Project Report on

SALES DATA ANALYSIS AND VISUALIZATION

For

Introduction to AI (AI101B)

By

Aanchal - 202410116100002
Devanshi Singhal- 202410116100060
Deepanshu Ruhela – 202410116100056
Dhwani Panchal -202410116100063

Session:2024-2025 (Semester II)

Under the supervision of

MR. APOORV JAIN (Assistant Professor)

KIET Group of Institutions, Delhi-NCR, Ghaziabad

DEPARTMENT OF COMPUTER APPLICATIONS
KIET GROUP OF INSTITUTIONS, DELHI-NCR, GHAZIABAD-201206

INTRODUCTION

In today’s data-driven business environment, organizations generate vast amounts of sales data. However, raw data alone does not provide much value unless it is analyzed and visualized effectively. Data visualization plays a crucial role in transforming raw sales figures into meaningful insights that can drive strategic business decisions. By leveraging data visualization techniques, businesses can detect patterns, monitor sales performance, and identify trends that influence profitability.

This project aims to utilize Python and its powerful data analysis libraries to visualize sales data, making it easier to interpret and act upon. With tools such as Pandas for data manipulation, matplotlib and seaborn for creating static visualizations, and plotly for interactive visualizations, this project demonstrates how data visualization enhances business intelligence.

The importance of sales data visualization extends beyond just monitoring performance. It allows companies to forecast trends, allocate resources efficiently, and understand customer purchasing behavior. Businesses can use these insights to make informed decisions regarding inventory management, marketing strategies, and sales operations. Additionally, sales data visualization can highlight seasonality effects, peak sales periods, and regional preferences, enabling businesses to optimize their strategies for better market penetration.

The project will follow a structured methodology, starting with data collection and preprocessing, followed by exploratory data analysis and visualization. By the end of this study, key insights into sales patterns will be derived, providing recommendations to improve overall business performance. The findings from this project will be beneficial for stakeholders who need a clear, visual representation of sales trends to enhance their decision-making processes.

METHODOLOGY

The methodology followed in this project ensures a structured approach to analyzing sales data through visualization. The key steps involved are as follows:

1. **Data Collection:** The first step is to gather relevant sales data. This may come from various sources such as company databases, publicly available datasets, or e-commerce transaction records. The dataset should include key attributes such as date, product, category, quantity sold, unit price, total sales, and regional information. The dataset is provided in **CSV (Comma-Separated Values) format**, containing information about sales transactions, including:
   - **Date** – The day the transaction took place.
   - **Product** – The name of the product sold.
   - **Units Sold** – The number of units sold on that day.
   - **Revenue** – The total revenue generated from the sales.

2. **Data Preprocessing:** Before proceeding with analysis, data preprocessing is essential to ensure accuracy and consistency. The preprocessing steps include:

   - Handling Missing Values**:** Checking for null or missing values and filling them using interpolation, mean substitution, or removal if necessary.
   - Formatting Dates**:** Ensuring that date columns are in proper date time format to facilitate time-series analysis.
   - Removing Duplicates**:** Identifying and eliminating duplicate records to maintain data integrity.
   - Creating Derived Metrics**:** Computing new fields such as total sales (Quantity Sold × Unit Price) for better insights.

3. **Exploratory Data Analysis (EDA):** EDA involves generating summary statistics and understanding data distributions through:

   - Descriptive Statistics**:** Mean, median, standard deviation, and correlation between sales variables.
   - Initial Visualizations**:** Creating histograms, box plots, and scatter plots to observe patterns and anomalies.
   - Outlier Detection**:** Identifying and handling extreme values that may distort the analysis.

4. **Visualization Development:** The core of this project lies in developing compelling visualizations. Different types of charts are used for specific insights:

   - Line Charts: Depicting sales trends over time.
   - Bar Graphs: Showing top-selling products and categories.
   - Pie Charts**:** Visualizing category-wise contribution to total sales.
   - Heat maps**:** Representing regional sales distribution.
   - Scatter Plots**:** Understanding correlations between variables.
   - Interactive Dashboards**:** Using Plotly to create dynamic and interactive charts for deeper insights.

