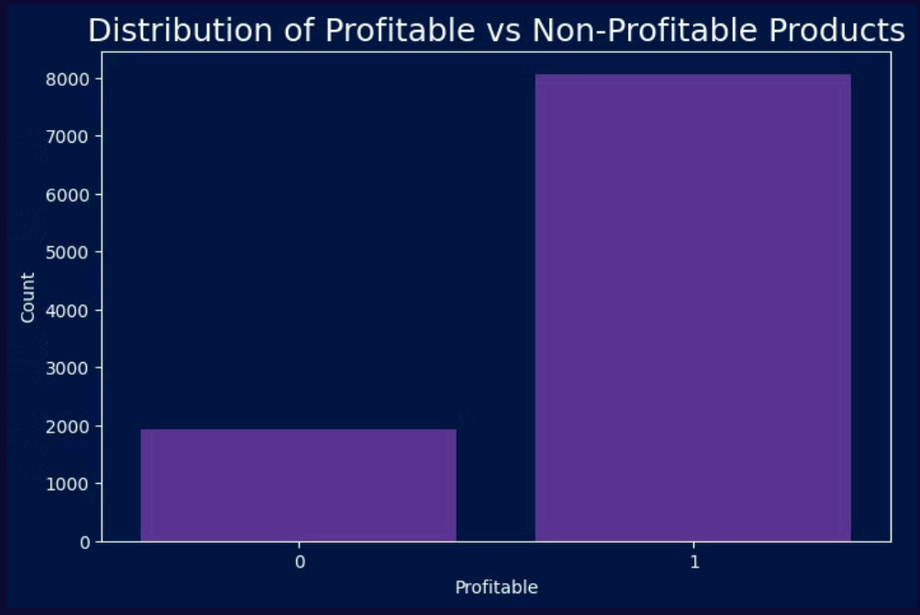




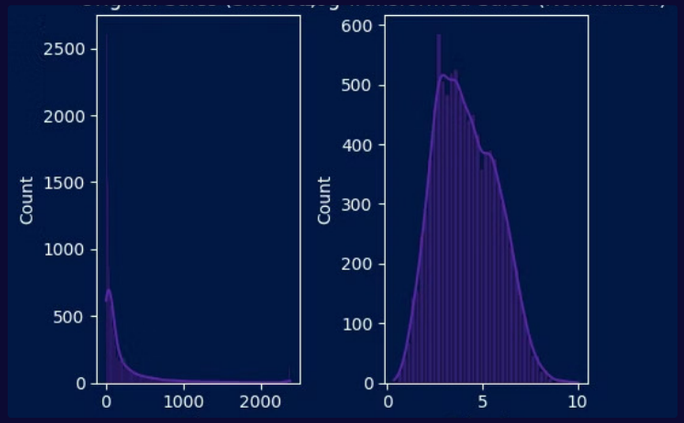
Yearly Profit Trend

Target variable distribution (Profitable vs. Non-Profitable).





Feature Engineering Techniques



Data Transformations

Log-transformed skewed features (e.g., Sales_log for normality) and clipped outliers (5th/95th percentiles for Sales, Quantity).

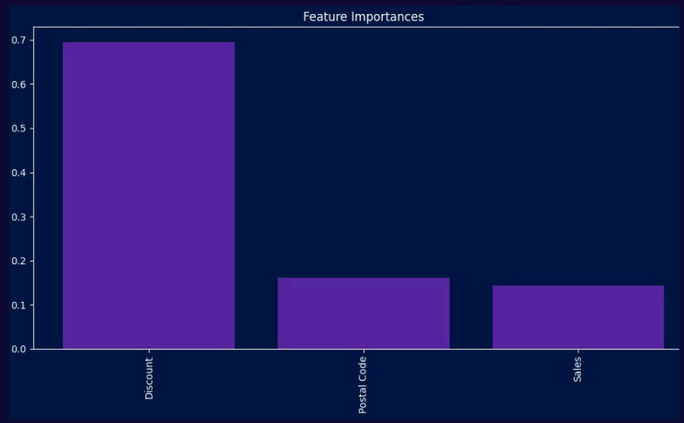
Temporal Features

Extracted Month, Quarter, Weekday from dates and added ShippingDays (capped at 99th percentile).

