EXPLORATORY DATA ANALYSIS ON COFFEE SALES DATA

DEBANJANA DAS

M.A. ECONOMICS (Specialization in Trade and Finance)

INDIAN INSTITUTE OF FOREIGN TRADE

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 details a comprehensive exploratory data analysis (EDA) of a coffee sales dataset to uncover key business insights. The project commenced with loading the dataset and conducting an initial quality assessment, which included checks for missing values and duplicate columns. The core methodology involved data pre-processing, such as converting data types, and feature engineering to extract 'Month' and 'Year' from date-time objects. To demonstrate robust data handling, the original dataset was augmented with 100 rows of synthetically generated data. The subsequent analysis focused on identifying temporal and product-based sales patterns. Key findings reveal that the highest average transaction values occur during the 'Night' period and that peak transaction amounts are concentrated in the months of March and April. Furthermore, the analysis identifies specific products, including Cappuccino, Cocoa, Hot Chocolate, and Latte, that are associated with the highest sales values.

2. Introduction

Analysing sales data is a cornerstone of modern retail strategy, providing businesses with the empirical evidence needed to make informed decisions. For a coffee shop, understanding purchasing patterns is critical for optimizing operations, from managing inventory levels and adjusting staffing schedules to targeting marketing promotions effectively. By dissecting transaction data, a business can identify its most popular products, busiest hours, and seasonal trends, thereby aligning its services more closely with customer demand. This project undertakes an exploratory data analysis to extract such actionable insights from a provided coffee sales dataset.

Training Overview:

1. Introduction to AI, ML, Deep Learning & Data Science: The course introduced AI tools for insights and analysis, and discussed common challenges in data science, such as dealing with large, complex, unstructured, dirty, noisy, incomplete, or null data. It covered the importance of data processing steps and the role of a Data Engineer within data science. Discussions also touched upon why Python is preferred over R in this field.
2. Python Fundamentals - Basic Concepts: Started with a foundational understanding of what a program is, involving algorithms, syntax, and instructions converted into machine code.
3. Python Fundamentals - Functions and Methods
4. Python Fundamentals - Control Flow and Error Handling
5. Python Fundamentals - Data Structures (Strings, Lists, Tuples, Dictionaries, Sets)
6. Python Fundamentals - Object-Oriented Programming (OOP)
7. Python Fundamentals - Modules and Libraries
8. NumPy Library
9. Pandas Library
10. Machine Learning - Core Concepts
11. Machine Learning - Linear Models
12. Machine Learning - Classification
13. Artificial Neural Networks (ANN) & Deep Learning
14. Artificial Intelligence - Large Language Models (LLMs)
15. Soft Skills / Professional Development through Communication

3. Project Objective

This section outlines the specific, measurable goals that guided the analytical process of this project. Each objective was designed to build upon the previous one, ensuring a systematic progression from raw data to actionable insights. These objectives served as the blueprint for the entire data exploration, from initial preparation to the final interpretation of results.

The key objectives of this project were:

1. To meticulously clean and pre-process the raw coffee sales data to ensure its integrity and suitability for analysis.
2. To engineer new, relevant features, such as 'Month' and 'Year', from existing data to enable deeper temporal analysis.
3. To demonstrate data augmentation by generating a synthetic dataset and integrating it with the original data, documenting the impact on the overall dataset structure.
4. To conduct a comprehensive exploratory data analysis to identify and visualize key sales patterns related to time (year, month, time of day) and product categories.

5. To derive actionable insights from the analysis regarding customer purchasing behaviour and product performance.

4. Methodology

This section provides a granular, step-by-step account of the data handling and analysis process employed throughout the project. The methodology was designed to be transparent and reproducible, ensuring the integrity of the findings. The process can be broken down into four distinct phases: initial assessment, pre-processing, data augmentation, and analysis using a defined set of tools.

4.1. Data Collection and Initial Assessment

The project utilized a pre-existing dataset provided in a CSV file named Coffe_sales.csv. Upon loading this data into a Pandas DataFrame, an initial assessment was performed to understand its structure and quality.

