**Lovely Professional University**



**Academic Task of INT-375**

**(Data Science Toolbox: Python Programming)**

**COMPUTER SCIENCE AND ENGINEERING**

**INT375-Data Science Toolbox: Python Programming**

**PROJECT REPORT**

(Project Semester January-April 2025)

**Title of the Project:**

"Comprehensive Sales Data Analysis Using Python: Statistical Evaluation and Visual Insights"

Submitted By: Ishita Yadav

Registration No. 12319348

Section: K23FK

Course Code: INT-375

**Under the Guidance of**

Dr.Karan Bajaj (32130)

**Discipline of CSE**

**Lovely School of Computer Science and Engineering**

**Lovely Professional University, Phagwara**

**Declaration**

I, Ishita Yadav , student of BTech (Computer Science and Engineering) under CSE 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: 12-04-2025

Registration No. 12319348 Ishita Yadav

**Certificate**

This is to certify that Ishita Yadav bearing Registration no. 12319348 has completed INT-375 “**Data Science Toolbox: Python Programming”** project titled, **"Comprehensive Sales Data Analysis Using Python: Statistical Evaluation and Visual Insights"** under my guidance and supervision. To the best of my knowledge, the present work is the result of his original development, effort and study.

**Dr. Karan Bajaj**

**(UID: 32130)**

**School of Computer Science and Engineering.**

Lovely Professional University

Phagwara, Punjab.

Date: 12-04-2025

**Acknowledgement**

I would like to express my sincere gratitude to my project guide for their valuable guidance, encouragement, and continuous support throughout the development of my project titled **"Comprehensive Sales Data Analysis Using Python: Statistical Evaluation and Visual Insights"**, undertaken as a part of the course **INT-375: Data Science Toolbox: Python Programming.**

I am thankful for the opportunity to work on a real-world problem, which allowed me to apply theoretical knowledge and enhance my practical skills in data analysis and programming. The insightful feedback and suggestions provided at various stages greatly helped in shaping the outcome of this project.

I would also like to acknowledge the resources, tools, and datasets that contributed to my research and analysis. Lastly, I extend my appreciation to my peers and faculty members who provided their support and encouragement during the project.

Ishita Yadav

Reg No. 12319348

**TABLE OF CONTENT**

|  |  |
| --- | --- |
| **Sr No.** | **Description** |
| **1** | **INTRODUCTION** |
| **2** | **Source of Dataset** |
| **3** | **EDA**  **Exploratory Data Analysis** |
| **4** | **Objectives** |
| **5** | **CONCLUSION** |
| **6** | **FUTURE SCOPE** |
| **7** | **References** |

**INTRODUCTION**

The dataset titled **"1000 Sales Records"** provides a rich overview of global sales transactions across various markets. It contains 1,000 individual sales records, each capturing essential business details such as the region and country of sale, the type of item sold, and the mode of sales channel—either online or offline. Additionally, the dataset logs the order priority, which ranges from Critical to Low, providing insight into operational urgency.

This dataset is particularly useful for financial and performance analysis, as it includes numerical data on units sold, unit price, unit cost, total revenue, total cost, and total profit. The inclusion of both order and shipping dates also allows for time-based analysis, such as delivery performance or seasonal trends. Given its structure and diversity, the dataset is ideal for conducting exploratory data analysis (EDA), visualizing sales metrics, evaluating profitability, and comparing performance across different channels, regions, and priorities.

**Source of Dataset**

**Link:**

[“CLICK HERE FOR THE URL OF THE DATASET”](https://excelbianalytics.com/wp/downloads-18-sample-csv-files-data-sets-for-testing-sales/)

**EDA**

**Exploratory Data Analysis**

**1. Dataset Overview:**

**The dataset contains 1,000 records and 13 columns, representing various aspects of sales transactions. These columns include a mix of geographic, categorical, and financial data, with key information on sales performance, item type, and order details. Key fields include:**

* **Categorical Data: Region, Item Type, Country, Sales Channel, Order Priority**
* **Numerical Data: Units Sold, Total Revenue, Total Cost, Total Profit**
* **Date Information: Order Date, Ship Date**

