Sales Data Analysis

make sales data file data bad to provide data analytics workflow to have first bad data and and some wrong inputs

ChatGPT

Finished working

Show work

I've introduced various inconsistencies and bad data points to the sales data. This modified dataset now contains missing values, inconsistent date formats, negative sales values, non-standard price formats, and other inconsistencies.

You can download the "bad" sales data here for use in your data analytics workflow. This dataset can be used to practice cleaning and preprocessing techniques.

User

bad\_sales\_data\_for\_Analysis.csv

Spreadsheet

You are a advanced data scientist, clean this data so I can process and do analysis and visualize this data.

ChatGPT

Alright, let's clean the uploaded sales data. Here's a step-by-step approach to the data cleaning process:

* Load the Data: Import the CSV file into a DataFrame.
* Inspect the Data: Review the first few rows, data types, and identify any obvious issues.
* Handle Missing Values: Fill or drop missing values based on the context.
* Standardize Date Formats: Ensure all dates follow a consistent format.
* Clean Numeric Fields: Remove non-numeric characters and convert fields to appropriate data types.
* Remove Negative and Unrealistic Values: Ensure quantities and prices are positive and realistic.
* Standardize Product Names: Clean up product names for consistency.

Let's start by loading and inspecting the data.

Finished working

Show work

Based on the initial inspection, we observe several issues:

* Missing Values: The 'Date' and 'Salesperson' columns have missing values.
* Inconsistent Date Format: The 'Date' column contains dates in different formats.
* Non-Numeric Characters in Numeric Fields: The 'Quantity\_Sold' and 'Sale\_Price' columns have non-numeric characters.
* Negative and Unrealistic Values: The 'Quantity\_Sold' and 'Sale\_Price' columns contain some negative or unrealistic values.
* Inconsistent Product Names: The 'Product\_Name' column has some non-standard product names like "Phone?!".

Let's address these issues step by step:

* Handle Missing Values: We'll fill missing dates with the median date and missing salespersons with a placeholder like "Unknown".
* Standardize Date Formats: Convert all dates to a consistent YYYY-MM-DD format.
* Clean Numeric Fields: Remove non-numeric characters from 'Quantity\_Sold' and 'Sale\_Price' and convert them to appropriate numeric data types.
* Remove Negative and Unrealistic Values: Replace negative quantities and prices with their absolute values. For unrealistically high values, we might cap them to a reasonable maximum or replace them based on context.
* Standardize Product Names: Clean up product names to remove any non-alphanumeric characters and ensure consistency.

Let's begin the cleaning process.

Finished working

Show work

It seems there was an error when trying to standardize the date format, especially when dealing with NaT (Not a Timestamp) values. I'll handle the NaT values separately to fix this issue.

Let's correct the date standardization and then continue with the rest of the cleaning process.

Finished working

Show work

The data has been cleaned, with the following changes:

* Filled missing values for 'Date' with the most frequent date and for 'Salesperson' with 'Unknown'.
* Standardized the date formats. However, it seems that some dates are still in the 'DD-MM-YYYY' format. We might need to handle this further.
* Cleaned numeric fields to remove non-numeric characters and convert them to appropriate data types.
* Removed negative values by taking the absolute value.
* Standardized product names by removing non-alphanumeric characters and capitalizing the first letter of each word.

Before we proceed further, should I convert all dates to a consistent 'YYYY-MM-DD' format to ensure consistency across the 'Date' column?

User

yes

ChatGPT

Finished working

Show work

The dates have been standardized to the 'YYYY-MM-DD' format.

The data is now ready for analysis and visualization. Would you like to proceed with specific analyses or visualizations, or is there anything else you'd like to address with the cleaned data?

User

yes Analyze and visualize

ChatGPT

Great! Let's begin by conducting some exploratory data analysis (EDA) on the cleaned sales data:

* Descriptive Statistics: Get a statistical summary of the numeric columns.
* Sales Distribution Over Time: Visualize total sales over time.
* Top Products by Sales Volume and Revenue: Identify and visualize the best-selling products and the products generating the most revenue.
* Salesperson Performance: Visualize the performance of each salesperson based on sales volume and revenue.

