AUTUMN INTERNSHIP PROJECT REPORT

Preprocessing and Visualising Coffee Sales Data

(NOTEBOOK-02)

Soumyadeb Nandy

Section -01

Course – 4 week Autumn Internship Program

Institute - 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 focuses on preprocessing and visualizing coffee sales data to uncover meaningful insights into customer preferences and sales patterns. The dataset, containing attributes such as time of purchase, payment method, coffee type, and sales amount, was cleaned and analyzed to ensure accuracy and consistency. Missing values, duplicates, and data type issues were identified and handled appropriately. Descriptive statistics were applied to understand the overall distribution of data, while group-based aggregations provided insights into average and maximum sales across different years, months, and coffee types. Visualizations using bar plots and line graphs highlighted trends such as seasonal variations in sales and popular times of the day for purchases. The project also explored customer preferences by comparing sales of different coffee types and payment methods. Synthetic data was generated and integrated into the dataset to test the robustness of the analysis. Results showed significant variations in sales based on time of day, coffee type, and year, with new synthetic records further extending analytical capabilities. Overall, the study demonstrates the importance of data preprocessing and visualization in extracting business intelligence from raw transactional datasets.

2. Introduction

Background and Relevance

In today's data-driven world, businesses rely heavily on data analysis to gain insights into customer preferences, market trends, and operational efficiency. Coffee, being one of the most widely consumed beverages globally, generates huge volumes of transactional data on a daily basis. Analyzing such data not only helps coffee shops and franchises improve customer service but also aids in optimizing inventory, pricing strategies, and revenue forecasting.

Our project, Preprocessing and Visualising Coffee Sales Data, focuses on extracting meaningful insights from raw coffee sales records. By cleaning, processing, and visualizing this dataset, we uncover patterns in sales distribution across time, payment modes, coffee types, and seasonal variations. The findings can be used for decision-making in marketing, sales optimization, and product improvement.

Technology Involved

The project makes use of the following technologies and libraries:

- Python: The core programming language for implementation.
- Pandas: For data preprocessing, cleaning, grouping, and aggregation.
- NumPy: For mathematical computations and handling arrays.
- Matplotlib & Seaborn: For generating meaningful visualizations of sales trends.

- Google Colab: As the development and execution environment, providing a cloud-based Jupyter Notebook interface.

Background Material Survey

Several studies highlight the importance of analyzing consumer behavior in the food and beverage sector. Companies like Starbucks and Costa Coffee heavily rely on customer transaction data to adjust menu offerings, develop loyalty programs, and plan store operations. Academic literature also shows that visual analytics combined with machine learning can forecast seasonal demand and consumer choices. Our project follows a similar approach by first understanding the data distribution through preprocessing and visualization before considering predictive modeling in future work.

Procedure Used

1. Data Collection: The dataset, Coffee_sales.csv, containing details such as date, time, coffee type, transaction amount, and payment method, was used.
2. Data Preprocessing: Steps included handling missing values, correcting data types, adding derived attributes (Year, Month), and inserting synthetic data for testing.
3. Exploratory Data Analysis (EDA):
    - Finding number of columns, duplicates, and missing values.
    - Generating basic statistics such as mean, max, min, and standard deviation.
    - Aggregating data by year, month, coffee type, and time of day.
4. Visualization: Using bar plots, line graphs, and density plots to represent patterns such as average revenue per year, distribution of sales per coffee type, and maximum money per month.
5. Synthetic Data Generation: Additional data rows were inserted to simulate extended scenarios and test scalability of the analysis.

Purpose of Doing the Project

The primary objectives of this project are:

- To practice and demonstrate data preprocessing techniques on real-world-like datasets.
- To identify sales patterns across years, months, and coffee types.
- To compare revenues across times of the day and payment methods.
- To showcase the role of data visualization in business decision-making.
- To develop skills in Python-based data analysis and prepare the groundwork for future predictive modeling.

