AUTUMN INTERNSHIP PROJECT REPORT

Preprocessing and Visualising Coffee Sales Data

(NOTEBOOK-02)

Banhiva Roy

(Section 01)

4-week Autumn Internship Programme | Government College of Engineering and Leather Technology

Period of Internship: 25th August 2025 - 19th September 2025

Report submitted to: IDEAS – Institute of Data Engineering, Analytics and Science Foundation, ISI Kolkata

1. Abstract :

This project is to analyze the different statistical parameters of a coffee sales dataset. It performs **Exploratory Data Analysis (EDA)** on a coffee sales dataset using Python libraries viz. **Pandas, NumPy, Matplotlib, and Seaborn**. Statistical summaries and groupings are applied to understand sales behavior across **years, months, coffee types, and times of the day**. Visualizations, including line plots and bar charts, are created to highlight patterns such as the peak sales months, most profitable coffee varieties, and revenue trends. Overall, the project demonstrates how Python-based EDA can uncover meaningful insights into consumer behavior and business performance in the coffee retail domain.

2. Introduction :

The project focuses on exploring a coffee sales dataset to derive actionable insights that can support business decisions.
In today's data-driven world, businesses rely heavily on analysing sales data to understand customer preferences, seasonal trends, and revenue patterns.
Analysing this data helps in identifying popular coffee products, high-demand time periods, and long-term business opportunities.

- **Technology Involved →**
  Pandas for data cleaning, manipulation, and aggregation.
  NumPy for numerical operations and efficient array handling.
  Matplotlib and ***Seaborn*** for creating meaningful data visualizations.

These libraries collectively provide a robust framework for statistical analysis and exploratory data analysis (EDA).
Exploratory Data Analysis (EDA) is a fundamental step in any data science workflow. It involves summarizing dataset characteristics and visualizing patterns before applying predictive models or advanced analytics.
For retail businesses, EDA has been particularly valuable in studying customer behavior, optimizing inventory, and maximizing sales performance.
During the learning programme of the internship, we covered the following topics:

- Python basics → Variables, Loops, Operators, Lists, Tuples, Strings
- Functions, Classes and Recursion → Fibonacci series, Armstrong number etc.
- NumPy → Initializing matrices, 2D matrices, manipulating for making operations,
- Pandas → Data Frame, working with datasets – filtering, grouping, merging
- Introduction to ML → Supervised, Unsupervised learning and their methods

3. Project Objective :

Data Cleaning & Preparation – Convert raw coffee sales data into a structured format by handling missing values, checking duplicates, and standardizing datatypes (e.g., converting dates)

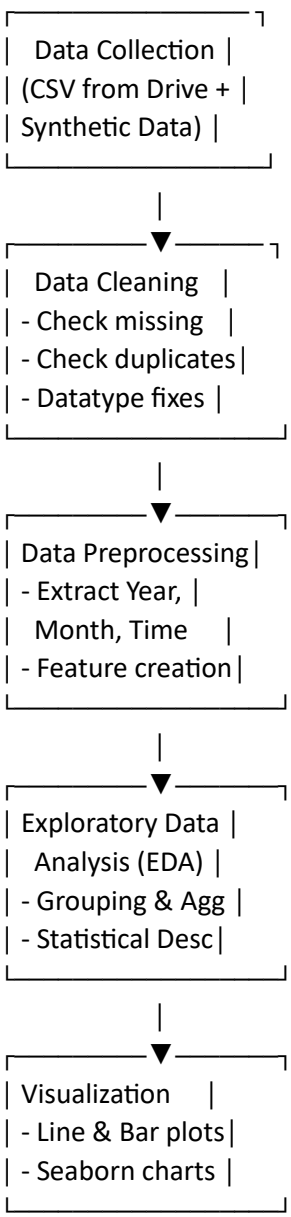
Statistical Exploration – Compute key descriptive statistics (mean, max, variance) to understand sales performance across different time periods and products.