- **Initial Dimensions:** The original dataset contained **3547 rows** and **11 columns**.
- **Initial Data Quality Checks:**
  - **Duplicate Columns:** A check for duplicate columns confirmed that **0 duplicate columns** were present.
  - **Missing Values:** An initial scan for missing data revealed that there were **0 missing values** across all columns, indicating a clean starting point for the analysis.

4.2. Data Pre-processing and Feature Engineering

To prepare the data for temporal analysis, several pre-processing and feature engineering steps were executed.

1. **Data Type Conversion:** The Date column, initially loaded as an object (string) data type, was converted to a proper datetime format using the pd.to_datetime function. This transformation is essential for performing date-based calculations and aggregations.
2. **Feature Engineering:** Two new features were engineered from the transformed Date column to facilitate granular temporal analysis:
   - A Month column was created by extracting the month number from each date.
   - A Year column was created by extracting the year from each date.

4.3. Data Augmentation

To demonstrate a common data handling technique, the original dataset was augmented with synthetically generated data.

- **Synthetic Data Generation:** A total of **100 rows** of synthetic data were generated. The logic for generation was tailored to the data type of each column:
  - For categorical columns (object type), new data points were randomly sampled from the set of existing unique values.

- For numerical columns (int64, float64), random numbers were generated within the minimum and maximum range of the original data.

- **Data Combination:** The original dataframe (coffee_data) was concatenated with the new synthetic dataframe (synthetic_data) to create a final, combined dataframe named combined_coffee_data.
- **Final Dimensions:** The final combined dataset had dimensions of **3647 rows** and **13 columns**. The column count increased by two due to the addition of the engineered Month and Year columns.

4.4. Tools and Libraries Used

The analysis was conducted entirely within the Python ecosystem, relying on a standard set of open-source libraries:

- **Python:** The primary programming language used for scripting the analysis.
- **Pandas:** The core library for data loading, manipulation, cleaning, and aggregation.
- **NumPy:** Used for underlying numerical computations and random number generation.
- **Matplotlib & Seaborn:** Employed in tandem to generate clear and informative static data visualizations.

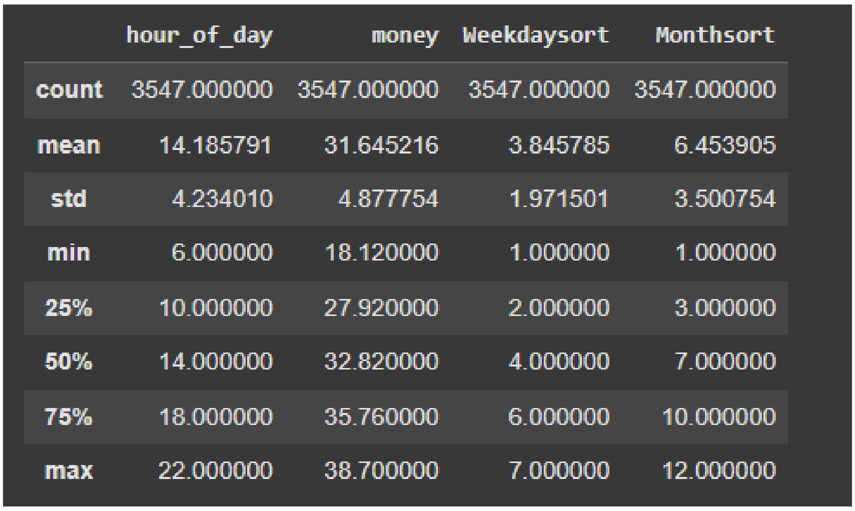
These methods and tools provided a robust framework for transforming the raw data into the analytical results presented in the following section.

5. Data Analysis and Results

This section presents the key findings from the exploratory data analysis conducted on the combined dataset. The results are structured to provide a multi-faceted view of the sales data, beginning with high-level statistics and moving into more detailed temporal and product-focused analyses.

