DATA SCIENCE TOOLBOX: PYTHON PROGRAMMING PROJECT REPORT

(Project Semester January-April 2025)

CLOTHING STORES DATA

Submitted by

Name: Shankar Sai Krishna Allumolu Registration No: 12312843

Programme and Section: Data Science, K23EU

Course Code**:** INT375

Under the Guidance of

Assistant Professor: Dr.Tanima Thakur(UID:23532)

Discipline of CSE/IT

**Lovely School of Computer Science**

**Lovely Professional University, Phagwara**

****

# CERTIFICATE

This is to certify that Shankar Allumolu bearing Registration no. 12312843 has completed INT375 project titled, “Clothing Stores Data” under my guidance and supervision. To the best of my knowledge, the present work is the result of her original development, effort and study.

Signature and Name of the Supervisor Designation of the Supervisor

School of Computer Science Lovely Professional University Phagwara, Punjab.

Date:13-04-2025

# DECLARATION

I, Shankar Allumolu, student of Data Science, under CSE/IT Discipline at, Lovely Professional University, Punjab, hereby declare that all the information furnished in this project report is based on my own intensive work and is genuine.

Date: 13-04-2025

Registration No. 12312843 Signature



# ACKNOWLEDGEMENT

I would like to express my heartfelt gratitude to my project guide, **Tanima Thakur**, for her invaluable guidance,constructive feedback, and constant encouragement throughout this project, *“Sales and Profit Analysis of Clothing Stores using EDA”*.I am sincerely thankful to the **Department of Computer Science and Engineering, Lovely Professional University**, for providing the necessary academic support, tools, and resources that enabled me to complete this project successfully. I would also like to acknowledge the creators of the dataset, whose publicly available data made this analysis possible and insightful. A special thanks to my peers for their meaningful discussions and collaborative spirit, which greatly enhanced my learning process. Lastly, I express my deepest appreciation to my family and friends for their continuous motivation, understanding, and unwavering support during the course of this project.

Name: Shankar Allumolu

Registration number: 12312843

**Table of Contents**

1. Introduction
2. Source of Dataset
3. EDA Process
4. Analysis on Dataset
   1. Introduction
   2. General Description
   3. Specific Requirements, Functions, and Formulas
   4. Analysis Results
   5. Visualization
5. Conclusion
6. Future Scope
7. References

# INTRODUCTION

In the competitive landscape of modern retail, data-driven decision-making has become essential for sustained business growth and customer satisfaction. The clothing and apparel industry, being one of the most dynamic sectors, experiences continuous changes in consumer preferences, seasonal demand patterns, and regional buying behavior. Retailers are increasingly relying on data analytics to optimize inventory, forecast trends, and enhance profitability.

The study of sales and profit trends within clothing stores offers valuable insights to business analysts, marketing strategists, and inventory managers. Understanding which product categories perform best, which regions drive the most revenue, and how discount strategies influence profitability can guide critical business decisions. With access to detailed transactional data including order dates, product categories, profit margins, and regional sales distribution, companies can craft more effective, targeted strategies.

This project investigates a dataset containing clothing store sales information, focusing on exploratory data analysis (EDA) to uncover trends and patterns in sales, profit, and product performance. It aims to answer key questions such as: Which regions contribute most to the business? What are the top-performing products? How does discounting impact profitability? What seasonal or monthly sales trends exist?

Using a combination of data cleaning, statistical exploration, and data visualization tools, this report transforms raw transactional data into meaningful business intelligence. The visual interpretation of data through charts and graphs enhances understanding and allows for intuitive insights even when analyzing complex relationships.

The significance of this analysis extends beyond academic interest. It provides practical implications for retail decision-makers looking to boost operational efficiency, align marketing campaigns with consumer behavior, and strategically plan for peak seasons. As the retail sector becomes more digital and data-centric, such insights play a pivotal role in driving revenue and enhancing customer experience.

This report is organized to cover the methodology, data processing steps, visual analysis, and business-oriented conclusions. By leveraging the power of Python libraries like Pandas, Matplotlib, and Seaborn, this project demonstrates how exploratory analysis can serve as the foundation for smarter retail strategies.

**Source of Dataset**

* File Name: Clothing\_Stores\_Orders
* Source: <https://docs.google.com/spreadsheets/d/1aAFTdAJOmoNW6uf6tSCQwje7T-dTVF41/edit?usp=drive_link&ouid=102635536571206128877&rtpof=true&sd=true>

**EDA Process (Exploratory Data Analysis Process)**

Exploratory Data Analysis (EDA) is an essential phase in any data science project, offering critical insights into the structure, distribution, and key characteristics of a dataset. In this project, EDA was performed using Python, with the help of libraries such as **pandas** for data manipulation and **matplotlib** and **seaborn** for visualization. The dataset under analysis comprises transactional records of a clothing retail business, including information on order dates, sales, discounts, product categories, profit margins, and regional distribution.