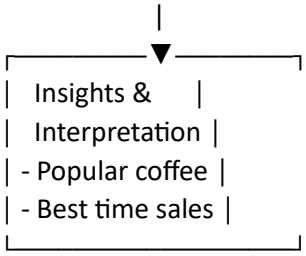
Trend & Pattern Analysis – Identify seasonal trends, peak sales months, and the distribution of sales across years, months, and times of the day.

Product Performance Evaluation – Compare sales performance across different coffee types to determine the most profitable and popular products.

Data Visualization for Insights – Use Matplotlib and Seaborn to create clear visualizations (line plots, bar charts, density plots) that highlight customer behavior and business trends.

4. Methodology :





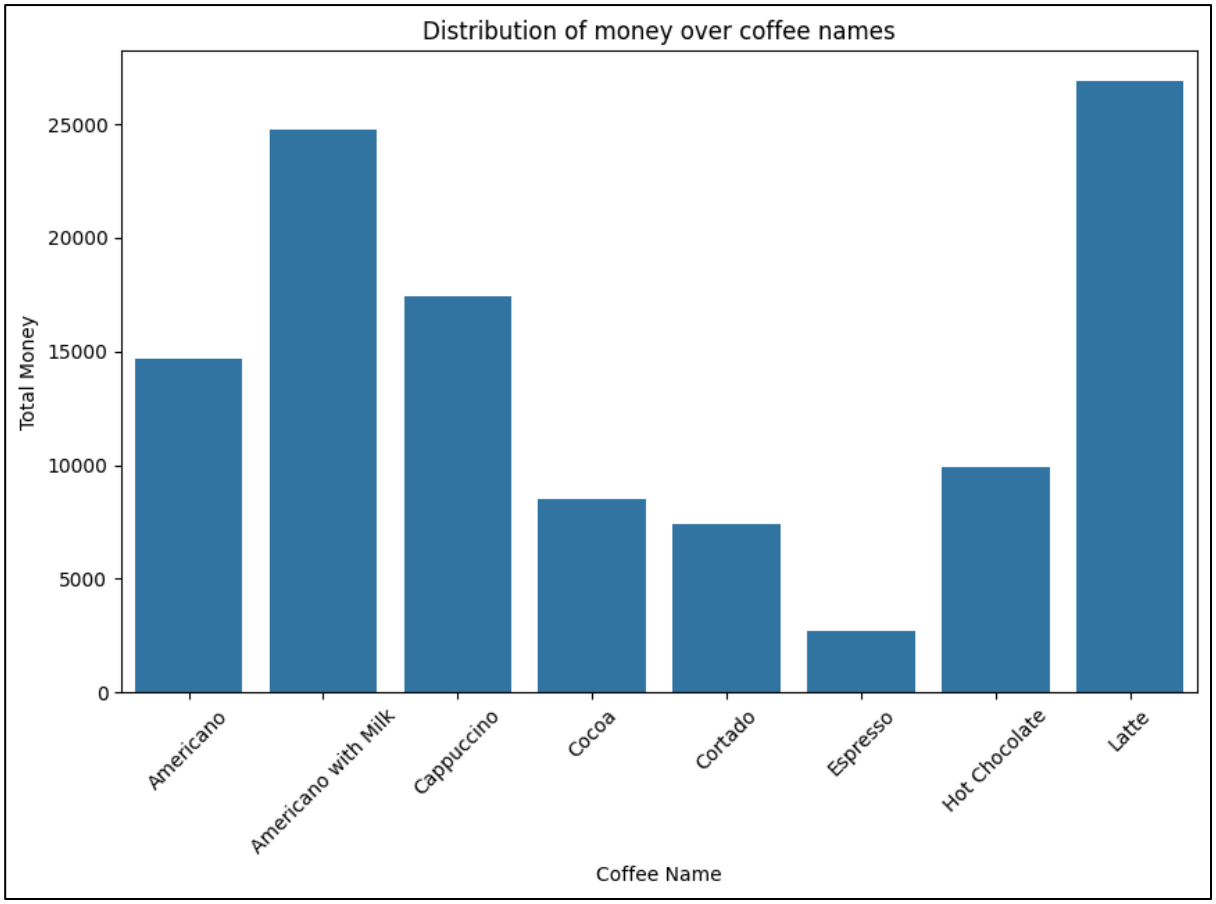
5. Data Analysis and Results :

