Preprocessing and Visualizing Coffee Sales Data

Ashaf Hasan Rami,

B.Sc. B.Ed. Economics | Indian Institute of Technology Bhubaneswar

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 report presents a data analysis project focused on preprocessing and visualizing coffee sales data to uncover sales patterns and customer preferences. The raw data sets contained attributes such as transaction time, hour of day, cash type, money, coffee name, date and month name. Preprocessing involved handling missing values, removing duplicates, converting date and time to usable formats and after that extracting the months and years out of that usable format. Descriptive statistics and exploratory visualization, including bar charts (using seaborn's bar plot function), line graphs (using seaborn's line plot function), were employed to identify monthly sales trends, year wise trend and popular coffee types. The results highlight key trends, including peak sales time of the day, average money for each year, distribution of money over months, and distribution of money over coffee names. The study demonstrates how systematic preprocessing combines with statistical visualization techniques can reveal actionable insights.

2. Introduction

a) **Background:**
Coffee is one of the most widely consumed beverages worldwide, and its sales generate large volumes of transactional data. This data includes information such as purchase time, product type, quantity, and revenue, which, if properly analysed, can reveal customer preferences and market trends. However, raw data is often incomplete, inconsistent, or unstructured, making data preprocessing a crucial step before meaningful analysis can take place. Furthermore, effective data visualization plays a key role in transforming complex datasets into clear and interpretable insights for decision-making.

b) **Relevance of the Study:**
In the competitive retail sector, understanding sales trends is essential for improving inventory management, revenue forecasting, and targeted marketing strategies. For coffee retailers, analysing sales data can highlight peak purchase times, the popularity of specific products, and seasonal demand fluctuations. These insights support data-driven business decisions, making this study relevant to both academia and industry.

c) Technology Involved

The project employs Python as the core programming language due to its strong data science ecosystem. The following tools and libraries were used:

Pandas for data preprocessing and feature engineering.

NumPy for numerical operations.

Matplotlib and Seaborn for statistical visualization of sales patterns.

Jupyter Notebook and Google Colab as the development environment for interactive coding and analysis.

c) **Procedure Used:**
The methodology followed in this project included:

1. Data Collection – Accessing the raw coffee sales dataset.

2. Preprocessing – Handling missing values, removing duplicates, converting date and time formats, and creating derived features (Years, Months).

3. Exploratory Data Analysis (EDA) – Using statistical summaries and visualizations to identify trends and anomalies.

4. Visualization – Employing bar plots and line plots to uncover patterns such as peak sales hours, seasonal variations, and product preferences.

d) Purpose of the Study:

The purpose of this project is to demonstrate how systematic preprocessing and visualization can transform raw coffee sales data into actionable business intelligence. By uncovering sales patterns and customer behaviour, the study aims to aid in operational planning, marketing strategies, and customer engagement. It also serves as a practical demonstration of data analysis techniques for academic and professional applications.

List of topics that I received training on during the first two weeks of internship:

1. Basics of Python: Data, Variables, Loop, Data Structures, Class, Functions.
2. Object Oriented Programming
3. Numpy, Pandas
4. Overview of Machine Learning, Regression Analysis and LLM Fundamentals
5. Communication Skills.

3)Project Objective

1. To preprocess the coffee sales dataset by cleaning, handling missing values, and generating new features such as years and months, ensuring data reliability.

2. To illustrate sales patterns and customer preferences through exploratory data analysis (EDA) and visualization techniques, such as line plots and bar charts.

3. To analyse money variations across different time periods (time of the day, monthly) and identify the most profitable product categories.

4. To demonstrate the role of data visualization in simplifying complex datasets and making insights more accessible for decision-making.

5. To highlight the business relevance of data analytics by showing how processed sales data can inform inventory management, marketing strategies, and customer engagement.

(No hypothesis testing or sample survey was conducted in this project; the analysis is based on secondary sales data.)

4) Methodology

1. Data Collection:
   The project is based on a secondary dataset of coffee sales records, which includes fields such as transaction date, time, coffee type, money against the coffee sold. The dataset was provided in CSV format and served as the raw input for analysis. No primary survey was conducted, and therefore, no questionnaire or sampling methodology was involved.
2. Tools and Technologies Used:
   To carry out the analysis, the following tools and technologies were employed:
   a) Python as the core programming language.
   b) Pandas for data preprocessing and cleaning.
   c) NumPy for numerical operations and calculations.
   d) Matplotlib and Seaborn for data visualization.
   e) Jupyter Notebook / Google Colab as the development environment for interactive coding and documentation.