**2. Data Inspection and Initial Review:**

* **The dataset was loaded using pandas, followed by an initial inspection using df.head() and df.tail() to understand its structure and content.**
* **Using df.info() and df.describe(), it was confirmed that there are no missing or null values, and an initial review of statistical properties (e.g., average unit price, revenue, and profit range) was conducted.**

**3. Categorical Data Exploration:**

* **Unique Values: The categorical columns, including Region, Item Type, Country, Sales Channel, and Order Priority, were examined for unique values and frequency distributions.**
* **Frequency Distribution: Visualizations such as bar charts and pie charts were created to analyze the distribution of categories, helping to identify the most common regions, item types, and order priorities.**

**4. Numerical Data Exploration:**

* **Units Sold, Total Revenue, Total Cost, and Total Profit were analyzed to assess the spread and variability of numerical values.**
* **Descriptive statistics were used to examine metrics such as mean, median, and standard deviation, providing insights into the central tendency and dispersion of the financial metrics.**

**5. Date and Time Analysis:**

* **Date columns like Order Date and Ship Date were converted into datetime objects for time-based analysis.**
* **A new column, Shipping Delay, was created by calculating the difference between Order Date and Ship Date, providing insights into the shipping efficiency.**

**6. Data Quality Check:**

* **Duplicate Records: The dataset was checked for any duplicate records to ensure data cleanliness and consistency.**

**7. Correlation Analysis:**

* **A correlation heatmap was generated to evaluate the relationships between key financial metrics, such as Total Revenue, Total Cost, and Total Profit.**
* **This helped identify any strong correlations that may provide deeper insights into sales performance.**

**8. Data Visualization:**

**Several visualizations were created to better understand key trends and relationships in the data:**

* **Histograms: Displayed the distribution of numerical fields like Units Sold and Total Revenue.**
* **Boxplots: Helped identify outliers in fields like Total Cost and Total Profit.**
* **Pie Charts: Showed the breakdown of categories like Sales Channel and Order Priority.**
* **Heatmaps: Visualized correlations and patterns between the financial metrics and categorical variables.**

**9. Key Insights and Observations:**

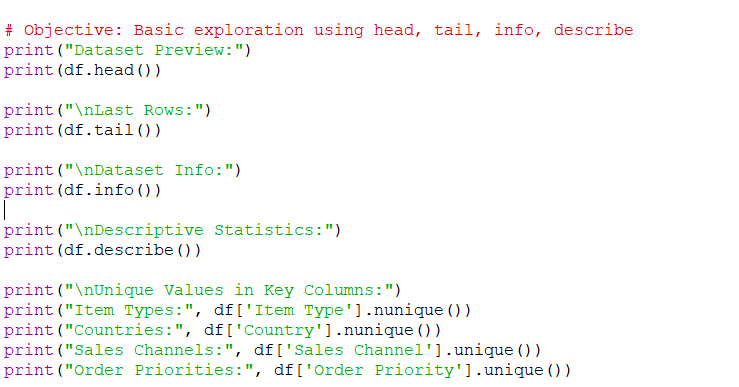
* **Item Popularity: Certain items have higher sales volume, which is visible through the distribution of Units Sold.**
* **Regional Profitability: Regions such as North America and Europe showed higher profitability compared to others.**
* **Sales Channel Impact: The Online Sales Channel generated the highest revenue, indicating its effectiveness as a sales platform.**
* **Order Priority: Orders marked as High Priority tend to have higher revenue and profit, highlighting their importance in business operations.**

**Objectives**

Objective 1:

Data Exploration and Summary Statistics: Generate summary statistics for numeric columns. Basic exploration using head, tail, info, describe

Coding Part:



Implementation:

A screenshot of a computer

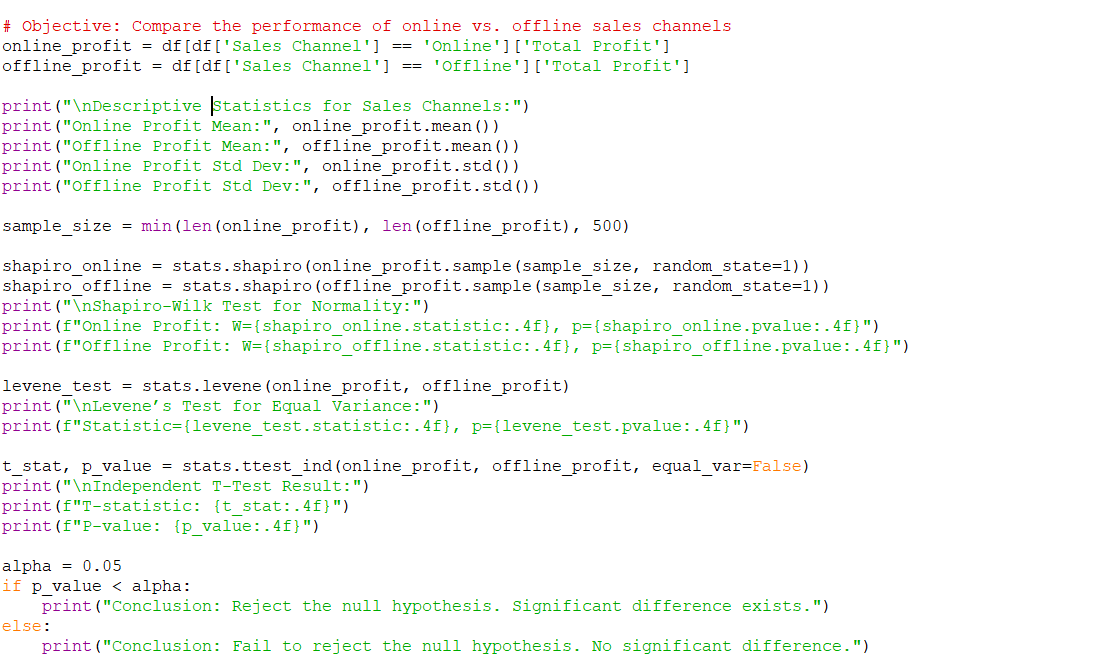
AI-generated content may be incorrect.

A screenshot of a computer

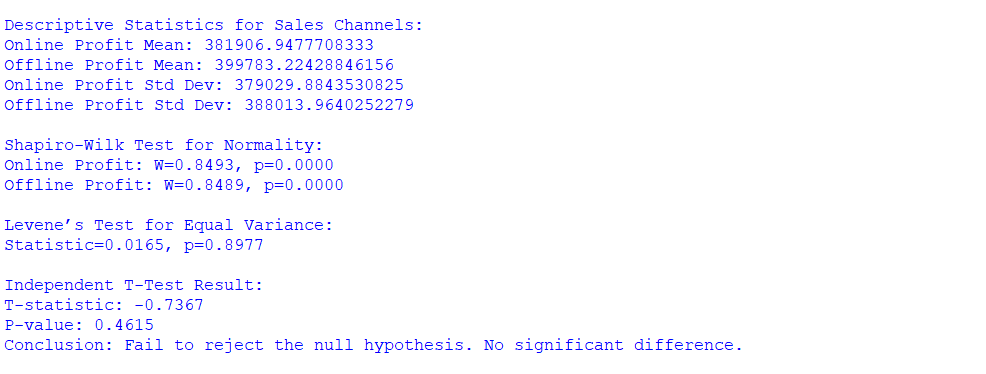
AI-generated content may be incorrect.

Objective 2:

Objective: Compare the performance of online vs. offline sales channelsCoding Part:



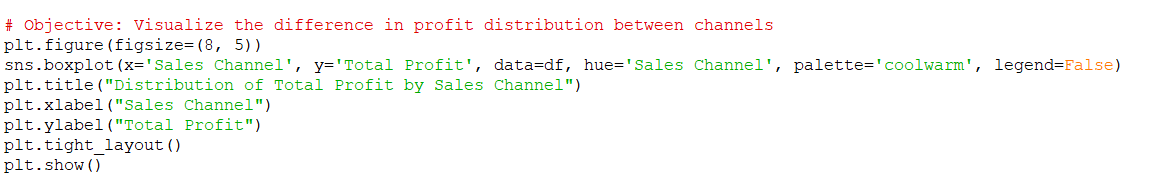
Implementation:



Objective 3:

Objective: Visualize the difference in profit distribution between channels

Coding Part:



Implementation:

A screenshot of a graph

AI-generated content may be incorrect.