Data Set Overview :

Parameter	Value
Number of Columns	11
Duplicate Columns	None
Missing Values	0
Data Type Adjustments	Converted Date to datetime; created Month, Year

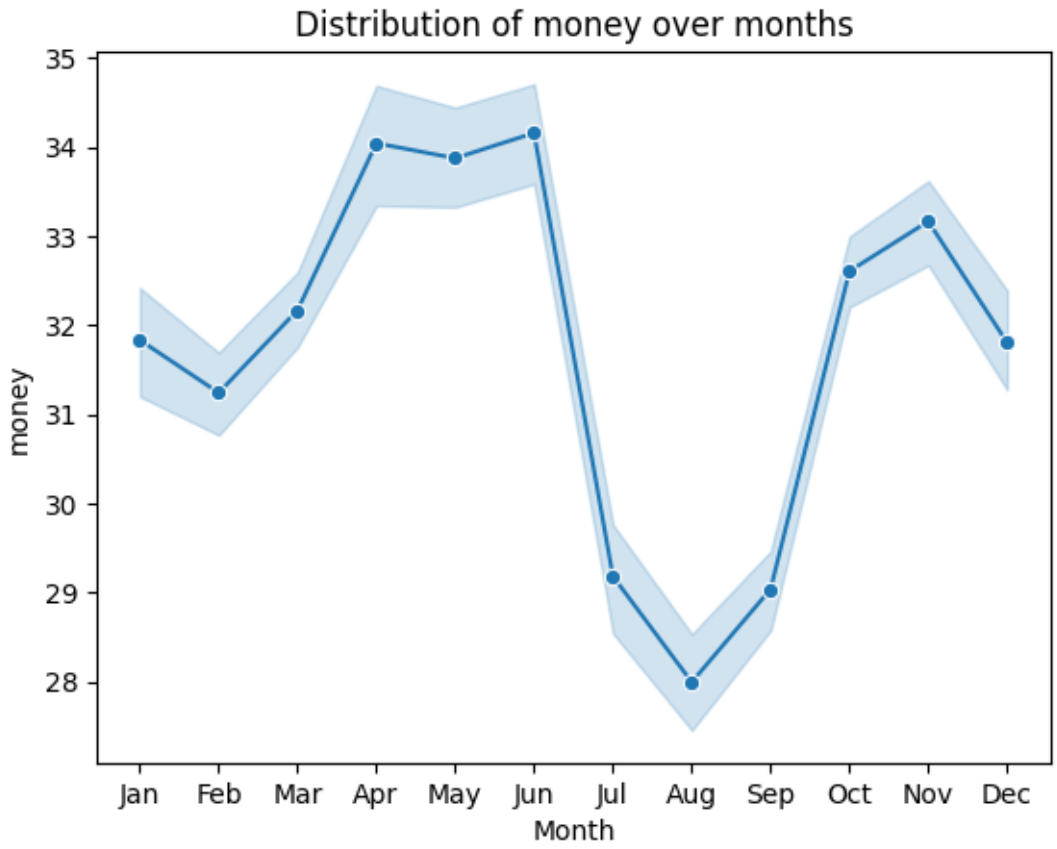
Coffee Type Distribution:

Coffee Name	Count
Americano with Milk	809
Latte	757
Americano	564
Cappuccino	486
Cortado	287
Hot Chocolate	276
Cocoa	239
Espresso	129