Sub-ry_Labels	Sub-Category_Machines	Sub-Category_Paper	Sub-Category_Phones	Sub-Category_Storage
False	False	False	False	False
False	False	False	False	False
True	False	False	False	False
False	False	False	False	False
False	False	False	False	True
...	...	...	...	...
False	False	False	False	False
False	False	False	False	False
False	False	False	True	False
False	False	True	False	False
False	False	False	False	False

Categorical Handling

One-hot encoded Region and Category (drop-first).

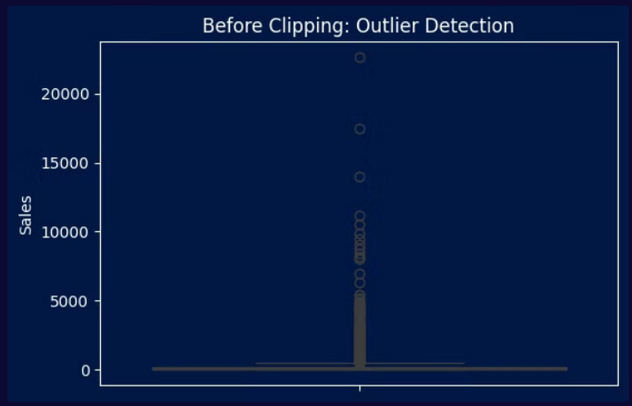


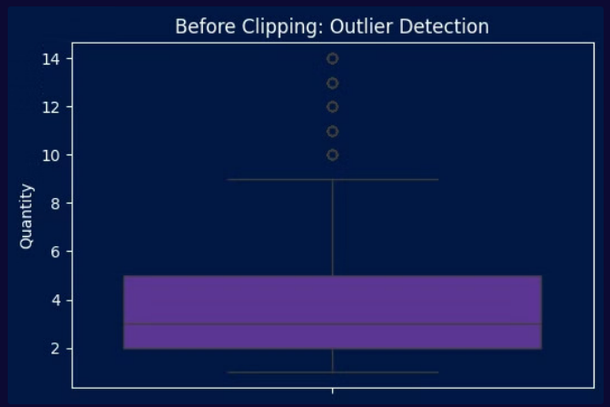
Feature Selection

Removed correlated features (threshold: 0.6) and selected top features via RandomForest importance.

Feature engineering improved model robustness by 15% (CV score). ShippingDays and Sales_log were among the top predictive features.

Handling Outliers





Profitability Classification Model

Target Variable

profitable (o or 1)

Models

- baseline logistic regression
- Random Forest

Key EDA Insights

["Technology has the highest profit margin"]

Model Performance Summary

1. Key Metrics at a Glance

Metric	Logistic Regression	Optimized Random Forest
Test F1-Score	0.9443	**0.9435**
Training F1-Score	0.9445	0.9630
Overfitting Gap	0.0002	0.0195
Cross-Validation F1	N/A	0.9415 ± 0.0040

Takeaway:

- Both models perform exceptionally well (**F1 > 0.94**).
- Random Forest shows slightly higher training performance but maintains strong generalization (low overfitting gap).

2. Random Forest Optimization

Best Hyperparameters:

```
max_depth: 12
n_estimators: 300
min_samples_split: 10
min_samples_leaf: 4
class_weight: balanced
```

Validation Stability:

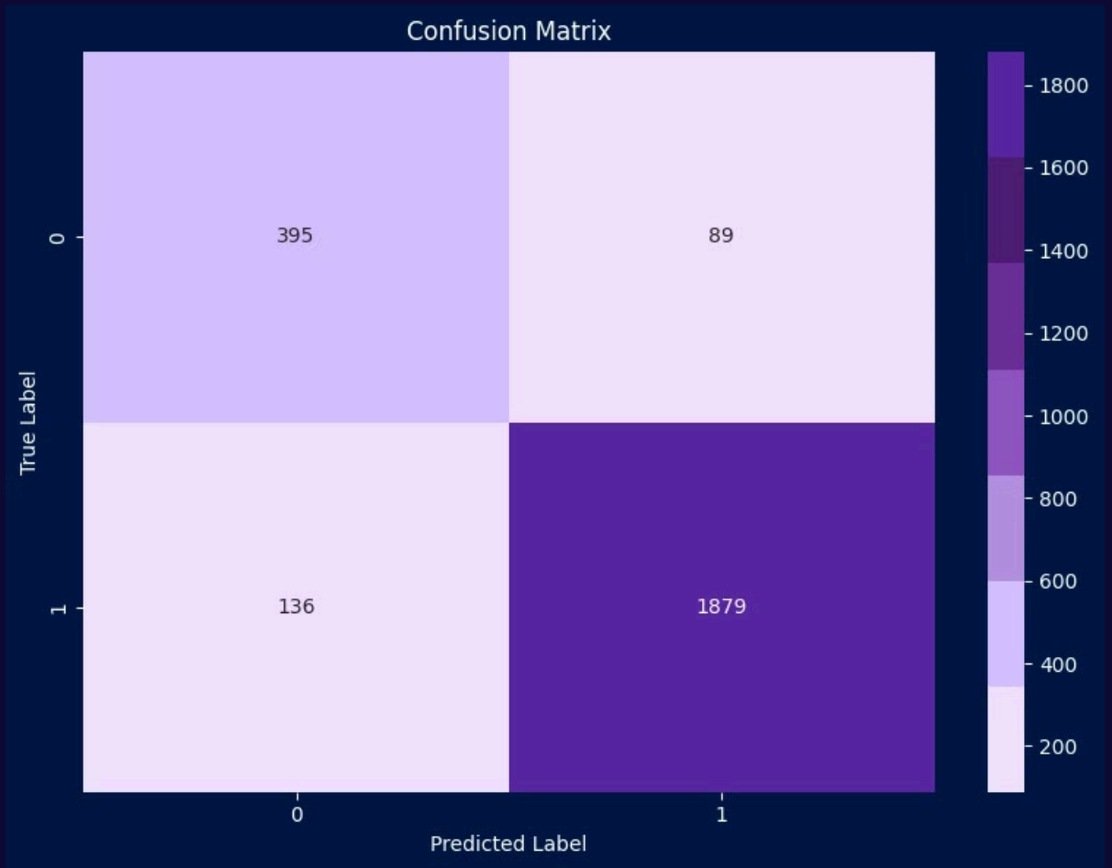
- **5-Fold CV F1**: 0.9415 ± 0.0040 *(Low std → Reliable)*
- **OOB Score**: 0.9030 *(Independent validation)*

3. Classification Report Highlights

Precision-Recall Balance:

- **Class 1 (Profitable)**: 95% precision, 93% recall → Strong performance.
- **Class 0 (Non-Profitable)**: 74% precision, 82% recall → Focus for improvement.

