

# AMAZON SALES FORECASTING AND ANALYSIS :

TRENDS ,INSIGHTS,AND PERFORMANCE

**Designed and Developed by**

1. **M. SAI RAM 2211CS010379**
2. **P.NIKHITHA 2211CS010438**
3. **M. VINAY 2211CS010367**
4. **G.KISHORE 2211CS010209**

**Guided by**

**Mr.G.Raju**

**Department of Computer Science & Engineering** **MALLA REDDY UNIVERSITY, HYDERABAD**

**2025-2026**



# CERTIFICATE

This is to certify that this is the Application development lab record entitled “**AMAZON SALES DATA ANLAYSIS**”, submitted by **M. SAI RAM (2211CS010379), P.NIKHITHA (2211CS010438), M.VINAY (2211CS010367), G. KISHORE (2211CS010209)** B.Tech III year II semester,Department of CSE during the year 2024-25. The results embodied in this report have not been submitted to any other university or institute for the award of any degree or diploma.

**Internal Guide Dean-CSE**

**MR.G.RAJU Dr.SHAIK MEERAVALI**

**External Examiner**



# DECLARATION

I declare that this project report titled **AMAZON SALES DATA ANLAYSIS**submitted in partial fulfillment of the degree of B. Tech in CSE is a record of original work carried out by me under the supervision of **MRS.G.RAJU** and has not formed the basis for the award of any other degree or diploma, in this or any other Institution or University. In keeping with the ethical practice in reporting scientific information, due acknowledgements have been made wherever the findings of others have been cited.

**Signature**

**M. SAI RAM (2211CS010379) P.NIKHITHA (2211CS010438)**

**M. VINAY (2211CS010367) G.KISHORE (2211CS010209)**

**ACKNOWLEDGEMENT**

We would like to express our gratitude to all those who extended their support and suggestions to come up with this software. Special Thanks to our mentor **MRS.G.RAJU** whose help and stimulating suggestions and encouragement helped us all time in the due course project development.

We sincerely thank our Dean **Dr. SHAIK MEERAVALI** for his constant support and motivation all the time. A special acknowledgement goes to a friend who enthused us from the backstage. Last but not the least our sincere appreciation goes to our family who has been tolerant, understanding our moods and extending timely support.

**M.SAI RAM (2211CS010379) P.NIKHITHA (2211CS010438)**

**M. VINAY (2211CS010367) G.KISHORE (2211CS010209)**

# ABSTRACT

The **Amazon Sales Data Analysis** project is an interactive dashboard that leverages **data visualization and predictive analytics** to analyze Amazon sales trends. Developed using **Dash, Plotly, Pandas, and Machine Learning algorithms**, the application provides insights into **revenue, profit, units sold, product performance, and regional sales variations**. It integrates **predictive modeling** to forecast future sales and evaluate model performance using various statistical metrics.

The application **preprocesses raw sales data**, handling missing values, converting date formats, and structuring numerical fields like **total revenue, total cost, profit, and unit price**. Users can filter data dynamically by **product category and region**, gaining customized insights through interactive **time-series analysis, revenue distribution, product price trends, and profitability comparisons**. Additionally, the dashboard highlights **order processing times, cost outliers, and sales patterns based on order priority**, helping businesses refine their pricing and logistics strategies.

To enhance predictive analytics, the project employs **Machine Learning techniques**, including **Linear Regression** for **sales forecasting and revenue prediction**. The **Lasso and Least Angle Regression (LARS) methods** are explored for feature selection and model optimization. Model performance is quantified using metrics such as **Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), R² score, Root Mean Squared Logarithmic Error (RMSLE), and Mean Absolute Percentage Error (MAPE)**. **Kernel Density Estimation (KDE) plots** are also utilized to visualize data distributions and model residuals.

The project further compares multiple models using **PyCaret**, an automated machine learning library, to determine the most effective approach for **sales trend prediction and anomaly detection**. By integrating **real-time data processing, interactive visualizations, and predictive analytics**, this project provides **data-driven insights to improve sales strategies, enhance profitability, and optimize inventory management** in the e-commerce sector.

# TABLE OF CONTENT

**DESCRIPTION PAGE NUMBER**

[CERTIFICATE ii](#_TOC_250019)

[DECLARATION iii](#_TOC_250018)

ACKNOWLEDGEMENTS iv

[ABSTRACT v](#_TOC_250017)

[LIST OF FIGURES viii](#_TOC_250016)

[Chapter 1 Introduction 1](#_TOC_250015)

* 1. [Introduction 1](#_TOC_250014)
  2. [Problem Statement 2](#_TOC_250013)
  3. [Objective of Project 3](#_TOC_250012)
  4. [Goal of Project 5](#_TOC_250011)

[Chapter 2 Problem Identification 6](#_TOC_250010)

* 1. Existing System 6
  2. Proposed System 7

[Chapter 3 Requirements 9](#_TOC_250009)

* 1. [Software Requirements 9](#_TOC_250008)
  2. [Hardware Requirements 10](#_TOC_250007)

Chapter 4 Design and Implementation 13

* 1. [Design 13](#_TOC_250006)
  2. [Implementation 16](#_TOC_250005)

[Chapter 5 Code 21](#_TOC_250004)

* 1. Source Code 21
  2. [Screenshot of Application 29](#_TOC_250003)

[Chapter 6 Results & Conclusion 36](#_TOC_250002)

* 1. Results 36
  2. [Conclusion 37](#_TOC_250001)

[REFERENCES 38](#_TOC_250000)

## LIST OF FIGURES

|  |  |  |
| --- | --- | --- |
| **FIGURE** | **TITLE** | **PAGE NUMBER** |
| 4.1.1 | **Architecture Design** | 15 |
| 4.2.1 | **Class Diagram** | 18 |
| 4.2.2 | **Block Diagram** | 18 |
| 4.2.3 | **Activity Diagram** | 19 |
| 4.2.4 | **UseCase Diagram** | 19 |
| 4.2.5 | **StructureDiagram** | 20 |
| 4.2.6 | **FlowChart** | 20 |
| 5.2.1 | **Circuit Diagram** | 29 |
| 5.2.2 | **Enabling Wifi ,Hotspot and Blue** | **tooth**  29 |
| 5.2.3 | **Wifi Ready** | 30 |
| 5.2.4 | **Switching to Bluetooth** | 30 |
| 5.2.5 | **App UserInterface** | 30 |
| 5.2.6 | **Enter Access Code** | 31 |
| 5.2.7 | **Commond for Light On** | 31 |
| 5.2.8 | **Light On** | 31 |
| 5.2.9 | **Commond for Light off** | 32 |
| 5.2.10 | **Light Off** | 32 |
| 5.2.11 | **Commond for Fan On** | 32 |
| 5.2.12 | **Fan On** | 33 |
| 5.2.13 | **Commond for Fan Off** | 33 |
| 5.2.14 | **Fan Off** | 33 |
| 5.2.15 | **Switching to Sensors** | 34 |
| 5.2.16 | **Testing Gas Sensor** | 34 |
| 5.2.17 | **Testing LDR Sensor** | 34 |
| 5.2.18 | **Server** | 35 |

# CHAPTER – 1 INTRODUCTION

### Introduction

In today's competitive e-commerce landscape, data-driven decision-making is crucial for businesses to maximize revenue and optimize sales strategies. The **Amazon Sales Data Analysis** project is designed to provide **insights into sales trends, revenue distribution, and product performance** through interactive visualizations and predictive analytics. This project utilizes **Dash, Plotly, Pandas, and Machine Learning (ML) techniques** to analyze historical sales data, forecast future trends, and evaluate sales performance across different regions and products.

The primary objective of this project is to help businesses and analysts **understand key sales metrics**, including **total revenue, profit margins, units sold, and regional sales patterns**. By leveraging **data preprocessing, visualization, and predictive modeling**, the project enables users to identify **high-performing products, seasonal sales trends, and potential inefficiencies in order processing**.

To enhance predictive capabilities, the project incorporates **Machine Learning algorithms** such as **Linear Regression, Lasso Regression, and Least Angle Regression (LARS)** for **sales forecasting and revenue prediction**. Performance evaluation is conducted using statistical metrics like **Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), R² score, Root Mean Squared Logarithmic Error (RMSLE), and Mean Absolute Percentage Error (MAPE)**. The **PyCaret library** is also used to compare multiple models and automate predictive analytics.

Additionally, **Kernel Density Estimation (KDE) plots** are utilized to analyze data distributions, while **anomaly detection techniques** help identify irregularities in sales data. The dashboard allows users to filter data dynamically based on **products and geographical regions**, providing **customized insights** through interactive visualizations.

By integrating **real-time data processing, advanced visualizations, and predictive analytics**, this project serves as a valuable tool for **business intelligence and decision-making**. The insights generated can help **improve sales strategies, optimize pricing models, forecast demand, and enhance overall operational efficiency** in the ecommerce domain.

### 1.2 Problem Statement

In the highly competitive e-commerce industry, businesses face challenges in understanding sales trends, predicting future revenue, and optimizing sales strategies. With vast amounts of transactional data generated daily, organizations often struggle to extract meaningful insights that can drive decision-making. Amazon, as one of the largest online marketplaces, requires efficient data analysis tools to monitor sales performance, identify profitable products, and forecast demand.

This project aims to address the following key problems:

1. **Lack of Sales Trend Visibility** – Businesses need to track sales patterns, revenue fluctuations, and product performance across different time periods and geographical regions. Without clear insights, they risk making uninformed decisions.
2. **Inaccurate Sales Forecasting** – Traditional reporting methods often fail to predict future sales accurately. The absence of predictive analytics leads to inventory mismanagement, stockouts, or overstocking, impacting profitability.
3. **Difficulty in Identifying High-Performing Products and Regions** – Companies struggle to pinpoint which products contribute most to revenue and which regions generate the highest sales. This lack of insight makes it difficult to allocate resources effectively.
4. **Inability to Detect Anomalies and Outliers** – Unusual sales spikes or drops can indicate market trends, pricing issues, or fraudulent activities. Without an efficient anomaly detection system, businesses may overlook critical risks and opportunities.
5. **Need for an Automated and Scalable Analysis Framework** – Manual data analysis is time-consuming and prone to human errors. Organizations require a real-time, automated dashboard that simplifies data exploration and enhances decision-making through interactive visualizations and predictive modeling.

### 1.3 Objective of Project

The **Amazon Sales Data Analysis** project aims to provide an **interactive, data-driven solution** for analyzing and predicting sales trends. The primary goal is to help businesses extract meaningful insights from historical sales data to enhance **decision-making, forecasting, and operational efficiency**.

The specific objectives of the project are:

**1.Analyze Sales Trends** – Identify **patterns in revenue, profit, and units sold** over time to understand fluctuations in Amazon sales performance.

**2.Develop an Interactive Dashboard** – Create a **user-friendly web application** using **Dash and Plotly** to visualize sales data dynamically, allowing users to filter insights based on **products and regions**.

**3.Predict Future Sales** – Implement **Machine Learning models (Linear Regression, Lasso Regression, and Least Angle Regression - LARS)** to forecast **future revenue and product demand** based on historical trends.

**4.Evaluate Model Performance** – Quantify model accuracy using statistical metrics such as **Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), R² score, Root Mean Squared Logarithmic Error (RMSLE), and Mean Absolute Percentage Error (MAPE)**.

**5.Compare Multiple ML Models** – Utilize **PyCaret**, an automated machine learning library, to compare different predictive models and determine the most accurate approach for sales forecasting.

**6.Detect Anomalies in Sales Data** – Identify **outliers and irregularities** in sales, cost, and order processing using **Kernel Density Estimation (KDE) plots and anomaly detection techniques**.

**7.Improve Business Decision-Making** – Enable businesses to **optimize inventory management, refine pricing strategies, and allocate resources effectively** by leveraging data-driven insights.

### 1.4 Goal of Project

The goal of the Amazon Sales Data Analysis project is to develop an interactive, data-driven dashboard that enables businesses to analyze, visualize, and predict sales trends, revenue patterns, and product performance. By leveraging data analytics and machine learning models, the project aims to help businesses make informed, data-backed decisions to optimize sales strategies, improve revenue forecasting, and enhance overall operational efficiency.

This project focuses on analyzing historical sales data by processing and cleaning raw datasets to extract meaningful insights. It identifies patterns in revenue, profit, and sales trends over time to provide a clear understanding of business performance. The system is designed as a web-based interactive dashboard using Dash and Plotly, allowing users to explore data dynamically by filtering sales records based on product category, region, and time period.

To enhance predictive analytics, the project implements machine learning techniques such as Linear Regression, Lasso Regression, and Least Angle Regression (LARS) to forecast future sales and revenue based on historical trends. Model performance is evaluated using statistical metrics including Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), R² score, Root Mean Squared Logarithmic Error (RMSLE), and Mean Absolute Percentage Error (MAPE). The project also integrates PyCaret, an automated machine learning library, to compare different models and select the most accurate approach for sales forecasting.

Beyond predictive modeling, the system provides insights to assist in business decision-making by identifying high-performing products, analyzing regional sales variations, and detecting anomalies in revenue trends. These insights help businesses optimize inventory management, refine pricing strategies, and enhance market positioning. By integrating real-time data processing, interactive visualizations, and machine learning-based predictive analytics, this project serves as a valuable tool for businesses seeking to improve profitability, forecast future sales trends, and streamline their overall sales operations in the competitive e-commerce landscape.

# CHAPTER 2 PROBLEM IDENTIFICATION

### Existing System

The current approach to **Amazon sales data analysis** in many businesses relies on **manual reporting, basic spreadsheets, and traditional business intelligence tools**. While these methods provide fundamental insights, they are often **time-consuming, inefficient, and lack predictive capabilities**.

#### **Challenges in the Existing System:**

**1.Static Reporting & Limited Insights** – Many businesses use **Excel or simple dashboards** that provide historical sales data but lack **interactive filtering, real-time updates, or deep analytical insights**.

**2.Lack of Predictive Analytics** – Traditional systems rely on **descriptive analytics**, which only analyze past sales trends. Without **Machine Learning (ML)-based forecasting**, businesses struggle to make data-driven future sales predictions.

**3.No Automated Model Comparison** – Existing systems typically do not **automate model selection or performance evaluation**. Companies manually experiment with models without understanding which one provides the most accurate forecasts.

**4.Inefficient Anomaly Detection** – Identifying **outliers in revenue, order delays, or pricing** is difficult with basic tools. Businesses may overlook **unexpected spikes, fraudulent activities, or sudden sales drops** due to the lack of automated anomaly detection.

**5.Limited Interactivity & Customization** – Most sales analysis tools do not allow dynamic **filtering by product category, region, or sales period**, making it hard for businesses to explore specific insights efficiently.

**6.Scalability Issues** – Traditional tools struggle to **handle large datasets** efficiently, leading to **slow processing times and difficulty scaling analysis** as data grows over time.