3. Data Preprocessing Steps:

The raw dataset required systematic preprocessing to ensure reliability and usability.

The steps included:

1. Importing the dataset using google.colab’s files command: However, pandas can also be used for it but as importing data set by using this command makes the available at the colab’s workspace. Using pandas made the file readable.

2. Initial inspection of dataset shape, column names, and data types: inspected for missing values and duplicate columns or rows and none of them were found to be missing duplicated respectively.

3. Our next step could have been handling missing values by either removing incomplete rows or imputing values where appropriate and. Removing duplicates to ensure accuracy in analysis but as we did not have and duplicated rows or column nor did we have and missing values so it was unnecessary.

4. Figuring out basic statistics of the data: Descriptive statistics: mean, median, maximum, minimum, and standard deviation of sales and money.

5. Checking for the data types of the variables in the data set.

6. Converting date and time columns into appropriate datetime formats for temporal analysis.

7. Extracting day, month, and year from date.

8. Visualization methods:

a) Line plots for time-series trends (monthly sales & coffee wise money distribution).

b) Bar charts for comparing year wise money density.

9) Code Availability: GitHub

(No machine learning model was developed in this project, as the focus was on EDA and visualization.)

Data Analysis and Results

1) Descriptive Analysis:

The dataset consists of 3,547 observations across four key variables: Hour of Day, Money, Weekday Sort, and Month Sort. A summary of descriptive statistics is provided below**.**

	Hour of Day	Money	Weekday Sort	Month Sort
Count	3547.000000	3547.000000	3547.000000	3547.000000
Mean	14.185791	31.645216	3.845785	6.453905
Std	4.234010	4.877754	1.971501	3.500754
Min	6.000000	18.120000	1.000000	1.000000
25%	10.000000	27.920000	2.000000	3.000000
50%	14.000000	32.820000	4.000000	7.000000
75%	18.000000	35.760000	6.000000	10.000000
max	22.000000	38.700000	7.000000	12.000000

Hour of Day:

- The observed hours range from 6:00 to 22:00, covering typical daily activity periods.
- The mean hour is 14.18 (≈2 PM), suggesting that activity (or data collection) is concentrated around the afternoon.
- The distribution is fairly spread out (Std = 4.23), with most activity between 10:00 (25th percentile) and 18:00 (75th percentile).

Money:

- The monetary variable ranges between 18.12 and 38.70, with an average value of 31.65.
- The 50th percentile (median) is 32.82, which is slightly higher than the mean, suggesting a slight left-skew (a few lower values pull the mean down).
- The variability is moderate (Std = 4.87).

2) Exploratory Data Analysis (EDA):

a) Average Money for Each Year:

- The average monetary value in **2025 is slightly lower** than in 2024, showing a marginal decrease of about **0.35 units**.

- While the difference is small, it may indicate a minor downward shift or fluctuation in the observed values between the two years.
- Given the relatively stable averages across both years, the overall distribution of Money appears **consistent**, without drastic year-to-year variation.

Year	Money
2024	*31.737634*
2025	*31.390011*

b) Maximum Money for Each Month:

- **Peak Values (March–April 2024):**
  - The highest observed value of Money is **38.70**, recorded in both **March and April 2024**.
  - This indicates a **seasonal peak in early spring** (Q1–Q2).
- **High Values (May–July 2024):**
  - The following three months (May–July 2024) show slightly lower but still elevated maximum values (**37.72**).
  - This suggests a sustained **high-activity period during late spring to midsummer**.

Year	Month	Month Name	Money
2024	**3**	**Mar**	**38.70**
	4	**Apr**	**38.70**

Year	Month	Month Name	Money
	5	May	37.72
	6	Jun	37.72
	7	Jul	37.72
	8	Aug	32.82
	9	Sep	35.76
	10	Oct	35.76
	11	Nov	35.76
	12	Dec	35.76
2025	1	Jan	35.76
2025	2	Feb	35.76
2025	3	Mar	35.76