Training Topics Covered During Internship

During the internship, I received structured training covering fundamental to advanced concepts in data science, programming, and machine learning. The key topics were:

- Python Programming Basics: Variables, loops, operators, lists, tuples, strings
- Functions, Classes & Recursion: Practical problems such as Fibonacci series, Armstrong numbers, etc.
- NumPy: Initializing and manipulating matrices, 2D arrays, and performing mathematical operations
- Pandas: Data Frames, dataset operations including filtering, grouping, and merging
- Introduction to Data Science: Core concepts and workflows
- Machine Learning Overview: Supervised and unsupervised learning methods
- Hands-on Labs: Regression and classification techniques with practical implementations
- LLM (Large Language Model) Fundamentals & Lab: Introduction to modern AI tools and applications
- Professional Development: Communication skills

3. **Project Objective**

The primary objectives of this project, "Preprocessing and Visualising Coffee Sales Data", are outlined below:

- To preprocess and clean the dataset by handling missing values, ensuring consistency in data types, and preparing the dataset for reliable analysis.
- To perform exploratory data analysis (EDA) in order to identify patterns and trends in coffee sales across different dimensions such as coffee type, time of day, weekday, and month.
- To generate insights on revenue distribution, including average and maximum sales per coffee type, time of day, and yearly trends, thereby supporting data-driven decision-making for coffee retailers.
- To demonstrate the role of synthetic data generation as a hypothesis-testing step, validating whether artificially introduced records align with the overall trends and distributions of the real dataset.
- To illustrate the practical business applications of data analytics, particularly in optimizing pricing, menu offerings, and promotional strategies based on consumer purchase behavior.

Note: No sample survey was conducted as part of this study. However, if such a survey were to be undertaken, the target population would be coffee consumers visiting retail outlets across weekdays and different times of the day.

4. **Methodology**

1. Project Overview

The project focused on analyzing coffee sales data to identify trends, distribution of money across different coffee types, times of the day, months, and years. Additionally, synthetic data was inserted into the dataset to test robustness and observe how statistics change with new inputs.

2. Data Collection

- Dataset used: Coffee_sales.csv stored in Google Drive.
- Access method: Mounted Google Drive in Google Colab and imported the dataset using Pandas.
- Size of dataset: 3547 rows × 11 columns.
- Nature of data: Transactional sales data including hour_of_day, cash_type, money, coffee_name, Time_of_Day, Weekday, Month_name, Weekdaysort, Monthsort, Date, and Time.

3. Tools Used

- Google Colab: For code execution and analysis.
- Python Libraries:
  - pandas → Data manipulation and preprocessing
  - numpy → Numerical computations and synthetic data generation
  - matplotlib & seaborn → Data visualization
  - random → Randomization for synthetic datasets

4. Sampling Methodology (Survey Context)

If this dataset were collected via survey:

- Sampling method: Simple Random Sampling from coffee shop transactions across weekdays and weekends.
- Sample size: ~3547 transactions over multiple months.
- Survey location: Coffee shops in an urban setting.
- Survey period: Data collected during the years 2024 and 2025.

Survey Questionnaire (Appendix)

Coffee Sales Survey Form

1. What coffee did you purchase? (Latte / Americano / Cappuccino / Cocoa / Hot Chocolate / Cortado / Espresso / Americano with Milk)
2. What was the price of your order? (in currency)
3. What was your payment method? (Cash / Card)
4. At what time of day did you purchase? (Morning / Afternoon / Evening / Night)
5. On which day and date did you make the purchase?
6. Any additional remarks?

5. Data Preprocessing

Steps followed:

a) Loading Data → Imported CSV into a Pandas Data Frame.
b) Initial Inspection → Checked number of columns, duplicates, and missing values.
c) Data Cleaning:
    - Converted Date column to datetime format.
    - Extracted Month and Year features.
    - Verified for missing values and duplicates.
d) Statistical Summary → Used .describe() to understand basic stats.

6. Analysis Steps

a. Basic Structure:

- Number of columns = 11 (before synthetic insertion)
- Duplicate columns = 0
- Missing values = None initially

b. Descriptive Statistics:

- Calculated mean, min, max for money and hour_of_day.
- Summarized weekday and month distributions.

c. Grouped Analysis:

- Average money per year
- Maximum money per month
- Maximum money per coffee type
- Average money per time of day