Objective 4:

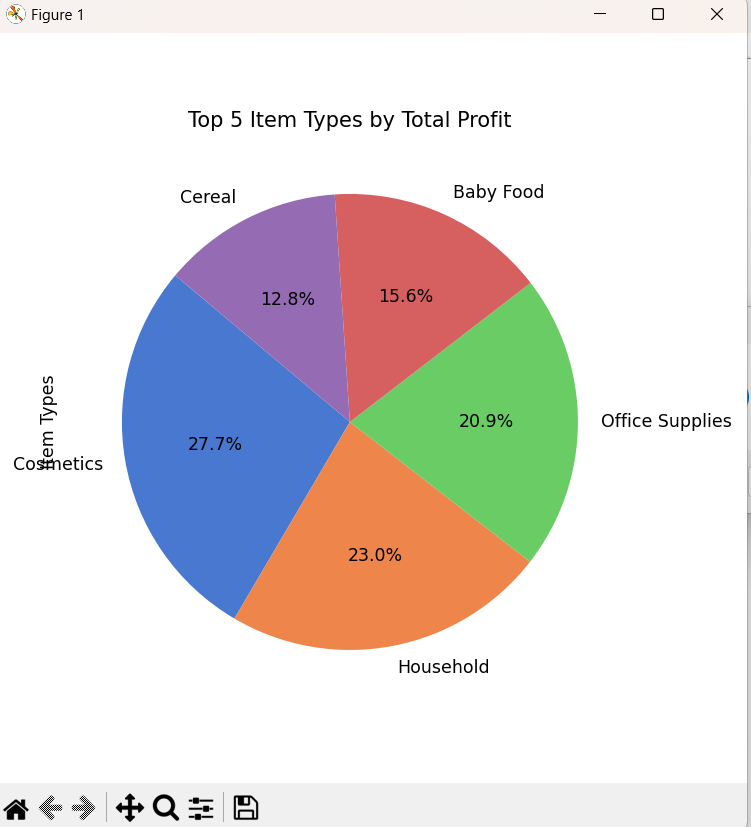
Objective: Determine which item type generates the highest total profit

Coding Part:

A close up of text

AI-generated content may be incorrect.

Implementation:



Objective 5:

Objective: Identify the most profitable and least profitable countries

Coding Part:

A computer code on a white background

AI-generated content may be incorrect.

Implementation:

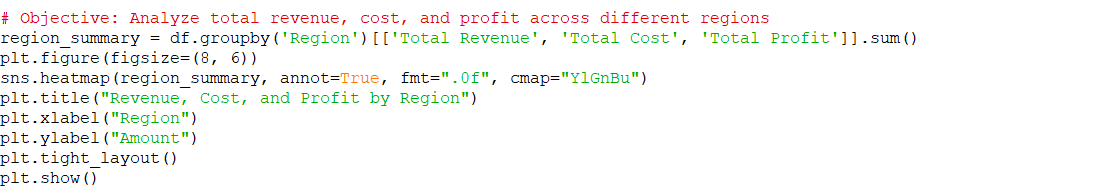
A screen shot of a graph

AI-generated content may be incorrect.

Objective 6:

Objective: Analyze total revenue, cost, and profit across different regions

Coding Part:



Implementation:

A screenshot of a computer

AI-generated content may be incorrect.

Objective 7:

Objective: Study the impact of order priority on total revenue and profit

Coding Part:

A close-up of a white background

AI-generated content may be incorrect.

Implementation:

A screenshot of a graph

AI-generated content may be incorrect.

Objective 8:Objective: Scatter plot for Total Revenue vs Total Profit

Coding Part:

A white background with colorful text

AI-generated content may be incorrect.

Implementation:



**CONCLUSION**

The **sales dataset** provides a comprehensive view of sales transactions, covering various dimensions like product type, sales region, revenue generation, and operational factors such as shipping delays and order priority. By analyzing the data, we can conclude:

* **Sales Performance**: The dataset reveals the impact of **regions** and **sales channels** on overall sales performance. Certain regions (such as North America and Europe) and online sales channels show significantly higher revenue and profit.
* **Product and Item Trends**: There are distinct patterns in the types of products sold, with certain **item types** consistently performing better than others in terms of **units sold**. This can guide inventory and marketing strategies.
* **Financial Insights**: The dataset enables a detailed examination of financial metrics like **total revenue**, **total cost**, and **total profit**, highlighting the relationship between sales volume, pricing, and profitability. It suggests that increasing revenue while managing costs is key to improving profits.
* **Operational Efficiency**: The inclusion of **shipping delays** gives insight into logistical efficiency, which is an important factor for customer satisfaction and order fulfillment. Minimizing shipping delays could lead to enhanced customer experience and repeat business.
* **Priority and Urgency**: Orders marked with **high priority** tend to generate more revenue, underlining the importance of prioritizing certain orders to meet customer demands and ensure timely deliveries.

