Advanced Python Project (2025/2026)
Supermarket Sales Analysis

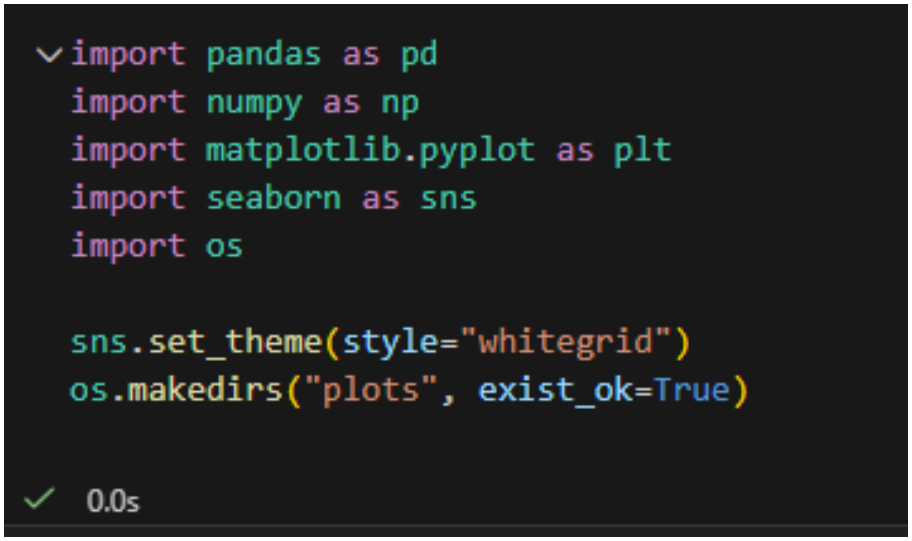
Student Name: Ruaa Hussin

Instructor: Dr. Ali Azwai – Dr. Ala Abuthawabeh

Date: 8/1/2026

Objective

The objective of this project is to apply Python data analysis tools such as Pandas, NumPy, Matplotlib, and Seaborn to analyze the Supermarket Sales dataset. The goal is to clean the data, explore patterns, visualize trends, and extract meaningful insights from real-world sales data.

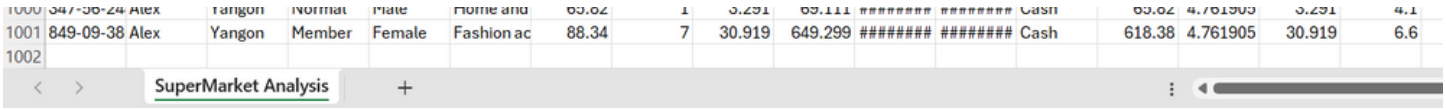


```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import os

sns.set_theme(style="whitegrid")
os.makedirs("plots", exist_ok=True)
```

✓ 0.0s

Dataset Description

The Supermarket Sales dataset contains 1000 transaction records with 17 attributes, including branch, customer type, gender, product line, payment method, sales, gross income, rating, and date. The dataset represents real sales transactions and is suitable for exploratory and statistical analysis.



Load & Inspect Data

The dataset was loaded using Pandas. Initial inspection was performed using head(), info(), and describe() to understand the structure, data types, and basic statistics of the dataset.

```
df = pd.read_csv("SuperMarket Analysis.csv")

print("First 5 Rows:")
display(df.head())

print("\nInfo:")
df.info()

print("\nDescribe (Numerical):")
display(df.describe())

print("\nMissing Values:")
display(df.isnull().sum())
```

✓ 0.0s

Data Cleaning

The Date column was converted to datetime format to allow time-based analysis. The dataset was checked for missing values, and no significant missing data was found. Column names were carefully verified to avoid errors during analysis and visualization.

```
df["Date"] = pd.to_datetime(df["Date"])
df["Time"] = pd.to_datetime(df["Time"], format="%I:%M:%S %p").dt.time

print("Duplicates:")
print(df.duplicated().sum())

df.to_csv("cleaned_supermarket_sales.csv", index=False)
print("Cleaned data saved to cleaned_supermarket_sales.csv")
```

✓ 0.0s

Exploratory Data Analysis

The analysis shows that sales transactions are almost equally distributed across the three branches. Member customers appear more frequently than normal customers. Female customers slightly outnumber male customers. E-wallet is the most commonly used payment method.