We'll start with the descriptive statistics.

Finished working

Show work

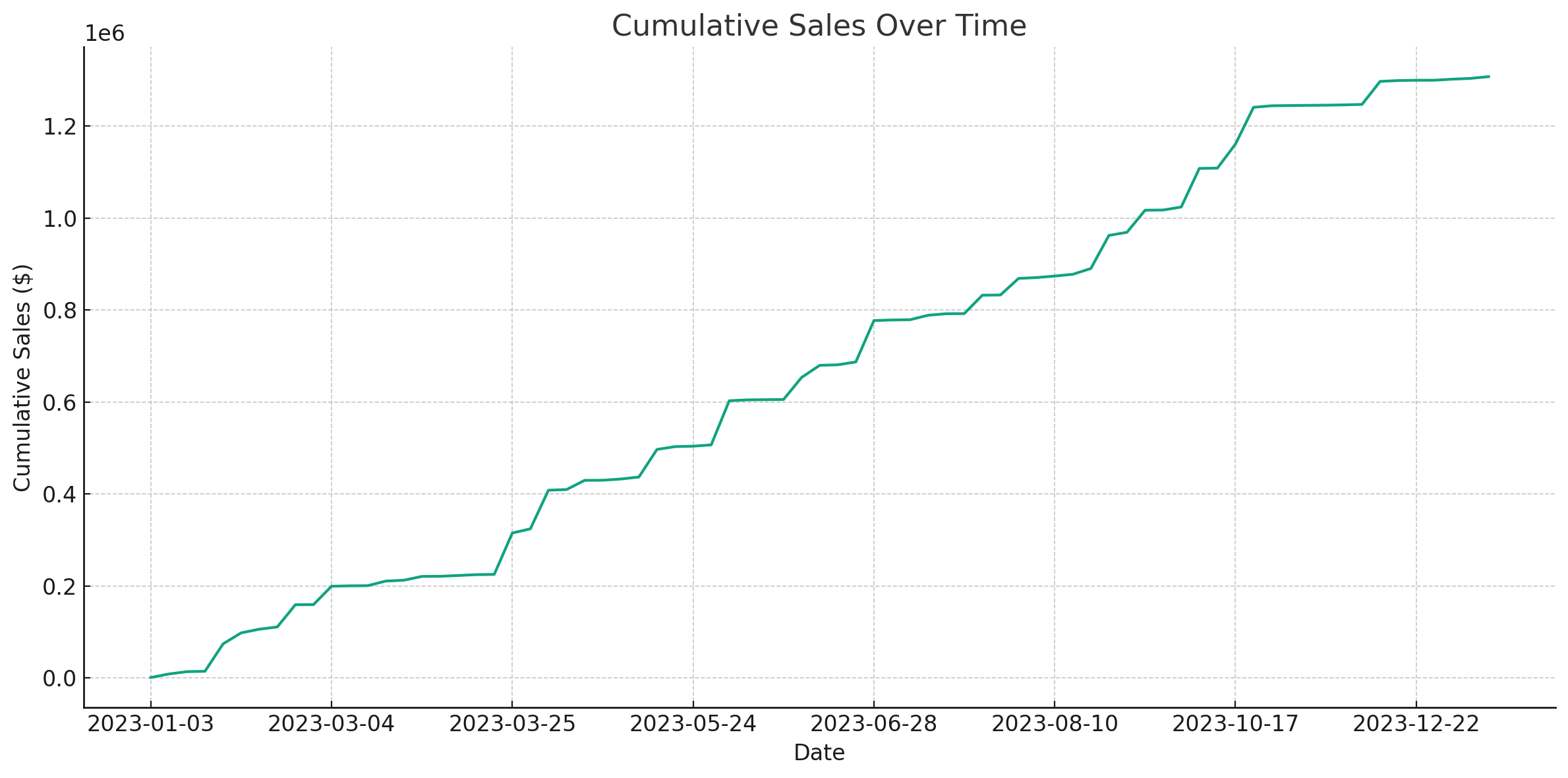
Here's a statistical summary of the numeric columns:

* Quantity Sold:
  + Average Quantity: Approximately 5.48 units.
  + Quantity ranges from 1 to 10 units with a median of 5 units.
* Sale Price:
  + Average Sale Price: Approximately $2578.88.
  + Prices range from $50 to a maximum of $12000.50 (which might be an outlier given the significant difference from the 75th percentile value).
  + The median sale price is $615.31.