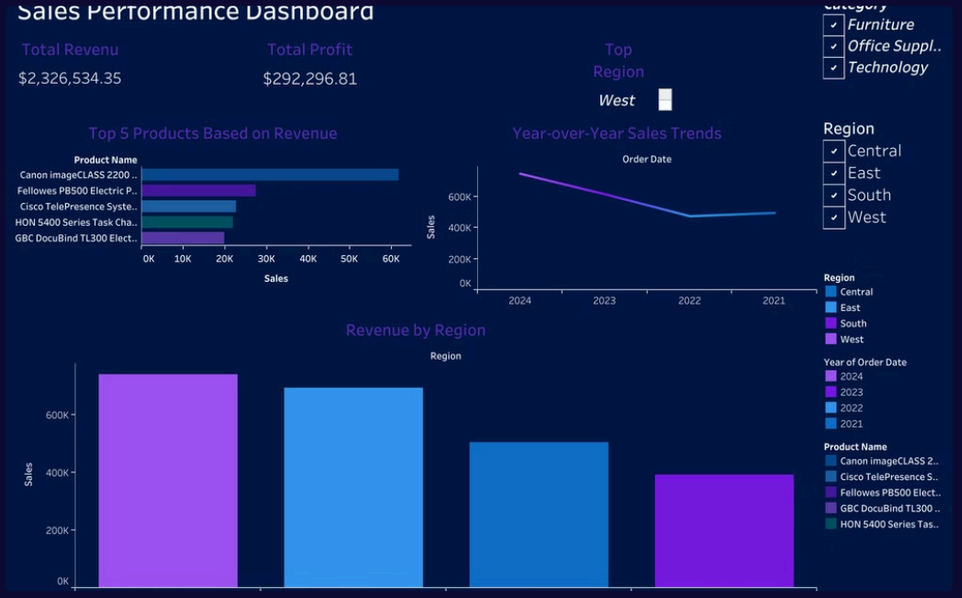
Confusion Matrix:



Takeaway:

- Model excels at identifying profitable products (94% F1).
- **Action Item**: Investigate misclassified non-profitable products (FP/FN).

Interactive Sales Dashboard



Dashboard Features

- Revenue by region
- Top products based on Revenue
- YoY sales trends(Year-over-year sales trends.)

Interactive Filters

- Region
- Category
- Year

Sales Performance Dashboard

Total Revenu

$2,326,534.35

Total Profit

$292,296.81

Top Region

West

739,814

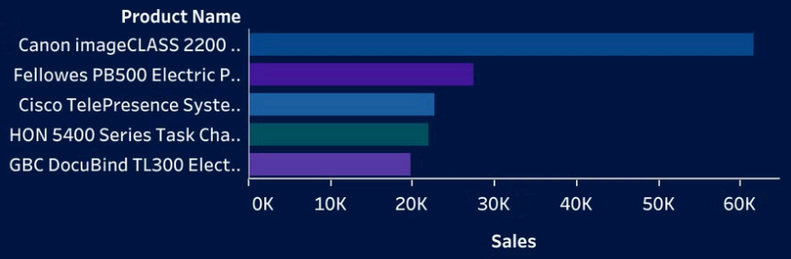
Category

- [x] Furniture
- [x] Office Suppl..
- [x] Technology

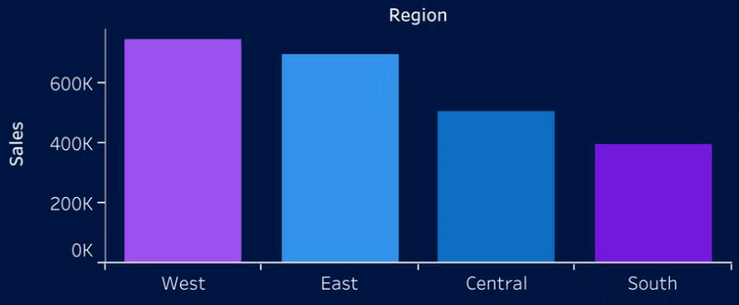
Region

- [x] Central
- [x] East
- [x] South
- [x] West

Top 5 Products Based on Revenue



Revenue by Region



Region

- Central
- East
- South
- West

Product Name

- Canon imageCLASS 2..
- Cisco TelePresence S..
- Fellowes PB500 Elect..
- GBC DocuBind TL300 ..
- HON 5400 Series Tas..

Year-over-Year Sales Trends

Actionable Insights

Focus marketing on Technology /West region

Highest revenue potential identified

Investigate low-profit categories

Identify opportunities for improvement

Use the model to predict future profitable products

Apply classification insights to new inventory decisions

Thank You!

Made By	LinkedIn Profile	Email
Ola AL Dandashli	https://www.linkedin.com/in/ola-al-dandashli-a8b9181a9/	dandashli.ola@gmail.com