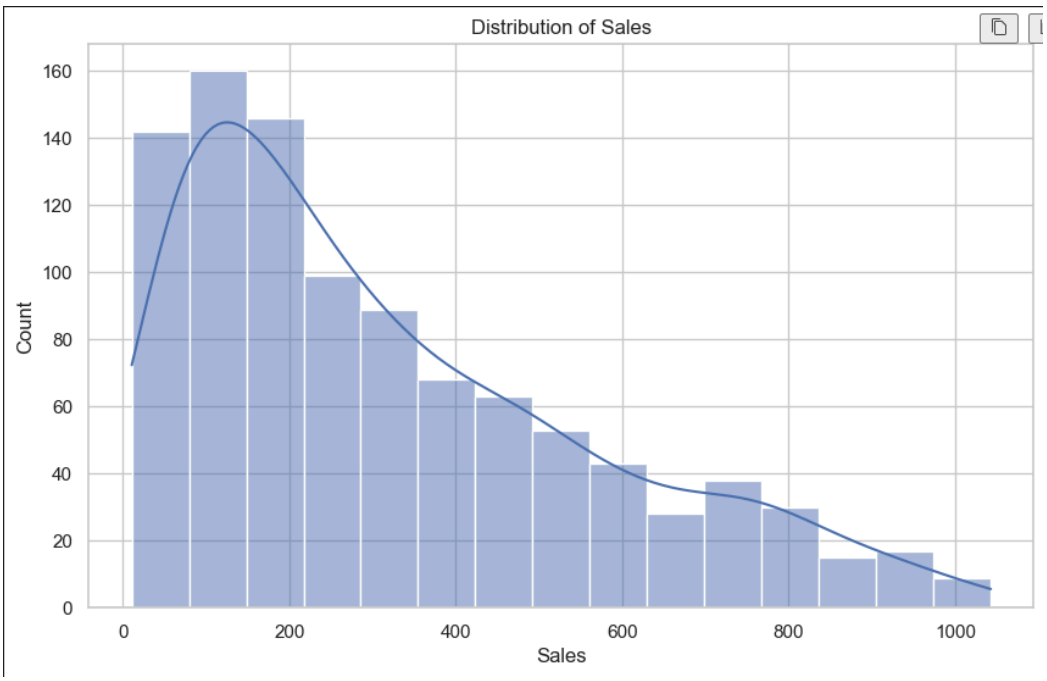
```python
df_clean = pd.read_csv("cleaned_supermarket_sales.csv")
df_clean["Date"] = pd.to_datetime(df_clean["Date"])

stats_cols = ["Branch", "Customer type", "Gender", "Payment"]
print("Basic Statistics for Categorical Columns")
for col in stats_cols:
    print(f"\nFrequency for {col}:")
    print(df_clean[col].value_counts())

print("\nNumerical Statistics")
num_cols = ["Sales", "Rating", "gross income"]
display(df_clean[num_cols].agg(["mean", "median", "max", "min"]))
```

Numerical Statistics

The average sales value is approximately 323, with a maximum exceeding 1000. The average customer rating is close to 7, indicating generally good customer satisfaction. Gross income values vary depending on the transaction and product line.