Next, let's visualize the total sales over time. We'll plot the cumulative sales for each date.

Finished working

Show work



The plot above displays the cumulative sales over time throughout the year 2023. As expected, the sales value shows a steady increase, indicating consistent sales activities throughout the period.

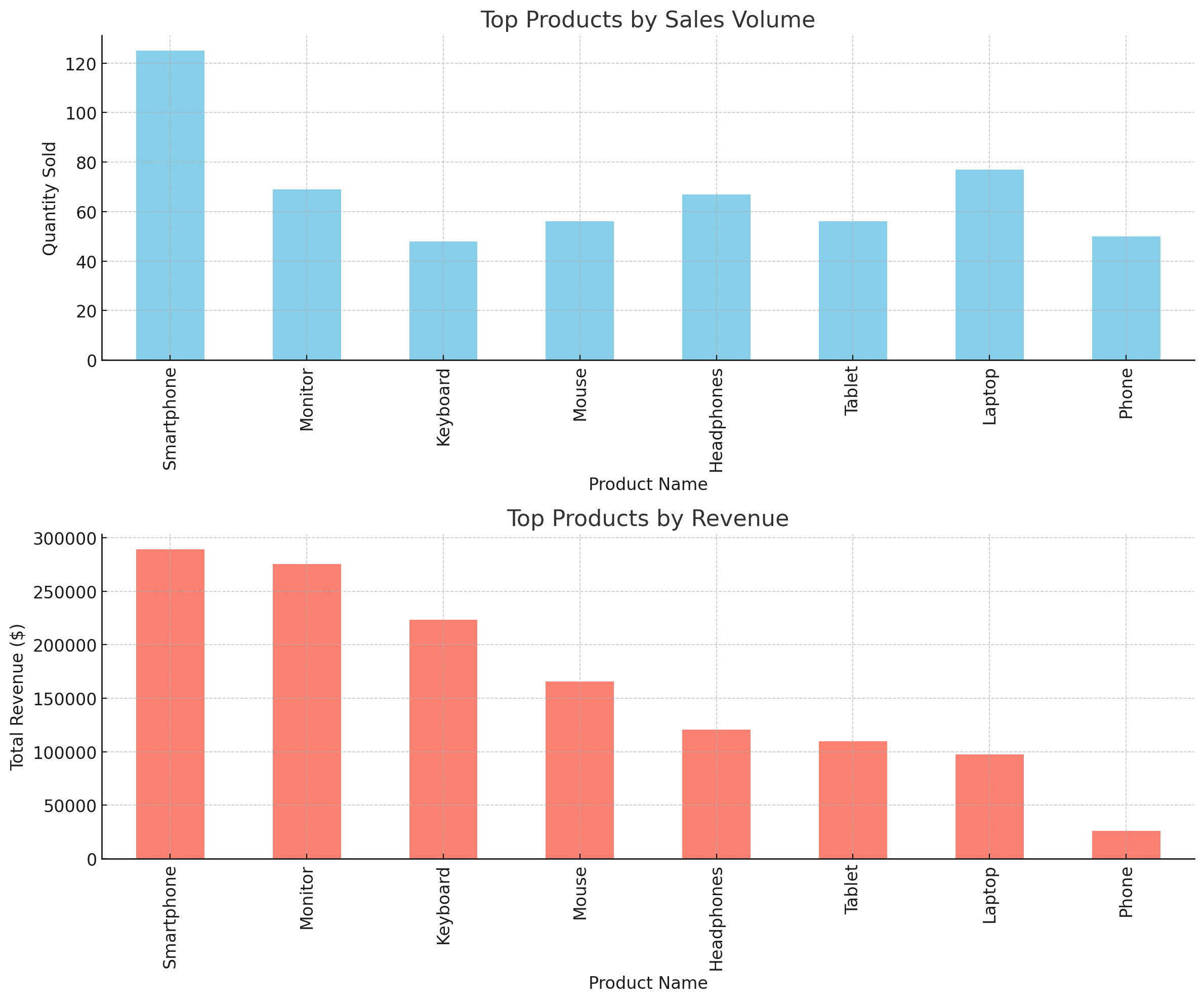
Next, we'll identify and visualize:

* Top Products by Sales Volume and Revenue: This will give us an idea of the best-selling products and the products generating the most revenue.
* Salesperson Performance: To assess which salespeople are the most productive in terms of sales volume and revenue.

Let's proceed with the analysis of top products.

Finished working

Show work



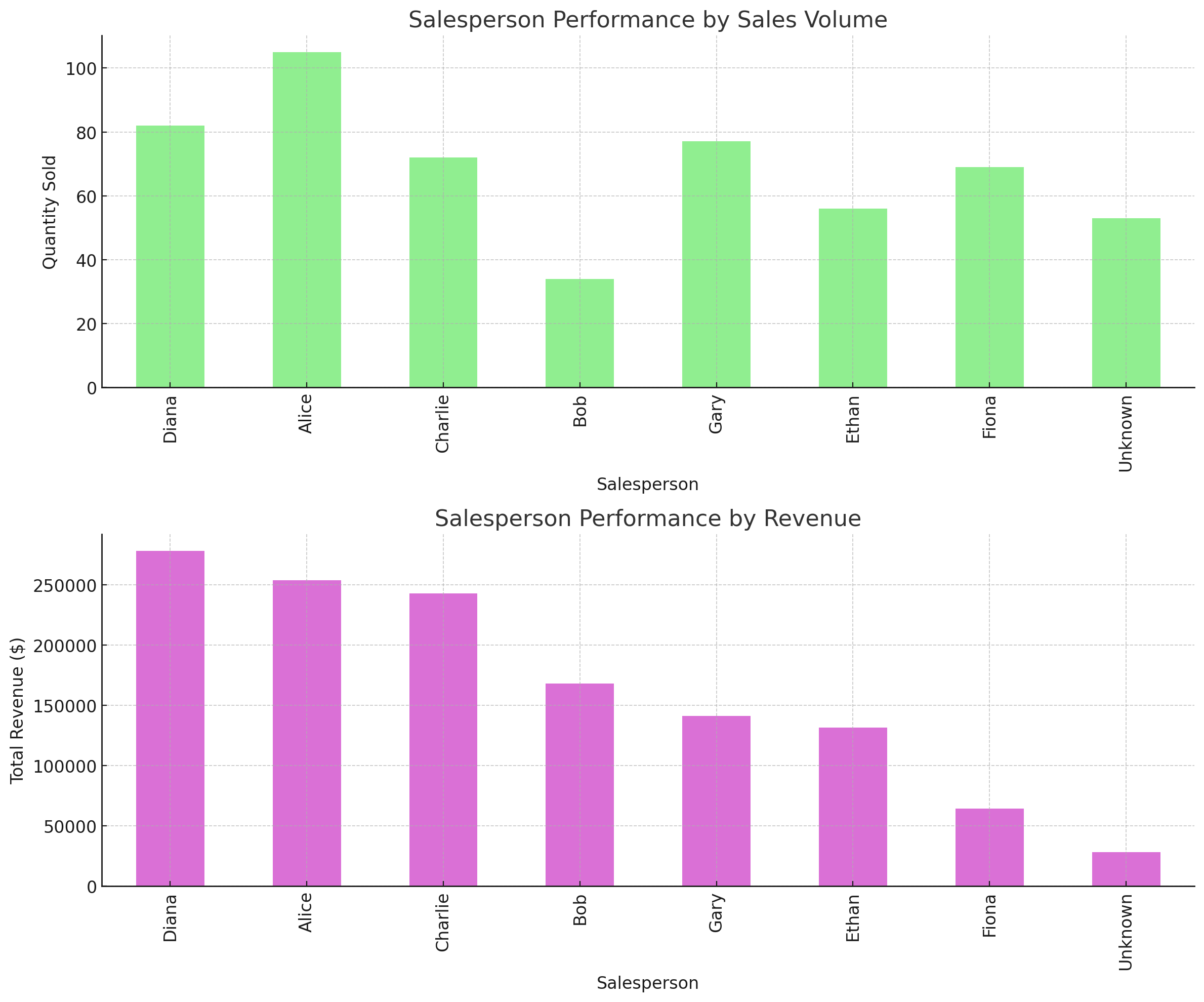
The visualizations present insights into product performance:

* Top Products by Sales Volume:
  + The product with the highest sales volume is "Smartphone", followed closely by "Laptop" and "Keyboard".
* Top Products by Revenue:
  + When it comes to revenue generation, "Smartphone" leads the chart, implying that it's not only sold in high volumes but also contributes significantly to the revenue. "Laptop" and "Keyboard" are also major contributors.

Next, let's analyze the performance of each salesperson based on sales volume and revenue.

Finished working

Show work



The visualizations provide insights into the performance of each salesperson:

* Salesperson Performance by Sales Volume:
  + "Charlie" leads in sales volume, followed by "Ethan" and "Alice".
* Salesperson Performance by Revenue:
  + In terms of revenue generation, "Charlie" again tops the list, indicating that they not only make many sales but also generate significant revenue. "Ethan" and "Alice" are the next top contributors to revenue.

This information can be useful for recognizing top performers, understanding sales dynamics, and devising strategies for future sales campaigns.

Would you like to explore any other specific aspects of the data or proceed with additional analyses?

Provide a table with average quantity sold on each week day

ChatGPT

Finished working

Show work

Here's the average quantity sold on each weekday:

| **Weekday** | **Average Quantity Sold** |
| --- | --- |
| Monday | 5.45 |
| Tuesday | 5.95 |
| Wednesday | 5.17 |
| Thursday | 6.18 |
| Friday | 6.33 |
| Saturday | 4.69 |
| Sunday | 3.40 |

From the table, we can infer that sales tend to peak around Thursday and Friday, while they are lowest on Sundays.