The EDA for this project was conducted in a systematic manner through the following steps:

**3.1. Data Collection**

The dataset used for this analysis was obtained from a simulated retail sales record system representing a clothing store. It consists of detailed order-level data, covering multiple attributes such as order ID, order date, product name, category, region, sales, quantity, discount, and profit. The goal of analyzing this dataset is to derive actionable business insights that can help understand product performance, customer buying behavior, and profitability trends across various regions.

**3.2. Data Cleaning and Preparation**

Before initiating the analytical phase, the dataset underwent extensive cleaning and preparation. This step was crucial to ensure the quality and reliability of the insights derived. The following tasks were performed:

* **Handling Missing Values**: All missing entries were identified and replaced with default values (such as 0) or flagged where necessary. Particular attention was paid to fields like Order Date, Sales, and Profit.
* **Removing Duplicates**: Duplicate rows that could distort the results were detected and eliminated.
* **Date Formatting**: The Order Date field was converted into datetime format to extract time-based features like Month and Year.
* **Feature Engineering**: Additional features such as Month and Year were derived from the order date to enable time-series analysis.
* **Data Type Conversion**: Categorical and numerical columns were formatted correctly to support sorting, grouping, and aggregation operations.

These cleaning procedures ensured the dataset was accurate, consistent, and ready for effective exploration and visualization.

**3.3. Data Exploration and Visualization**

With the dataset preprocessed, a variety of visual analyses were conducted to uncover hidden patterns and trends:

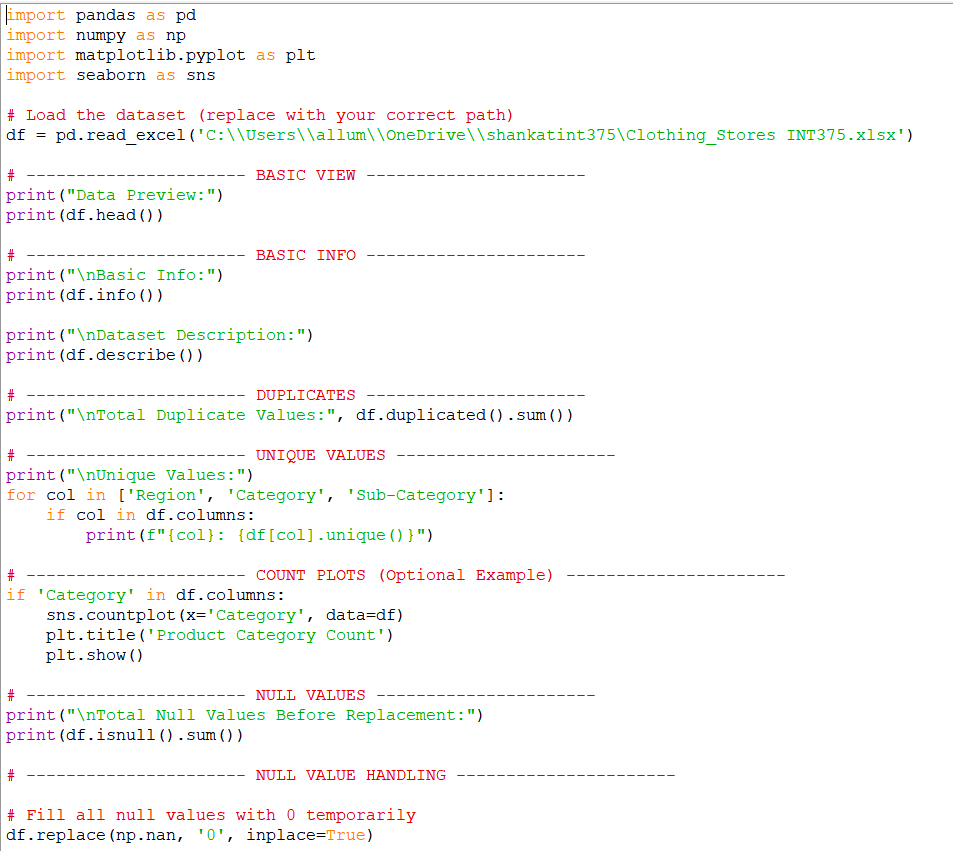
* **Line Plot**: Created to illustrate monthly sales trends across different years, helping identify seasonal demand or spikes.
* **Bar Charts**: Used to analyze total sales and profit distribution by region and to highlight the top-selling products.
* **Pie Chart**: Displayed the contribution of each product category to overall sales, offering a quick understanding of category performance.
* **Box Plot**: Showcased the distribution of profit margins across different product categories, revealing outliers and variations in earnings.
* **Heatmap**: Correlation heatmap plotted to examine relationships among numerical features like Sales, Profit, Discount, and Quantity.