### ****Need for an Improved System****

To overcome these limitations, the **Amazon Sales Data Analysis** project introduces an **interactive, machine learning-powered dashboard** that enhances sales analysis with:

* **Real-time visualization** of sales trends
* **ML-based sales forecasting** with accuracy evaluation
* **Automated model comparison** using PyCaret
* **Advanced anomaly detection** for fraud and pricing issues
* **Dynamic filtering and interactive exploration** of sales data

This **improved system** helps businesses **save time, enhance forecasting accuracy, and make more informed strategic decisions** in the competitive e-commerce industry.

### ****2.2 Proposed System****

The **Amazon Sales Data Analysis** project introduces an **advanced, interactive, and machine learning-driven solution** to overcome the limitations of existing sales analysis methods. The proposed system integrates **data visualization, predictive analytics, and anomaly detection** to provide businesses with deeper insights into **sales trends, revenue patterns, and product performance**.

#### **Key Features of the Proposed System:**

**1.Interactive and Dynamic Sales Analysis**

* The system provides an **interactive dashboard** built with **Dash and Plotly**, allowing users to explore **sales trends, revenue distribution, and product-wise performance** dynamically.
* Users can filter sales data by **region, product category, and date range** for customized analysis.

**2.Predictive Analytics Using Machine Learning**

* The system incorporates **Linear Regression, Lasso Regression, and Least Angle Regression (LARS)** to **predict future sales and revenue trends** based on historical data.
* **Train-test split methodology** is used to improve model accuracy.

**3.Model Performance Evaluation**

* The system quantifies **forecasting accuracy** using statistical metrics such as **Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), R² score, Root Mean Squared Logarithmic Error (RMSLE), and Mean Absolute Percentage Error (MAPE)**.
* The **PyCaret library** is integrated to **automate model comparison**, ensuring the best-performing model is selected.

**4.Anomaly Detection for Sales Data**

* **Kernel Density Estimation (KDE) plots** and other statistical techniques help identify **outliers in sales, cost, and order processing times**.
* Anomalies in revenue trends can indicate **fraud detection, pricing errors, or unexpected market changes**.

**5.Automated and Scalable Data Processing**

* The system efficiently processes **large datasets** using **Pandas and NumPy**, ensuring **fast computation and scalability**.
* It automatically cleans data, handles missing values, and standardizes numerical features for **better model accuracy**.

# CHAPTER – 3

# REQUIREMENTS

### Software Requirements:

* **Operating System:** Windows 10/11, macOS, or Linux
* **Python Version:** 3.7 or higher

**Development Environments:**

* **Jupyter Notebook** (for analysis & ML modeling)
* **VS Code**

### ****Python Libraries & Dependencies****

✅ **Data Processing & Analysis:**

* pandas – For handling structured data
* numpy – For numerical operations

✅ **Data Visualization:**

* matplotlib – For plotting charts and graphs
* seaborn – For statistical data visualization
* plotly – For interactive dashboards

✅ **Machine Learning & Predictive Analytics**

* **Regression Models for Sales Forecasting**
  + scikit-learn → **Linear Regression, Lasso Regression, Least Angle Regression (LARS)**

**Model Performance Evaluation Metrics**

* + **Mean Absolute Error (MAE)**
  + **Mean Squared Error (MSE)**
  + **Root Mean Squared Error (RMSE)**
  + **R² Score** (Coefficient of Determination)
  + **Root Mean Squared Logarithmic Error (RMSLE)**
  + **Mean Absolute Percentage Error (MAPE)**

**Automated Model Selection**

* + pycaret → **For comparing multiple ML models and selecting the best one**

### ****Dataset Requirements****

* **Dataset Format:** .csv (Comma-Separated Values)
* **Dataset Name:** Amazon Sales data.csv
* **Required Columns:**
  + Order Date, Ship Date
  + Total Revenue, Total Cost, Profit, Units Sold, Unit Price
  + Region, Country, Product

### Hardware Requirements:

* **Processor:** Intel Core i3 / AMD Ryzen 3 or equivalent
* **RAM:** 4GB (for basic data processing & visualization)
* **Storage:** 10GB free disk space

# CHAPTER – 4

**DESIGN AND IMPLEMENTATION**

### Design

### 

### The design of the **Amazon Sales Data Analysis** project follows a structured approach that integrates **data processing, visualization, and machine learning models** into an interactive and user-friendly system. The system architecture consists of three key components: **data preprocessing, dashboard development, and machine learning integration**.

**1.Data Preprocessing**

* Load the **Amazon Sales data.csv** file using Pandas.
* Handle missing values and ensure consistency in numerical fields like **Total Revenue, Total Cost, Profit, and Units Sold**.
* Convert **Order Date and Ship Date** to datetime format for time-based analysis.
* Extract additional features such as **Year, Month, and Processing Time (Ship Date - Order Date)**.

**2.Dashboard Design (Frontend & Visualization)**

* Built using **Dash and Plotly** for an interactive web-based user interface.
* Users can filter data dynamically based on **Product Categories, Regions, and Order Dates**.
* Key Visualizations:
  + **Sales Trends Over Time:** Line chart showing total revenue trends.
  + **Regional Sales Analysis:** Bar chart displaying revenue distribution across different regions.
  + **Country-Wise Revenue Comparison:** Visualization to compare total sales by country.
  + **Profit vs. Units Sold:** Scatter plot showing product-wise profitability.
  + **Unit Price Distribution:** Box plot to analyze product pricing trends.
  + **Order Processing Time Analysis:** Histogram to evaluate delivery efficiency.

**3.Machine Learning Integration (Sales Forecasting)**

* Implemented **Linear Regression, Lasso Regression, and Least Angle Regression (LARS)** for predicting future sales.
* Split the dataset into **training and testing sets** using the train-test split method.

Evaluated model performance using statistical metrics:

* + **Mean Absolute Error (MAE)**
  + **Mean Squared Error (MSE)**
  + **Root Mean Squared Error (RMSE)**
  + **R² Score (Coefficient of Determination)**
  + **Root Mean Squared Logarithmic Error (RMSLE)**
  + **Mean Absolute Percentage Error (MAPE)**

Used **PyCaret** to compare multiple regression models and select the most accurate one.

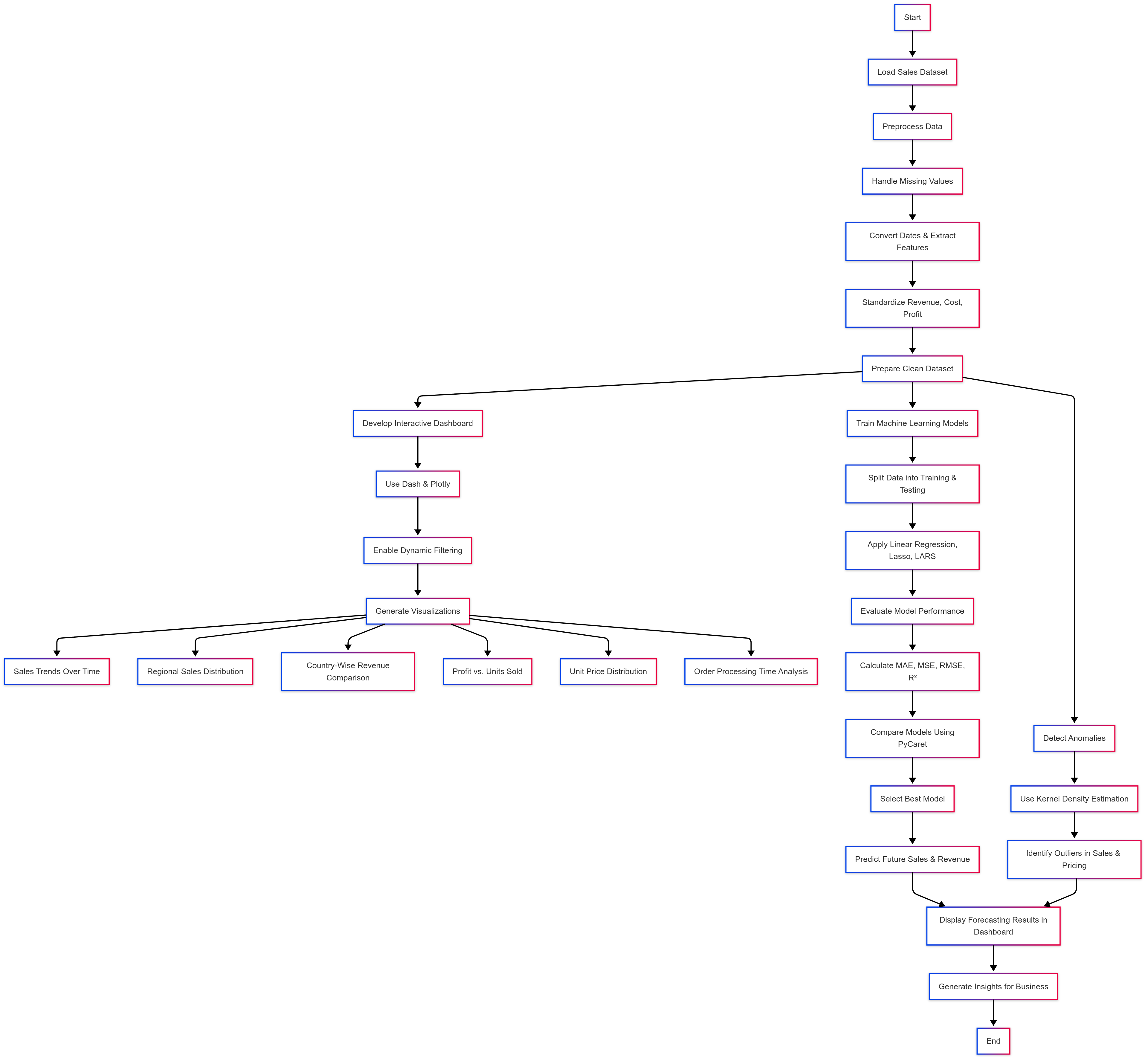
**4.Anomaly Detection & Statistical Analysis**

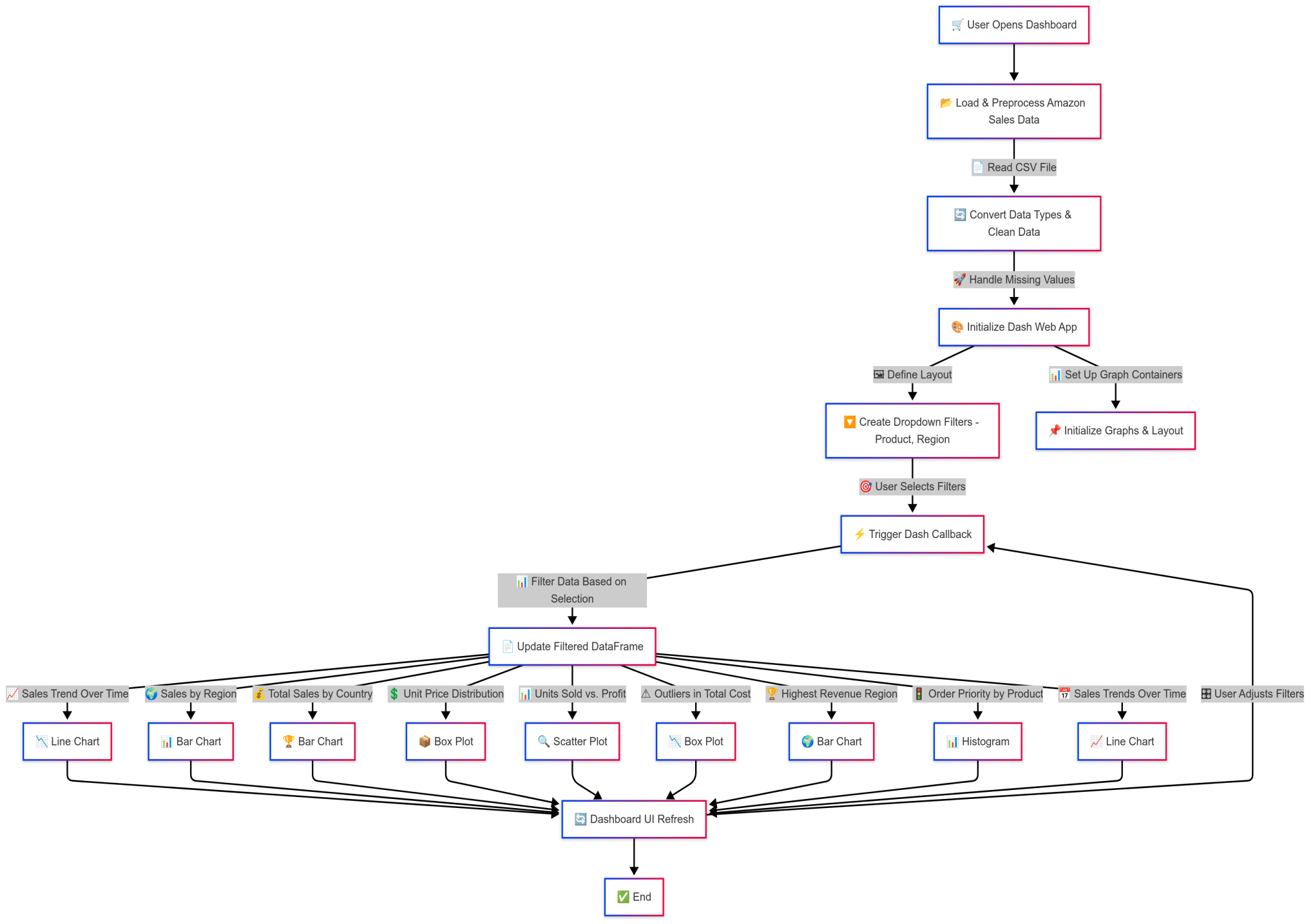
* Kernel Density Estimation (KDE) used to detect **outliers in sales and pricing data**.
* Identifies **unexpected revenue fluctuations, cost anomalies, and order delays** for business insights.