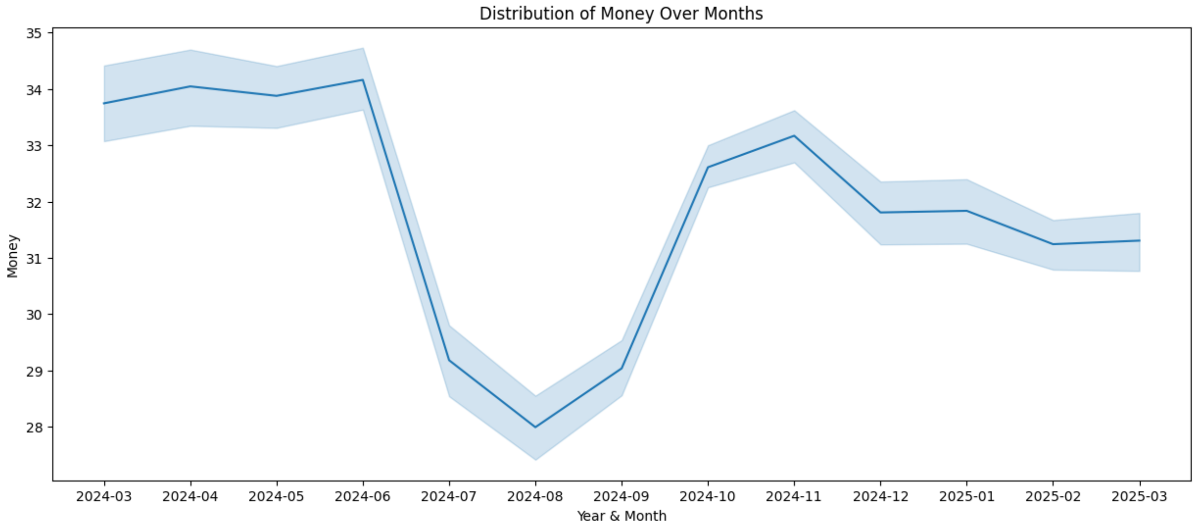
- Moderate Values (August–December 2024):
  - From August to December 2024, maximum Money stabilises around 32.82–35.76.
  - This indicates relatively moderate fluctuations in the second half of the year.
- Stable Early 2025:
  - For January–March 2025, the maximum values remain constant at 35.76, suggesting a plateau or stabilization compared to the higher peaks of 2024.
- Seasonal Pattern:
  - The pattern highlights a clear rise in maximum values during March–July, followed by a decline and stabilisation in later months.
  - This could point to seasonal effects or demand cycles, depending on the context of the dataset.

c) Distribution of Money Over Months:

- Summer Months (Mar-Jun):

- Money values start at a moderate level and show a gradual increase, peaking in **Jun (33.5-34.5)**.
- This suggests that the first quarter experiences higher spending activity.

- Autumn Months (Jul-Aug):
  - A sharp decline over these months can be seen.
  - This indicates a low spending over coffee by consumers.

