Sales Data Analytics Project

Submitted by: Disha Sharma

Tools Used: Python, Pandas, NumPy, Matplotlib
Platform: Jupyter Notebook

Abstract

This project focuses on performing Exploratory Data Analysis (EDA) on a retail sales dataset to uncover meaningful business insights. Using Python and data analysis libraries such as Pandas and Matplotlib, the dataset was cleaned, preprocessed, and analyzed to understand sales trends, product performance, and regional sales distribution.

The analysis includes category-wise sales comparison, year-wise sales trend evaluation, identification of top-performing sub-categories, and region-wise sales analysis. Visualizations were used to support the findings and enhance interpretability.

The results show that the Technology category generates the highest revenue, sales increased significantly after 2016, and the West region leads in overall sales. This project demonstrates how data analytics techniques can be applied to real-world datasets to support data-driven business decision-making.

1. Introduction

Data analytics plays a vital role in modern businesses by enabling organizations to analyze historical data and derive actionable insights. Sales data analysis helps businesses understand customer preferences, product demand, and regional performance, which are essential for strategic planning and growth.

Exploratory Data Analysis (EDA) is an important step in the data analytics process. It involves summarizing datasets, identifying patterns, detecting anomalies, and visualizing relationships within the data. EDA helps analysts gain a deeper understanding of the dataset before applying advanced analytical or predictive models.

The objective of this project is to analyze a retail sales dataset using Python-based data analytics techniques to identify sales trends, top-performing product categories and sub-categories, and regional sales performance.

2. Dataset Description

The dataset used in this project contains retail sales transaction data collected over multiple years. It consists of **9,800 records** and **18 attributes**.

Key Attributes

- Order Date
- Ship Date
- Sales
- Category
- Sub-Category
- Region
- Customer Segment
- Product Name

The dataset represents sales across different product categories and geographical regions, making it suitable for exploratory sales analysis.

3. Tools and Technologies Used

The following tools and technologies were used in this project:

- **Python** – programming language used for analysis
- **Pandas** – data manipulation and preprocessing
- **NumPy** – numerical operations
- **Matplotlib** – data visualization
- **Jupyter Notebook** – interactive development environment

4. Methodology

The project followed a structured data analytics workflow:

4.1 Data Loading

The dataset was loaded into the Jupyter Notebook using the Pandas library. Encoding and parsing issues were handled to ensure successful data ingestion.

4.2 Data Cleaning

- Missing values in the Postal Code column were handled.
- Date columns were converted into proper datetime format.
- Mixed date formats were resolved.
- Duplicate records were removed to ensure data consistency.

4.3 Feature Engineering

New features were created from the Order Date column:

- Year
- Month

These features enabled time-based analysis of sales trends.

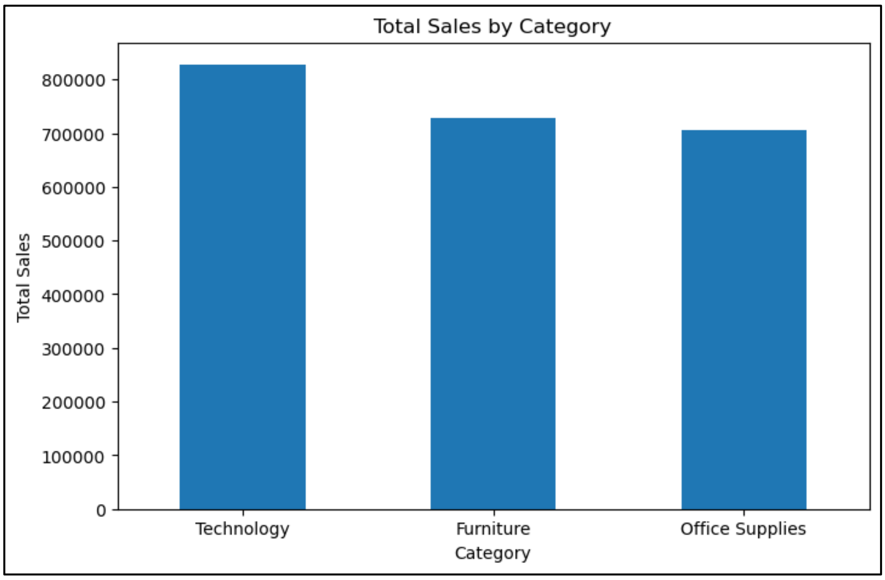
4.4 Exploratory Data Analysis

Exploratory analysis was conducted to:

- Compare sales across categories
- Analyze sales trends over time
- Identify top-performing sub-categories
- Evaluate regional sales performance

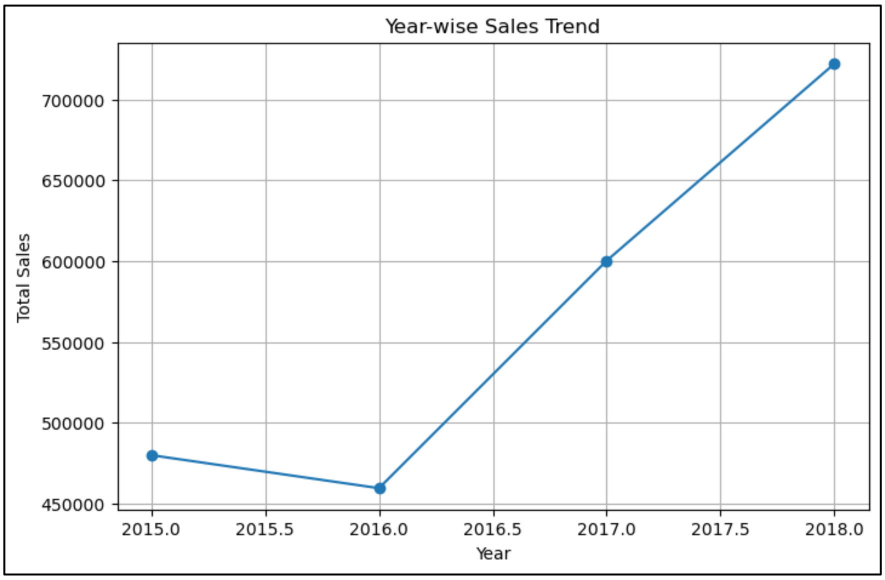
5. Exploratory Data Analysis and Visualizations

5.1 Category-wise Sales Analysis



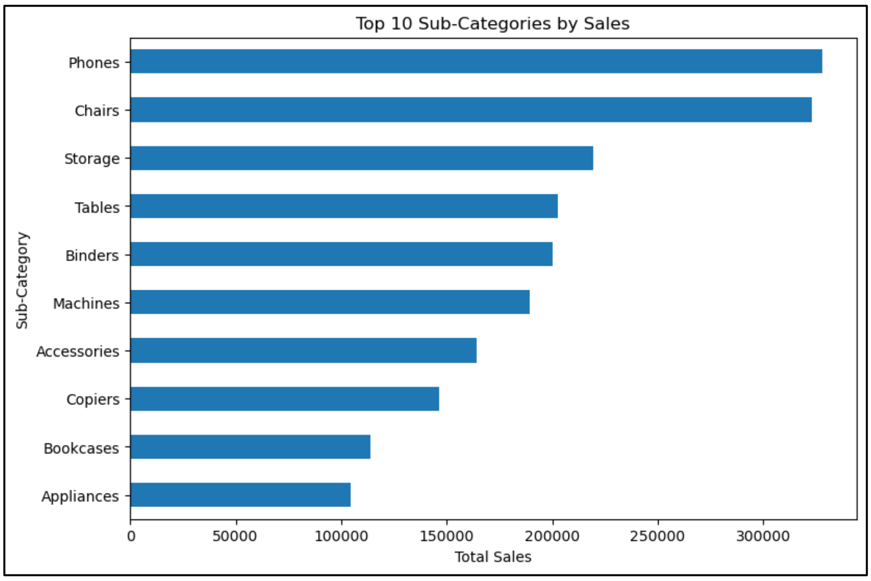
Explanation:
The analysis compares total sales across different product categories. The Technology category generates the highest revenue, followed by Furniture and Office Supplies. This indicates strong demand for technology-related products and their significant contribution to overall sales.