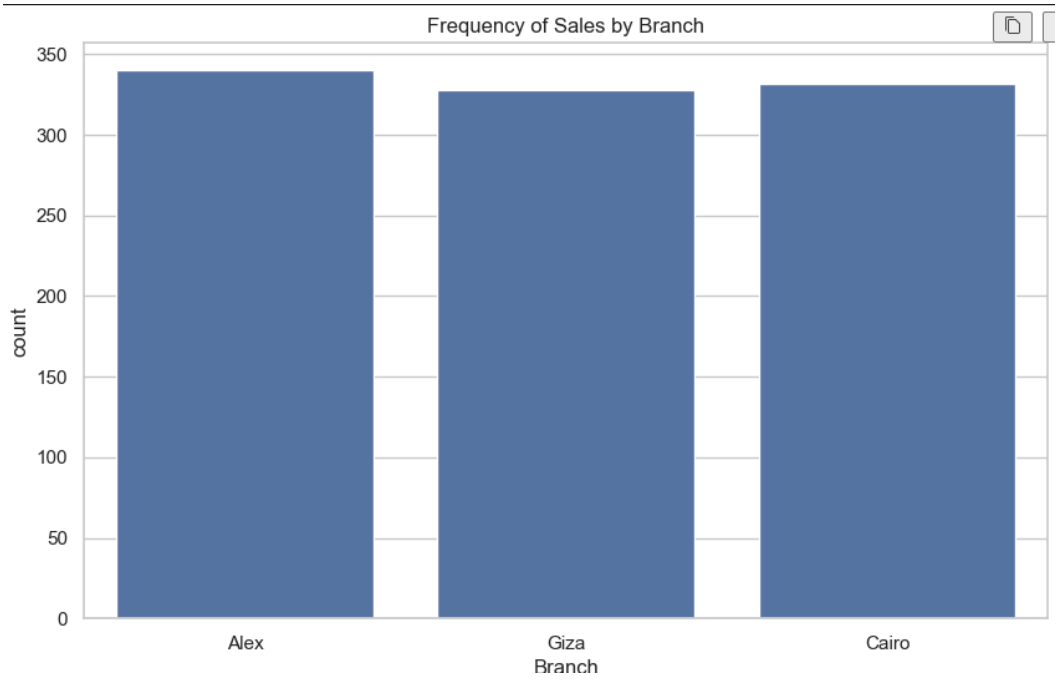
Visualizations

The histogram shows that most sales values are concentrated in the lower to mid range, with a few high-value transactions acting as outliers.



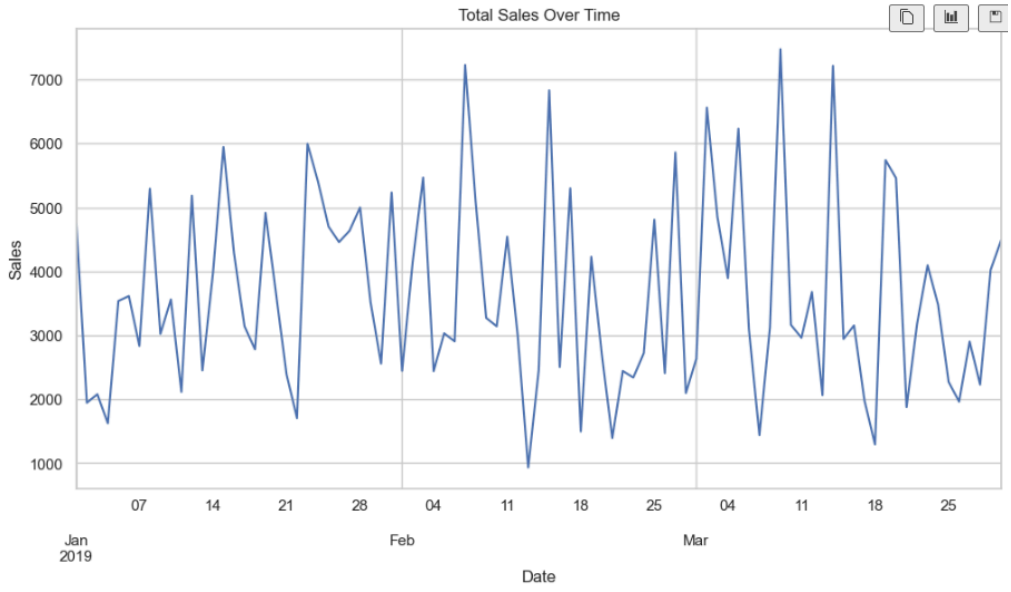
Branch Frequency

The bar chart indicates a balanced number of transactions across all branches.