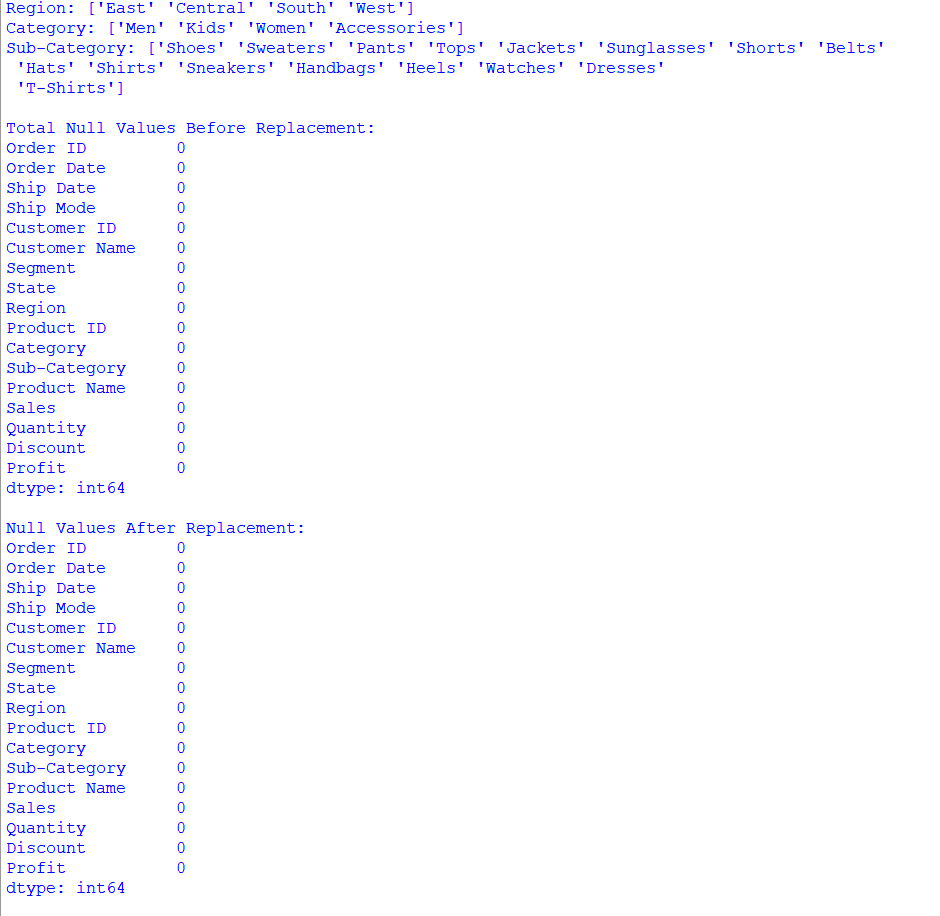
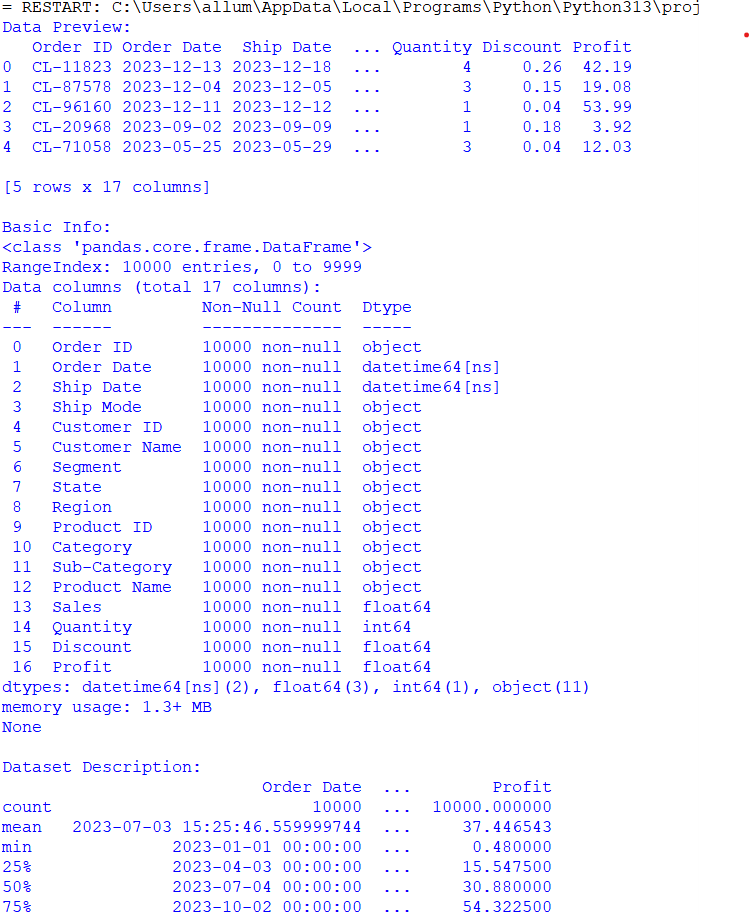
These visualizations played a vital role in understanding customer preferences, evaluating product and region-level performance, and identifying potential areas for strategic improvements.

**Conclusion of EDA**

The EDA phase provided significant insights into sales performance, consumer buying trends, and the effectiveness of discount strategies. It also helped identify high-performing regions and product lines. These insights lay a strong foundation for more advanced analytics, such as predictive modeling, demand forecasting, and targeted marketing strategies.

By combining rigorous data cleaning with visually intuitive analysis, this project showcases how data science can transform raw sales data into meaningful business intelligence that drives decision-making in the retail sector.