5.1. Descriptive Analysis

A descriptive statistical summary provides a foundational understanding of the numerical columns within the dataset. The key statistics for transaction value (money), time of day (hour_of_day), and sorting keys (Weekdaysort, Monthsort) are presented below.



	hour_of_day	money	Weekdaysort	Monthsort
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

5.2. Temporal Analysis

The analysis of sales data over time reveals important trends regarding yearly, monthly, and daily performance.

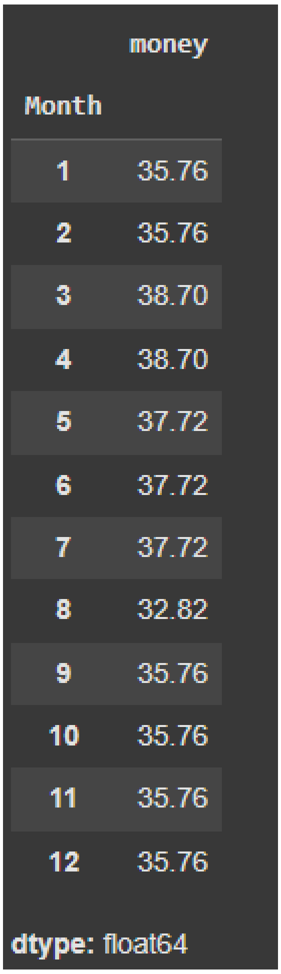
Yearly Trends

When grouped by year, the average transaction value shows slight variation between the two years present in the dataset.

- The average transaction value in **2024** was **31.62**.
- The average transaction value in **2025** was **31.31**.

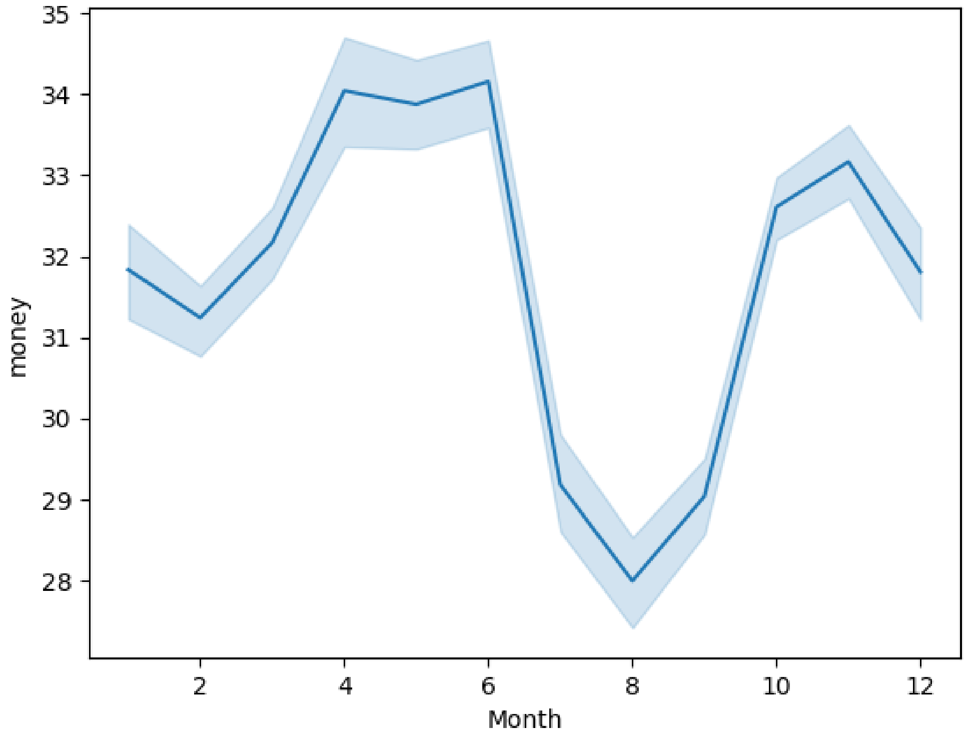
Monthly Sales Patterns

Analysis of sales by month indicates seasonal peaks in transaction values. The highest single transaction values were recorded in the early spring.



