**1. Introduction**

**Sales forecasting is a crucial process that enables businesses to make data-driven decisions about inventory management, marketing strategies, and financial planning. By analyzing historical sales data, companies can predict future trends and optimize resource allocation.**

**In this report, I will walk through the process of collecting, exploring, and preparing the dataset for predictive modeling. I will explain the steps taken, highlight key findings, and discuss potential areas for improvement. this document will provide you with a clear understanding of the process and the insights gained**.

**2. Data Collection**

**2.1 Dataset Overview**

The dataset (**Supermarket sales**) consists of **9,800 records** and **18 features**. It contains various details about customer orders, including:

**Order Date & Ship Date:** These timestamps are crucial for understanding seasonal trends and shipping efficiency.

**Sales:** The total revenue from each order. This is our primary target variable for forecasting.

**Category & Region:** Helps in understanding which product types and geographical locations contribute most to sales.

**Customer Segment:** Identifies whether the purchase was made by a consumer, corporate client, or a small business.

**2.2 Data Issues Identified**

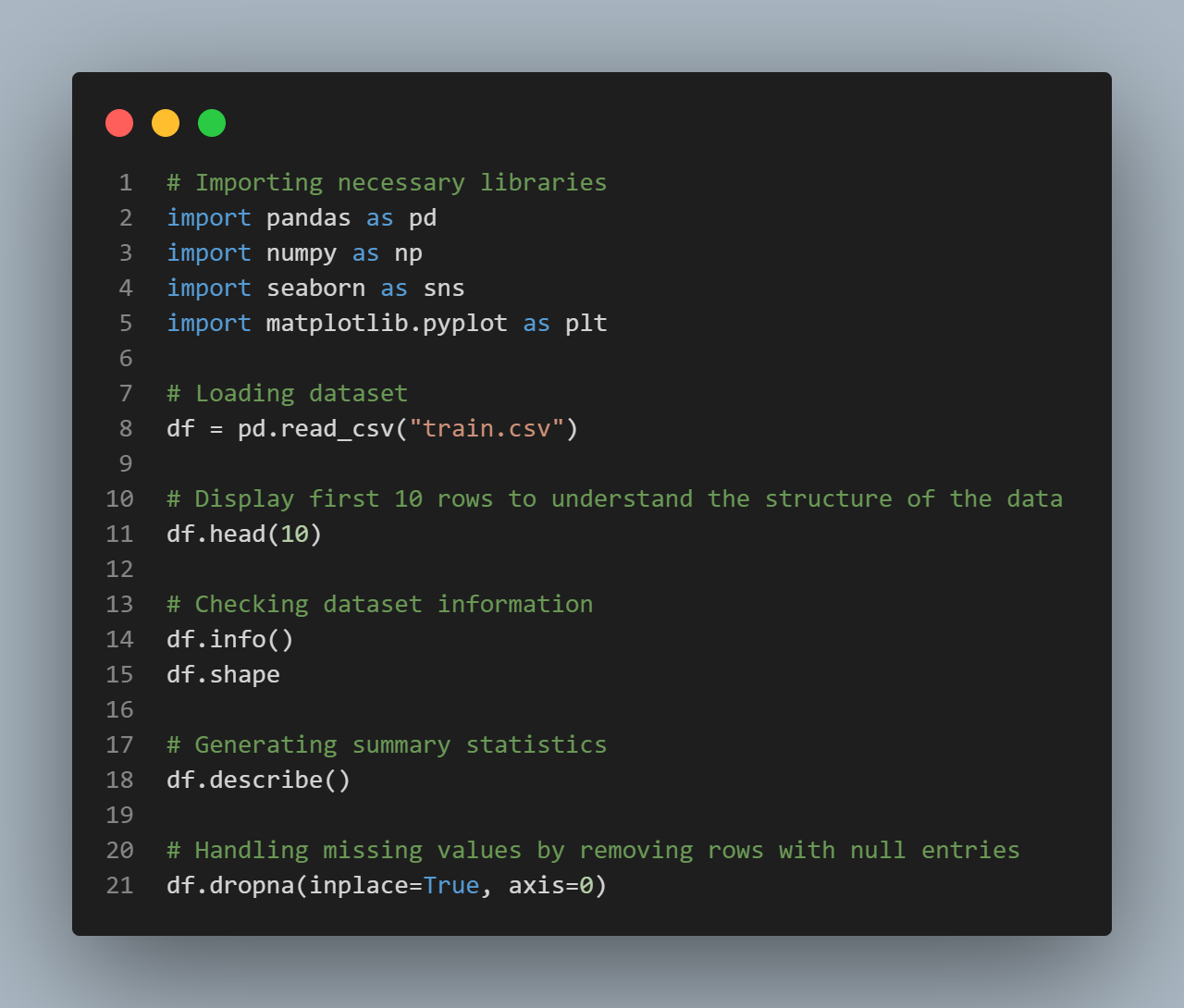
Before diving into analysis, I performed an initial quality check on the dataset and found some common issues:

**Missing Values:** The Postal Code column had 11 missing values. To maintain consistency, these records were removed.

**Data Types:** Order Date and Ship Date were stored as text, which makes time-based analysis difficult. These fields were converted into datetime format.

**Outliers:** The Sales column had extreme values, with some orders exceeding **$22,000**, which could distort predictions. Further outlier analysis was required.

**Code:**

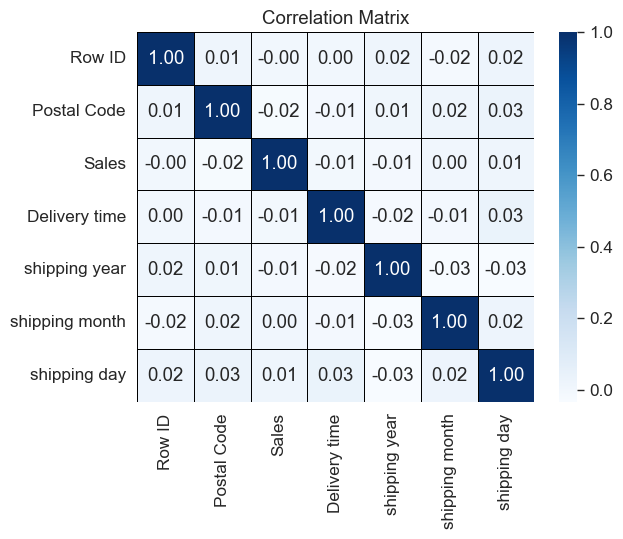
****

**3. Exploratory Data Analysis (EDA)**

**3.1 Correlation Analysis**

**To identify patterns in the dataset, I generated a correlation matrix to see how different features relate to Sales.**

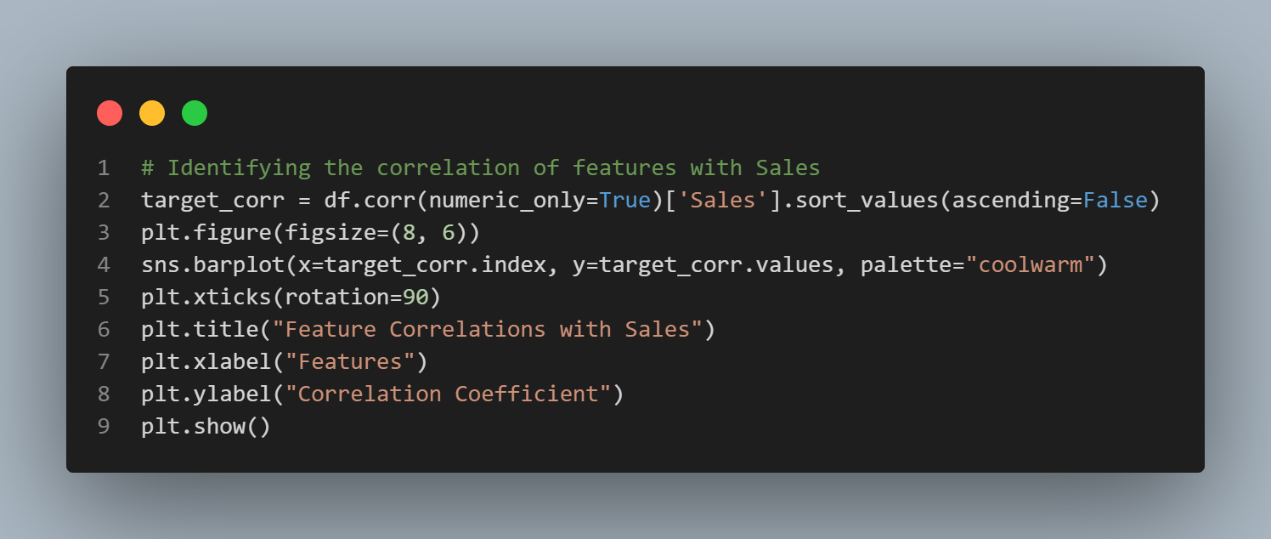