5. **Insight Extraction:** Once visualizations are developed, the next step is to analyze the key findings:

   - Identifying peak sales months and seasonal patterns.
   - Determining which products and regions contribute the most to sales.
   - Analyzing customer buying behavior based on trends.
   - Understanding external factors influencing sales, such as promotions or economic conditions.

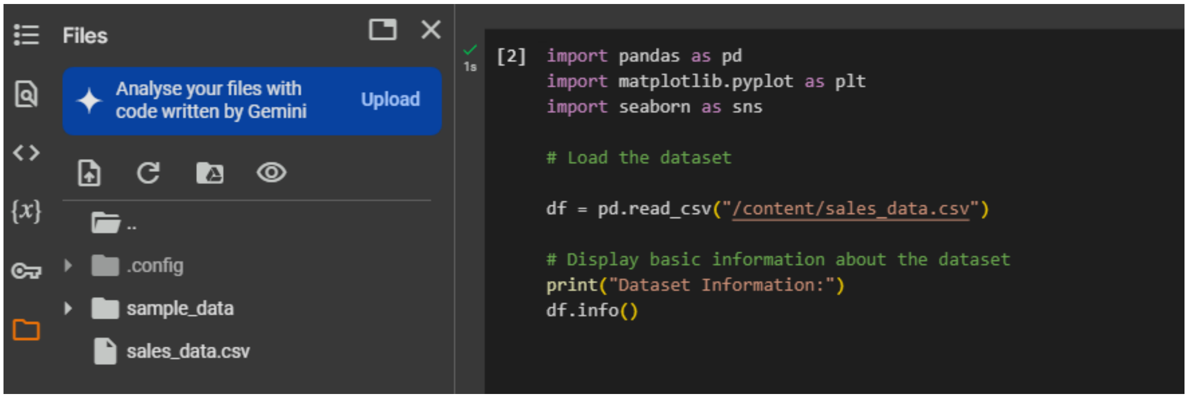
6. **Report Generation:** The final step is compiling the findings into a structured report. This includes:

   - Summarizing key insights derived from visualizations.
   - Providing actionable recommendations based on the analysis.
   - Suggesting future improvements and potential areas for further research.

By following this systematic methodology, the project ensures a comprehensive approach to sales data visualization, making it an effective tool for business decision-making..

PYTHON CODE

```
#import libraries

import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns



# Load the dataset

df = pd.read_csv("/content/sales_data.csv")
```

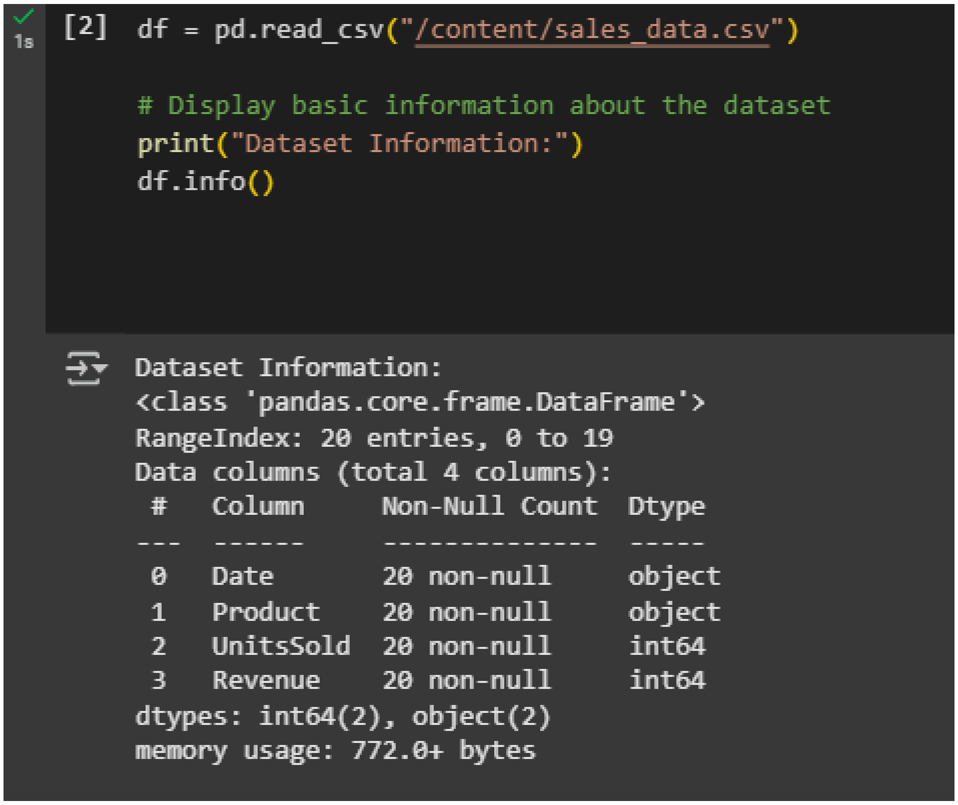


```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

# Load the dataset

df = pd.read_csv("/content/sales_data.csv")

# Display basic information about the dataset
print("Dataset Information:")
df.info()
```