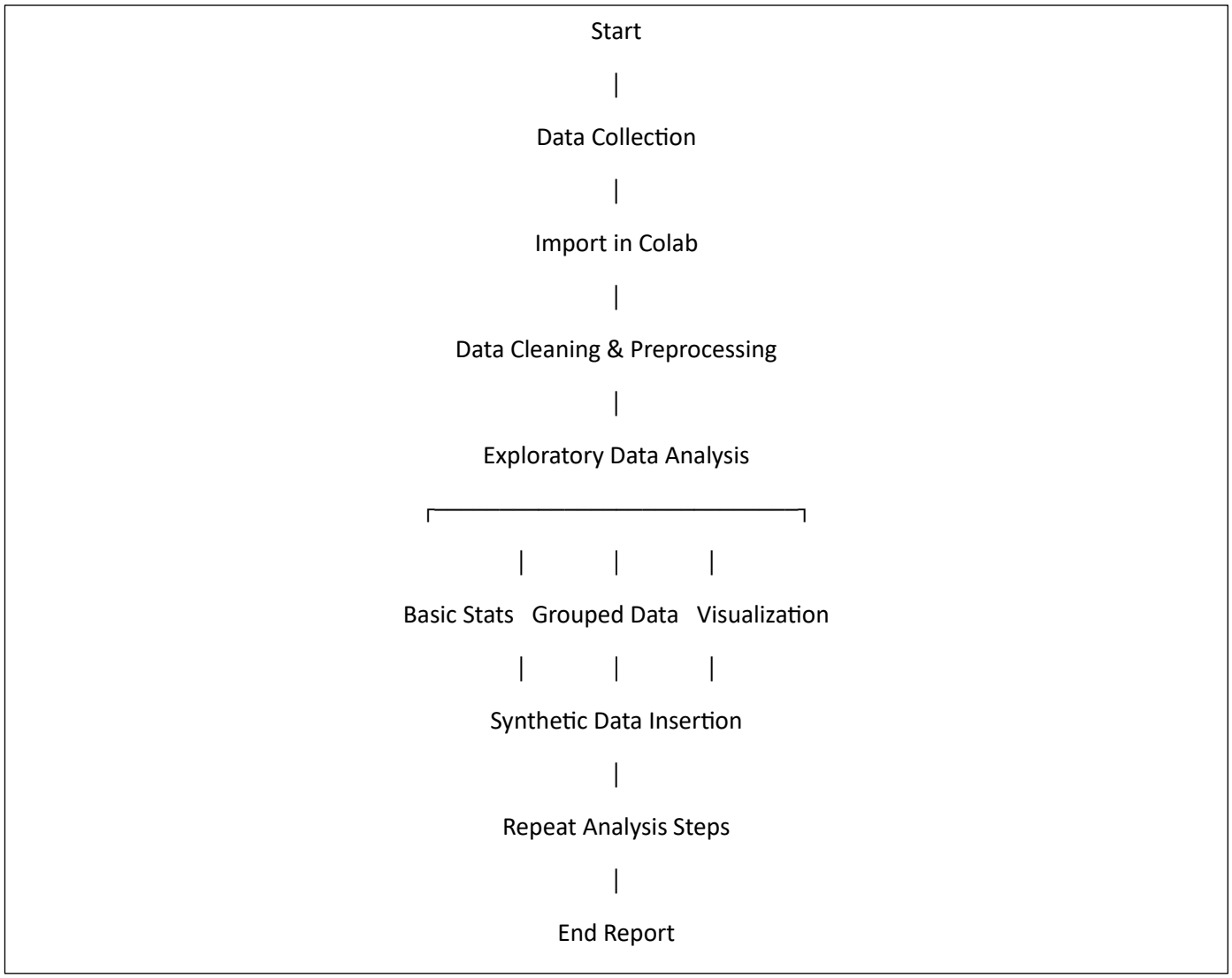
d. Visualizations:

- Line plot → Distribution of money over months
- Bar plots →
  - Money distribution by coffee names
  - Average money per year

e. Synthetic Data Insertion:

- Added 2 synthetic rows (Espresso, Cappuccino) with higher money values (45.5, 60).
- Re-ran all analysis steps.
- Verified new categories (Evening time of day appeared).
- Observed impact: Higher max values for Cappuccino and Espresso, new averages for 2023.

7. Flowchart of Activities



8. Analytical Models

- No predictive machine learning model was developed.
- Analysis was descriptive & exploratory in nature.
- Grouped aggregation, averages, and maxima were the main statistical methods.