Maximum Money for Each Month **:**

Month	**Max Money**
Jan	35.76
Feb	35.76
Mar	38.70
Apr	38.70
May	37.72
Jun	37.72
Jul	37.72
Aug	32.82
Sep	35.76
Oct	35.76
Nov	35.76
Dec	35.76



Average Money by Time of Day:

Time of Day	Average Money
Morning	30.42
Afternoon	31.64
Night	32.89

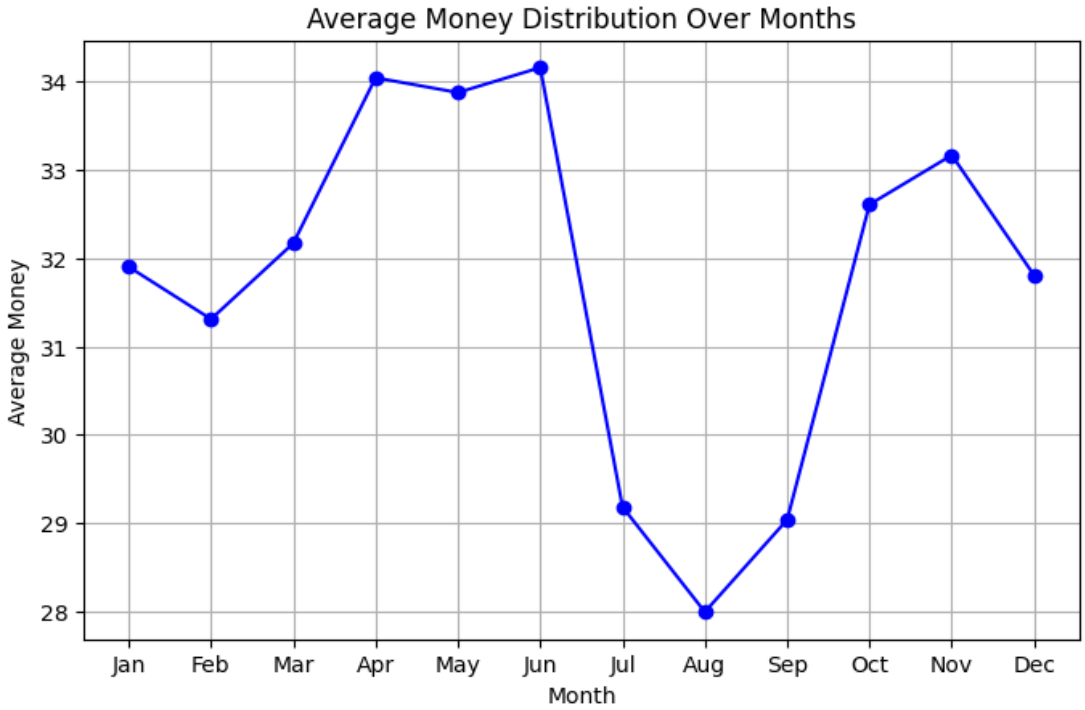
Visualizing with synthetic data :

This section examines the coffee sales dataset with 13 well-structured columns, containing no missing or duplicate values. Synthetic data was included to enrich analysis. Key findings highlight yearly and monthly spending patterns, maximum revenue per coffee type, and time-of-day variations. Visualizations of trends and distributions provide deeper insight into customer behavior and sales performance.