Overall, the dataset offers valuable insights for optimizing sales strategies, improving operational efficiency, and understanding the key factors driving revenue and profitability. It provides a solid foundation for data-driven decision-making in sales and logistics management.

**FUTURE SCOPE**

While the current analysis provides a solid foundation, there are several potential areas for future exploration and enhancement:

1. **Predictive Analytics:**
   * By leveraging machine learning models, we can predict future sales trends based on historical data. This could involve predicting the **total revenue**, **units sold**, or even **shipping delays**, helping the business make informed decisions on inventory management and marketing strategies.
2. **Customer Segmentation:**
   * Further analysis can be conducted to segment customers based on their purchasing behavior. Clustering techniques (e.g., K-means) could help identify groups of customers with similar preferences, enabling targeted marketing and personalized promotions.
3. **Sentiment Analysis:**
   * If customer feedback or reviews are available, **sentiment analysis** can be applied to gauge customer satisfaction and correlate it with product sales. This can provide deeper insights into product performance and customer experience.
4. **Supply Chain Optimization:**
   * The analysis of **shipping delays** can be expanded by incorporating factors like warehouse location, delivery routes, and carrier performance. This would help optimize supply chain processes, reduce costs, and improve delivery times.
5. **Dynamic Pricing Strategy:**
   * By analyzing pricing data in conjunction with **units sold**, **total cost**, and **total revenue**, dynamic pricing models can be developed. These models can help adjust prices in real-time based on market conditions, demand, and competition, leading to optimized profits.
6. **Sales Forecasting with Time Series Analysis:**
   * Time series analysis can be applied to the dataset to forecast future sales trends, considering seasonality, trends, and external factors (such as holidays or promotions). This would assist in more accurate financial planning and demand forecasting.
7. **Sales Channel Performance:**
   * A deeper dive into each **sales channel**'s performance could help understand which channels (online vs offline) are more profitable or efficient. By analyzing factors like conversion rates, customer acquisition cost, and return on investment (ROI), businesses can allocate resources more effectively.
8. **Geospatial Analysis:**
   * The addition of **geographic data** (if available) can enhance insights through **geospatial analysis**. This could involve analyzing sales performance by geographic location, understanding regional preferences, or optimizing the supply chain by identifying areas with higher sales potential.
9. **Real-time Data Integration:**
   * The integration of **real-time data** from sales transactions, inventory, and customer interactions could provide a more dynamic and up-to-date view of business performance, enabling quicker decision-making and more adaptive strategies.
10. **Product Lifecycle Analysis:**
    * Tracking product performance over time could allow for the identification of **product lifecycle stages**. This would help determine when products are at their peak, when they should be discounted, or when they may need to be phased out.

By addressing these areas, businesses can unlock deeper insights, enhance operational efficiency, and ultimately improve profitability and customer satisfaction. The future scope is vast, with significant opportunities to drive growth through advanced analytics and data-driven strategies.

**REFERENCES:**

1.Dataset Link:

<https://excelbianalytics.com/wp/downloads-18-sample-csv-files-data-sets-for-testing-s>

2.Python Libraries:

* **NumPy Developers. (2023).** *NumPy: Fundamental package for scientific computing with Python*. Retrieved from: <https://numpy.org/>
* **pandas development team. (2023).** *Pandas: Python Data Analysis Library*. Retrieved from: <https://pandas.pydata.org/>
* **Matplotlib Developers. (2023).** *Matplotlib: Visualization with Python*. Retrieved from: <https://matplotlib.org/>
* **Seaborn Developers. (2023).** *Seaborn: Statistical data visualization*. Retrieved from: <https://seaborn.pydata.org/>
* **SciPy Community. (2023).** *SciPy: Scientific Computing Tools for Python*. Retrieved from: <https://scipy.org/>