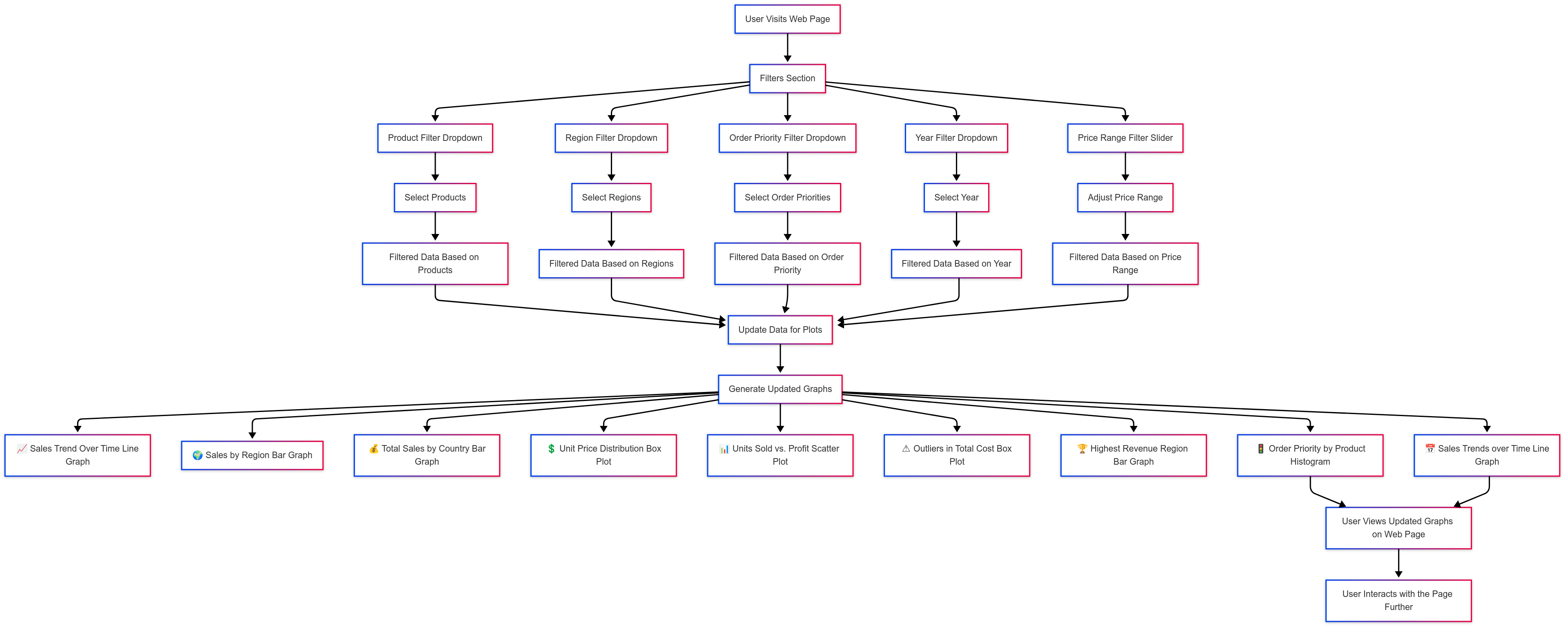
**System Architecture**

****

**Fig.4.1.1:Architecture Design**

****

****



**4.2 Implementation**

· **Data Preprocessing**

* Format **Order Date** and **Ship Date**.
* Extract **Year, Month, Processing Time**.
* Handle missing values and standardize **Total Revenue, Total Cost, and Profit**.

· **Dashboard Development**

* Built using **Dash and Plotly**.
* **Dropdown filters** for selecting products and regions.
* **Real-time graph updates** based on user input.

· **Machine Learning Integration**

* Models used: **Linear Regression, Lasso Regression, and LARS**.
* **Predicts future sales and revenue** trends.
* Evaluated using **MAE, MSE, RMSE, and R² score**.
* **PyCaret** automates model comparison.

· **Visualization & User Interaction**

* **Sales Trends Over Time**
* **Regional & Country-Wise Revenue**
* **Profit vs. Units Sold**
* **Unit Price Distribution**
* **Order Processing Time Analysis**

· **Anomaly Detection & Insights**

* **Kernel Density Estimation (KDE)** detects unusual sales trends.
* Identifies **pricing outliers, revenue spikes, and operational inefficiencies**.

· **Conclusion**

* **Dynamic sales dashboard** with **real-time filtering and ML forecasting**.
* **Helps businesses optimize pricing, inventory, and revenue strategies**.

**UML DIAGRAMS:**

A diagram of a flowchart

AI-generated content may be incorrect.

**Fig.4.2.1:Class Diagram**

A screenshot of a computer

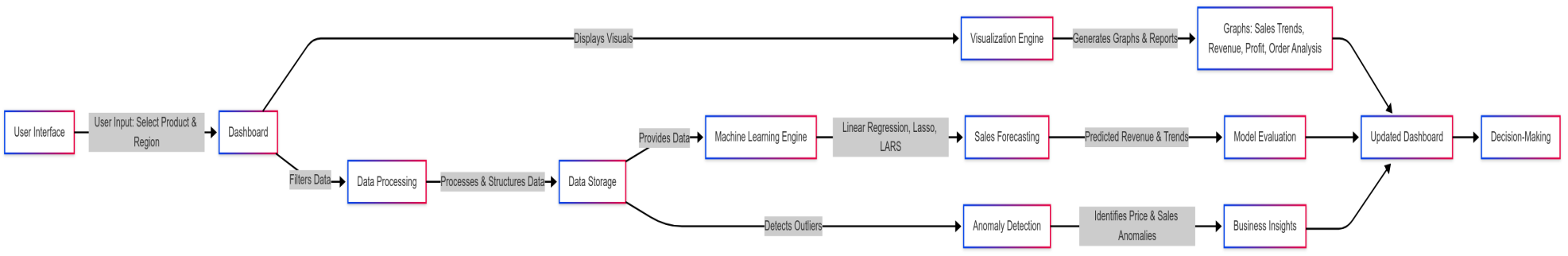
AI-generated content may be incorrect.

**Fig.4.2.2:Sequence Diagram**

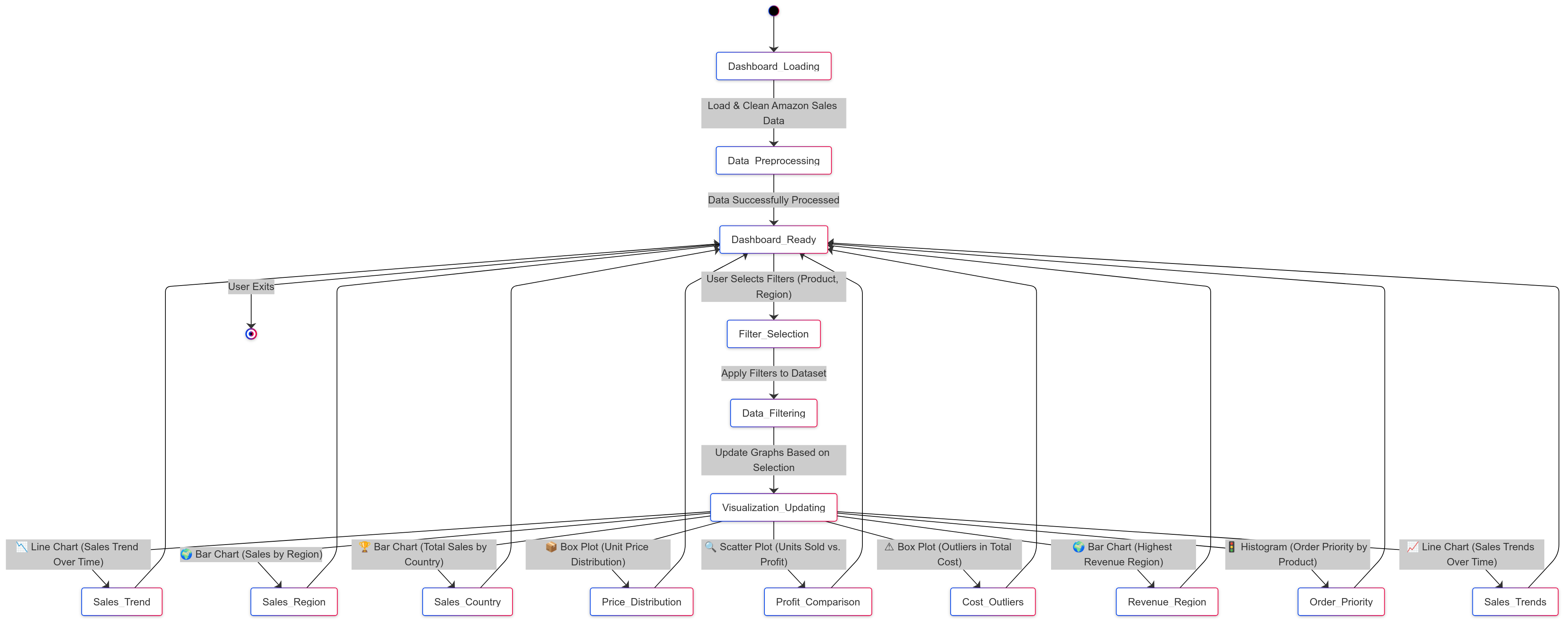
A screenshot of a computer

AI-generated content may be incorrect.

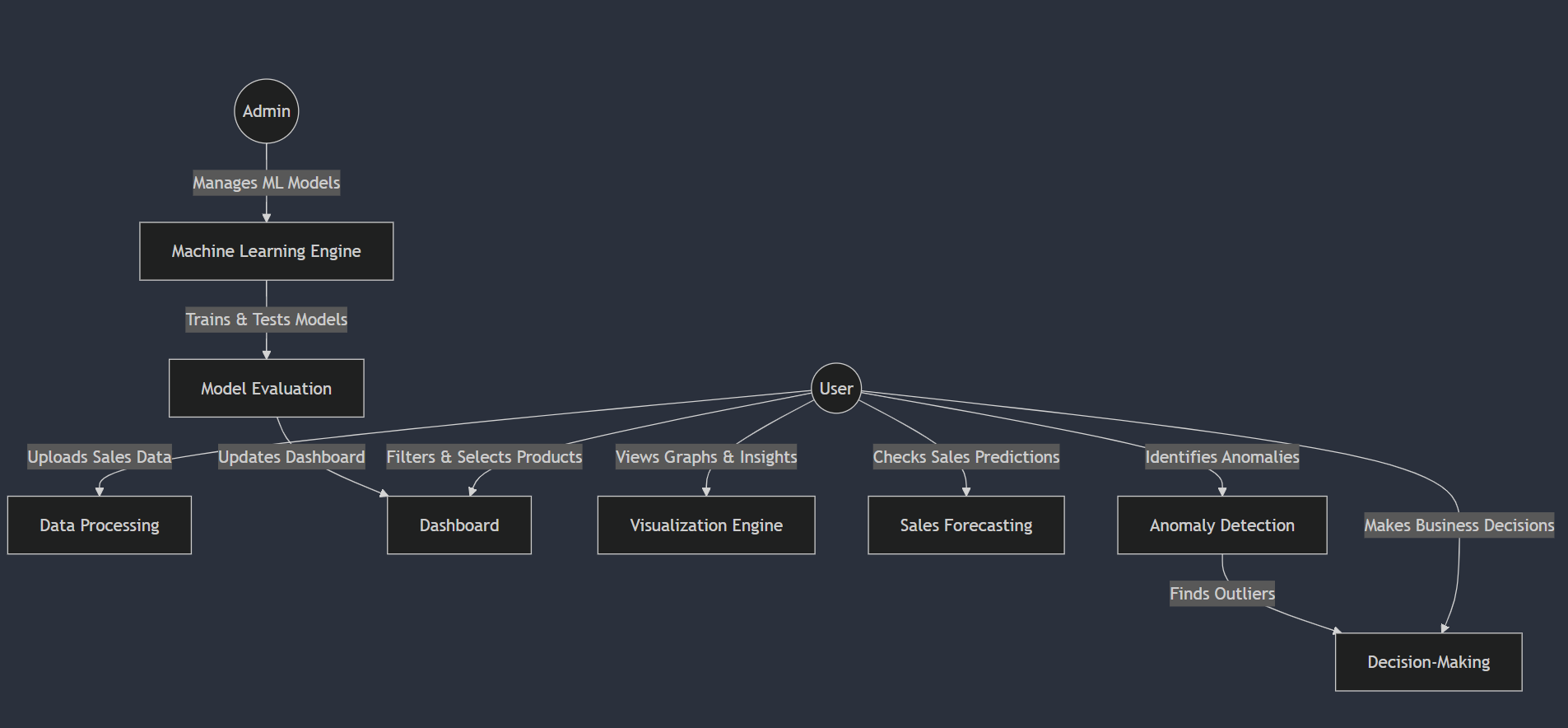
**Fig.4.2.3:State Diagram**

****

**Fig.4.2.4:Block Diagram**

****

**Fig.4.2.4:Activity Diagram**



**Fig.4.2.4:Use Case Diagram**

### Source Code

**App.py**

import dash

from dash import dcc, html, Input, Output

import pandas as pd

import plotly.express as px

from dash\_extensions.enrich import DashProxy, ServersideOutputTransform

# ✅ Load and clean data

df = pd.read\_csv("Amazon Sales data.csv", encoding='latin-1')

df.columns = df.columns.str.strip()  # Remove extra spaces in column names

# ✅ Print columns to verify dataset structure

print("Columns in dataset:", df.columns.tolist())

# ✅ Convert date columns

df['Order Date'] = pd.to\_datetime(df['Order Date'], errors='coerce')

df['Ship Date'] = pd.to\_datetime(df['Ship Date'], errors='coerce')

df = df.dropna(subset=['Order Date'])

# ✅ Add derived columns

df['Year'] = df['Order Date'].dt.year

df['Month'] = df['Order Date'].dt.month\_name()

df['Processing Time'] = (df['Ship Date'] - df['Order Date']).dt.days

df['Total Revenue'] = pd.to\_numeric(df['Total Revenue'], errors='coerce').fillna(0)

df['Total Cost'] = pd.to\_numeric(df['Total Cost'], errors='coerce').fillna(0)

df['Profit'] = pd.to\_numeric(df['Profit'], errors='coerce').fillna(0)

df['Units Sold'] = pd.to\_numeric(df['Units Sold'], errors='coerce').fillna(0)

df['Unit Price'] = pd.to\_numeric(df['Unit Price'], errors='coerce').fillna(0)

# ✅ Initialize Dash app

app = DashProxy(\_\_name\_\_, transforms=[ServersideOutputTransform()])

# ✅ Updated Layout with Correct CSS

app.layout = html.Div([

    html.H1("📊 Amazon Sales Data Analysis", style={'text-align': 'center', 'margin-bottom': '20px'}),

    # Filters Section

    html.Div([

        dcc.Dropdown(

            id='product-dropdown',

            options=[{'label': i, 'value': i} for i in df['Product'].dropna().unique()],

            multi=True,

            placeholder="Select Products",

            style={'width': '200px'}

        ),

        dcc.Dropdown(

            id='region-dropdown',

            options=[{'label': i, 'value': i} for i in df['Region'].dropna().unique()],

            multi=True,

            placeholder="Select Regions",

            style={'width': '200px'}

        ),

        dcc.Dropdown(

            id='order-priority-dropdown',

            options=[{'label': i, 'value': i} for i in df['Order Priority'].dropna().unique()],

            multi=True,

            placeholder="Select Order Priority",

            style={'width': '200px'}

        ),

        dcc.Dropdown(

            id='year-dropdown',

            options=[{'label': str(i), 'value': i} for i in df['Year'].unique()],

            multi=True,

            placeholder="Select Year",

            style={'width': '200px'}

        )

    ], style={'display': 'flex', 'flex-wrap': 'wrap', 'gap': '10px', 'justifyContent': 'center', 'margin-bottom': '20px'}),