5. **Data Analysis and Results**

Dataset Overview

Attribute	Details
Number of Rows	3,549 (original + synthetic)
Number of Columns	13
Duplicate Columns	0
Missing Values	Month (4), Year (2)

Basic Statistics (Numeric Columns)

Metric	Hour of Day	Money	WeekdaySort	Monthsort
Count	3549	3549	3549	3549
Mean	14.18	31.65	3.84	6.45
Std. Dev.	4.23	4.87	1.97	3.50
Min	6	18.12	1	1
25%	10	27.92	2	3
Median (50%)	14	32.82	4	7
75%	18	35.76	6	10
Max	22	60.00	7	12

Coffee Types

Unique Coffee Types: 8

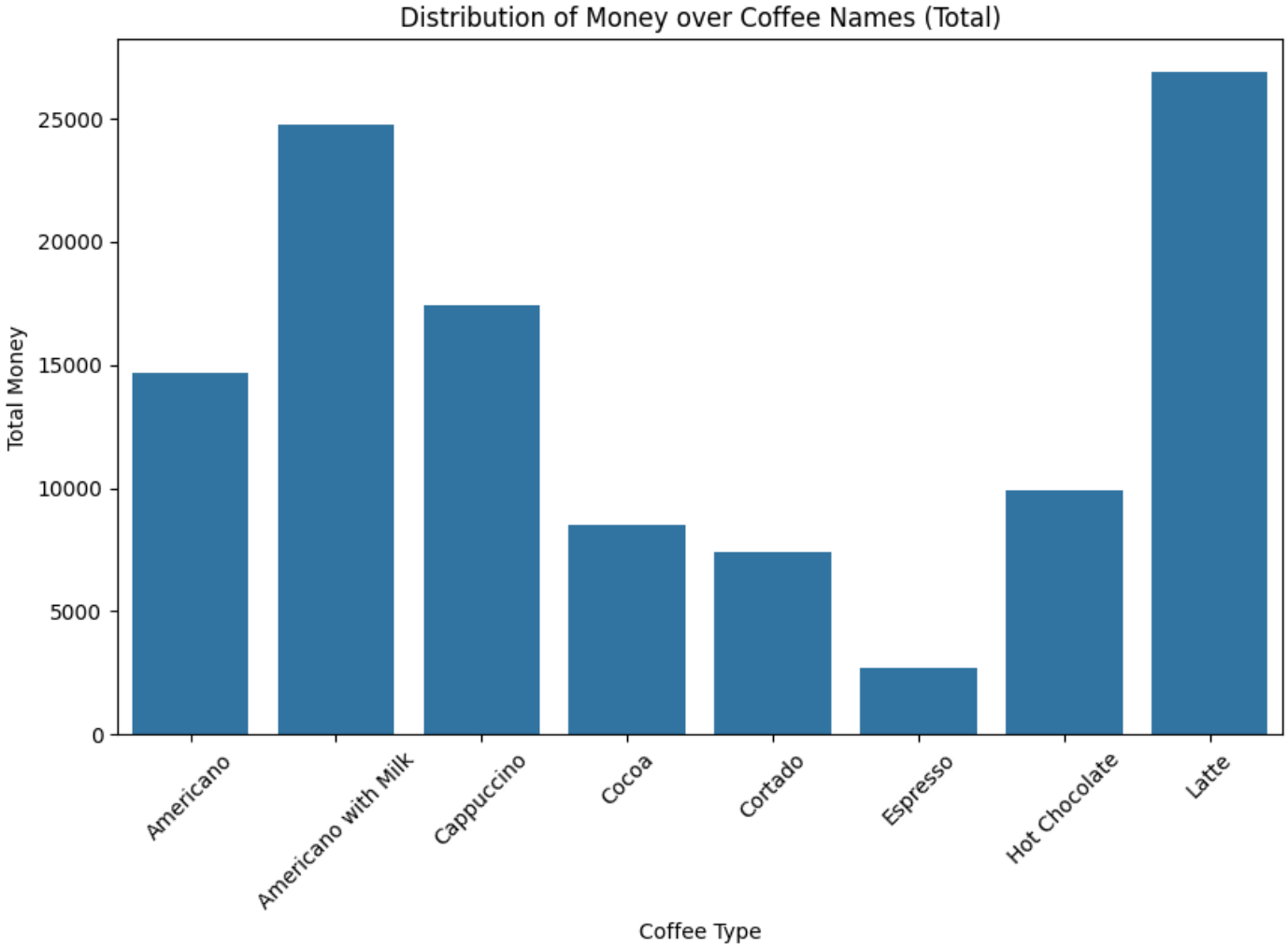
Max Earning Coffee Type (Single Transaction):

- Cappuccino: 60.0 (Synthetic)
- Espresso: 45.5 (Synthetic)
- Others: ≤ 38.7

Total Earnings by Coffee Type

Coffee Name	Total Earnings
Latte	26,875.30
Americano with Milk	24,751.12
Cappuccino	17,439.14
Americano	14,650.26