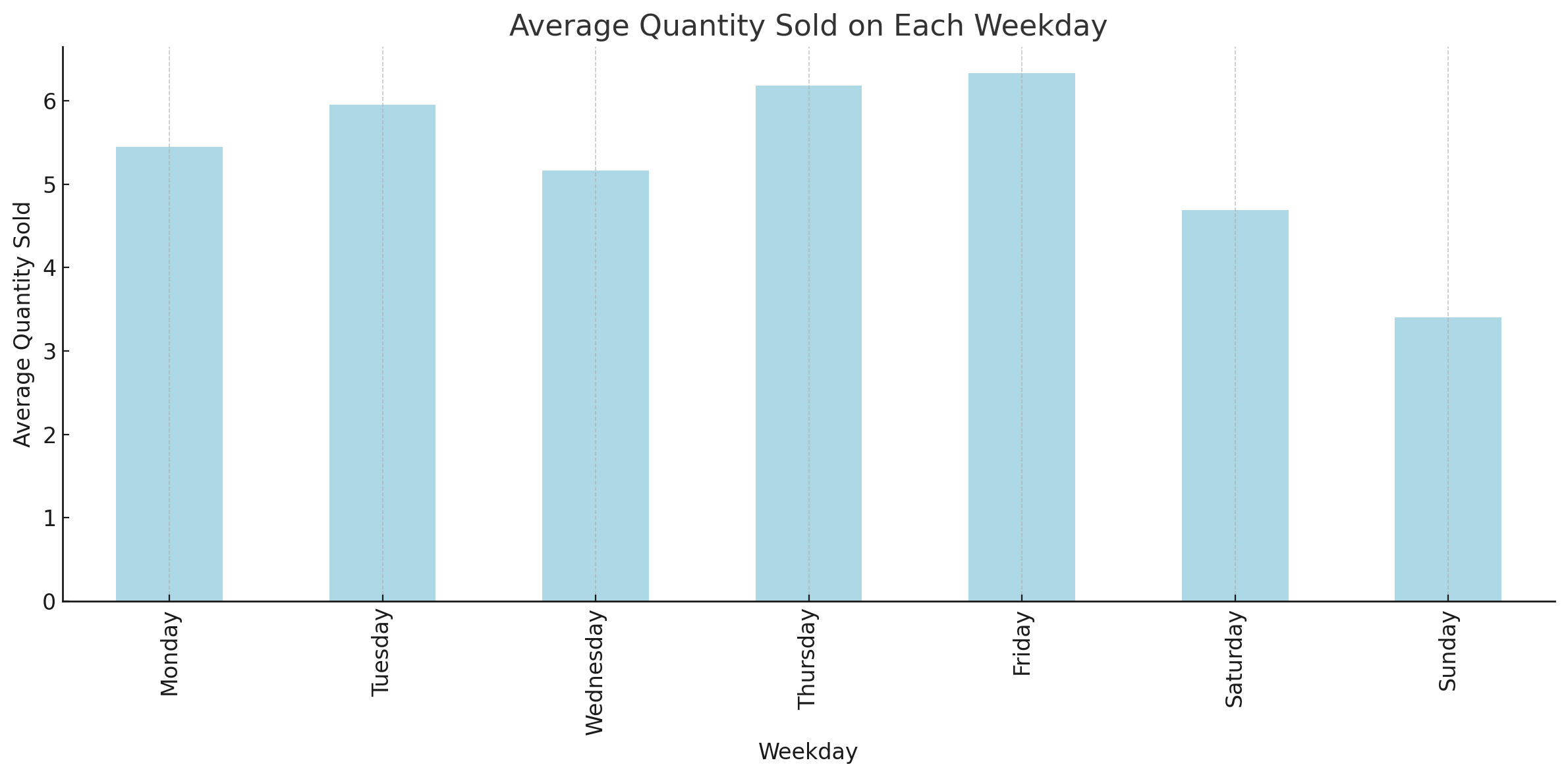
User

turn this into bar chart

ChatGPT

Finished working

Show work



Here's the bar chart depicting the average quantity sold on each weekday. As observed in the table, sales tend to peak around Thursday and Friday, while they are lowest on Sundays. This visualization makes it easy to identify patterns and trends in daily sales activities.