    # ✅ Price Range Slider (CSS Fixed)

    html.Div([

        html.Label("Filter by Unit Price:", style={'font-weight': 'bold', 'margin-bottom': '10px'}),

        dcc.RangeSlider(

            id='price-range-slider',

            min=df['Unit Price'].min(),

            max=df['Unit Price'].max(),

            step=5,  # Step value for smoother selection

            marks={i: f"${i}" for i in range(int(df['Unit Price'].min()), int(df['Unit Price'].max()) + 1, 50)},

            value=[df['Unit Price'].min(), df['Unit Price'].max()],  # Default full range

        )

    ], style={'width': '60%', 'margin': 'auto', 'padding-bottom': '20px'}),

    # ✅ Graph Containers with Grid Layout

    html.Div([

        dcc.Graph(id='sales-trend'),

        dcc.Graph(id='sales-by-region'),

        dcc.Graph(id='total-sales-by-country'),

        dcc.Graph(id='unit-price-distribution'),

        dcc.Graph(id='units-sold-profit'),

        dcc.Graph(id='total-cost-outliers'),

        dcc.Graph(id='highest-revenue-region'),

        dcc.Graph(id='order-priority-item'),

        dcc.Graph(id='order-date-trends')

    ], style={'display': 'grid', 'grid-template-columns': 'repeat(auto-fit, minmax(400px, 1fr))', 'gap': '20px', 'justify-content': 'center'})

])

# ✅ Callback Function for Filtering

@app.callback(

    [Output('sales-trend', 'figure'),

     Output('sales-by-region', 'figure'),

     Output('total-sales-by-country', 'figure'),

     Output('unit-price-distribution', 'figure'),

     Output('units-sold-profit', 'figure'),

     Output('total-cost-outliers', 'figure'),

     Output('highest-revenue-region', 'figure'),

     Output('order-priority-item', 'figure'),

     Output('order-date-trends', 'figure')],

    [Input('product-dropdown', 'value'),

     Input('region-dropdown', 'value'),

     Input('order-priority-dropdown', 'value'),

     Input('year-dropdown', 'value'),

     Input('price-range-slider', 'value')]

)

def update\_plots(selected\_products, selected\_regions, selected\_priorities, selected\_years, price\_range):

    filtered\_df = df.copy()

    # Apply filters

    if selected\_products:

        filtered\_df = filtered\_df[filtered\_df['Product'].isin(selected\_products)]

    if selected\_regions:

        filtered\_df = filtered\_df[filtered\_df['Region'].isin(selected\_regions)]

    if selected\_priorities:

        filtered\_df = filtered\_df[filtered\_df['Order Priority'].isin(selected\_priorities)]

    if selected\_years:

        filtered\_df = filtered\_df[filtered\_df['Year'].isin(selected\_years)]

    if price\_range:

        filtered\_df = filtered\_df[(filtered\_df['Unit Price'] >= price\_range[0]) & (filtered\_df['Unit Price'] <= price\_range[1])]

    if filtered\_df.empty:

        return tuple(px.scatter(title="No data available") for \_ in range(9))

    # Generate updated graphs

    sales\_trend = px.line(filtered\_df, x='Order Date', y='Total Revenue', title='📈 Sales Trend Over Time', color='Region')

    sales\_by\_region = px.bar(filtered\_df.groupby('Region', as\_index=False)['Total Revenue'].sum(), x='Region', y='Total Revenue', title='🌍 Sales by Region')

    total\_sales\_by\_country = px.bar(filtered\_df.groupby('Country', as\_index=False)['Total Revenue'].sum(), x='Country', y='Total Revenue', title='💰 Total Sales Revenue by Country')

    unit\_price\_distribution = px.box(filtered\_df, x='Product', y='Unit Price', title='💲 Unit Price Distribution')

    units\_sold\_profit = px.scatter(filtered\_df, x='Units Sold', y='Profit', color='Product', title='📊 Units Sold vs. Profit')

    total\_cost\_outliers = px.box(filtered\_df, y='Total Cost', title='⚠ Outliers in Total Cost')

    highest\_revenue\_region = px.bar(filtered\_df.groupby('Region', as\_index=False)['Total Revenue'].mean(), x='Region', y='Total Revenue', title='🏆 Highest Revenue Region')

    order\_priority\_item = px.histogram(filtered\_df, x='Product', color='Order Priority', title='🚦 Order Priority by Product')

    order\_date\_trends = px.line(filtered\_df.groupby('Order Date', as\_index=False)['Total Revenue'].sum(), x='Order Date', y='Total Revenue', title='📅 Sales Trends over Time')

    return sales\_trend, sales\_by\_region, total\_sales\_by\_country, unit\_price\_distribution, units\_sold\_profit, total\_cost\_outliers, highest\_revenue\_region, order\_priority\_item, order\_date\_trends

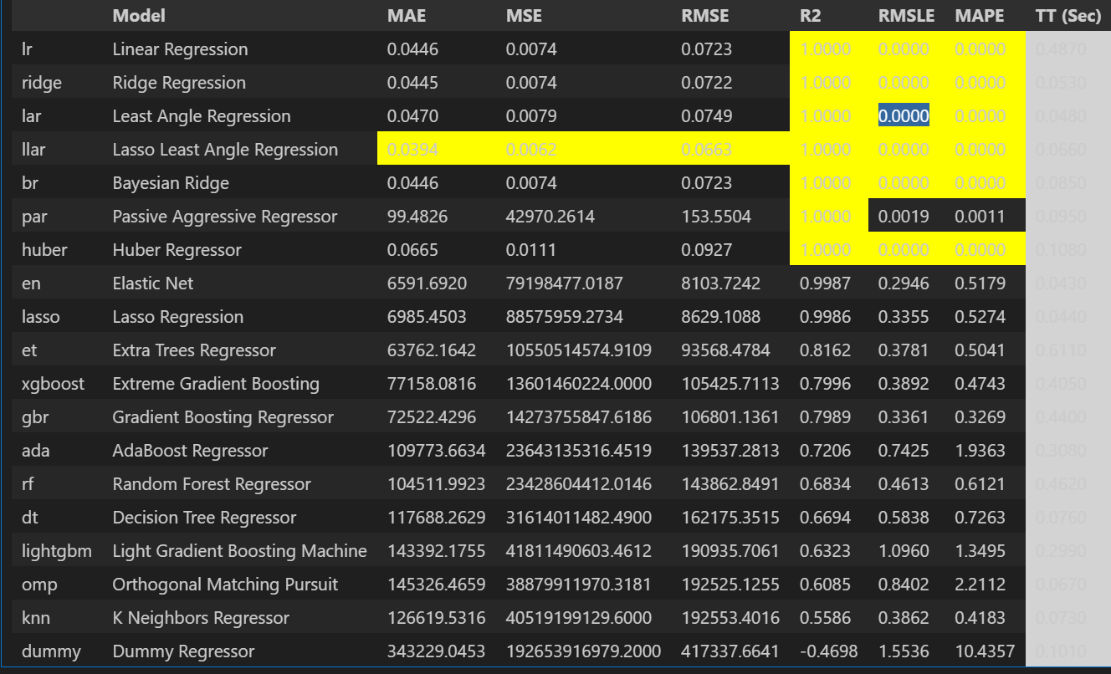
# ✅ Run the Dash app

if \_\_name\_\_ == '\_\_main\_\_':

    app.run\_server(debug=True)

**Amazon\_Sales\_Data\_Analysis\_Final.ipynb**

# Comparing multiple models

compare\_models()****

**#Exploratory Data Analysis**

**###Checking Outliers in the dataframe**

def detect\_outliers(dataframe, column):

    threshold = 2     ## 2rd standard deviation

    mean = np.mean(column)

    std = np.std(column)

    outliers = []

    for i, value in enumerate(column):

        z\_score = (value - mean) / std

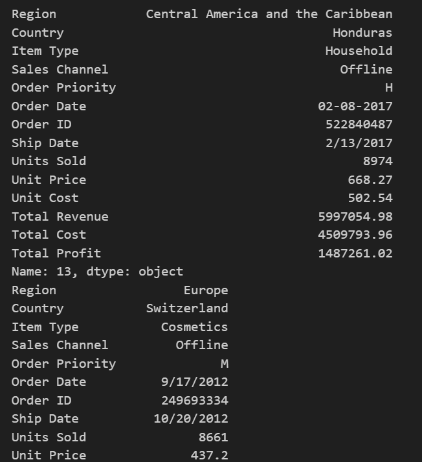
        if np.abs(z\_score) > threshold:

            outliers.append(i)

            print(dataframe.loc[i])

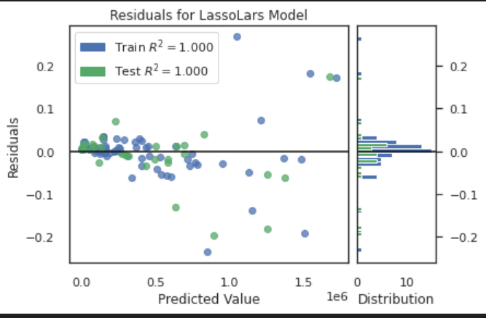
    return outliers

outliers = detect\_outliers(df, df["Total Profit"])

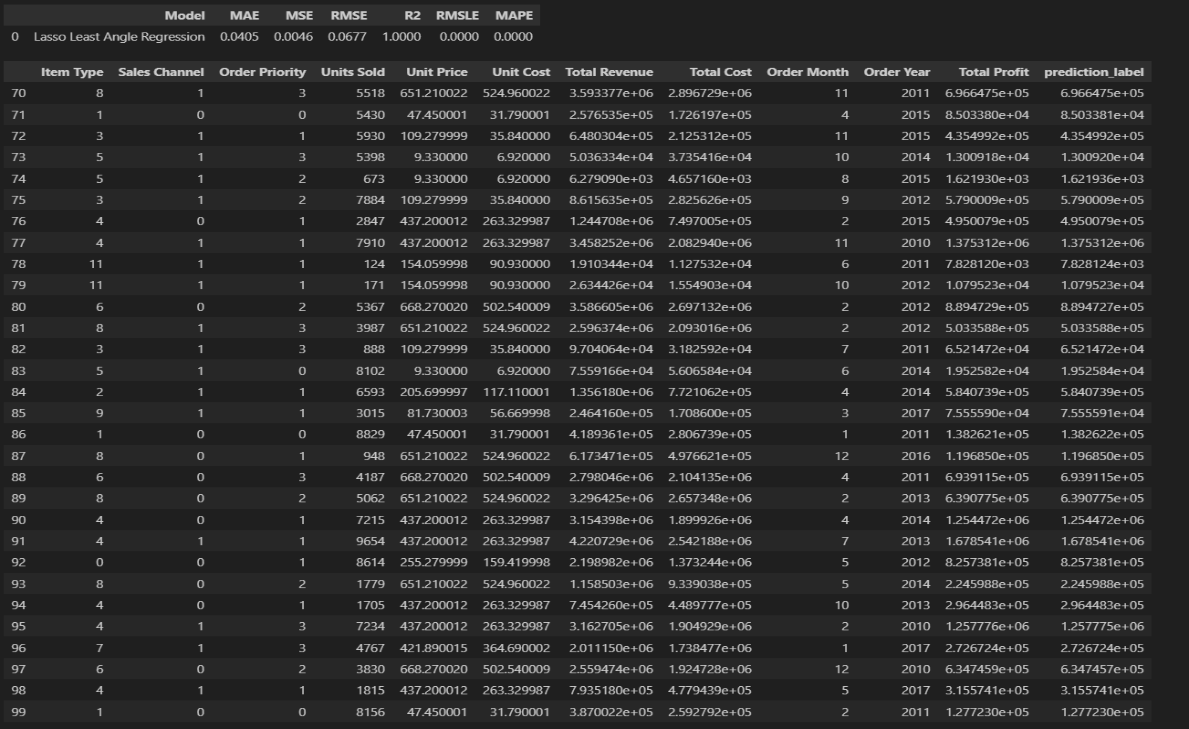
****

# Plot residuals for Lasso Least Angle Regression based trained model

plot\_model(tuned\_llar\_model)

****

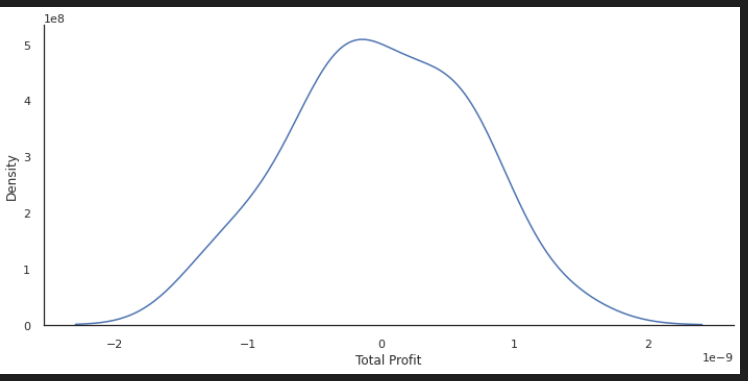
# Predicting total profit using Lasso Least Angle Regression based trained model

****predict\_model(tuned\_llar\_model)

# Creating kernel density estimate plot

import seaborn as sns

sns.displot(reg\_pred - y\_test,kind='kde', height=5, aspect=2)

****

**# Finding Accuracy percentage on the bases of r2 score**

from sklearn.metrics import r2\_score

score = r2\_score(reg\_pred,y\_test)

# Calculate the percentage of accuracy

accuracy\_pct = score \* 100

print("Accuracy: {:.2f}%".format(accuracy\_pct))

Accuracy: 88.41%

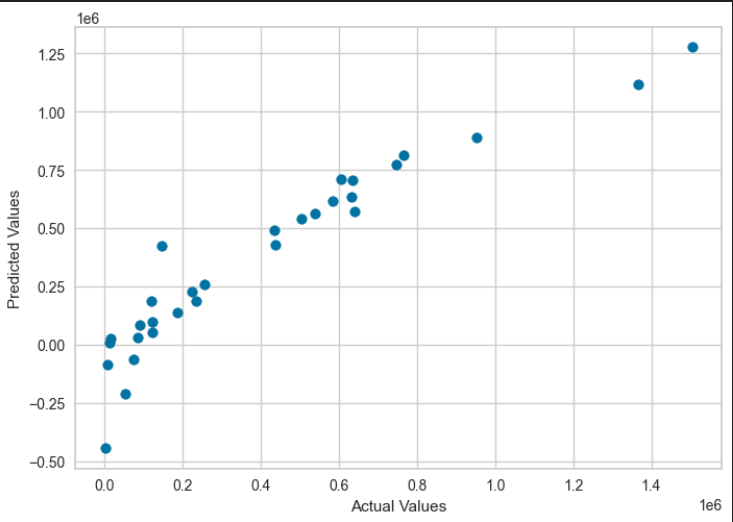
**# Plot the predicted values against the actual values to visualize how well the model** is fitting the data.