# 4. ANALYSIS ON DATASET

# Objectives:

# 1.create a line chart to show Monthly Sales Trend

# 2. Create a Bar plot to show Sales and Profit by Region

# 3. Pie chart Category-wise Sales

# 4. Create a Box Plot to show Profit by Category

# 5. Top 5 Products by Sales using bar chart

# 6. Heatmap of correlation b/w sales, profit, discount, quantity

This section presents key findings drawn from the exploratory data analysis and visualizations conducted on the Clothing Store dataset. The analysis includes insights into monthly sales trends, regional performance in terms of both sales and profit, category-wise contributions, and the impact of discounts on profitability. Furthermore, it explores customer purchasing behavior by identifying top-performing products and visual correlations between key business metrics. Through a combination of statistical summaries and graphical representations, this study provides actionable insights to support better decision-making in marketing, inventory management, and sales strategies.

**4.1 Line Chart – Monthly Sales Trend**

**4.1.1 Introduction**

This line chart illustrates the overall monthly sales performance across different years, providing insights into business seasonality and long-term growth.

**4.1.2 General Description**

The data was grouped by both year and month to calculate the total monthly sales. The results were plotted with a time series line to observe trends.

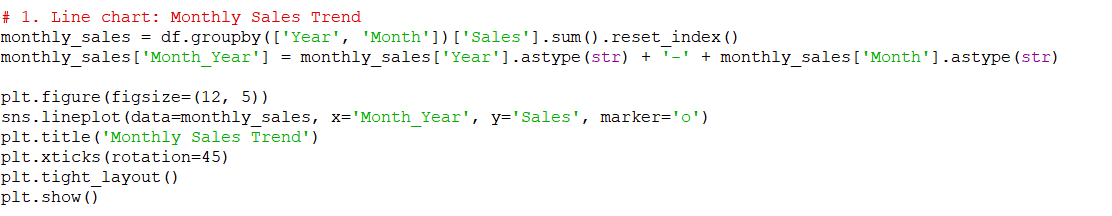
**4.1.3 Specific Requirements**

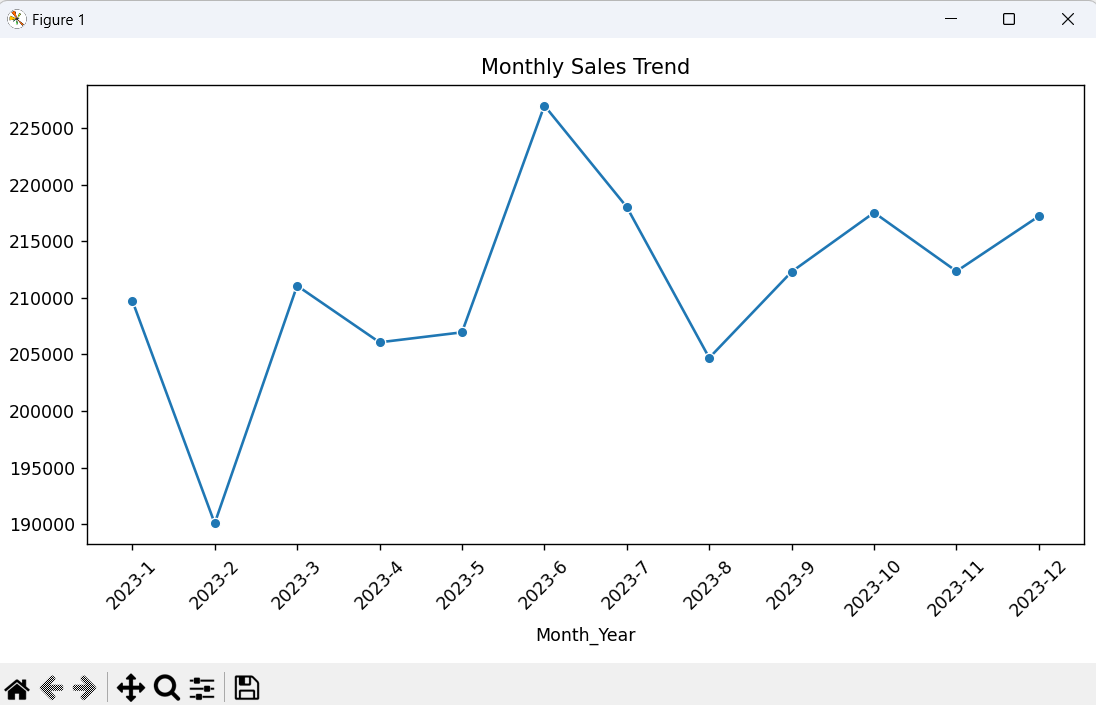
* **Libraries**: pandas, matplotlib.pyplot, seaborn
* **Grouping Formula**: monthly\_sales = df.groupby(['Year', 'Month'])['Sales'].sum().reset\_index()
* Combined into a Month\_Year field for visualization.

**4.1.4 Analysis Results**

The graph reveals spikes in sales during specific months, possibly due to promotional events or seasonal shopping trends. Sales increased in late months, showing a year-end sales boost.