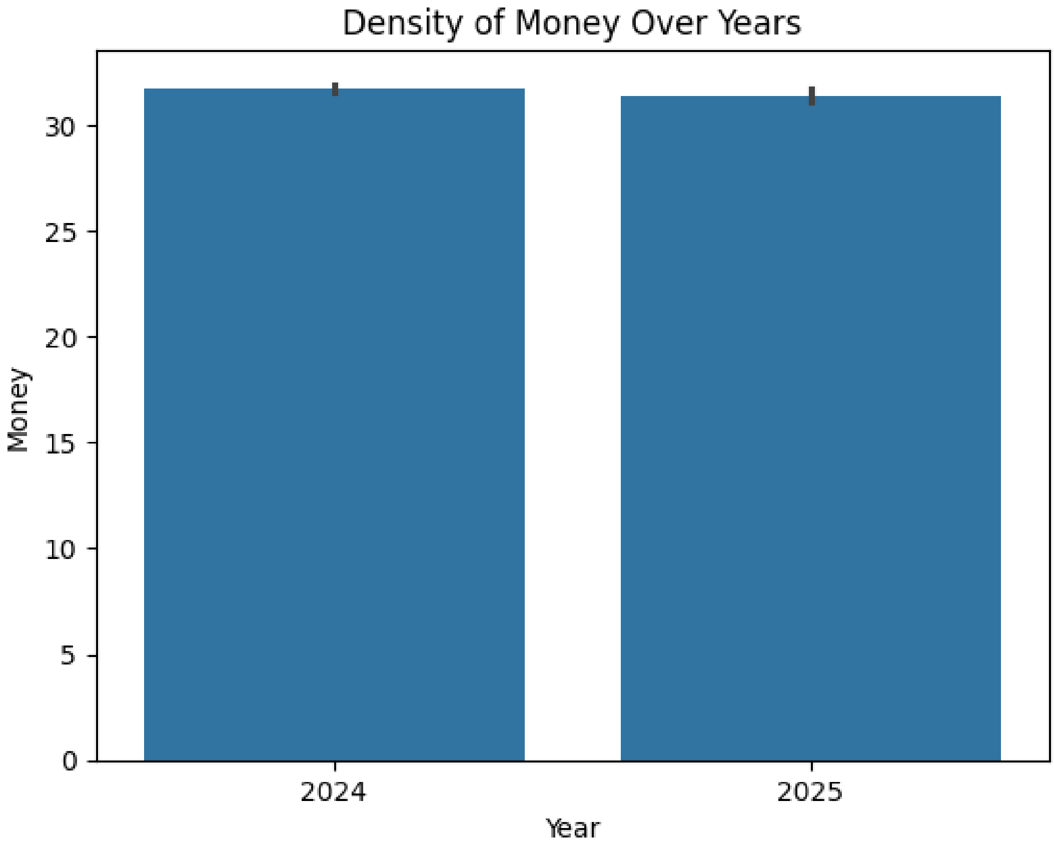


- Late Summer–Autumn (Aug–Oct):
  - After the decline again the spending on coffee is increasing.
  - This dip and recovery could be attributed to seasonal effects or reduced demand during this period and then increasing demand.

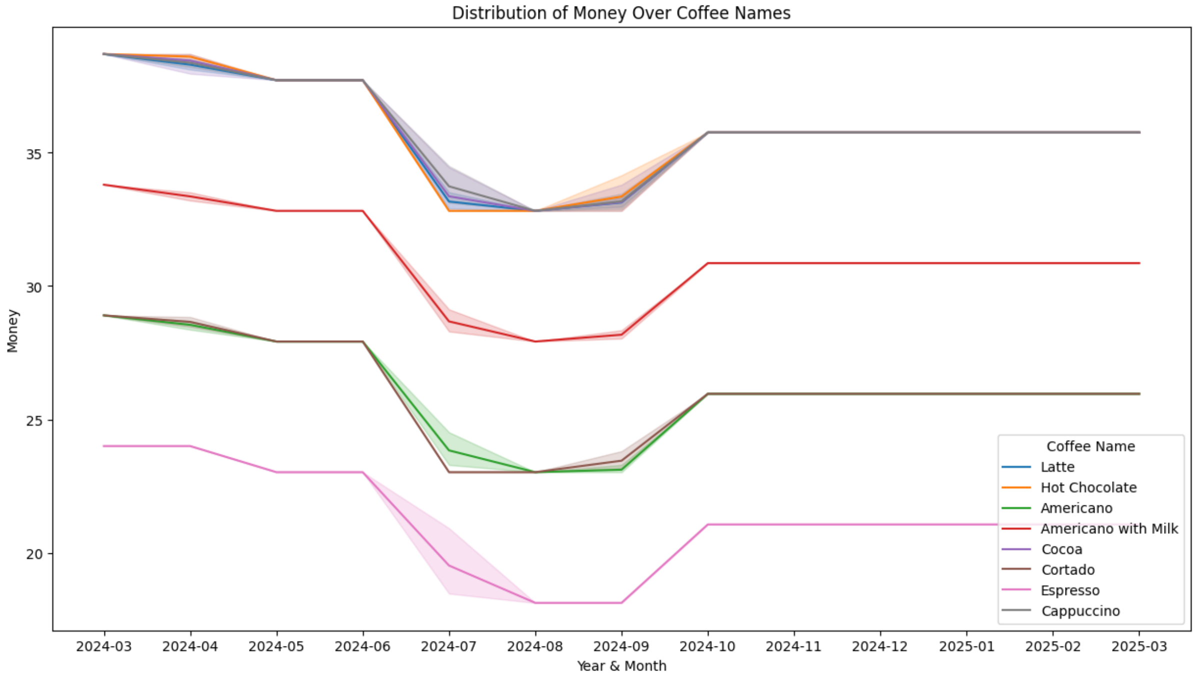
- Year-End (Nov–Dec):
  - A recovery can be seen in consumer spending over coffee but after late October it again seems to be declining at a lower rate this period’s level is consistent but slightly lower level compared to the March–July peak.