	money
Month	
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

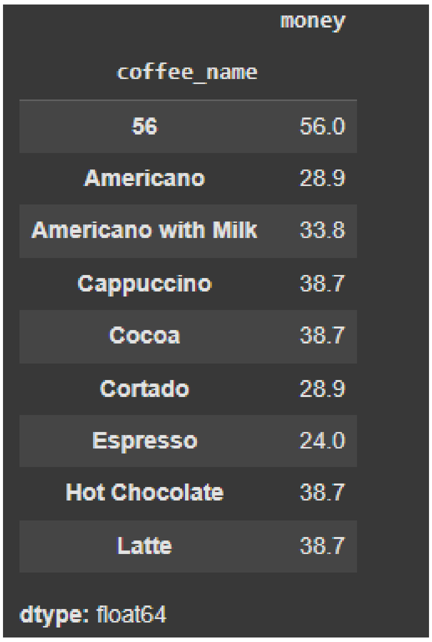
dtype: float64



The line plot above illustrates the fluctuation in transaction values across the months, with clear peaks observed in March and April, suggesting a potential seasonal increase in high-value purchases.

Intra-day Sales

The average transaction value varies depending on the time of day, with the highest average occurring in the evening.



	money
coffee_name	
56	56.0
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

dtype: float64

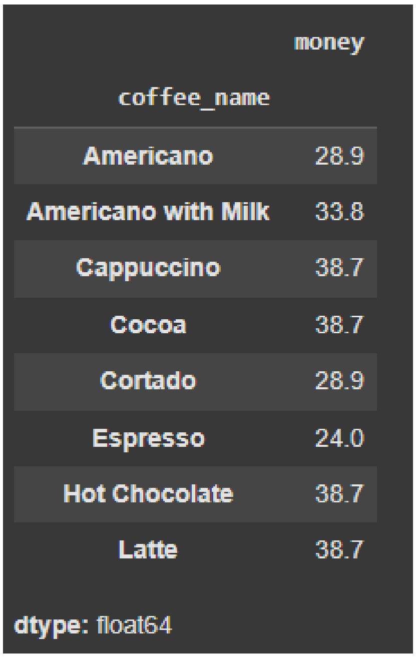
This finding indicates that customers tend to make higher-value purchases during the 'Night' period. This suggests a shift in purchasing behaviour, possibly indicating that evening customers are more inclined towards larger orders, premium beverage options, or dessert-style coffees, compared to the quicker, more functional purchases typical of the morning.

5.3. Product-Based Analysis

The dataset includes sales data for **8 unique coffee types**. While maximum transaction value can highlight products with a high price ceiling, a complete picture requires understanding both central tendency and distribution.

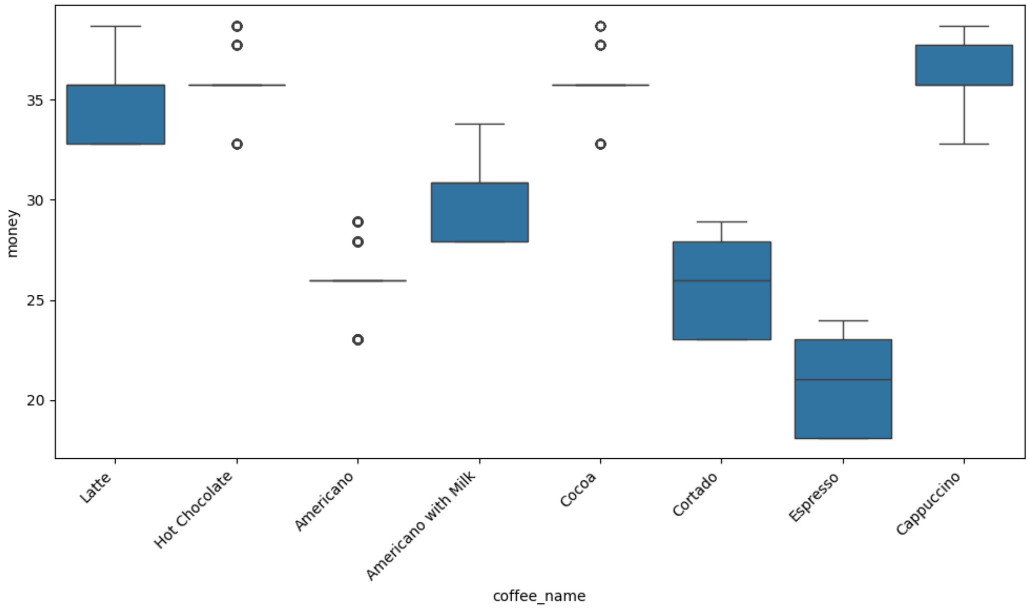
The table below details the maximum transaction value reached for each coffee name.