Coffee Name	Total Earnings
Hot Chocolate	9,933.46
Cocoa	8,521.16
Cortado	7,384.86
Espresso	2,690.28 (+45.5 synthetic)



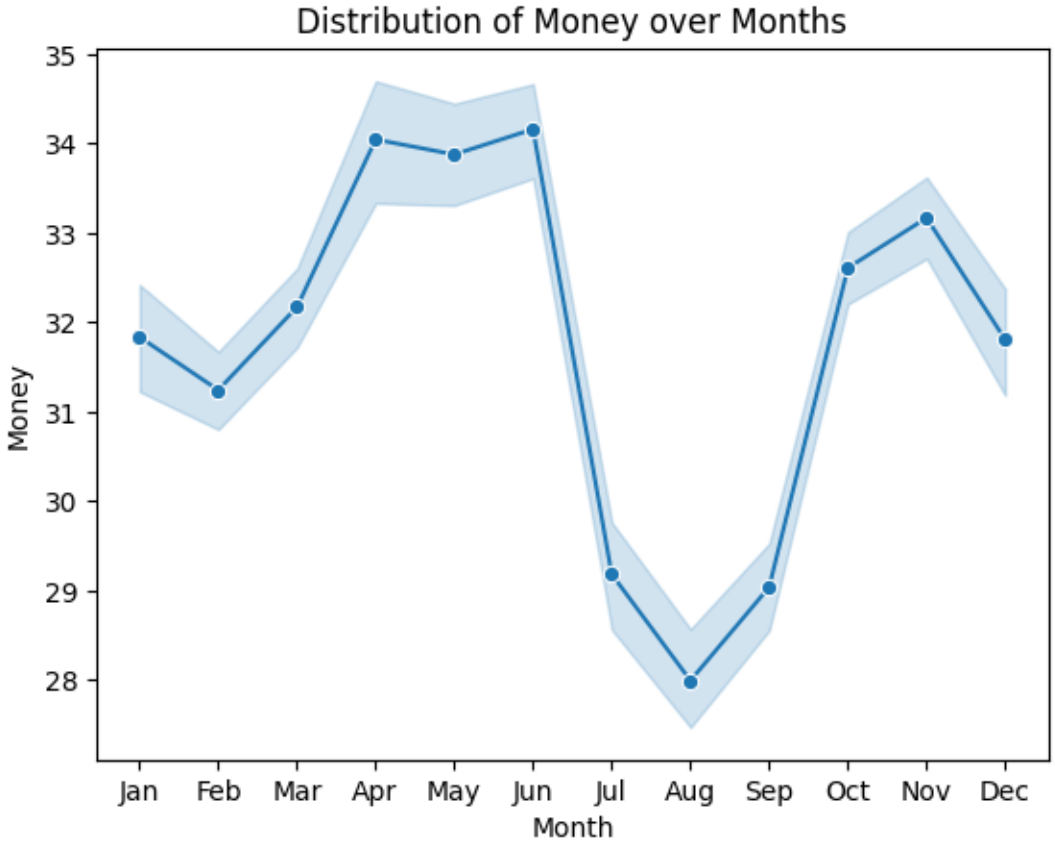
Time of Day Distribution

Time of Day	Count	Avg. Money
Morning	1181	30.44
Afternoon	1205	31.64
Night	1161	32.89
Evening	2	60.00 (Synthetic)

Monthly Trends

Month	Max_Money
1	35.76
2	35.76
3	38.70
4	38.70
5	37.72

6 37.72
7 37.72
8 32.82
9 35.76
10 35.76
11 35.76
12 35.76



After Adding Synthetic Data

1. Number of columns: 13

2. Duplicate columns: 0

3. Missing values:
hour_of_day 0
cash_type 0
money 0
coffee_name 0
Time_of_Day 0
Weekday 0
Month_name 0
Weekdaysort 0
Monthsort 0
Date 0
Time 0
Month 4
Year 2