**Code:**

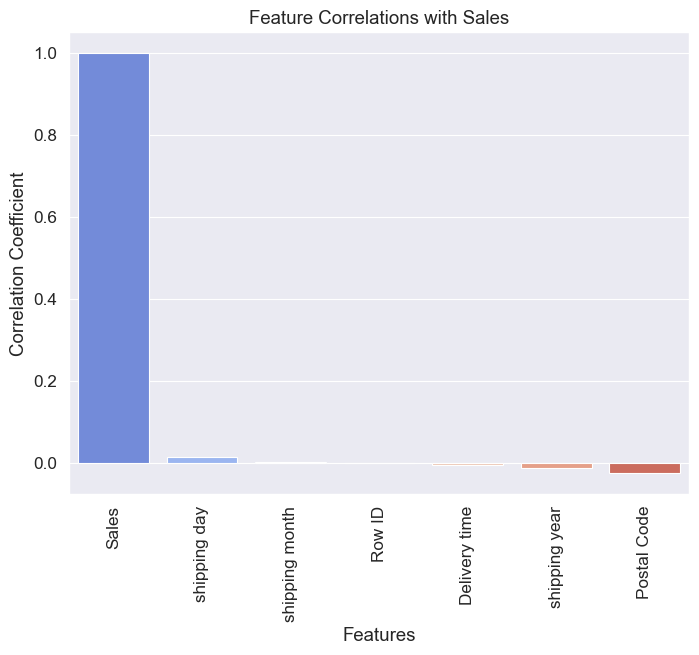
**A computer screen shot of a black screen

AI-generated content may be incorrect.Scheme:**

***This visualization helps identify which features have the strongest impact on sales. For example, discount rates or product categories may be more relevant than customer demographics.***

**Code:**

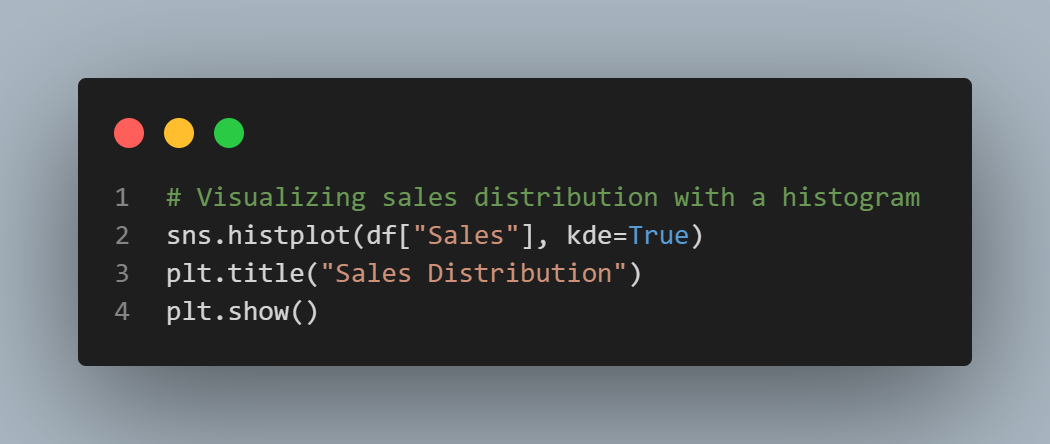
****

**Scheme :**

**3.2 Sales Distribution**

One of the first things I wanted to check was how sales are distributed across all transactions. This helps detect patterns such as skewness and outliers.

