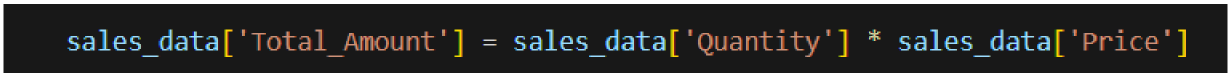
Week 1 - Data Analysis for Business Insights

Working as a data analyst, I analyzed a sales dataset of a retail company ShopEase. The analysis includes cleaning, exploring, and uncovering the hidden trends of the sales dataset.

1. Cleaned Dataset

- I started with importing the necessary libraries I need: pandas, matplotlib, and seaborn. I then loaded my sales dataset using pd.read_excel() and analyzed the first view rows using sales_data.head().
- While I was inspecting the data entries, I found two potential risks:
  - Total_Amount column has a missing value.
  - Customer_ID [C001 and C002] had appeared twice claiming they had repeat purchases on different dates.
- Next, I inspected columns, rows, data types, information and statistical measures using .columns(), .shape(), .dtypes(), .info(), .describe() to understand the data and the formats. I changed the date format using pd.to_datetime() to make it compatible for visualization.
- To confirm no missing values are there I used sales_data.isnull().sum(). To mitigate the risk of missing value in 'Total_Amount' column I applied logic and filled the value as the product of price and quantity [Total_Amount = Quantity × Price].



```
sales_data['Total_Amount'] = sales_data['Quantity'] * sales_data['Price']
```

- Furthermore, to check duplicates I used sales_data.duplicated().sum() and confirmed no duplicates were present.
- Now, the data was clean and ready for further analysis. I saved the cleaned sales data as 'sales_data_cleaned' using .to_csv as it is simple and easy to analyze.

2. Exploratory Data Analysis

- As previously mentioned, I confirmed the repeat purchase of duplicate customer ids. While it may be considered an error but the unique transactions on different dates confirmed the purchase as deliberate. Of all reasons, one can be satisfactory products. I kept these entries to preserve customer behavior insights.

- I then explored the categorical entries using sales_data_cleaned.unique() and .value_counts(), which revealed:

    - **Categories**: Electronics. Clothing, and Books
    - **Products:** Laptop, Smartphone, Headphones, Tablet, Book, Shoes, T-Shirt, Smartwatch
    - **Payment Methods:** Credit Card, Debit Card, Cash, PayPal
    - **Regions:** North, South, East, and West

- Moving on, I examined the monthly total amount over a period of time. To accomplish this, I grouped total sales with the date column, and sorted data by sales. I discovered that:

    - January is the best-selling month with highest total sales [2100].
    - March was the lowest performing month, with a decrease in sales [325].

- I analyzed category-level performance and found that electronics had the highest total selling revenue, followed by clothing and books..
- Then, I grouped products by quantity and total_amount to calculate total sales revenue. I discovered that

    - Books are sold more popular than other products [9 sold].
    - Smartphone have the largest sales volume [3000].

- To examine numeric relationships, I generated a correlation matrix using .corr(). I found:

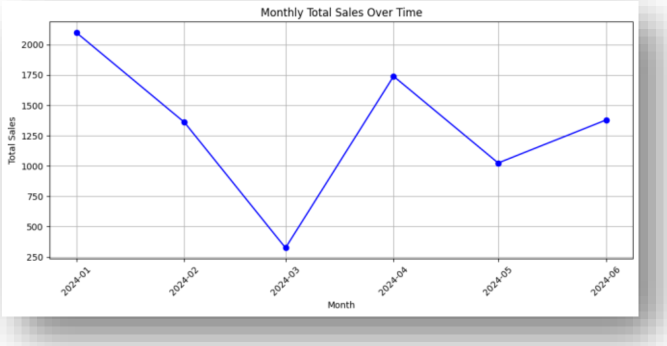
    - Total_Amount has strong positive correlation with Price (0.89).
    - Quantity showed a moderate negative correlation with Price (-0.42), indicating that higher-priced items were purchased in smaller quantities.

3. Visualizations

I used both **Python (matplotlib & seaborn)** for quick visual analysis and **Power BI** for an interactive dashboard.

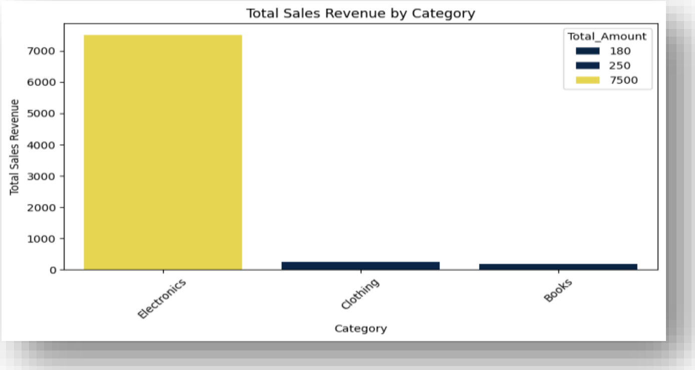
Python-Based

1. **Monthly Total Sales Over Time**



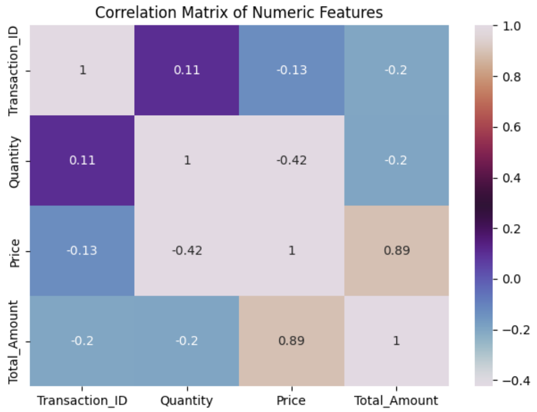
I chose a line plot for showcasing monthly sales over a period. January has the highest peak showing high total sales. Sales eventually dropped in February and hit the lowest in March.