import matplotlib.pyplot as plt

plt.scatter(y\_test, reg\_pred)

plt.xlabel('Actual Values')

plt.ylabel('Predicted Values')

****plt.show()

**# Calculate the mean squared error (MSE) or root mean squared error (RMSE) to quantify the model's performance.**

from sklearn.metrics import mean\_squared\_error

import numpy as np

mse = mean\_squared\_error(y\_test, reg\_pred)

rmse = np.sqrt(mse)

print('MSE:', mse)

print('RMSE:', rmse)

MSE: 17994796525.40397

RMSE: 134144.68504344096

**import plotly.graph\_objs as go**

**import pandas as pd**

**from sklearn.linear\_model import LinearRegression**

**# Create a linear regression model**

reg = LinearRegression()

# Fit the model to the training data

reg.fit(X\_train, y\_train)

# Make predictions on the test set

y\_pred = reg.predict(X\_test)

# Calculate the R-squared score

r2 = reg.score(X\_test, y\_test)

# Ensure correct number of column names

expected\_columns = ['Item Type', 'Sales Channel', 'Order Priority', 'Units Sold',

                    'Unit Price', 'Unit Cost', 'Total Revenue', 'Total Cost',

                    'Order Month', 'Order Year']

# Match column names dynamically

if X\_train.shape[1] == len(expected\_columns):

    X\_train\_df = pd.DataFrame(X\_train, columns=expected\_columns)

else:

    column\_names = [f"Feature\_{i}" for i in range(X\_train.shape[1])]

    X\_train\_df = pd.DataFrame(X\_train, columns=column\_names)

# Extract coefficients

coef = reg.coef\_

# Create a DataFrame with feature importance

feature\_importances = pd.DataFrame({'Feature': X\_train\_df.columns, 'Importance': coef})

# Sort by importance

feature\_importances = feature\_importances.sort\_values('Importance', ascending=False)

# Create a bar chart of feature importance using Plotly

fig = go.Figure()

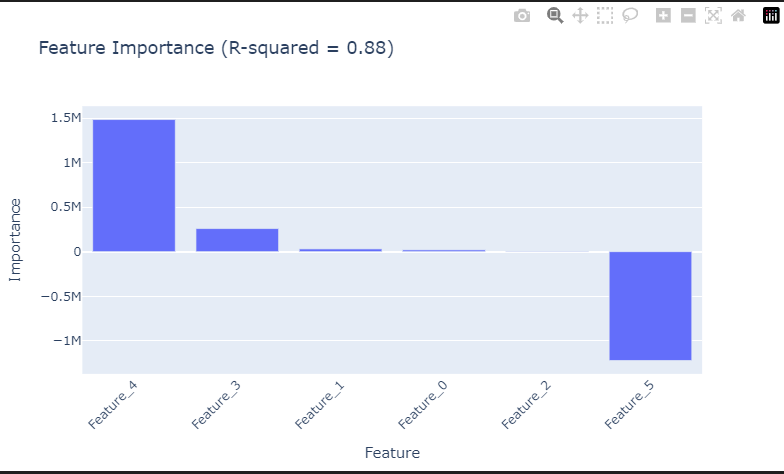
fig.add\_trace(go.Bar(x=feature\_importances['Feature'], y=feature\_importances['Importance']))

fig.update\_layout(title='Feature Importance (R-squared = {:.2f})'.format(r2),

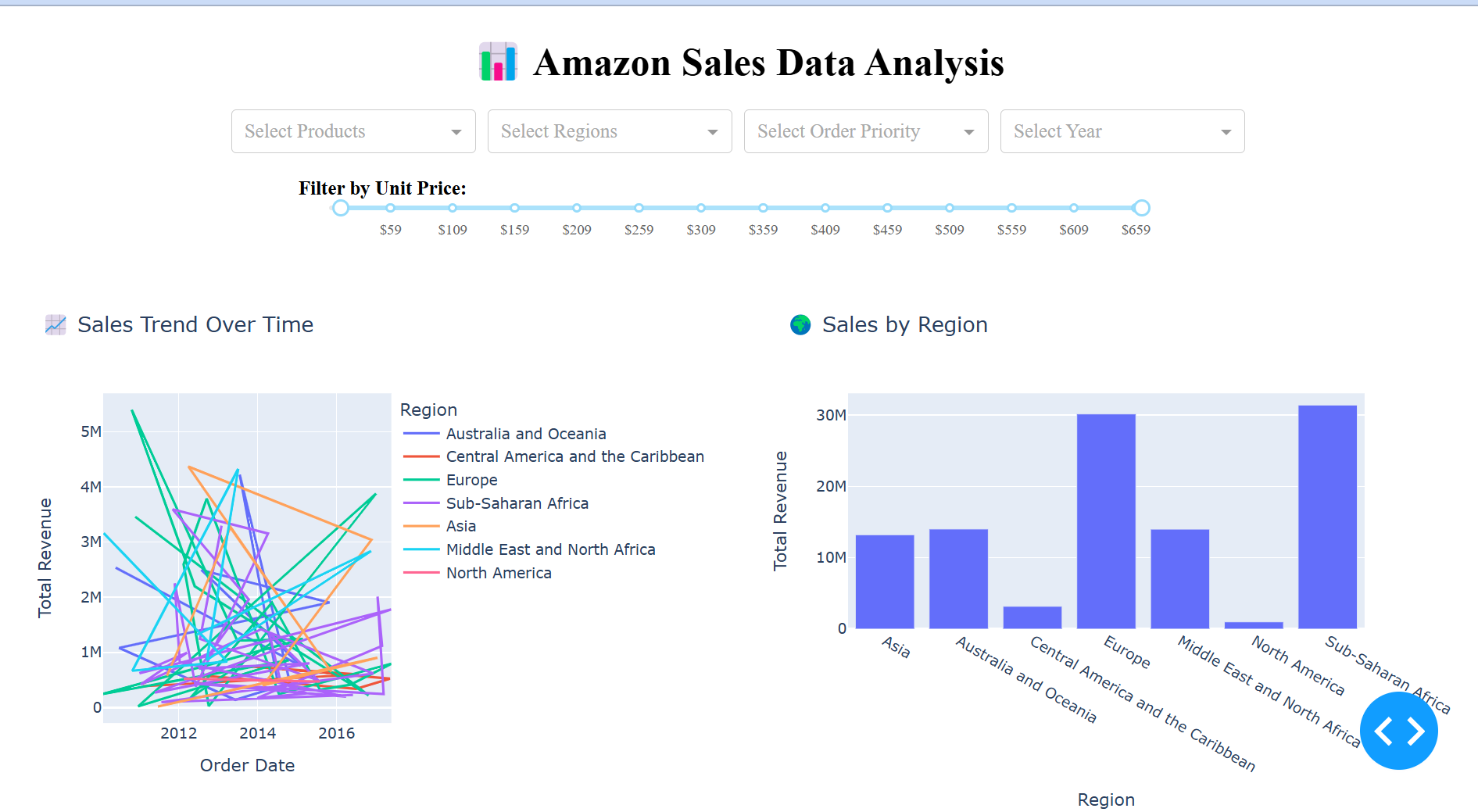
                  xaxis\_title='Feature',

                  yaxis\_title='Importance',

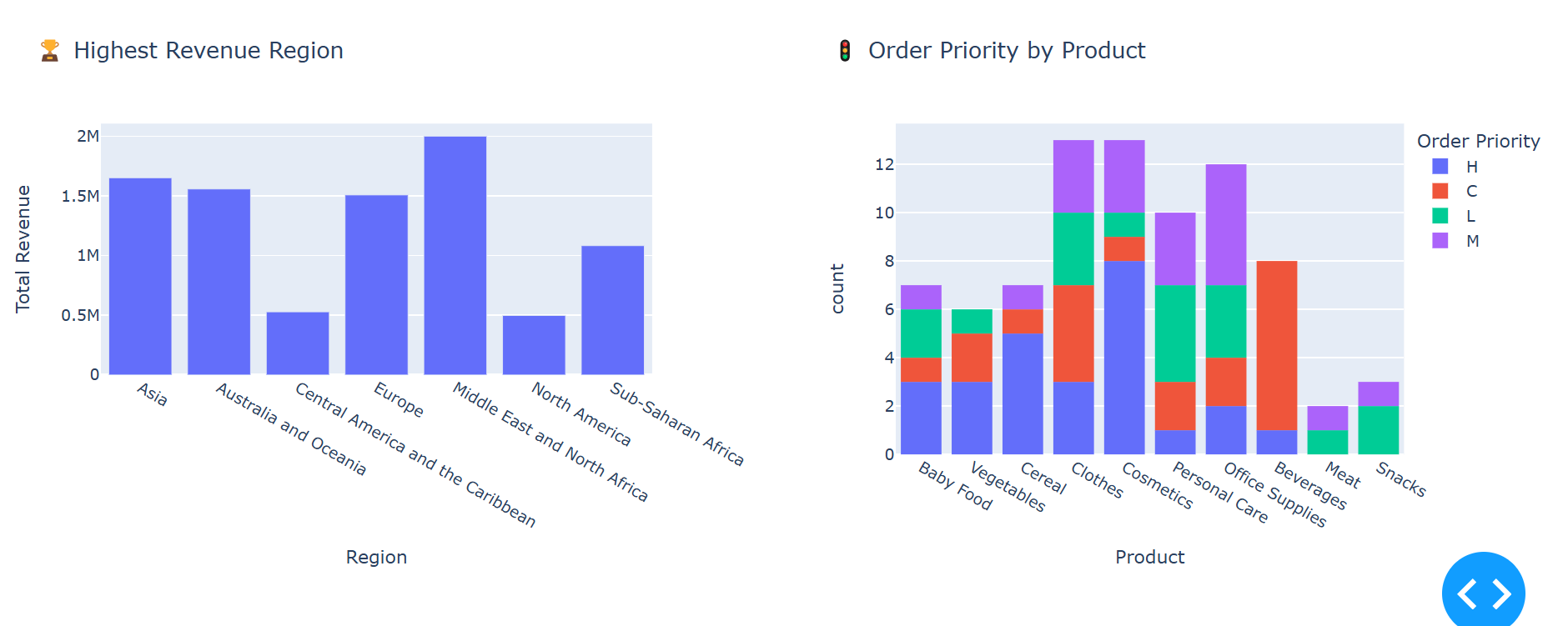
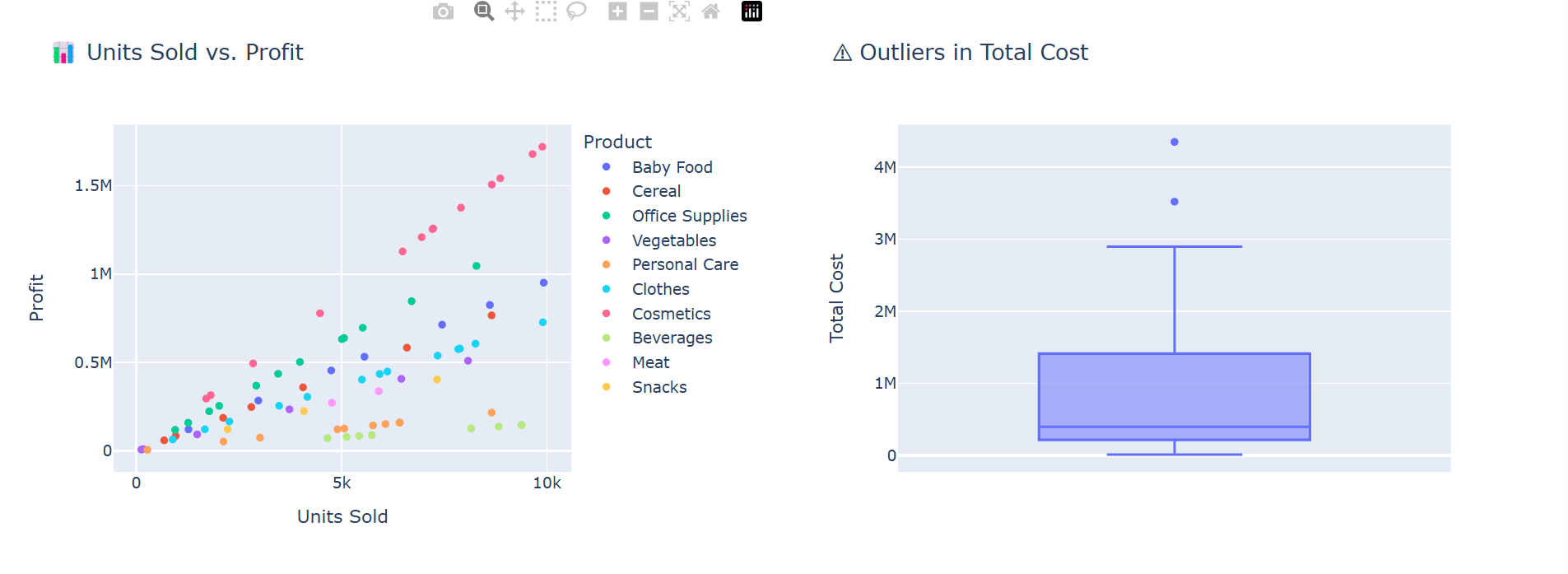
                  xaxis\_tickangle=-45)

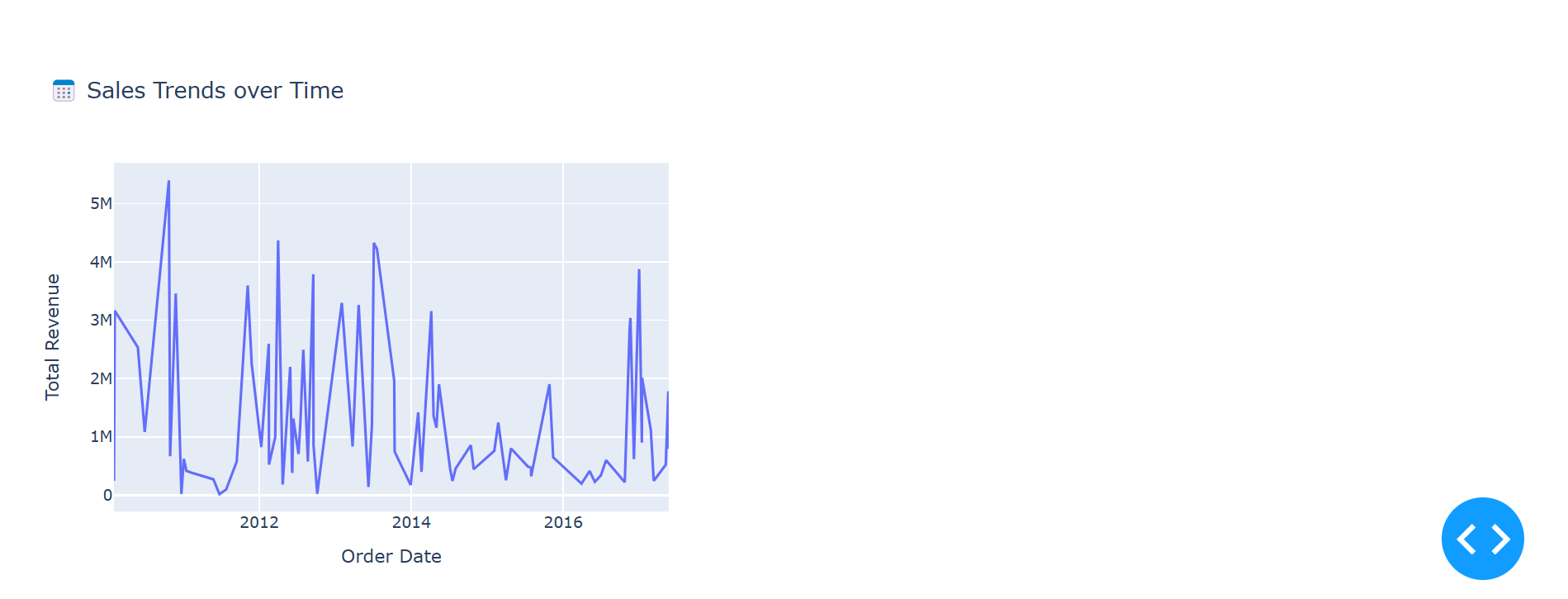
fig.show()