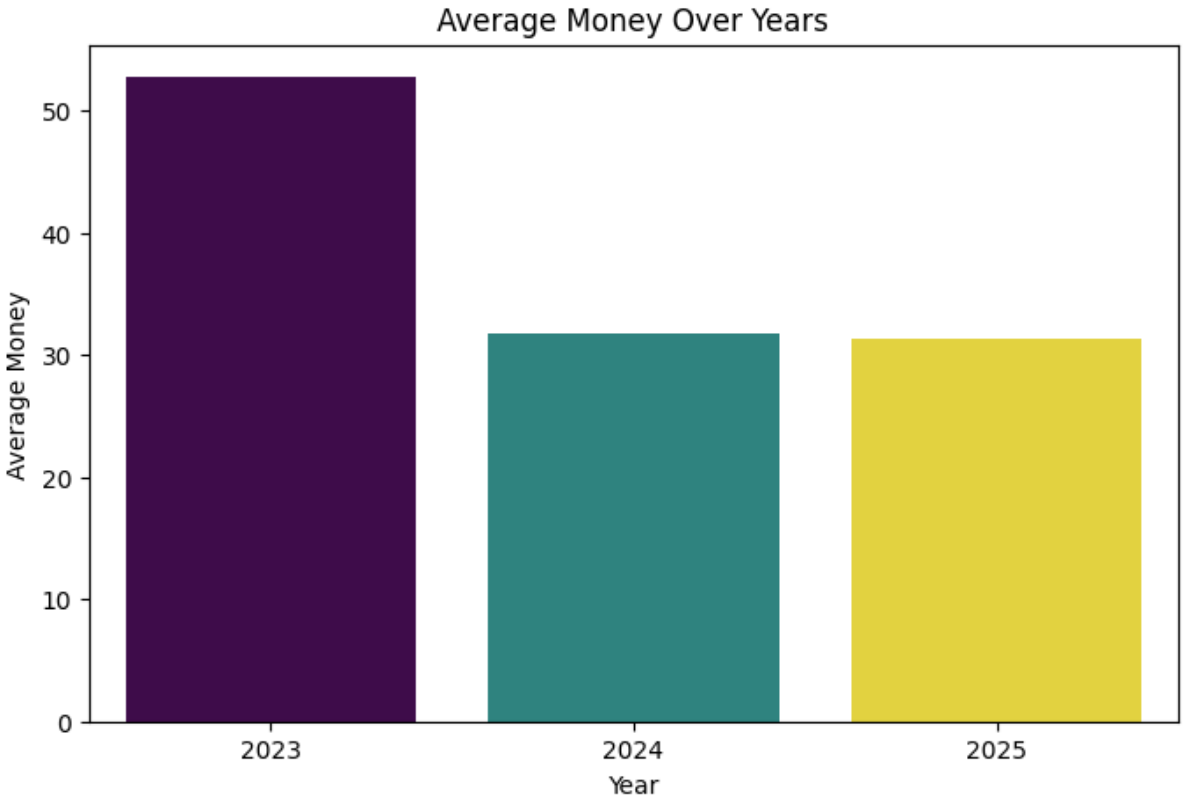
Here are the key analyzed results :