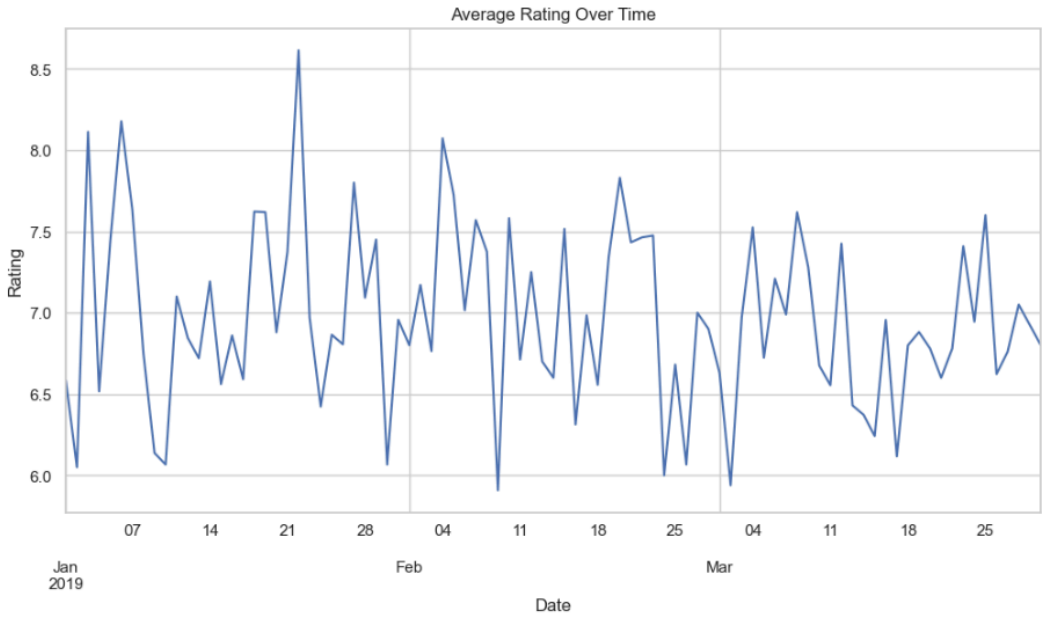
Total Sales Over Time

Sales over time fluctuate with noticeable peaks and drops, suggesting the influence of time-related factors.



Average Rating Over Time

Customer ratings remain relatively stable over time with minor fluctuations.

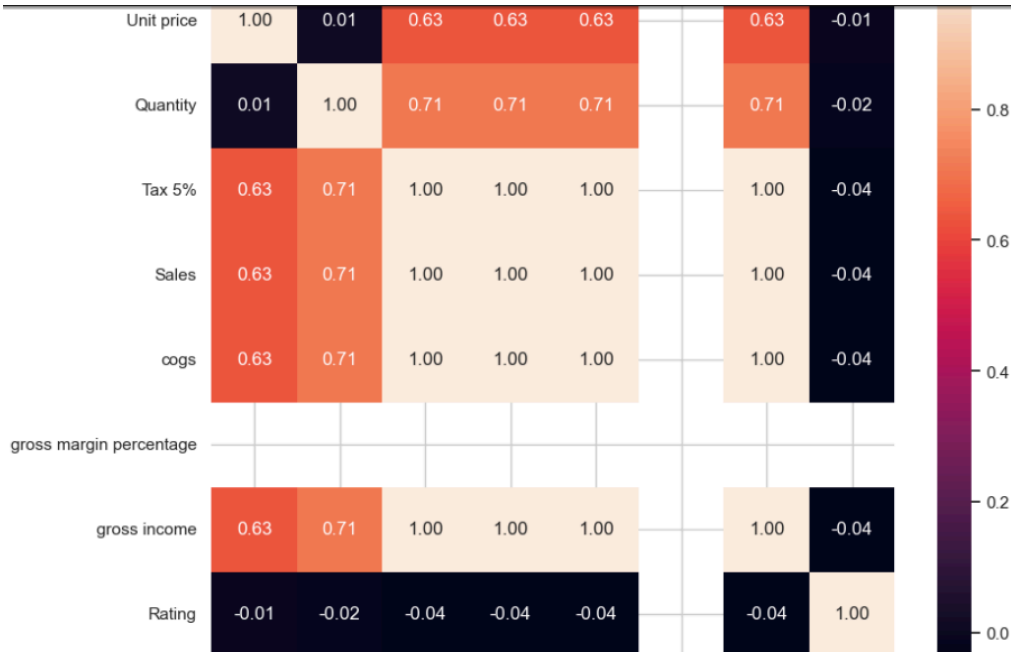


Sales vs Rating

The scatter plot suggests a weak relationship between sales value and customer rating.