2. **Total Sales Revenue by Category**



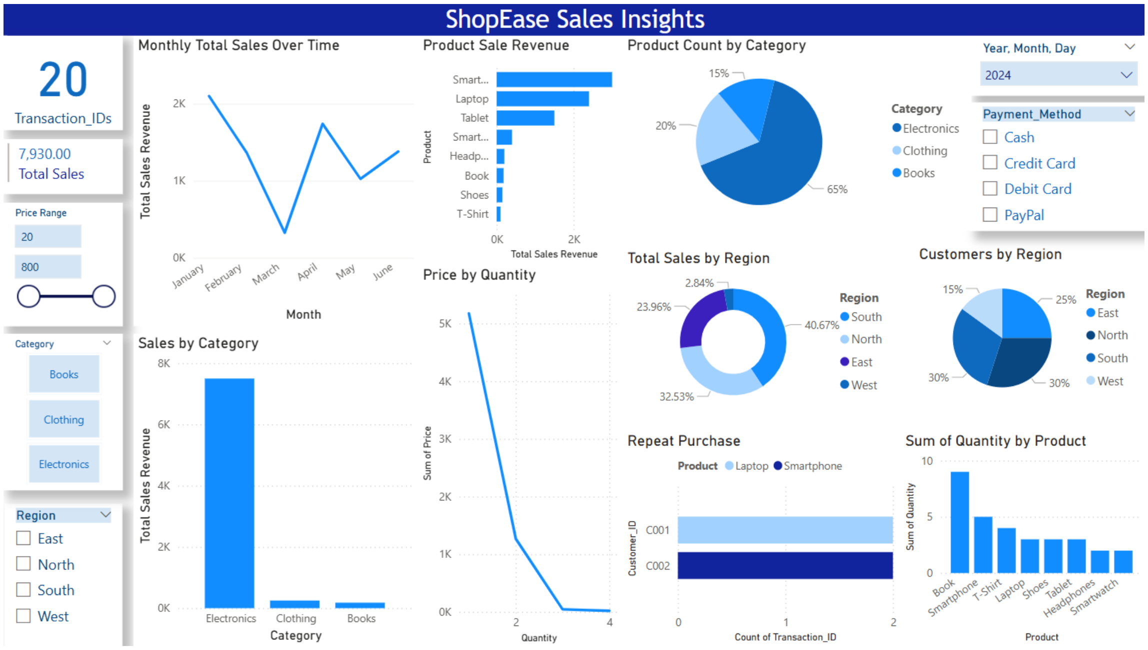
Here I used a bar chart which revealed that the **Electronics** category has the highest contribution to overall revenue (7500). Followed by Clothing and Books.

3. Correlation Matrix (Heatmap)



The numerical variables (Price, Quantity, Total_Amount) showed that Total_Amount strongly correlates with Price, while Quantity is **negatively correlated** with Price.

Power BI Dashboard – Final Data Insights



- ➢ **Best-Selling Month**: January 2024 had the highest sales primarily due to winter clearance sale, driving customers interest.
- ➢ **Sales Fluctuation**: March took the lowest hit as the sales in Q1 significantly dropped by 84.5%. It may be due to seasonal change such as an end to winter clearance

promotion in January. However, April experiences recovery with 435% increases in sales possibly due to spring promotions.

- **Top Product Category**: Electronics led both in revenue and volume, accounting for **65%** of product sales.
- **Best-Performing Region**: The **South region** generated the highest revenue (40.67%), followed by North (32.53%).
- **Customer Behavior**: Repeat purchases were observed from C001 and C002, showing signs of satisfaction and loyalty.
- **Payment Trends**: Most transactions were made using **Cash**, followed by Credit Card.
- **Price Sensitivity**: Higher-priced products tend to be purchased in lower quantities, as shown in the Price vs. Quantity chart.
- **Correlation Insight**: A strong positive correlation exists between Total_Amount and Price, but a moderate **negative correlation** between Quantity and Price.

4. Actionable Recommendations and Strategies to Boost Sales (Bonus)

- Focus **marketing campaigns** during peak sales months like January by launching seasonal promotions (Winter clearance sale) and ensuring adequate stock levels to meet increased demand.
- Increase **promotions** for high-performing categories such as Electronics. Consider offering loyalty discounts, free shipment or early-access deals for returning customers.
- **Address underperformance in March** by introducing targeted campaigns, such as:
    - **Theme-promotions** (e.g., discounts on books during International Women's Day).
    - **Cross-promotions** (e.g., bundle a book with a clothing item as part of a spring sale).
- **Explore regional expansion** by identifying and targeting underserved markets, particularly the **West region**, which had the lowest sales share.

5. Ethical Considerations

1. The dataset does not have any personal information like emails or phone numbers.
2. Anonymous customer ids used.

6. References

- Y, A. (2025, April 16). *Master Pandas with this cheatsheet for data analysis [Online forum post]*. Ajay Y. Posted on the Topic | LinkedIn. https://www.linkedin.com/posts/ajayyadav1996_pandas-cheatsheet-activity-7318148308424654849-3S6f?utm_source=social_share_send&utm_medium=android_app&rcm=ACoAABa--qwB1-26I875JIZq2tug9gcdRycGuWo&utm_campaign=copy_link
- Simplilearn. (2025, May 21). *Power BI Full Course 2025 | Power Bi Tutorial for Beginners | Power Bi Training | SimpliLearn* [Video]. YouTube. https://www.youtube.com/watch?v=hkbkErQvk3c