d) The Density of Money Over Years:

- Money density for both the years are nearly at the same level.
- However, year 2024 is slightly high in energy density but we can’t conclude anything over this small difference.



e) Distribution of Money Over Coffee Names:



- Most coffee names experience a sharp decline in money around June–July 2024, followed by a gradual recovery or stabilization towards early 2025.
- Latte, Hot Chocolate, Americano, Americano with Milk, Cappuccino: These start at a high level, drop mid-2024, then return to their initial levels by late 2024 and remain stable.
- Espresso: Maintains the lowest money values, dips significantly in June–July 2024, though it partially recovers thereafter.
- The sharp mid-2024 decline may indicate an external factor (e.g., pricing change, demand drop, seasonal influence) that impacted all coffee types simultaneously.

f) Number of Coffee Sold at Different Times of the Day:

- **Highest Sales in the Afternoon:**
  - The afternoon period records the largest number of coffee sales (**1205 cups**).
  - This suggests that customers are more likely to purchase coffee in the **post-lunch period**, possibly as an energy boost.
- **Morning Sales:**
  - Morning sales are slightly lower (**1181 cups**) but remain a significant portion of total sales.
  - This aligns with the common habit of drinking coffee as part of a **morning routine**.

Time of Day	Count
Afternoon	1205
Morning	1181
Night	1161

- Night Sales:
  - Nighttime sales are the lowest (1161 cups) but still relatively close to morning and afternoon figures.
  - This could indicate demand among people who work late hours or prefer coffee in the evening.
- Balanced Distribution:
  - Overall, the difference between the three time periods is small (less than 50 cups between the highest and lowest).
  - This shows a consistent demand for coffee across the day, with only slight variations.

g) Maximum Money from Coffee Name:

- Premium Group: Cappuccino, Cocoa, Hot Chocolate, and Latte all have the highest money values (38.7), likely representing top-tier or specialty menu items favoured for their ingredients or preparation style.
- Entry Level: Espresso is the lowest at 24.0, reflecting its basic, unadorned nature and appeal to budget or purist consumers.

Coffee Name	Money
Americano	28.9
Americano with Milk	33.8
Cappuccino	38.7
Cocoa	38.7
Cortado	28.9
Espresso	24.0
Hot Chocolate	38.7
Latte	38.7

h) Average Money Made at Different times of the Day:

- **Highest Average at Night:**
  - The **night period records the highest average money (32.89)**, suggesting that even though the number of cups sold at night is slightly lower, customers tend to spend **more per transaction**.
  - This could be due to a preference for larger or premium drinks in the evening.
- **Afternoon Sales:**
  - The afternoon shows a strong average (**31.64**), reflecting both a high volume of sales and consistent spending patterns.
  - This confirms the afternoon as a peak period in terms of overall revenue generation.
- **Morning Sales:**
  - Morning records the lowest average money (**30.42**).
  - This may reflect customer habits of choosing **smaller or standard coffees** during the morning rush rather than more expensive options.

Time of Day	Money
Afternoon	31.643187
Morning	30.422693
Night	32.890904

5) Conclusion

The analysis of the coffee sales dataset demonstrates that systematic data preprocessing and visualization techniques are highly effective for uncovering meaningful business insights from raw transactional data. By handling missing values, duplicates, and re-formatting date and time attributes, the project ensured reliability in the results and extracted important features for subsequent analysis.

Exploratory data analysis identified distinctive sales patterns, including seasonal peaks in the early spring and sustained high activity through midsummer, followed by moderate declines and stabilizations into the next year. The study revealed that premium coffee varieties such as Cappuccino, Cocoa, Latte, and Hot Chocolate consistently generate the highest monetary values, while Espresso retains its position as the most accessible option, aligning with common consumer preferences.

Afternoon hours marked the peak for total sales, but spending per transaction was highest during night hours, suggesting underlying differences in consumer buying behaviours based on time of day. The near-uniform money densities across years and moderate month-to-month variations suggest a stable market environment, with only minor shifts likely driven by external influences or seasonal effects.

The project demonstrates how leveraging comprehensive data analysis and visualization can reveal critical patterns within coffee sales, enabling businesses to make informed decisions that optimize resource allocation, target customer preferences, and respond effectively to evolving market trends.

6) Appendices

1. GitHub link - https://github.com/RamiBabu347/Autumn-Internship-Program-on-Data-Science-IDEAS-TIH-ISI-..git