**Code:**

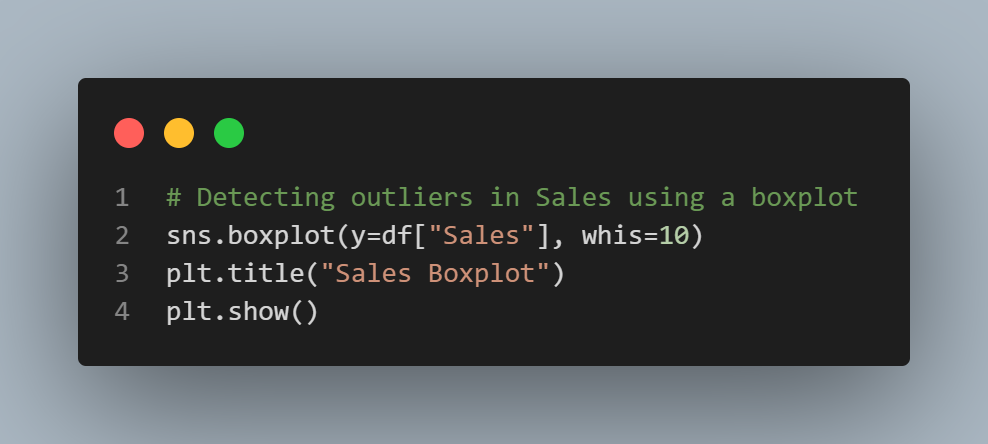
**Scheme** **:**

A graph with numbers and lines

AI-generated content may be incorrect.

From the histogram, I noticed that sales data is **highly right-skewed**, meaning there are many small-value sales and a few very high-value transactions.

**Code:**



A graph with lines and dots

AI-generated content may be incorrect.**Scheme:**

The boxplot confirms that there are significant outliers. This suggests that we might need to normalize the data using a **log transformation** to reduce skewness.

**3.3 Sales Trends Over Time**

Next, I wanted to explore how sales changed over the years. This helps identify whether there are seasonal trends or growth patterns.

**Code:**

**A computer screen shot of a black screen

AI-generated content may be incorrect.**

**Scheme:**

A graph showing the number of sales across years

AI-generated content may be incorrect.A line graph with numbers and a line

AI-generated content may be incorrect.