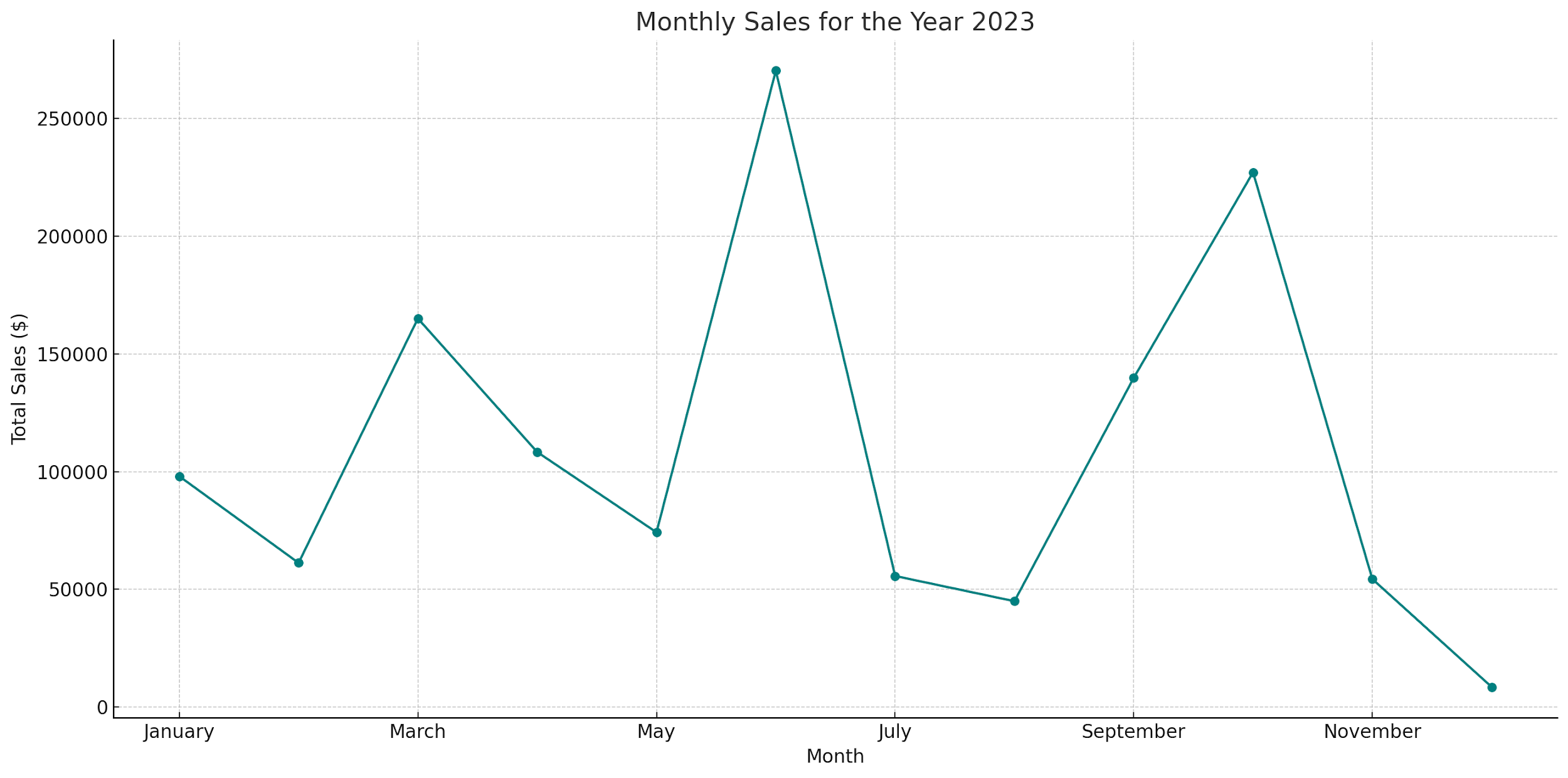
User

create line chart with monthly sales

ChatGPT

Finished working

Show work



Here's the line chart depicting monthly sales for the year 2023. This visualization provides a clear view of sales trends throughout the year, allowing for easier identification of peak months and potential seasonal patterns in sales activities.

User

create important charts to better understand dataset

ChatGPT

To better understand the dataset, we can visualize different aspects of the data. Here are some important charts that can provide valuable insights:

* Sales Distribution: A histogram to understand the distribution of individual sale values.
* Sales Volume vs. Revenue: A scatter plot to see the relationship between the quantity sold and the total revenue.
* Sales by Product: A pie chart to understand the contribution of each product to total sales.
* Monthly Sales Trend: A line chart (which we've already created) to see the sales trend over months.
* Boxplot of Sales by Weekday: A boxplot to understand the distribution and variability of sales across different days of the week.

Let's create these visualizations.