Notably, **Cappuccino, Cocoa, Hot Chocolate, and Latte** all reached the dataset's maximum transaction value of **38.70**. This suggests these items are either premium-priced or are frequently part of larger, multi-item orders. However, relying solely on maximum value can be misleading. A more nuanced view is provided by visualizing the distribution of transaction values.



	money
coffee_name	
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

dtype: float64



The box plot provides a visual summary of the price distribution for each coffee type. It effectively illustrates the median price, the typical price range (interquartile range), and any outliers. This visualization confirms that products like Cappuccino and Latte exhibit a wider price distribution and a *higher 75th percentile and maximum value compared to others like Espresso, reinforcing their position as premium offerings.*

These quantitative results provide the empirical foundation for the conclusions and strategic recommendations that follow.

6. Conclusion

This project successfully conducted an exploratory data analysis of a coffee sales dataset, proceeding from data cleaning and feature engineering to the identification of meaningful business patterns. The methodology employed a systematic approach to ensure data integrity and produced a series of clear, data-driven insights into sales performance over time and across different products.

Summary of Findings

The analysis yielded several critical insights that can inform business strategy:

- **Peak Sales Period:** The 'Night' period generates the highest average transaction value (€32.78), suggesting a customer tendency towards higher-value purchases in the evening. This could be an opportunity for upselling or promoting premium products.
- **Top-Value Products:** Products such as **Cappuccino, Cocoa, Hot Chocolate, and Latte** are associated with the highest single transaction values (€38.70) in the dataset, identifying them as key premium items or components of large orders.
- **Monthly Performance:** The analysis revealed that the highest single transaction values occurred in March and April, indicating a potential seasonal peak that could be leveraged for targeted marketing campaigns or promotions.

Recommendations:

Based on the findings of this exploratory analysis, several avenues for future work are recommended:

- **Predictive Modelling:** Given that the 'Night' period yields the highest average transaction value (€32.78), develop a time-series model to specifically forecast evening demand. This would enable optimized staffing and inventory management for high-margin products during these peak value hours. Furthermore, since March and April show the highest single transaction values (€38.70), a follow-up analysis could isolate the drivers of this seasonal peak, and a predictive model could be trained to anticipate this period, informing targeted marketing campaigns in the preceding months.
- **Advanced Customer Segmentation:** If more granular data becomes available (e.g., customer IDs), a deeper analysis could be performed to segment customers based on purchasing behaviour. This would allow for highly personalized marketing and loyalty programs.
- **Product Combination Analysis:** Investigate which products are most frequently purchased together (market basket analysis). This could uncover cross-selling opportunities and inform menu bundling strategies.

In conclusion, this project has demonstrated the immense value of exploratory data analysis in transforming raw sales data into a strategic asset for a retail coffee business.

7. Appendices

Github repository link: https://github.com/debanjana-das22/IDEAS-TIH-ISI-PROJECT-SUBMISSION.git