Display basic information about the dataset

print("Dataset Information:")

df.info()

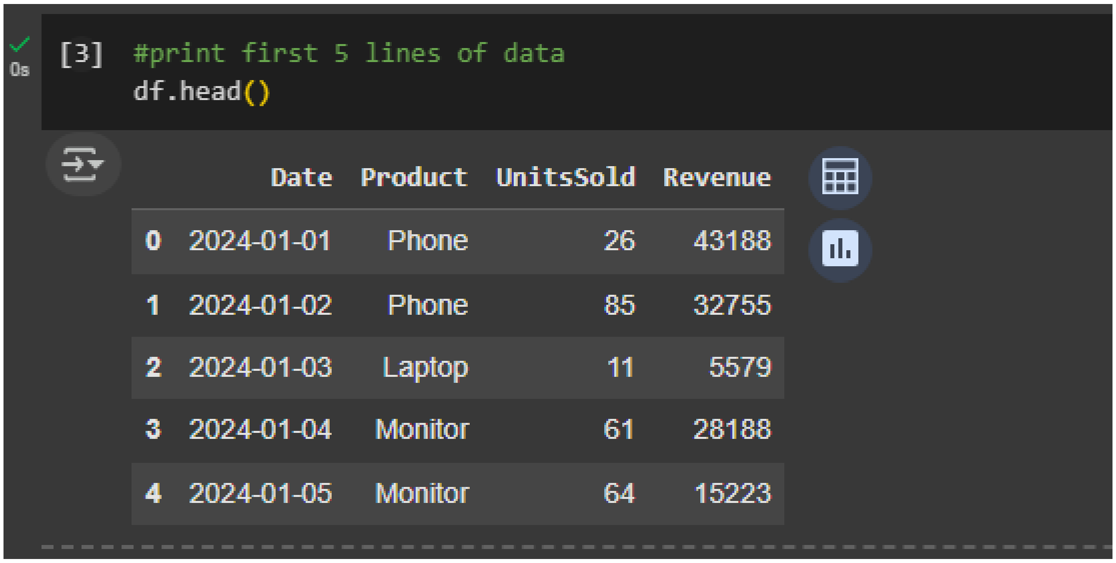


```
[2] df = pd.read_csv("/content/sales_data.csv")

    # Display basic information about the dataset
    print("Dataset Information:")
    df.info()
```

```
Dataset Information:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 20 entries, 0 to 19
Data columns (total 4 columns):
 #   Column     Non-Null Count  Dtype 
---  ------     --------------  ----- 
 0   Date       20 non-null     object
 1   Product    20 non-null     object
 2   UnitsSold  20 non-null     int64 
 3   Revenue    20 non-null     int64 
dtypes: int64(2), object(2)
memory usage: 772.0+ bytes
```