**4.1.5 Visualization**

****



**4.2 Bar Chart – Sales and Profit by Region**

**4.2.1 Introduction**

This bar chart compares total sales and profit across regions to identify the most profitable markets.

**4.2.2 General Description**

The dataset was grouped by region, and both sales and profit values were aggregated.

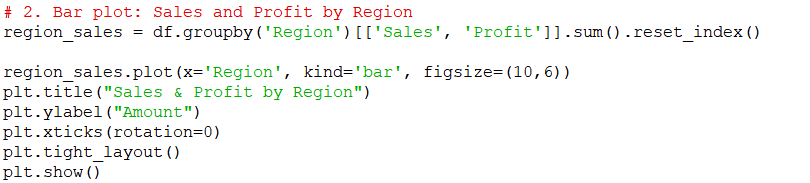
**4.2.3 Specific Requirements**

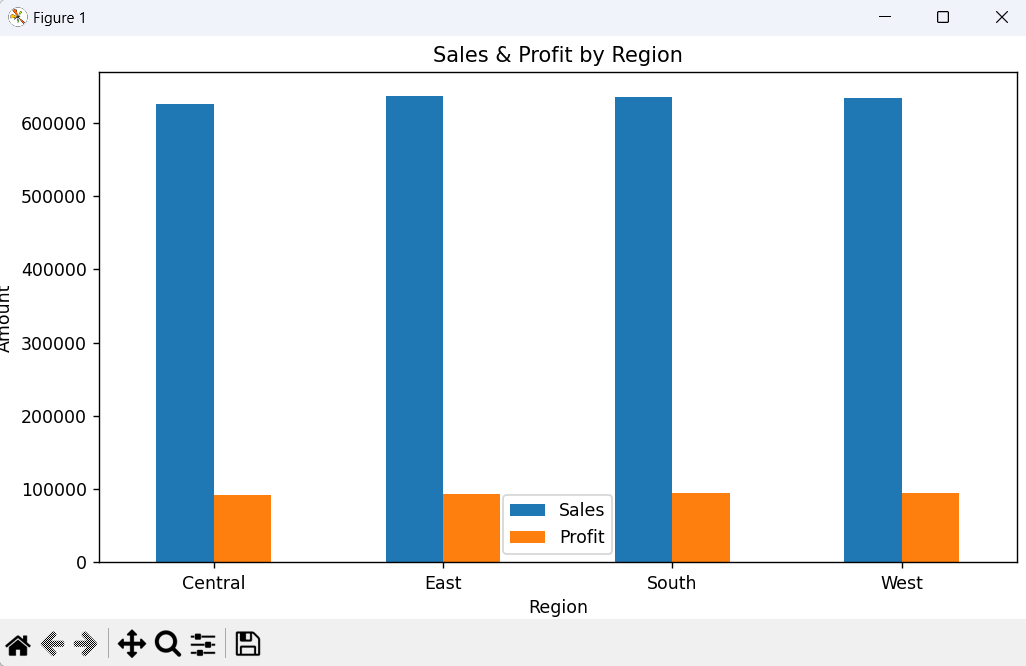
* **Grouping**: region\_sales = df.groupby('Region')[['Sales', 'Profit']].sum().reset\_index()
* **Plotting**: region\_sales.plot(x='Region', kind='bar')

**4.2.4 Analysis Results**

The Central and East regions perform strongly in both sales and profit, indicating better margins or higher consumer activity in those zones.

**4.2.5 Visualization**

****



**4.3 Pie Chart – Sales by Category**

**4.3.1 Introduction**

This pie chart breaks down total sales by product category to highlight top-selling categories.

**4.3.2 General Description**

Sales were grouped by product category and visualized as a pie chart to show proportional contribution.

**4.3.3 Specific Requirements**

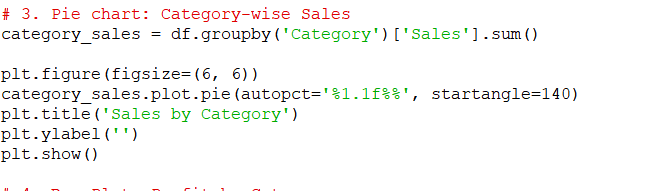
category\_sales = df.groupby('Category')['Sales'].sum()