### Screenshot of Application

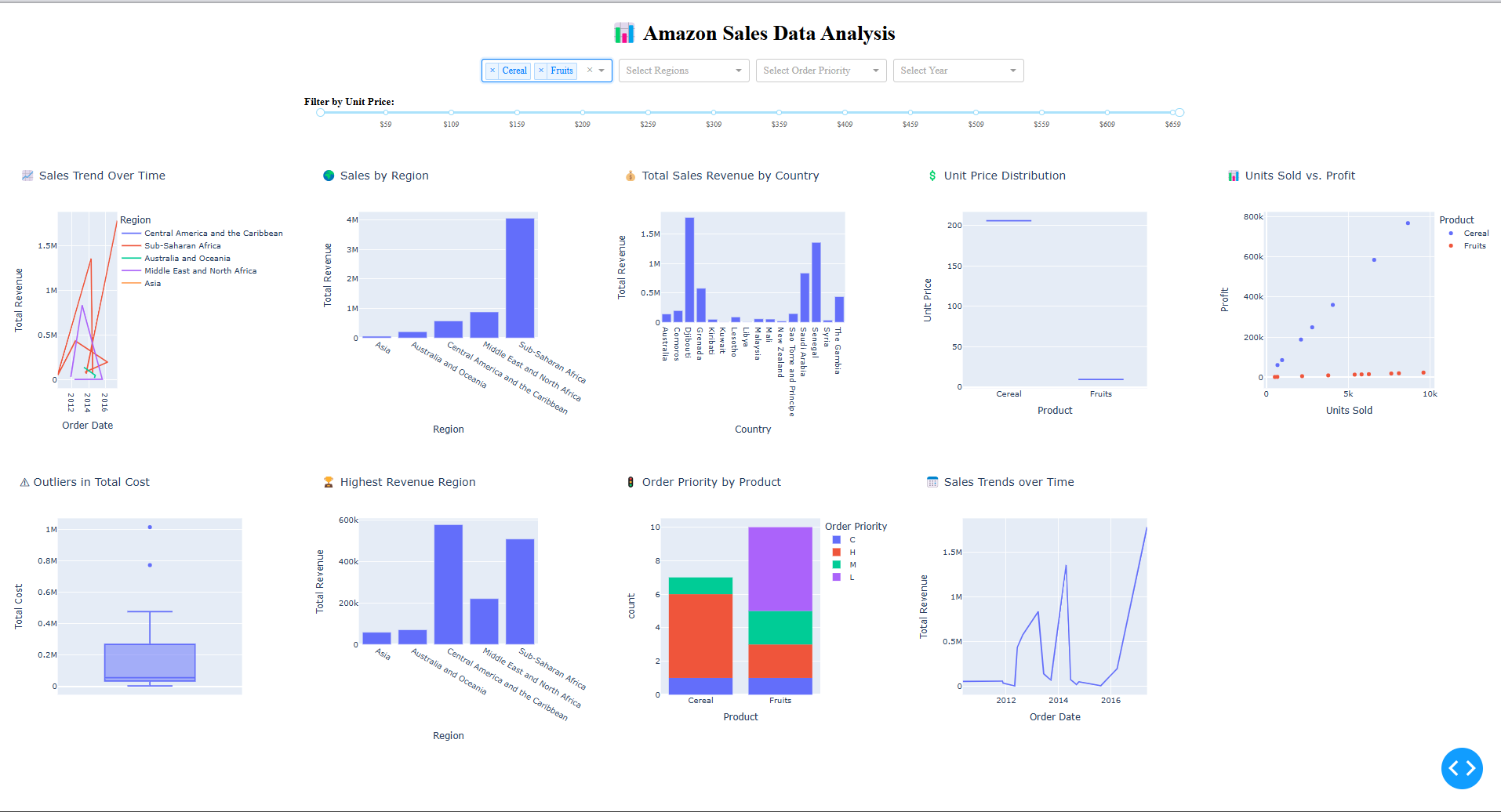


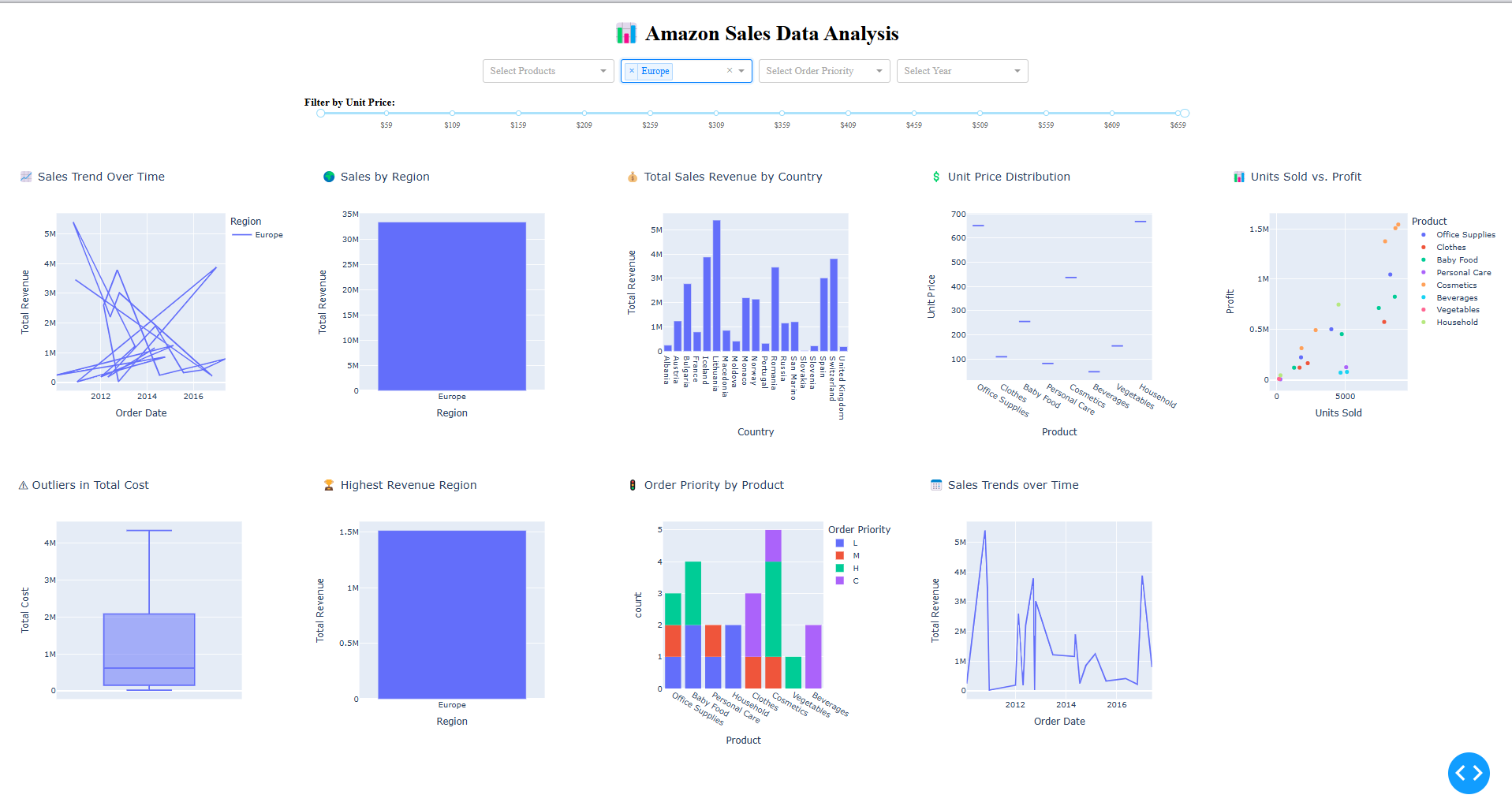


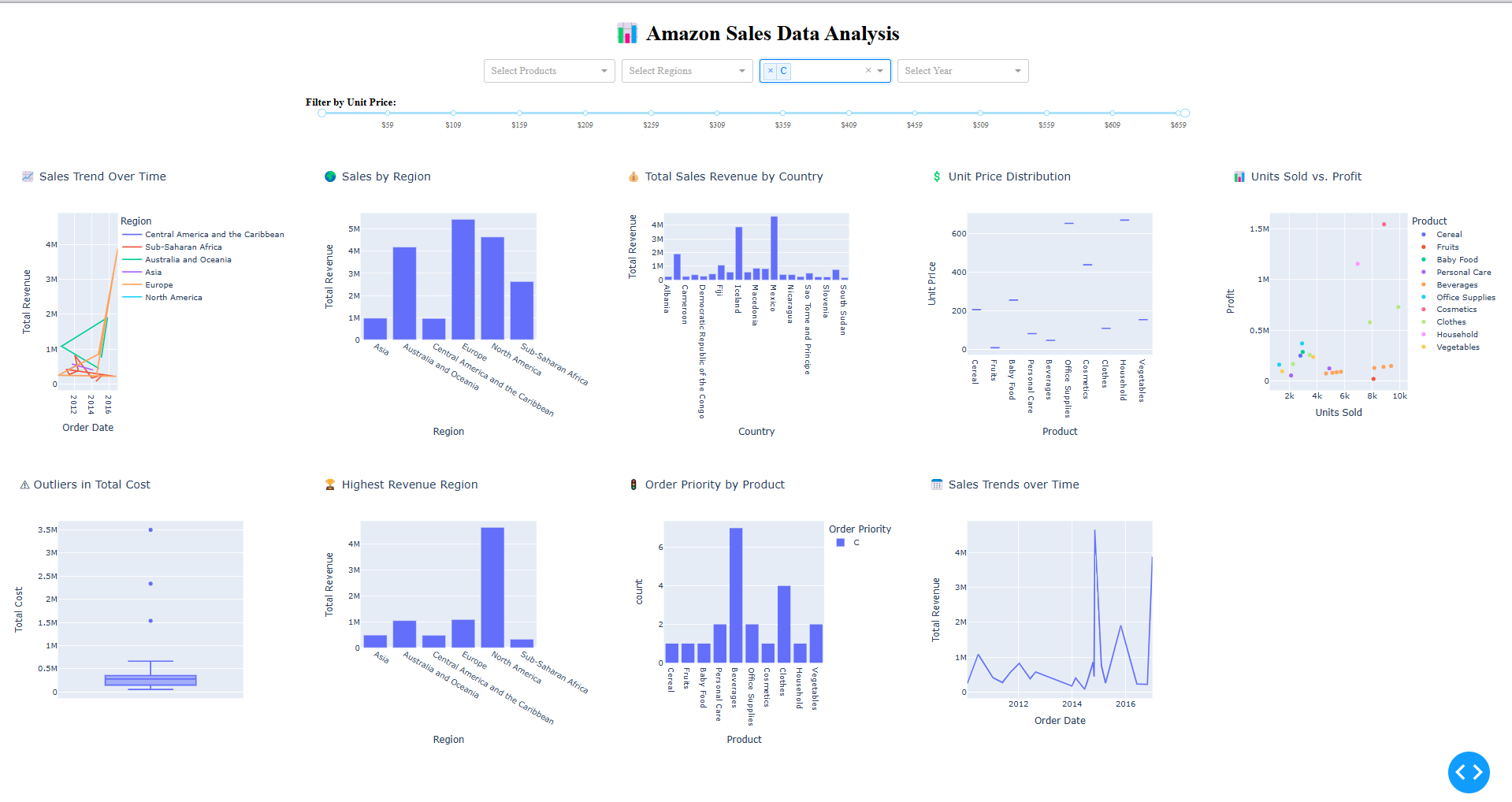


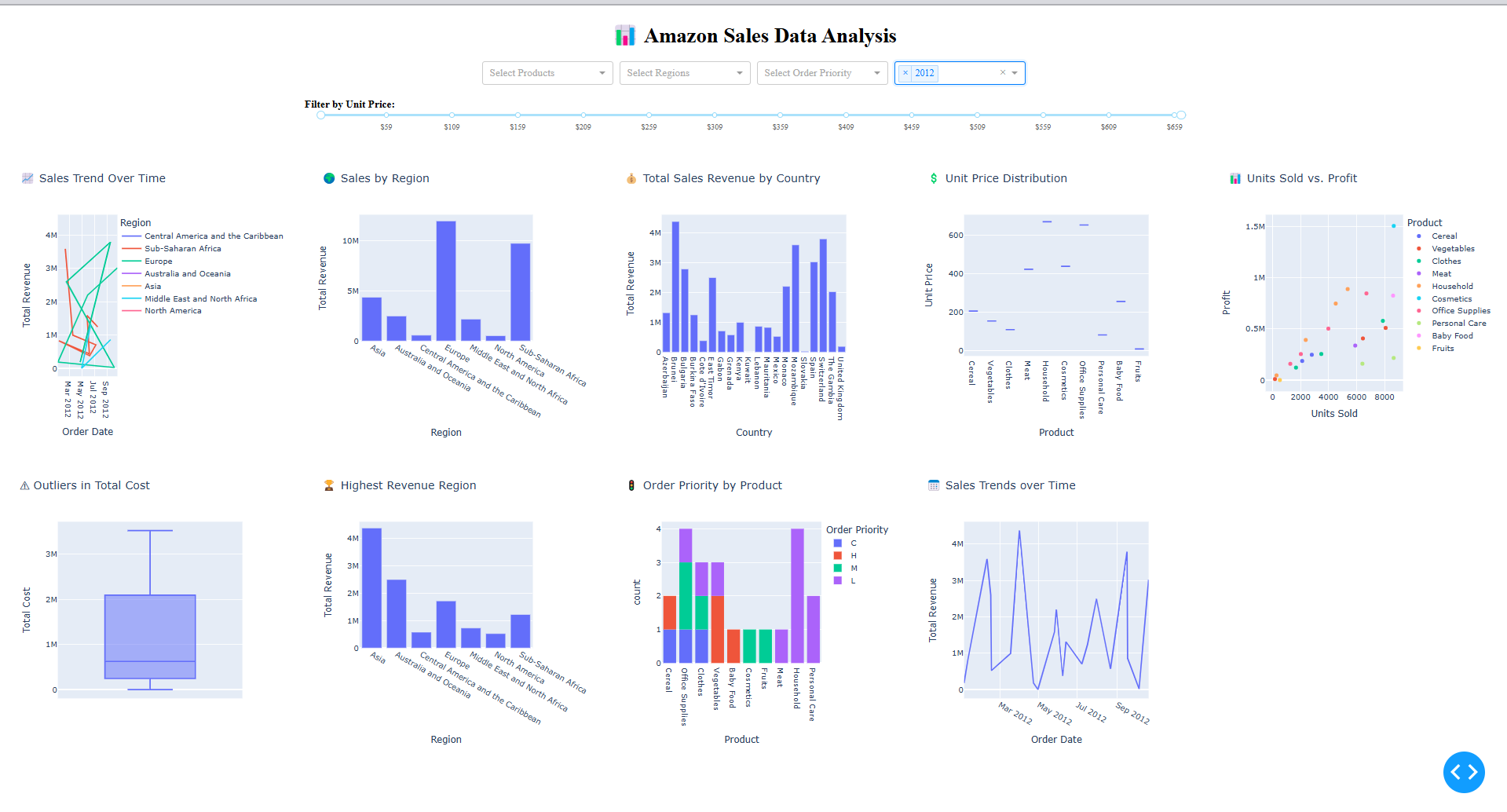


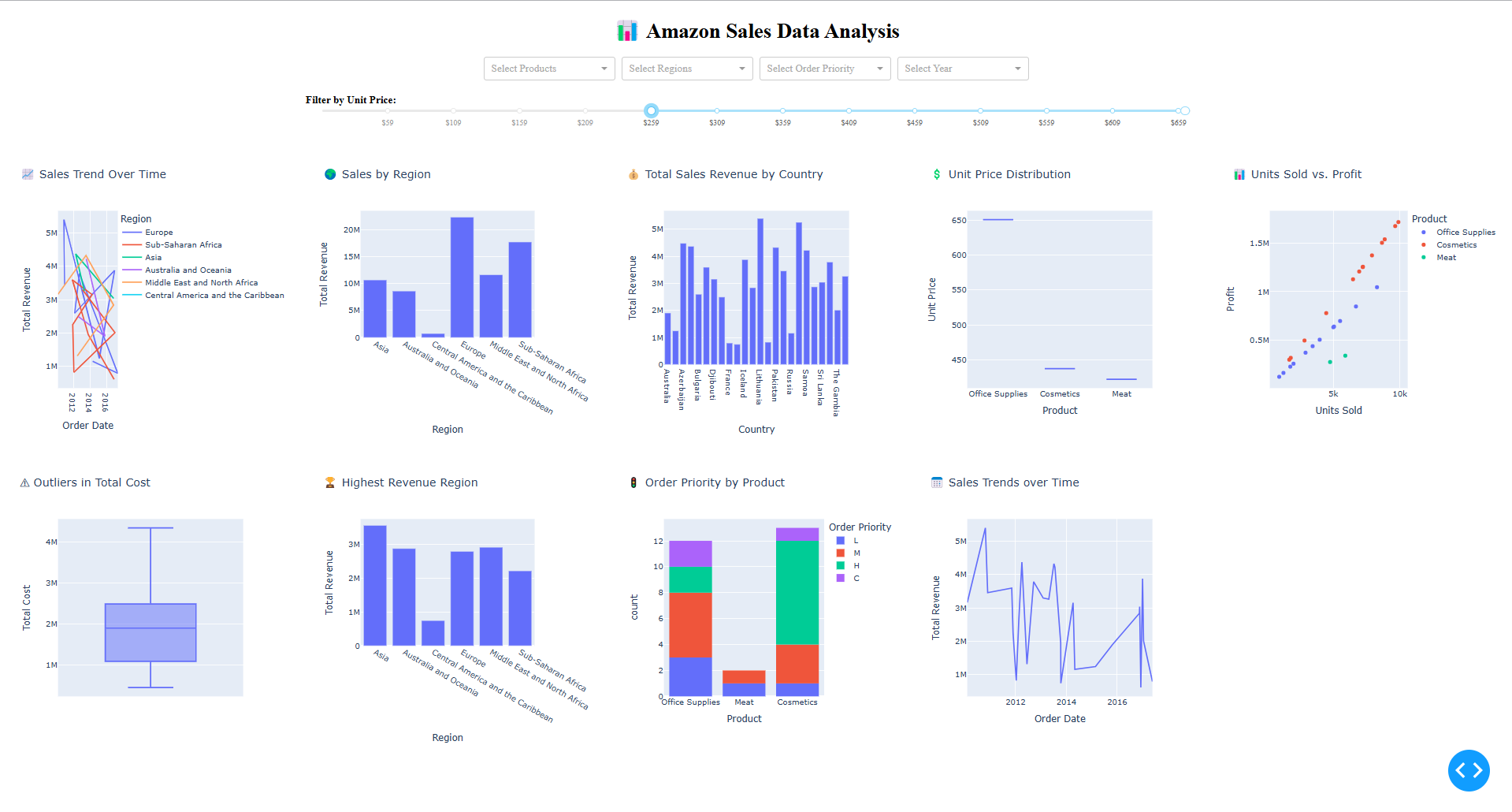
**Fig.5.2.1: Web Page**

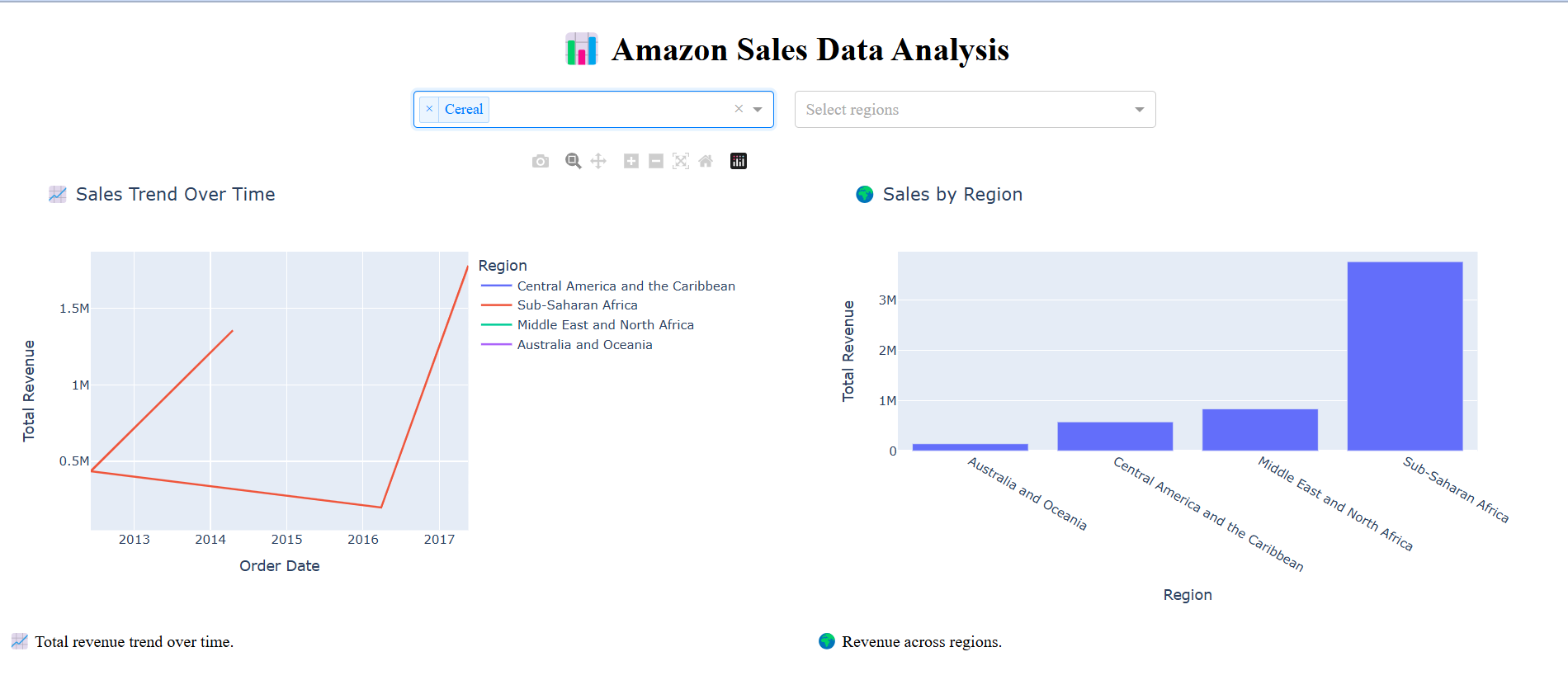
****

****

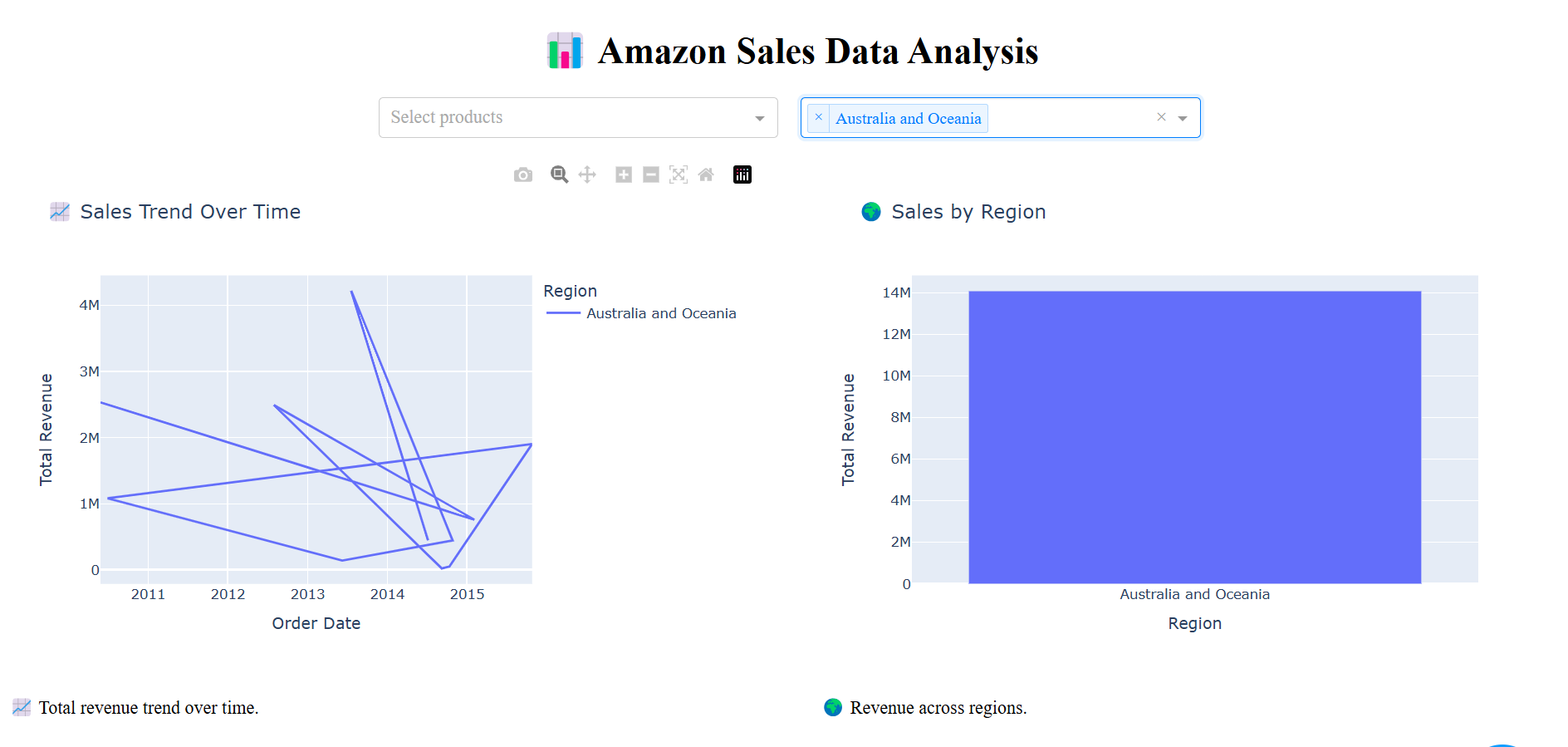
****

****

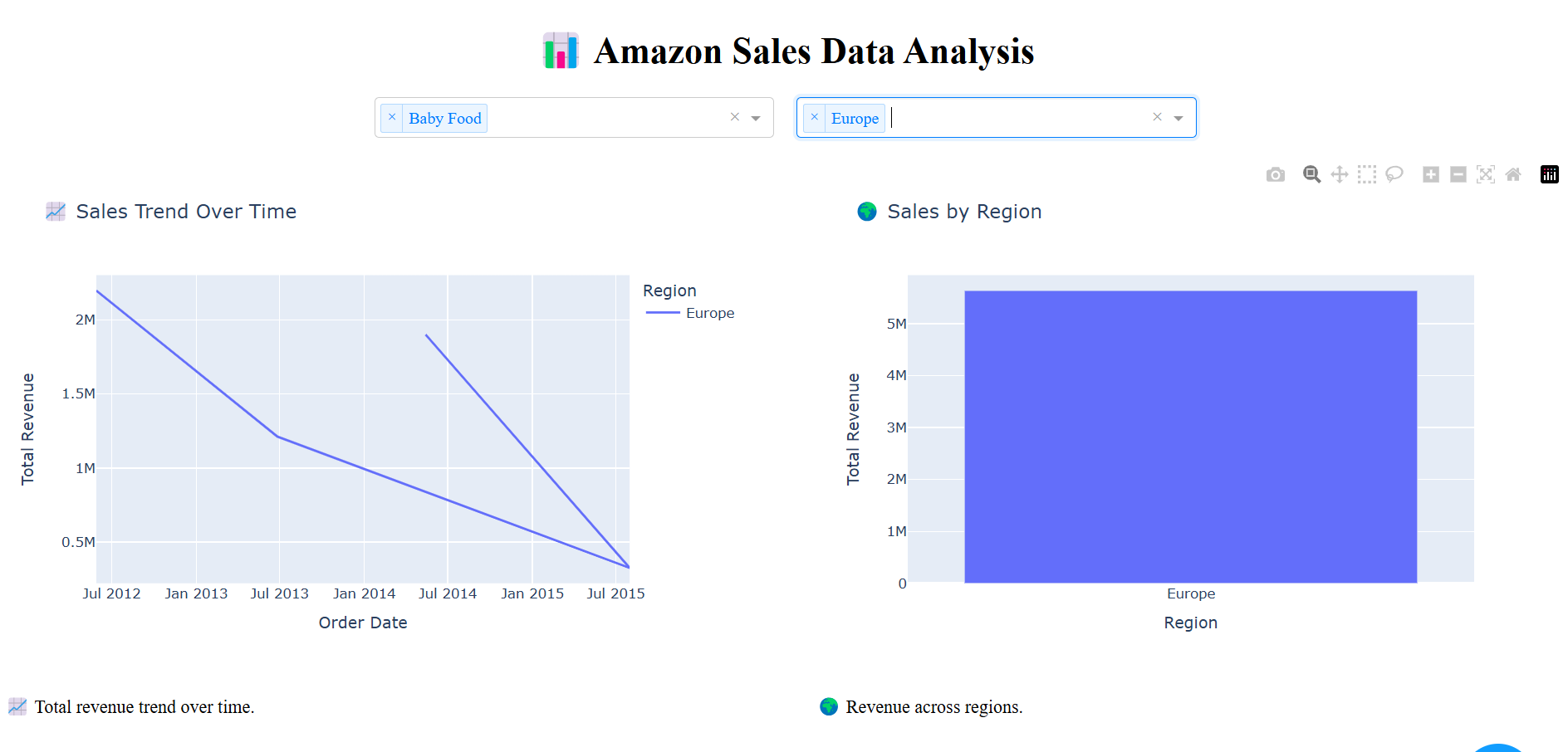
****

****

**Fig.5.2.2: Filtering products**

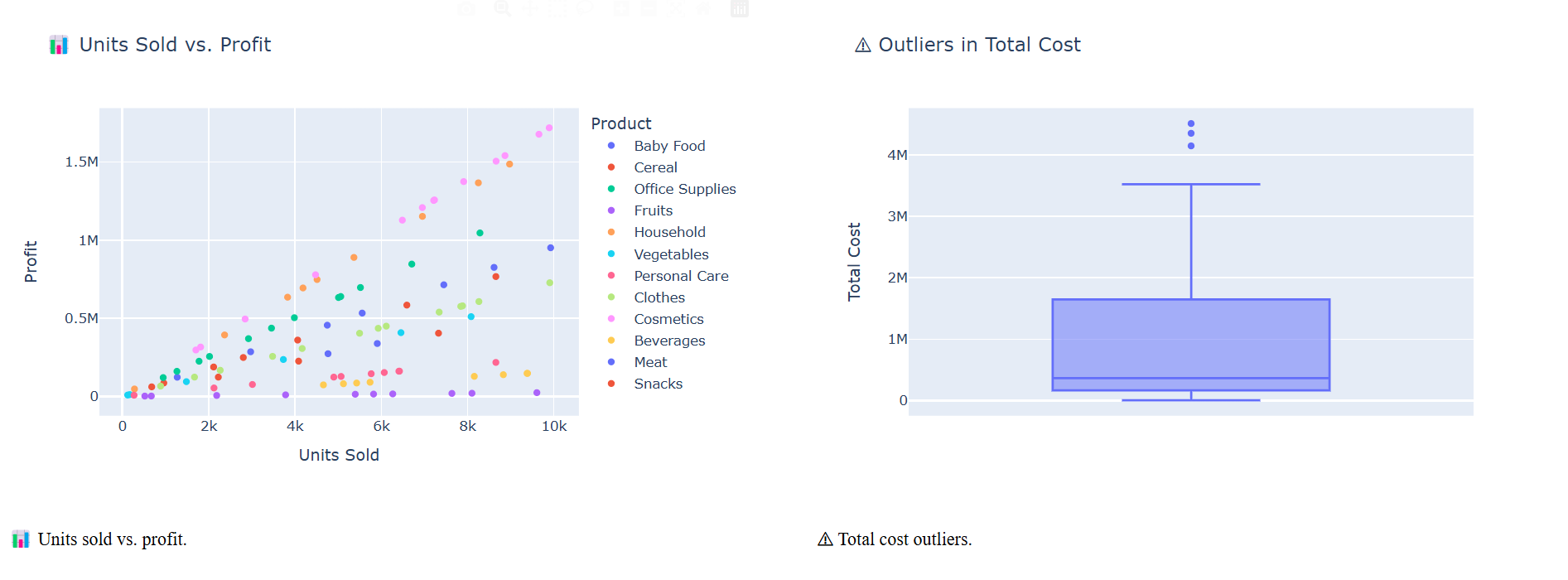
****

**Fig.5.2.3: Filtering Region**

****

**Fig.5.2.4: Filtering products and products**

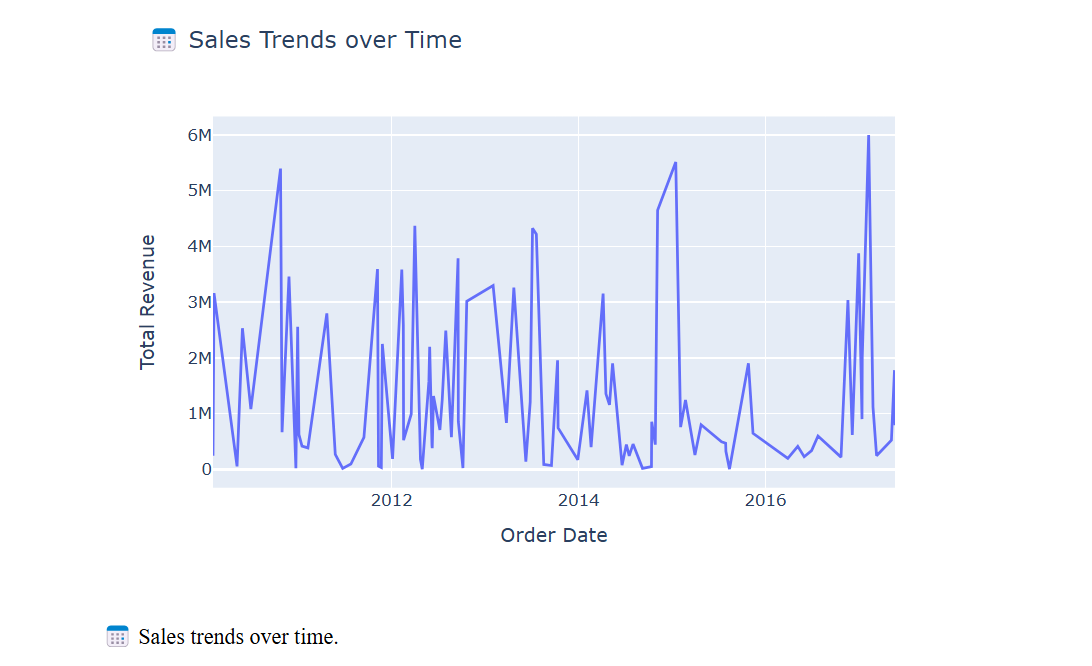
** Fig.5.2.5: Sales Revenue by Country and Unit Price Distrubution**

****

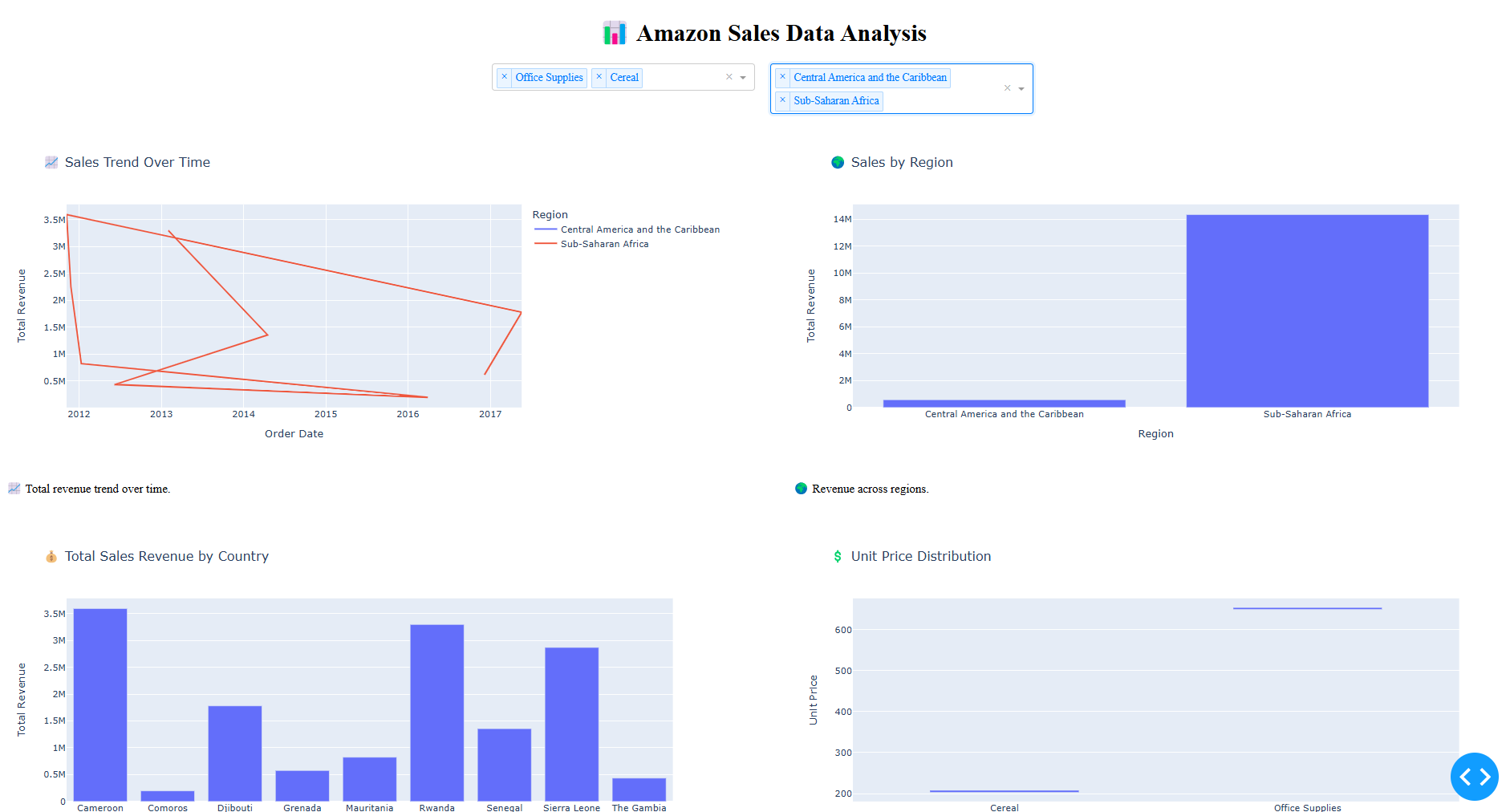
**Fig.5.2.6: Units sold vs Profit and Outliers in Total Cost**

****

**Fig.5.2.7: Highest Revenue and Order priority by product**

****

**Fig.5.2.8: Highest Revenue and Order priority by product**

****

**Fig.5.2.9: Selecting Multiple products and regions**

**CHAPTER – 6**

**RESULTS & CONCLUSION**

### Results

1. Interactive Dashboards: The project provides a dynamic and interactive dashboard that visualizes Amazon sales data. Users can explore different aspects of sales, such as trends over time, regional sales, unit prices, and profitability. This facilitates data-driven decision-making by presenting actionable insights at a glance.
2. Data Filtering: Users can filter the data based on various criteria such as product category, region, order priority, and year. This feature allows users to target specific insights, providing a more personalized view of the data relevant to their business needs.
3. Sales Trends: The dashboard effectively highlights trends in sales, revenue, and profit over time, helping businesses to identify patterns and forecast future sales. It supports decision-making around inventory, promotions, and other business strategies.