#print first 5 lines of data

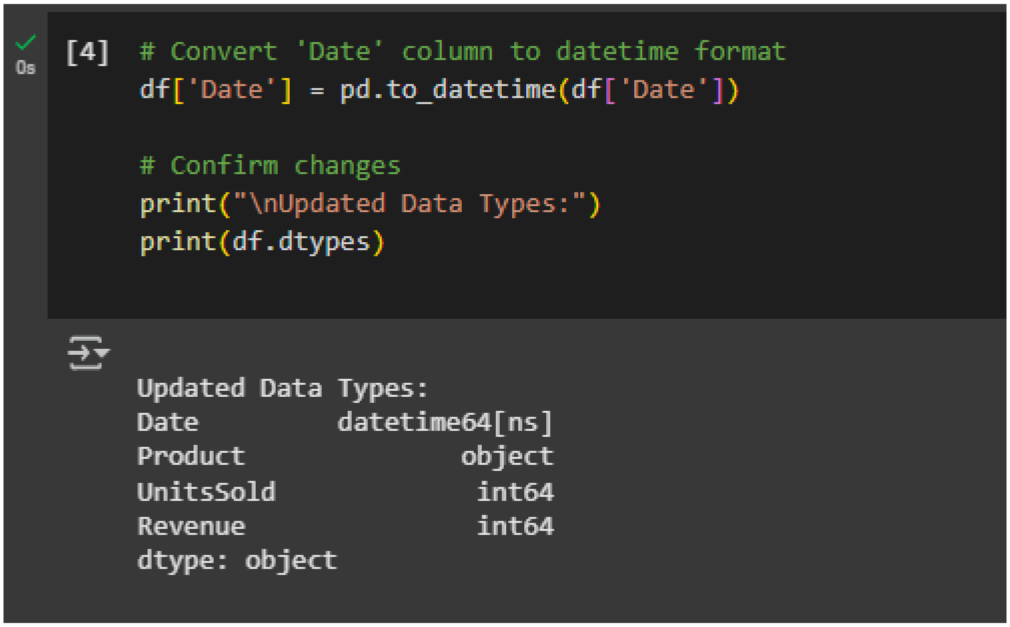
df.head()



```
[3] #print first 5 lines of data
    df.head()
```

	Date	Product	UnitsSold	Revenue
0	2024-01-01	Phone	26	43188
1	2024-01-02	Phone	85	32755
2	2024-01-03	Laptop	11	5579
3	2024-01-04	Monitor	61	28188
4	2024-01-05	Monitor	64	15223

```
# Convert 'Date' column to datetime format

df['Date'] = pd.to_datetime(df['Date'])

# Confirm changes

print("\nUpdated Data Types:")

print(df.dtypes)
```



```
# Convert 'Date' column to datetime format
df['Date'] = pd.to_datetime(df['Date'])

# Confirm changes
print("\nUpdated Data Types:")
print(df.dtypes)
```

```
Updated Data Types:
Date         datetime64[ns]
Product              object
UnitsSold             int64
Revenue               int64
dtype: object
```