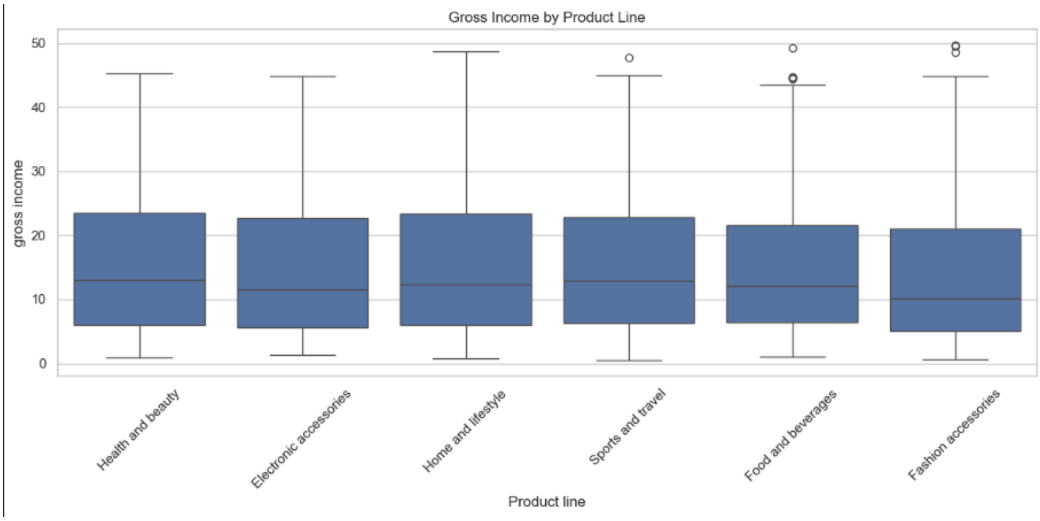
Correlation Heatmap

The heatmap shows a strong correlation between sales and gross income, while other correlations are weak.



Gross Income by Product Line

The boxplot shows differences in profitability across product lines and highlights income variability.



Advanced Questions

1-Giza branch generates the highest total revenue, likely due to higher average transaction values.

2-Members spend more on average than normal customers, indicating customer loyalty.

3-E-wallet is the most frequently used payment method due to convenience.

4-Food and Beverages has the highest average customer rating.

5-There is a very weak relationship between unit price and quantity purchased.

```
Q4: Average Rating by Product Line:
Product line
...
Total Gross Income: 15379.37
Average Tax (5%): 15.38

Product Line Statistics:
```

Output is truncated. View as a scrollable element or open in a text editor. Adjust cell output settings...

Product line	Sales	Quantity	gross income
Food and beverages	56144.8440	952	15.365310
Sports and travel	55122.8265	920	15.812630
Electronic accessories	54337.5315	971	15.220597
Fashion accessories	54305.8950	902	14.528062
Home and lifestyle	53861.9130	911	16.030331
Health and beauty	49193.7390	854	15.411572

```
Q1: Revenue by Branch:
Branch
Giza     110568.7065
Alex     106200.3705
Cairo    106197.6720
Name: Sales, dtype: float64
The branch with the highest revenue is: Giza

Q2: Average Spend by Customer Type:
Customer type
Member    335.742945
Normal    306.372379
Name: Sales, dtype: float64

Q3: Payment Method Usage:
Payment
Ewallet        345
Cash           344
Credit card    311
Name: count, dtype: int64
```

Conclusion

This project demonstrated how Python can be used to analyze real-world sales data and extract valuable business insights through statistical analysis and visualization.