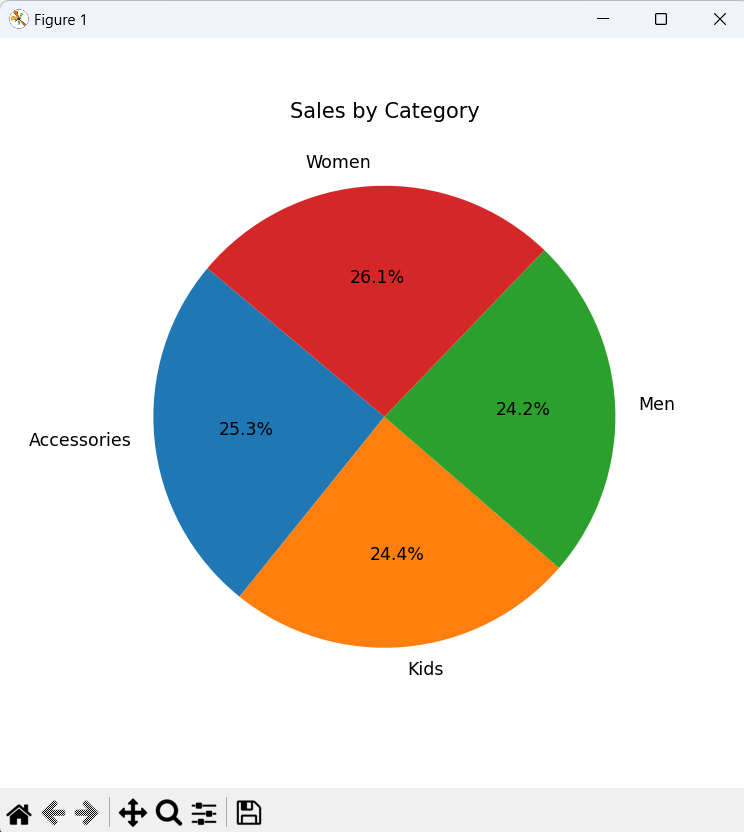
category\_sales.plot.pie(autopct='%1.1f%%', startangle=140)

**4.3.4 Analysis Results**

Clothing and Footwear emerged as dominant categories, indicating strong market preference and demand.

**4.3.5 Visualization**

****



**4.4 Box Plot – Profit by Category**

**4.4.1 Introduction**

This box plot compares profit distributions across product categories to assess performance variability.

**4.4.2 General Description**

The plot reveals outliers and spread in profits per category.

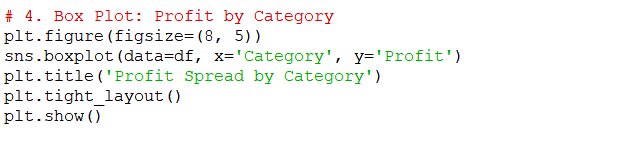
**4.4.3 Requirements**

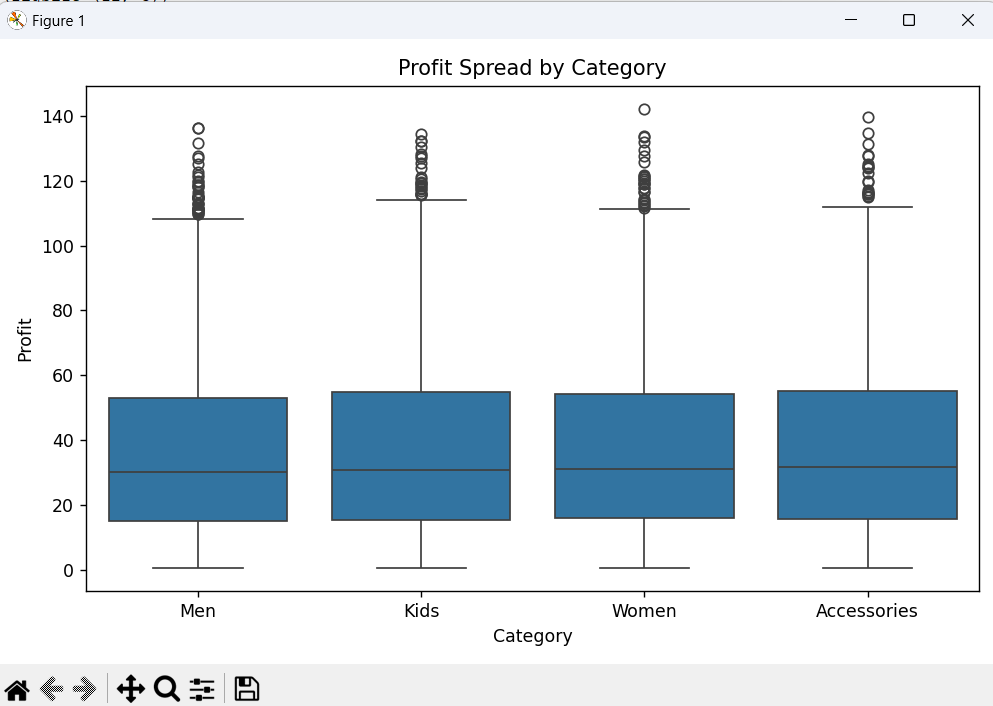
sns.boxplot(data=df, x='Category', y='Profit')

**4.4.4 Analysis Results**

Some categories have high variance, indicating inconsistent profitability. A few outliers suggest occasional high-margin transactions.

**4.4.5 Visualization**

****

****

**4.5 Bar Chart – Top 5 Products by Sales**

**4.5.1 Introduction**

This chart highlights the most popular products based on revenue, helping inventory and marketing strategies.

**4.5.2 General Description**

Products were grouped, sales summed, and the top five items identified and visualized.