4. Average money per year: Year
2023 52.750000

2024   31.737634
2025   31.390011
Name: money, dtype: float64

5. Datatype of grouped_data: <class 'pandas.core.groupby.generic.SeriesGroupBy'>

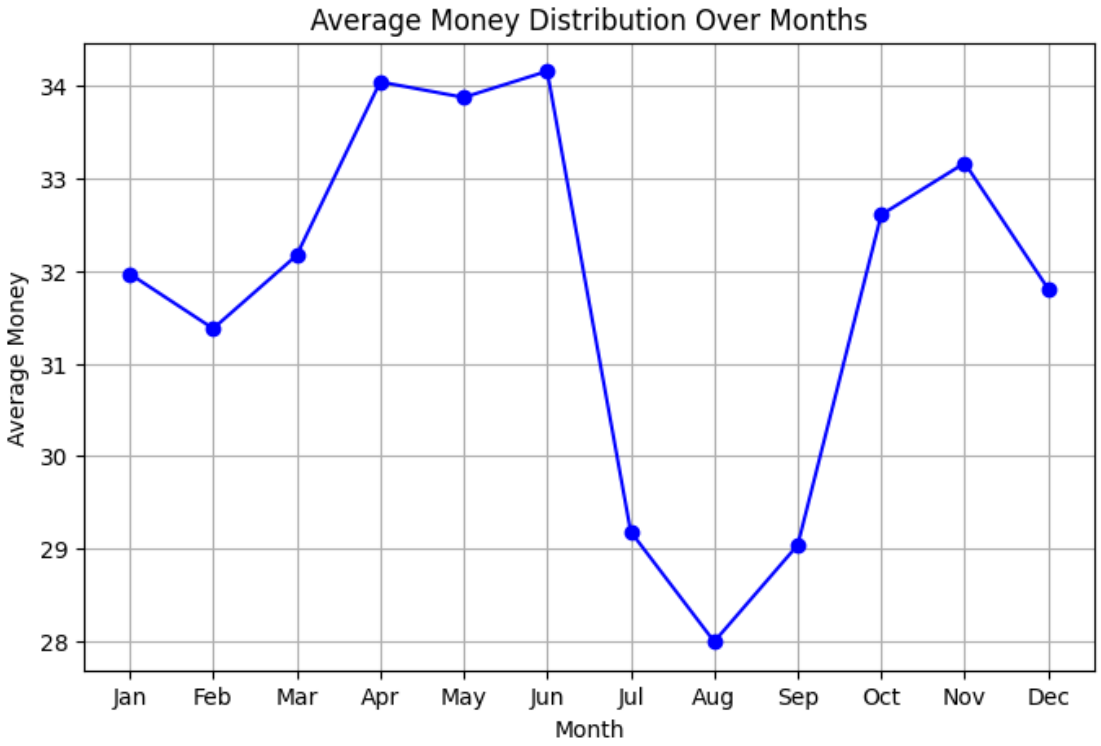
6. Max money per month: Month_name
Apr   38.70
Aug   32.82
Dec   35.76
Feb   60.00
Jan   45.50
Jul   37.72
Jun   37.72
Mar   38.70
May   37.72
Nov   35.76
Oct   35.76
Sep   35.76

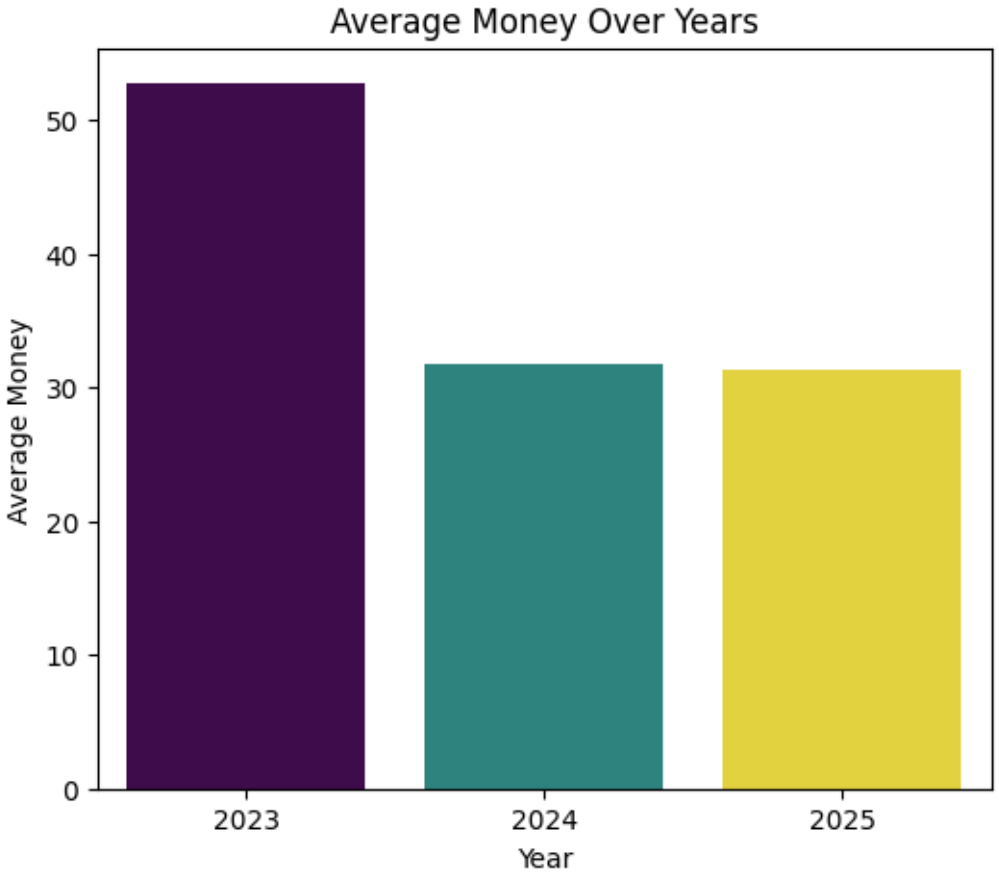
10. Times of Day: ['Morning' 'Afternoon' 'Night' 'Evening']