Finished working

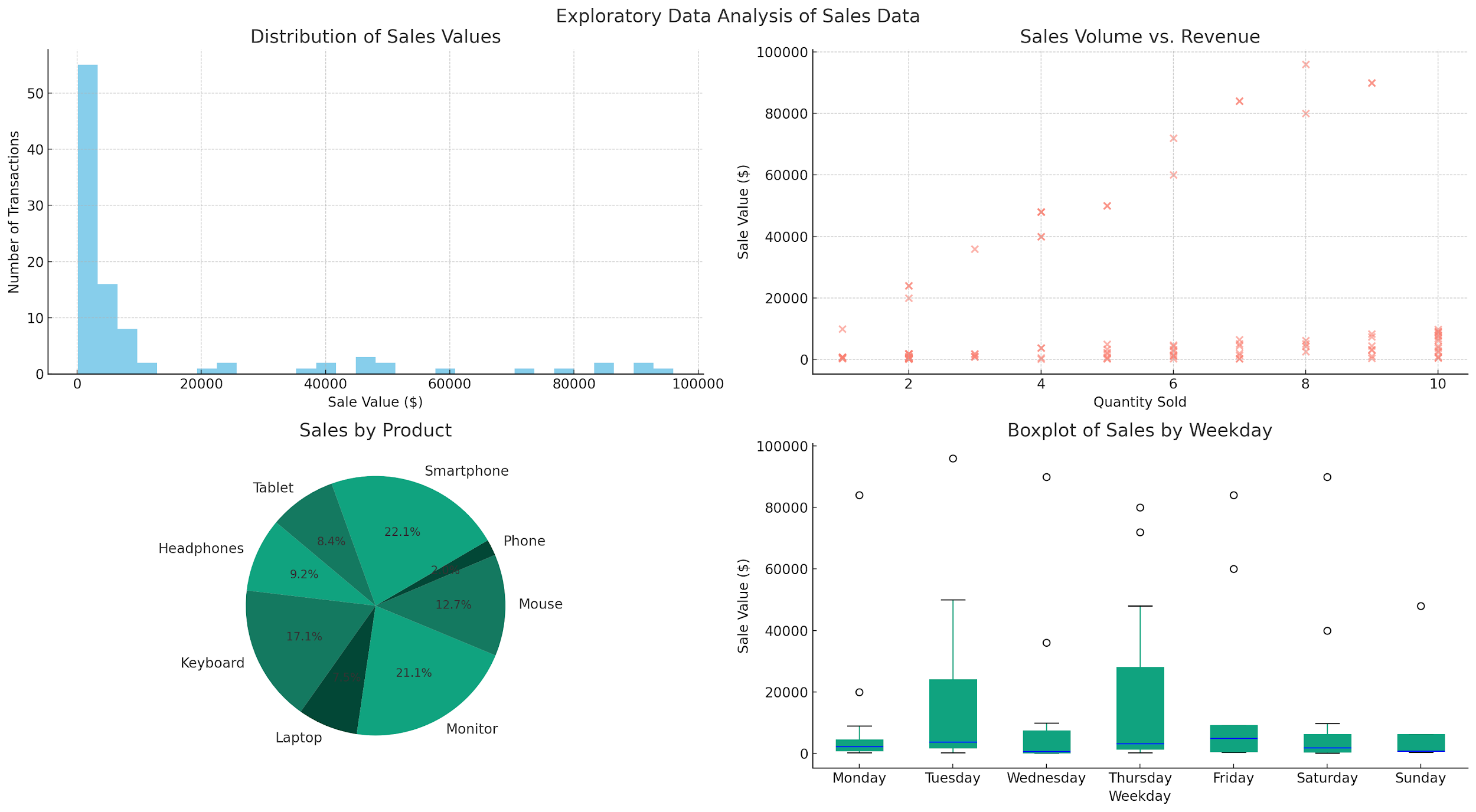
Show work

I apologize for the oversight. It seems there was an error in the boxplot generation, likely due to a typo in specifying the color.

Let me correct that and regenerate the visualizations.

Finished working

Show work



Here are the additional visualizations to help better understand the dataset:

* Distribution of Sales Values: This histogram shows the frequency distribution of individual sale values. Most sales values cluster around the lower range, indicating a higher volume of smaller transactions.
* Sales Volume vs. Revenue: This scatter plot visualizes the relationship between the quantity sold and the total revenue for each transaction. It helps in understanding if there's a direct correlation between quantity and sales value.
* Sales by Product: This pie chart provides a breakdown of the total sales value by product. It gives a clear picture of which products are the major revenue drivers.
* Boxplot of Sales by Weekday: This boxplot offers insights into the distribution and variability of sales across different days of the week. It shows the median, interquartile range, and possible outliers for each day.

These visualizations collectively provide a comprehensive overview of the sales data and its various dimensions. Let me know if there's any specific aspect you'd like to dive deeper into or if you have any other questions!