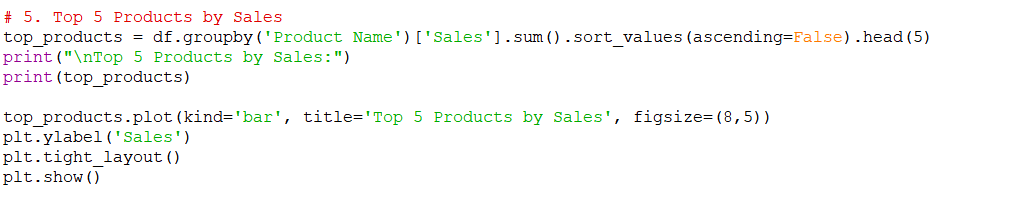
**4.5.3 Requirements**

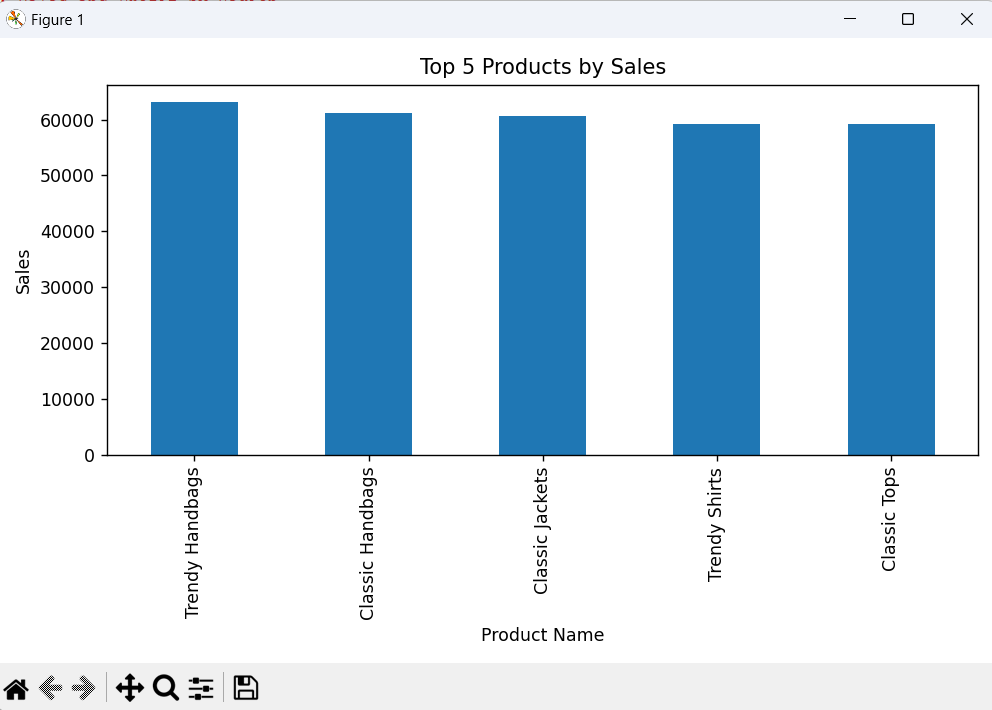
top\_products = df.groupby('Product Name')['Sales'].sum().sort\_values(ascending=False).head(5)

**4.5.4 Analysis Results**

Top-selling products, including Jackets and Jeans, contribute significantly to total revenue.

**4.5.5 Visualization**

****

****

**4.6 Heatmap – Correlation Between Sales, Profit, Discount, Quantity**

**4.6.1 Introduction**

This heatmap shows correlations between key numerical metrics to explore relationships.

**4.6.2 General Description**

The dataset’s numerical features were passed into a correlation matrix to detect patterns such as impact of discounts on profits.