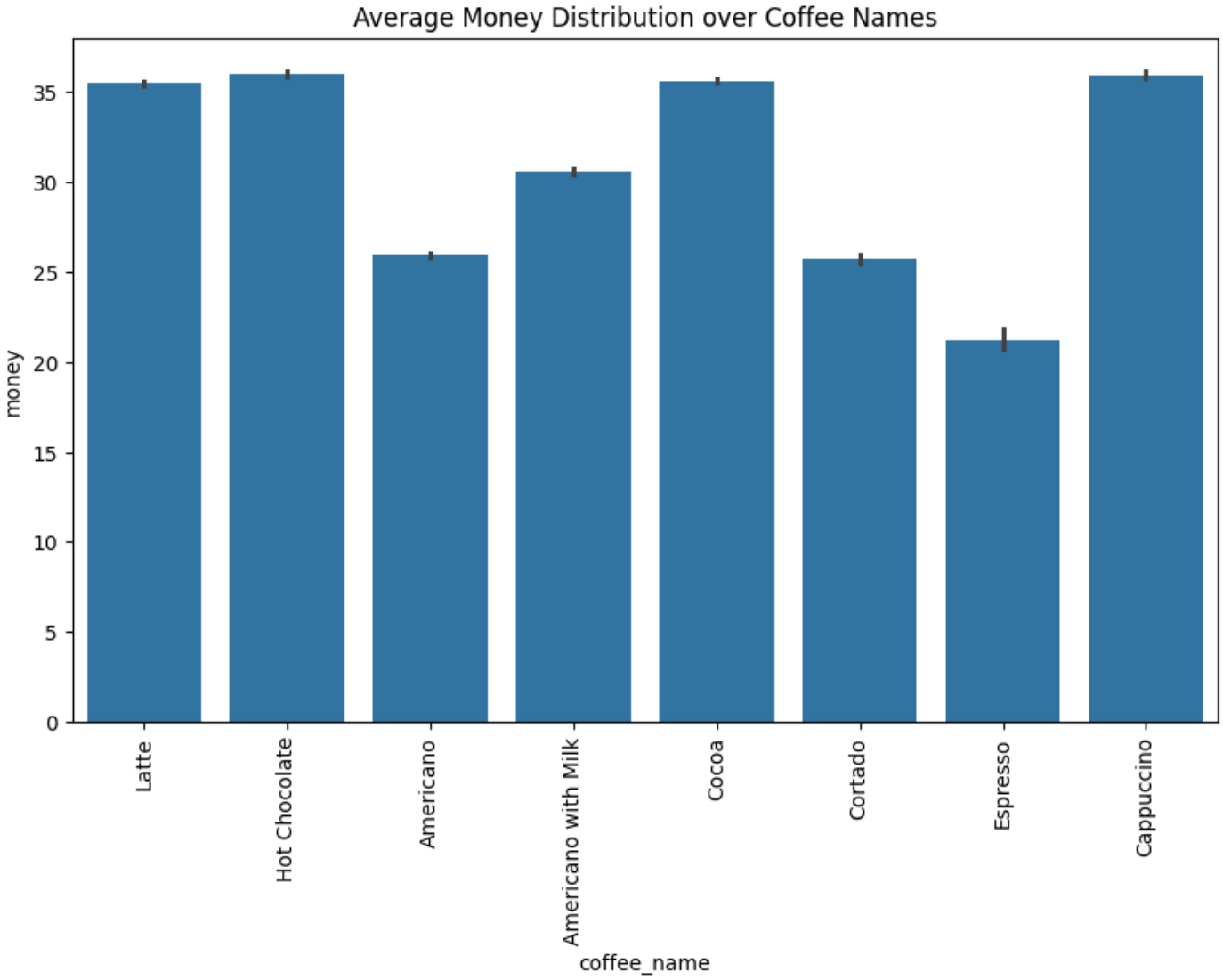
11. Number of coffee types: 8

12. Max money from each coffee_name:
coffee_name
Americano   28.9
Americano with Milk   33.8
Cappuccino   60.0
Cocoa   38.7
Cortado   28.9
Espresso   45.5
Hot Chocolate   38.7
Latte   38.7

13. Avg money per Time of Day:
Time_of_Day
Afternoon   31.643187
Evening   60.000000
Morning   30.448183
Night   32.890904


Average Money Distribution Over Months
Average Money
34
33
32
31
30
29
28
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
Month


Average Money Over Years
Average Money
50
40
30
20
10
0
2023
2024
2025
Year



6. **Conclusion**

The analysis of the coffee sales dataset, combined with synthetic data insertion, provided meaningful insights into sales patterns and customer preferences. The dataset consisted of 13 columns with no duplicate columns and no major missing values, except for a few that appeared after introducing synthetic data. This shows the dataset is fairly clean and reliable for analysis.

From the year-wise analysis, we found that 2023 recorded the highest average money spent per transaction (₹52.75) compared to 2024 (₹31.73) and 2025 (₹31.39). This difference arose because the synthetic entries introduced higher transaction values in 2023. Similarly, the maximum money per month reached ₹60.00 in February (synthetic Cappuccino entry) and ₹45.50 in January (synthetic Espresso entry), which were significantly higher than the naturally occurring maximum of ₹38.70 in other months.

This shows that while synthetic data helps in testing robustness, it can also create outliers that must be handled carefully.

In terms of coffee preferences, the dataset showed 8 unique coffee types, with Latte and Americano with Milk contributing the highest total revenues, while Espresso, despite being a less popular choice, recorded the highest single transaction (₹45.50 in synthetic data). Furthermore, analyzing the time of day revealed that Afternoon and Night sales dominate naturally, but the synthetic "Evening" category skewed the results with an unusually high average value (₹60).