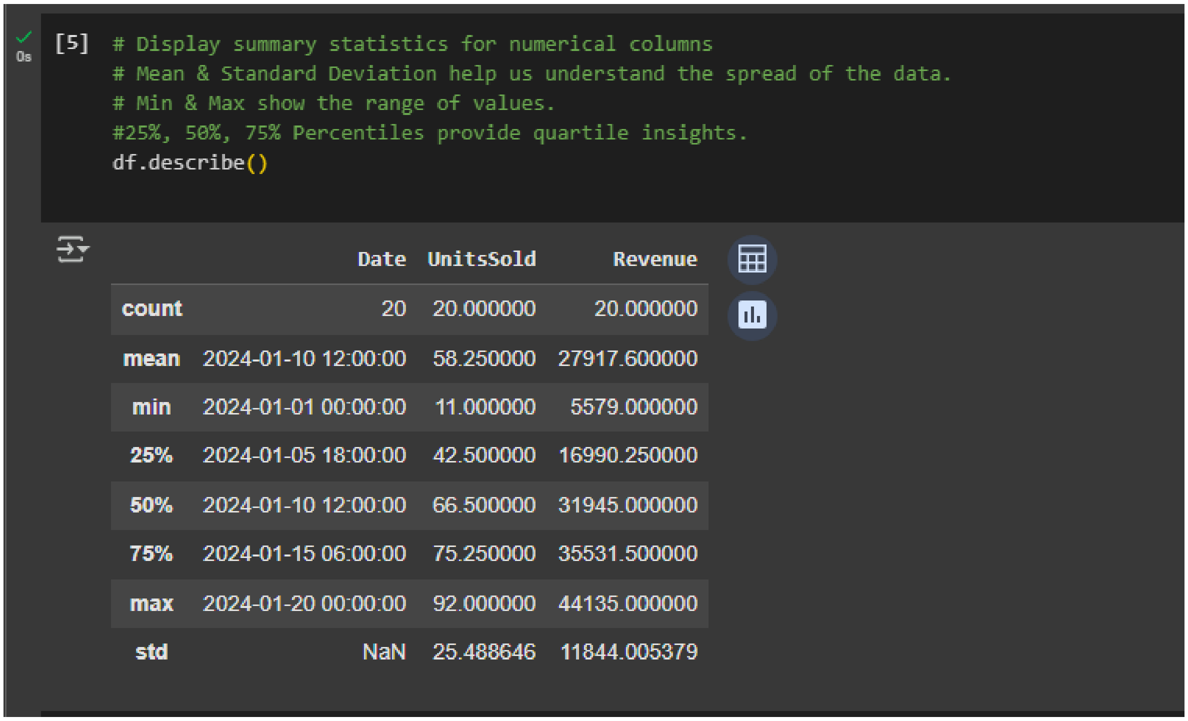
# Display summary statistics for numerical columns

# Mean & Standard Deviation help us understand the spread of the data.

# Min & Max show the range of values.

#25%, 50%, 75% Percentiles provide quartile insights.

df.describe()



```python
# Display summary statistics for numerical columns
# Mean & Standard Deviation help us understand the spread of the data.
# Min & Max show the range of values.
#25%, 50%, 75% Percentiles provide quartile insights.
df.describe()
```

```
                      Date  UnitsSold       Revenue
count                   20  20.000000     20.000000
mean   2024-01-10 12:00:00  58.250000  27917.600000
min    2024-01-01 00:00:00  11.000000   5579.000000
25%    2024-01-05 18:00:00  42.500000  16990.250000
50%    2024-01-10 12:00:00  66.500000  31945.000000
75%    2024-01-15 06:00:00  75.250000  35531.500000
max    2024-01-20 00:00:00  92.000000  44135.000000
std                    NaN  25.488646  11844.005379
```