- Total number of columns : 13
- Missing values : 0
- Number of duplicates : 0

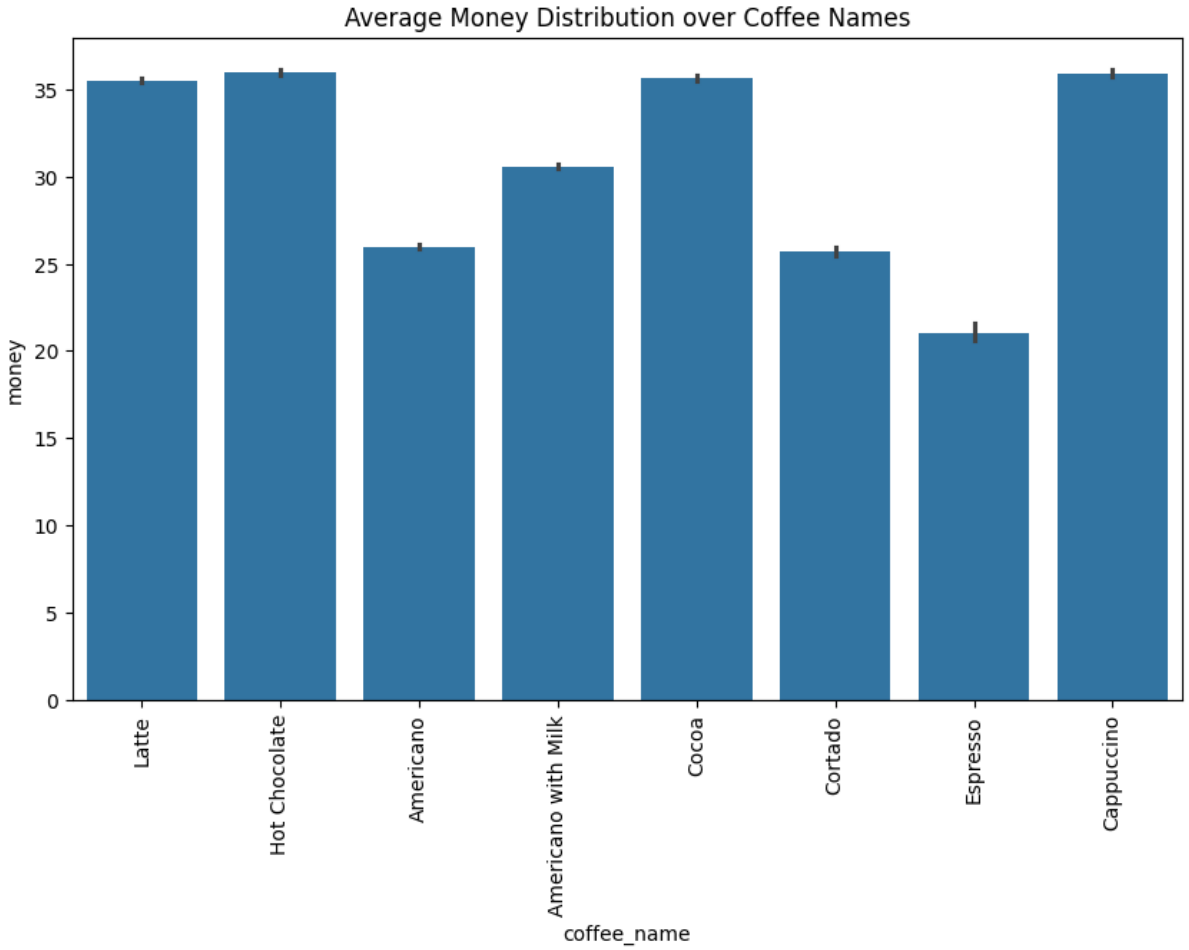
- **Average Money Distribution over months** :



- **Average Money Over Years** :



- **<u>Average Money Distribution over Coffee Names</u>** :



6. Conclusion :

The analysis of the coffee sales dataset using **NumPy, Pandas, Matplotlib, and Seaborn** provided meaningful insights into sales patterns, customer preferences, and time-based revenue generation. The findings revealed that *Americano with Milk* and *Latte* are the most popular beverages, highlighting a strong customer preference for milk-based coffees. Time of day analysis showed that sales were consistently higher at night compared to morning and afternoon, which indicates that customers tend to purchase more coffee during evening hours. Additionally, seasonal patterns emerged, with revenue peaking between March and May, suggesting that early spring months attract more sales.

Lastly, through this project, I have learned how to effectively use NumPy**,** Pandas**,** Matplotlib**,** and Seaborn for data cleaning, statistical analysis, and visualization. I successfully implemented these technologies to analyse the coffee sales dataset and derive meaningful insights.

7. APPENDICES

Appendix A : References

- Official Pandas Documentation: https://pandas.pydata.org/docs/
- NumPy Documentation: https://numpy.org/doc/
- Matplotlib Documentation: https://matplotlib.org/stable/index.html
- Seaborn Documentation: https://seaborn.pydata.org/

Appendix B : GitHub Link

- Repo Link : https://github.com/Banhivaroy/IDEAS_TIHub_Project_Submission

Appendix C : Supporting Documents

- Dataset Link:

Coffee_sales.csv