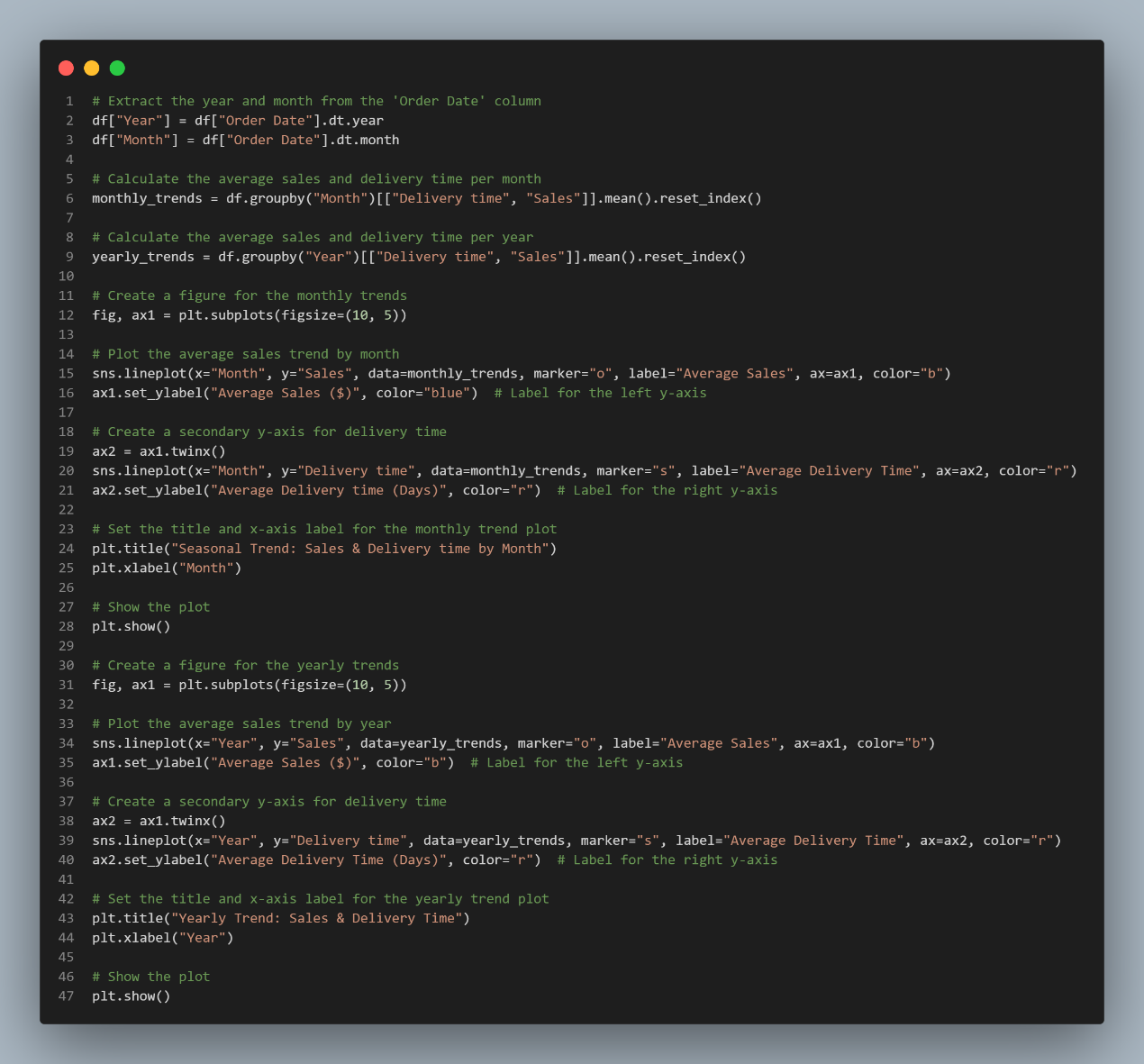
**3.3.1 Seasonal and Yearly Sales & Delivery Time Trends**

To better understand the relationship between sales and delivery times, I generated two visualizations:

The first chart illustrates the monthly trends in sales and delivery time, showcasing how these variables fluctuate throughout the year.

The second chart presents the yearly trends, helping identify long-term patterns in sales performance and shipping efficiency.These insights allow us to analyze seasonal peaks and yearly variations, which are crucial for forecasting and optimizing inventory management.

**Code:**



A graph of sales and delivery

AI-generated content may be incorrect.**Scheme:**

A graph showing the growth of sales and delivery time

AI-generated content may be incorrect.

**3.4 Sales by Category and Region**

To further break down the trends, I analyzed sales based on **product category** and **geographic region**.

**Code:** A screen shot of a computer

AI-generated content may be incorrect.

**Scheme:**

A graph of sales by category

AI-generated content may be incorrect.A graph of different colored bars

AI-generated content may be incorrect.

**3.5 Shipping Mode Analysis**

I also examined which shipping modes are used most frequently and whether they impact total sales.

**Code:**

**A screen shot of a computer program

AI-generated content may be incorrect.**

**Scheme:**

A graph of different colored bars

AI-generated content may be incorrect.

**4. Data Preprocessing**

**Data preprocessing is a crucial step to ensure that the dataset is clean, structured, and optimized for modeling. The following steps were performed:**

**4.1 Handling Missing Values**

**The dataset contained missing values in the Postal Code column.**

**These missing values were removed to maintain consistency.**

**4.2 Feature Engineering**

**Extracted Year and Month from the Order Date column to help in time-series analysis.**

**Created a Delivery Time feature by calculating the difference between Ship Date and Order Date.**

**4.3 Handling Outliers**

**Since the Sales column had extreme values, we applied log transformation to reduce the impact of outliers.**