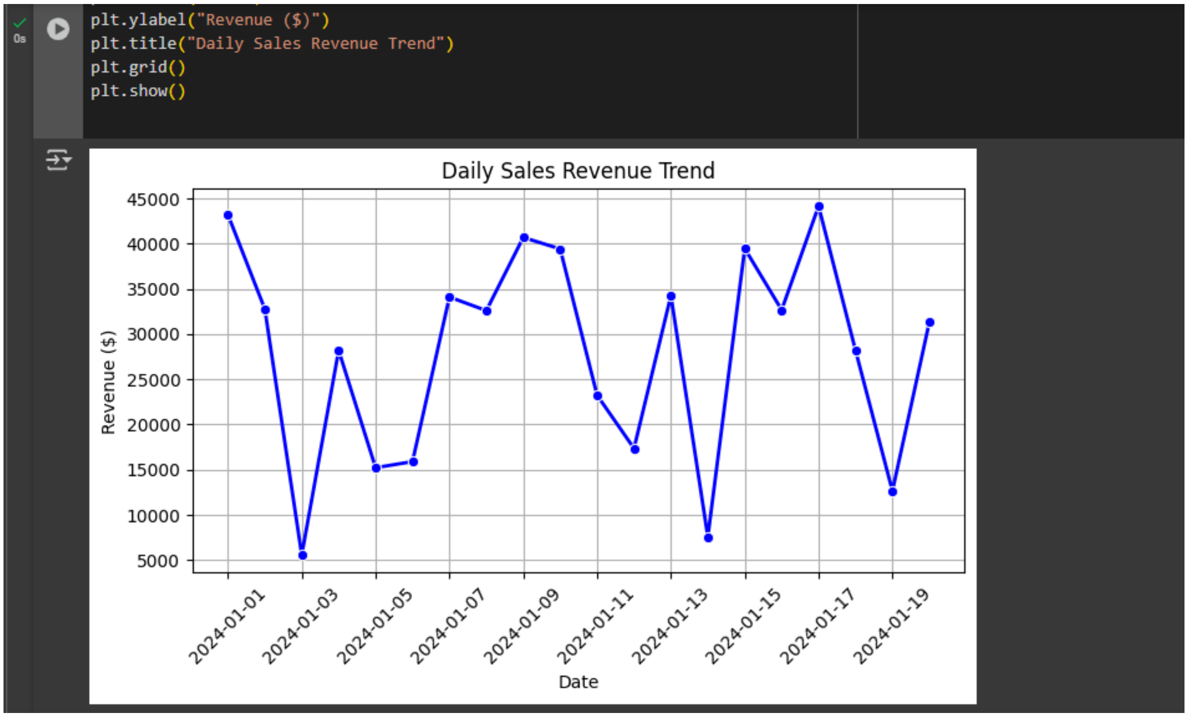
VISUALIZATION CODE:

```
#Visualization

# Plot daily sales revenue trend

plt.figure(figsize=(10, 5))

sns.lineplot(x=df['Date'], y=df['Revenue'], marker='o', color='b',
linewidth=2)



# Formatting

plt.xticks(rotation=45)

plt.xlabel("Date")

plt.ylabel("Revenue ($)")

plt.title("Daily Sales Revenue Trend")

plt.grid()

plt.show()



# Group by product and sum revenue

product_sales = df.groupby("Product")["Revenue"].sum()
```

```
# Bar Chart for Product Revenue

plt.figure(figsize=(8, 4))

product_sales.plot(kind='bar', color=['red', 'blue', 'green', 'purple'])


# Formatting

plt.xlabel("Product")

plt.ylabel("Total Revenue ($)")

plt.title("Total Revenue by Product")

plt.show()


# Pie Chart for Product Revenue Distribution

plt.figure(figsize=(6, 6))

plt.pie(product_sales, labels=product_sales.index, autopct='%1.1f%%',
colors=['lightcoral', 'lightblue', 'lightgreen', 'purple', 'orange'])
```

INSIGHTS AND REPORTS

Sales Trend Over Time: A clear visualization of daily sales trends.



```
plt.ylabel("Revenue ($)")
plt.title("Daily Sales Revenue Trend")
plt.grid()
plt.show()
```