4. Anomaly Detection: The system can identify anomalies in the data, such as unexpected cost outliers or sudden revenue dips. This feature helps businesses detect irregularities early, enabling prompt action to rectify issues and optimize operations.
5. Real-Time Updates: The dashboard dynamically updates visualizations based on user inputs, providing real-time insights as filters are applied. This ensures that businesses are always working with the most current data available.
6. Performance Evaluation: The predictive models used in the project show promising results, with machine learning techniques offering accurate forecasts of sales and profits. These insights help businesses predict future trends and optimize their strategies accordingly.
7. User Experience: The user interface is simple and intuitive, making it easy for users with varying levels of expertise to navigate and extract insights. This enhances the accessibility of the tool for non-technical users.

Overall, the project successfully integrates data analysis, machine learning, and interactive visualization to deliver a comprehensive solution for Amazon sales data analysis, helping businesses optimize strategies and make informed decisions.

### Conclusion

The **Amazon Sales Data Analysis** project successfully demonstrated how **data-driven insights** can enhance **sales performance, profitability, and operational efficiency** in the e-commerce industry. By leveraging **Python for data preprocessing** and **Power BI for visualization**, the project provided a **comprehensive understanding of sales trends, product performance, regional variations, and customer behavior**.

**Identified top-selling products and high-revenue regions**, helping businesses focus on the most profitable areas.  
**Analyzed sales trends and seasonal fluctuations**, allowing for better inventory and pricing strategies.  
**Evaluated sales channel performance**, emphasizing the importance of online platforms for revenue growth.  
**Explored customer buying behavior**, highlighting the role of **repeat customers** in sustained profitability.  
**Provided data-driven recommendations**, enabling businesses to optimize **stock management, marketing strategies, and sales distribution**.

**Future Scope:**

**Real-time Sales Monitoring** – Implement live dashboards for real-time decision-making.  
**Predictive Analytics** – Use **machine learning** to forecast future sales trends.  
**Automated Data Processing** – Reduce manual intervention with **automated ETL pipelines**.  
**Enhanced Customer Segmentation** – Improve personalized marketing strategies using advanced analytics.

By utilizing **data analytics and visualization**, businesses can **make informed decisions, optimize sales strategies, and stay competitive** in the rapidly growing e-commerce landscape.

# REFERENCES

[1]. Sharma, R., & Patel, D. (2022). "The Role of Data Analytics in E-commerce Growth." \*International Journal of Business Intelligence\*. [2]. McKinsey & Company. (2023). "Harnessing Data for E-commerce Success." \*McKinsey Insights\*. Retrieved from [https://www.mckinsey.com](https://www.mckinsey.com). [3]. Amazon Web Services. (2022)."ETL Best Practices for Large-Scale E-commerce Data." \*AWS Documentation\*. Retrieved from [https://aws.amazon.com](https://aws.amazon.com). [4]. Gupta, A., & Roy, S. (2021). "Sales Trend Analysis Using Power BI and Python." \*Proceedings of the Data Science Conference\*. [5]. Kamat, R., & Gupta, P. (2022). "Enhancing Sales Analytics with Power BI: A Case Study on E-commerce Data." \*International Journal of Business Analytics\*.