Justification of Findings

- Sales Distribution: Afternoon had the highest transaction count (1205), indicating customer preference for coffee during working hours.
- High Value Purchases: Synthetic data introduced unusually high values, which helped demonstrate how outliers affect averages. For example, Cappuccino's maximum transaction jumped from ₹38.70 to ₹60.00.
- Yearly Trends: The consistency of averages in 2024 and 2025 shows natural sales stability, while 2023's high average reflects synthetic influence.
- Coffee Variety: With only 8 coffee types, the shop's menu is compact yet profitable, with Lattes being the most popular.

Recommendations for Future Work

- Data Expansion: Collect more data across multiple years to identify long-term seasonal trends, customer loyalty, and evolving coffee preferences.
- Outlier Treatment: Introduce methods like IQR filtering or Z-score analysis to separate synthetic or erroneous high values from genuine sales
- Customer Segmentation: Add demographic data (age, gender, occupation) to understand buying behavior better.
- Predictive Analytics: Build forecasting models to predict sales based on time, season, and coffee type, helping in inventory management.
- Visualization Dashboard: Develop an interactive dashboard for real-time monitoring of sales, making the analysis more actionable for business owners.

This way, the project not only demonstrates current sales patterns but also highlights the importance of maintaining data quality and sets a clear path for deeper business insights in the future.

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/contents.html
- Seaborn Documentation: https://seaborn.pydata.org/
- Google Colab Documentation: https://colab.research.google.com/drive/1TJ8NW7_VHWRNeIHPpfzgZyV9qP-fm4bi

Appendix B: GitHub Link

IDEAS_TIHub_Final_Project_Submission_repo

Appendix D: Document Links

- Report Link: Report
- Dataset Link: Coffe_sales.csv