5.2 Year-wise Sales Trend



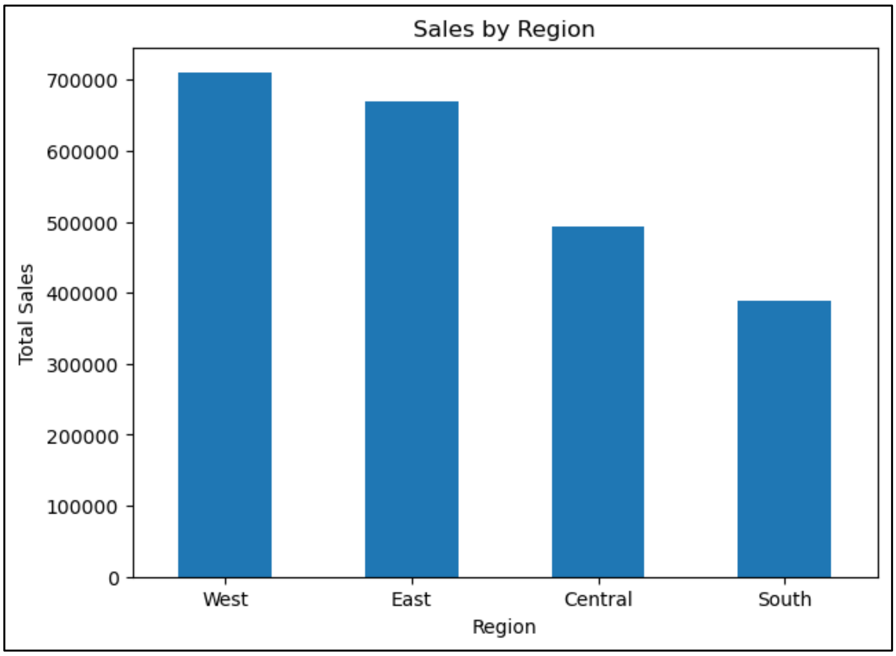
Explanation:
The year-wise sales trend shows a slight decline between 2015 and 2016. However, from 2016 onwards, sales increased steadily, reaching a peak in 2018. This trend suggests improved business performance and growth over time.

5.3 Top 10 Sub-Categories by Sales



Explanation:
Phones and Chairs are identified as the top-performing sub-categories in terms of total sales. Technology-related sub-categories dominate the top 10 list, reinforcing the importance of the Technology segment in driving revenue.

5.4 Region-wise Sales Analysis



Explanation:
The West region contributes the highest total sales, followed by the East region. The Central region shows moderate performance, while the South region records the lowest sales. This analysis highlights regional disparities and potential opportunities for expansion.

6. Key Findings

- Technology is the highest revenue-generating product category.
- Sales show strong growth from 2016 to 2018.
- Phones and Chairs are the top-performing sub-categories.
- The West region contributes the largest share of total sales.
- The South region presents potential for future growth.

7. Conclusion

This project demonstrates how exploratory data analysis can be used to derive valuable business insights from retail sales data. By analyzing sales patterns across categories, time periods, and regions, the project highlights areas of strong performance and opportunities for improvement. The findings can support strategic decision-making related to product focus, regional expansion, and sales optimization.

8. Future Scope

The project can be extended in the following ways:

- Sales forecasting using machine learning models
- Profit and discount analysis
- Customer segmentation and behavior analysis
- Dashboard development using Power BI or Tableau