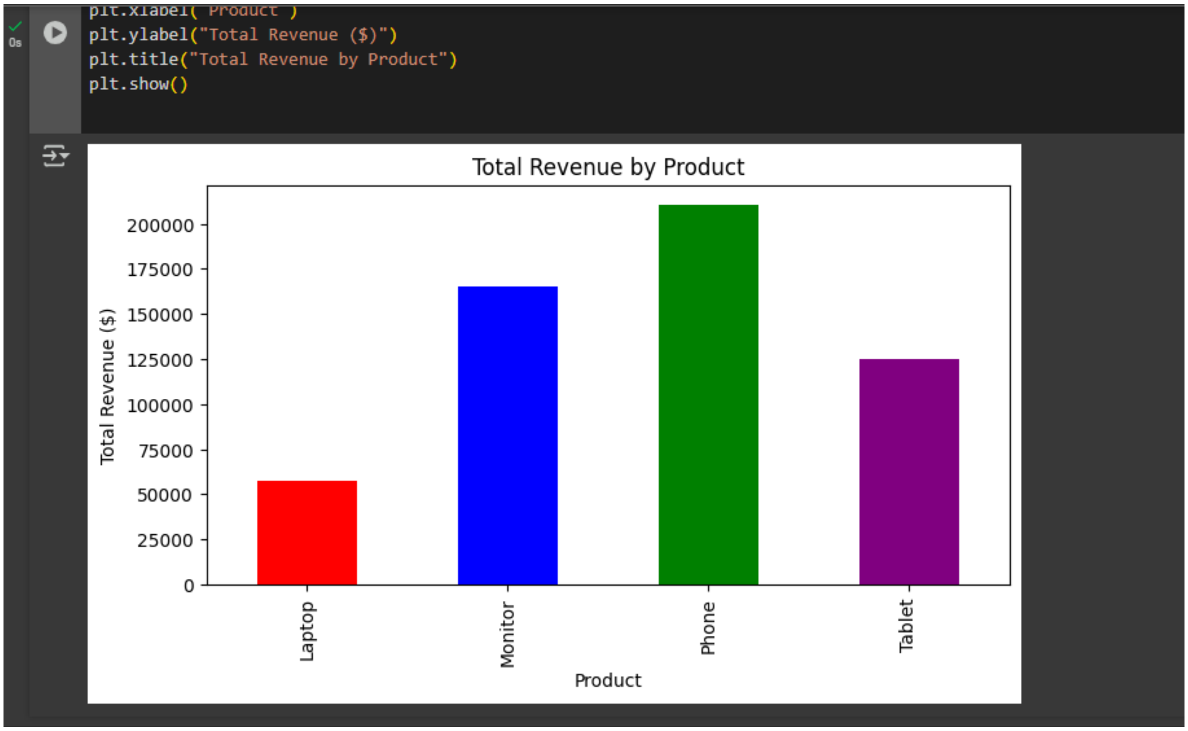
The graph shown is a line chart depicting the Daily Sales Revenue Trend over a period of time in January 2024. Below is a breakdown of its key components:

1. X-Axis (Date): Represents the dates in YYYY-MM-DD format, indicating the time series of sales revenue. The graph spans from January 1, 2024, to January 19, 2024, showing daily fluctuations in sales revenue.
2. Y-Axis (Revenue in $): Represents the daily sales revenue. The revenue ranges approximately from $5,000 to $45,000, highlighting significant variation in daily earnings.
3. Trend Analysis: The sales revenue experiences frequent fluctuations, indicating that sales are inconsistent across days. There are multiple peaks and dips, suggesting high and low sales periods. Notable Highs & Lows:
   - Highest revenue (~$45,000) on January 1 & January 16.
   - Sharp drop (~$5,000) on January 3 & January 18.
   - This could indicate factors such as demand fluctuations, marketing campaigns, seasonal effects, or external influences on sales.
4. Visualization Features: The blue line with markers helps in identifying individual sales data points clearly. Grid lines aid in better readability. Title & Labels enhance clarity, making it easier to interpret the trends.

Insights from the Graph:

- There is high volatility in daily sales, suggesting a need to investigate the causes of these fluctuations.
- Possible factors affecting sales include weekend vs. weekday trends, promotional events, product availability, or seasonal demand shifts.
- Further analysis is required to determine what drives high revenue days and how to reduce losses on low revenue days.

Revenue Based on Products sold: Providing recommendations based on insights obtained from the analysis.



```
plt.xlabel("Product")
plt.ylabel("Total Revenue ($)")
plt.title("Total Revenue by Product")
plt.show()
```

The given graph is a bar chart representing Total Revenue by Product. Below is a detailed breakdown:

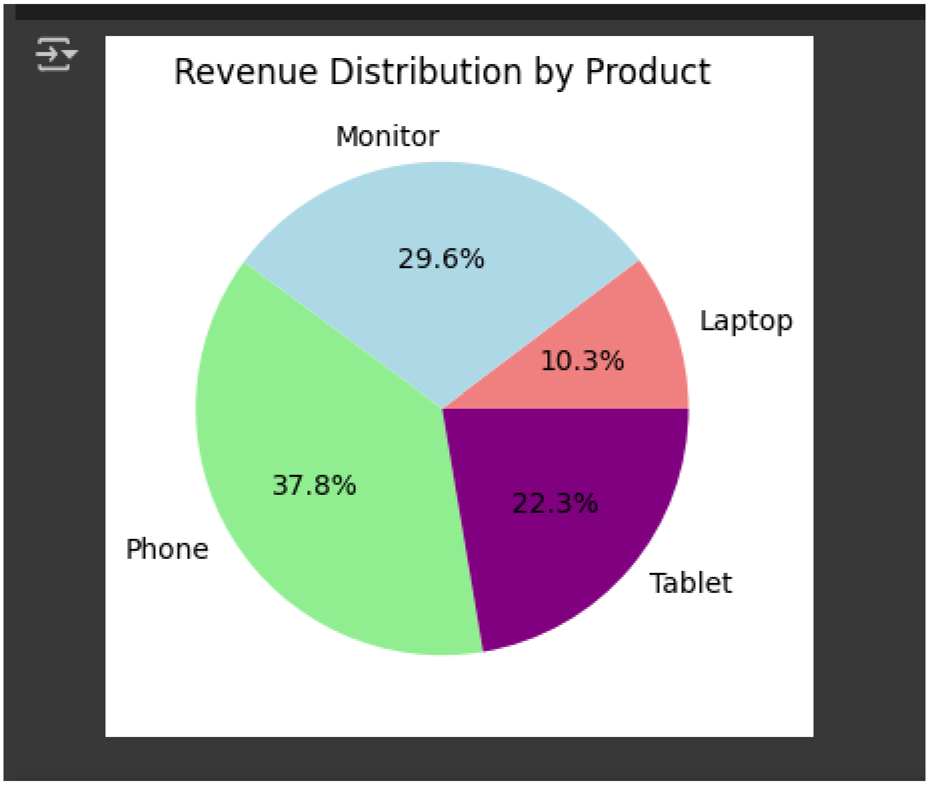
1. X-Axis (Product Categories): Displays four different product categories: Laptop, Monitor, Phone, and Tablet. Each bar corresponds to a specific product, showing its contribution to total revenue.
2. Y-Axis (Total Revenue in $): Represents the total revenue generated by each product. The range goes up to approximately $220,000, indicating the highest-earning product.

3. Key Observations:
    - **Phone** (Green Bar) generates the highest revenue (above $200,000), making it the most profitable product.
    - **Monitor** (Blue Bar) follows closely, contributing significantly to the total revenue.
    - **Tablet** (Purple Bar) has moderate revenue, lower than Phones and Monitors but still notable.
    - **Laptop** (Red Bar) generates the least revenue, indicating it is the lowest-selling or least profitable product.

Insights from the Graph:

- The Phone category is the primary revenue driver, suggesting a high demand or high-margin sales.
- Monitors and Tablets also contribute substantially, meaning they are essential for maintaining sales performance.
- Laptops have the lowest revenue, which may indicate lower demand, pricing issues, or the need for better marketing strategies.
- Businesses may consider promotional offers, pricing adjustments, or targeted marketing campaigns to improve Laptop sales.

Best-Selling Products: Identification of top-performing products and categories.



The given pie chart represents the Revenue Distribution by Product, visually breaking down the contribution of each product category to the total revenue.

1. Labels & Percentages:

- Phone (Green) – 37.8%: The highest revenue contributor, indicating strong sales performance.
- Monitor (Light Blue) – 29.6%: The second-largest share, showing a significant portion of the revenue.
- Tablet (Purple) – 22.3%: A moderate contributor to overall sales.
- Laptop (Red) – 10.3%: The least revenue-generating product in this dataset.

2. Business Implications:

- The company may focus more on Phone sales by increasing inventory, optimizing marketing, or launching new models.
- Monitors and Tablets can be further promoted to increase their revenue share.
- Laptops may require strategic improvements, such as pricing adjustments or bundling offers to boost sales.

Insights from the Graph:

- Phones dominate the revenue share, contributing more than one-third of total sales.
- Monitors and Tablets also play a major role, suggesting they are strong-selling product categories.
- Laptops have the lowest revenue contribution, which may indicate lower demand or higher competition in this segment.