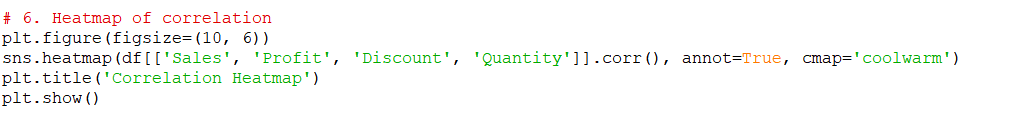
**4.6.3 Requirements**

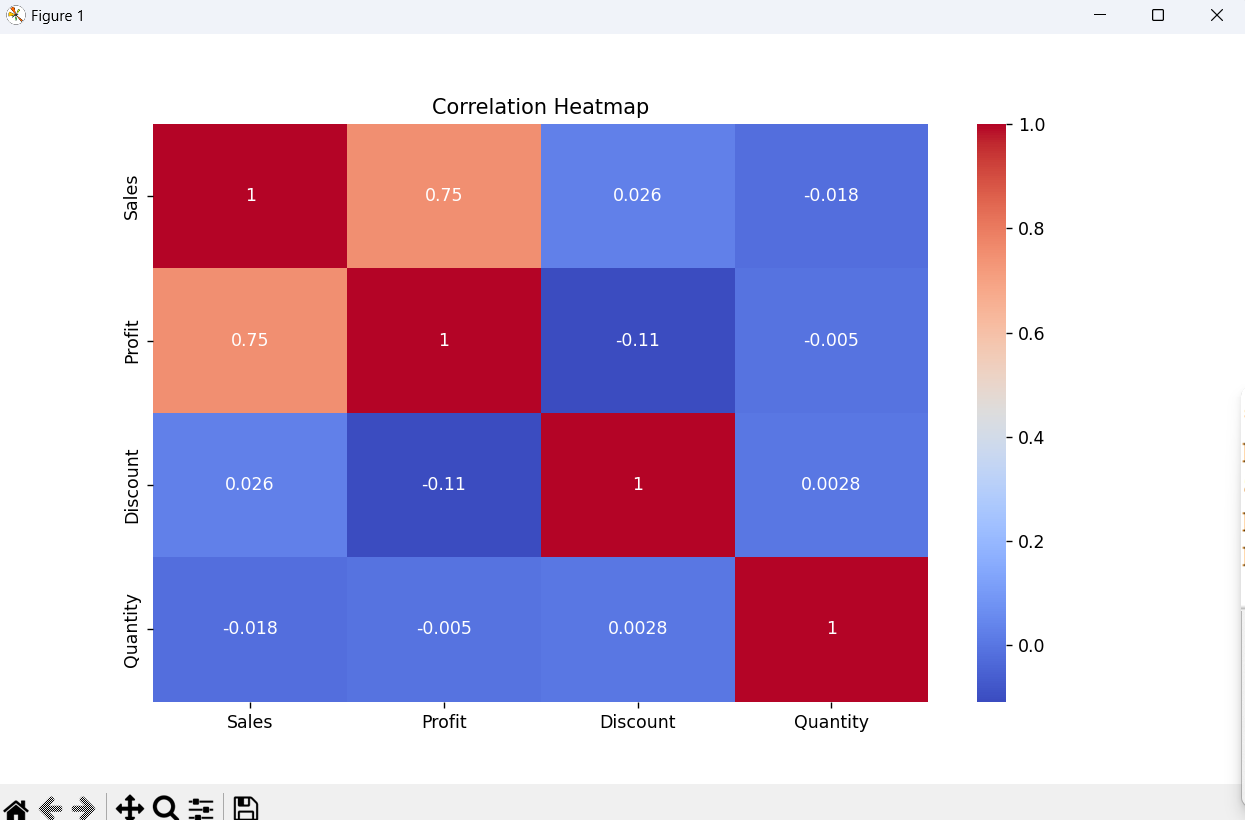
sns.heatmap(df[['Sales', 'Profit', 'Discount', 'Quantity']].corr(), annot=True, cmap='coolwarm')

**4.6.4 Analysis Results**

* Positive correlation between Sales and Profit.
* Negative correlation between Discount and Profit, suggesting discounting affects profit margins.

**4.6.5 Visualization**

****

****

# FUTURE SCOPE

This current analysis provides meaningful insights into sales trends, regional performance, category-wise sales, and profitability for a clothing retail business. However, there are several opportunities to enhance the analysis for deeper and more actionable insights in the future:

1. **Real-Time Sales Integration:**  
   Incorporate live or real-time transactional data to monitor sales trends dynamically and support quick decision-making for promotions and inventory adjustments.
2. **Customer Segmentation and Behavior Analysis:**  
   Use advanced analytics to segment customers based on purchasing behavior, frequency, and spending, enabling targeted marketing and personalized offers.
3. **Demand Forecasting:**  
   Apply machine learning models to forecast sales and inventory needs, helping in proactive inventory planning and reducing overstock or stockouts.
4. **Seasonal Trend Analysis:**  
   Analyze seasonality in clothing sales across regions and categories to optimize stock levels and promotional campaigns during peak and off-peak seasons.
5. **Recommendation Systems:**  
   Develop product recommendation systems to improve customer experience and increase average order value by suggesting related or popular products.
6. **Geospatial Expansion Planning:**  
   Extend the dataset with geographic data to identify high-performing locations and plan new store expansions in areas with high potential.
7. **Interactive BI Dashboards:**  
   Design interactive dashboards using tools like Tableau or Power BI to allow stakeholders to explore sales, profit, discount, and quantity metrics with dynamic filters and visualizations.

**In conclusion**, this Clothing Store Data Analysis project acts as a strong foundation for driving data-informed strategies in retail, optimizing operations, enhancing customer satisfaction, and scaling future growth in the competitive apparel market.

# REFERENCES

[1] W. McKinney, *Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython*, 2nd ed., O’Reilly Media, 2017.

[2] J. D. Hunter, “Matplotlib: A 2D graphics environment,” *Computing in Science & Engineering*, vol. 9, no. 3, pp. 90–95, May–Jun. 2007.

[3] M. Waskom et al., “Seaborn: Statistical data visualization,” *Journal of Open Source Software*, vol. 6, no. 60, p. 3021, 2021.

[4] NumPy Developers, “NumPy: The fundamental package for scientific computing with Python,” [Online]. Available: <https://numpy.org>. [Accessed: Apr. 10, 2025].

[5] Pandas Development Team, “Pandas Documentation,” [Online]. Available: <https://pandas.pydata.org>. [Accessed: Apr. 10, 2025].

[6] M. Lutz, *Learning Python*, 5th ed., O'Reilly Media, 2013.