PREPROCESSING AND VISUALISING COFFEE SALES DATA

RITTIKA SURAI

B Sc in Statistics (SISTER NIBEDITA GENERAL DEGREE COLLEGE FOR GIRLS')

SECTION 1

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 presents an **Exploratory Data Analysis (EDA)** of coffee-shop sales data using Python.
Key tasks included importing and cleaning data, deriving new date/time features, generating descriptive statistics, and visualising sales patterns with pandas, NumPy, matplotlib, and seaborn.
The work highlights how Python simplifies end-to-end data handling—from raw CSV files to meaningful insights.

2. Introduction

This project focuses on performing **Exploratory Data Analysis (EDA)** of coffee-shop sales using Python.
The goal is to demonstrate how raw transactional data can be transformed into meaningful insights through open-source tools. Retail businesses—like coffee shops—generate large volumes of time-stamped sales data, and understanding patterns such as seasonal trends, popular products, and revenue fluctuations is crucial for decision-making.

Technology Involved

Key Python libraries used include:

- **pandas** – data cleaning and manipulation
- **NumPy** – numerical computations
- **matplotlib** and **seaborn** – data visualisation

These tools enable fast exploration of CSV datasets, creation of derived columns (e.g., Year, Month), and generation of clear graphical summaries.

Background and Procedure

The analysis began with importing the coffee-shop dataset, checking for missing values or duplicates, and converting the Date column to a proper datetime format.
Next, the data was grouped to calculate averages and maxima of the money column, sales distribution was examined across months and years, and revenue patterns were visualised.
Finally, a synthetic dataset was generated to validate the workflow and demonstrate that the analysis approach scales to new data.

Purpose

The main purpose of the project is to illustrate a complete Python data-analysis pipeline—from loading real-world style data to generating actionable business insights—within a short internship timeframe.

Training Topics Covered

Before starting the project, the internship provided structured sessions covering:

- Python programming basics (variables, loops, and core data structures)
- Object-oriented concepts, functions, and classes
- Data analysis using NumPy and pandas
- Machine-learning fundamentals, including regression and classification
- Introduction to large language models (LLMs)

- Communication and presentation skills

3. Project Objective

The key objectives of this project are:

- **Conduct Exploratory Data Analysis (EDA)** on a coffee-shop sales dataset to reveal trends and patterns in revenue.
- **Apply Python data-handling techniques** such as cleaning, datetime conversion, and creation of Month and Year fields.
- **Identify business insights** including yearly averages, monthly peaks, and top-selling coffee products.
- **Demonstrate scalability** of the analysis by adding and analysing synthetic data alongside the real dataset.

This project focuses on descriptive analysis only; no survey or hypothesis testing was involved.

4. Methodology

The project was carried out as a step-by-step exercise in Python to answer the analysis questions provided during the internship.
The key steps were:

1. Data Source

- Used the **Coffee Sales CSV file** supplied as part of the internship materials.

- A second **synthetic dataset** of 100 rows was generated with Python to demonstrate the same analysis on simulated data.

2. Data Preparation

- Loaded the datasets into **Google Colab** using pandas.read_csv.
- Checked for missing values and duplicate columns.
- Converted the Date column to datetime format and derived **Month** and **Year** columns.
- Created a **Time_of_Day** column from the transaction time to group results by morning, afternoon, and night.

3. Analysis

- Answered each question using pandas groupby and aggregation functions (mean, max, value_counts, etc.).
- Calculated number of columns, missing values, average money per year, maximum money per month, number of coffee types, and average money by time of day.
- Generated line plots and bar charts with **Matplotlib** and **Seaborn** to visualize trends.

4. Synthetic Data Task

- Used pd.date_range and np.random.randn to create a 100-row synthetic time-series dataset.
- Repeated the same set of analyses on this dataset to verify the workflow.

5. Tools

- **Google Colab** notebook environment.
- Python libraries: pandas, numpy, matplotlib, and seaborn.

6. Code Repository

- The complete Colab notebook, datasets, and the two-minute demonstration video will be placed in a dedicated **GitHub repository** as required by the internship guidelines.
- The final report will include the public GitHub link so evaluators can directly access the notebook and supporting files.

5. Data Analysis and Results

5.1 Descriptive Analysis

Question / Metric	Real Dataset Result	Synthetic Dataset Result	Key Observation
Number of columns	**11**	**11**	Structure of both datasets is identical.
Missing values	**0**	**0**	Data was already clean; no imputation required.
Average money per year	grouped_data showed a **mean ≈ ₹31.6** (float64)	**mean ≈ ₹31.3**	Consistent annual averages across datasets.
Maximum money per month	Reindexed monthly max (Jan–Dec) showed peaks around **₹38.7**	Similar monthly max values **₹38.3–₹38.7**	Seasonal variation is minor; sales remain steady.
Number of coffee types	**8**	**8**	Good product variety maintained.
Average money by time of day	Morning **₹30.42**, Afternoon **₹31.64**, Night **₹32.89**	Morning **₹31.41**, Afternoon **₹31.31**, Night **₹31.21**	Night sales have the highest average in the real dataset; synthetic data shows balanced averages.

5.2 Visualizations

Attach or embed screenshots of the following plots generated in the Colab notebook:

- **Monthly Sales Trend (Line Plot):** Seasonal consistency with slight mid-year peaks.
- **Annual Money Density (Bar Plot):** Shows steady yearly performance.
- **Distribution by Coffee Name (Bar Plot):** Highlights the most popular coffee varieties.
- **Average Spend by Time of Day (Bar Plot):** Confirms higher night-time spending in the real dataset.

5.3 Inferential Analysis

- No formal hypothesis testing or predictive modeling was performed.
- Exploratory comparisons indicate that **time of day** has a small but noticeable effect on average spending.

5.4 Synthetic Dataset Results

- The 100-row synthetic dataset was generated using a random walk (np.cumsum(np.random.randn(100))) over a daily date range.
- Cleaning and analysis steps mirrored the real dataset, confirming the robustness of the workflow.

- Key metrics (means, maxima, and coffee-type counts) closely match the original data, demonstrating reproducibility.

5.5 Repository Link

The complete Colab notebook, raw CSV files, and this report will be uploaded to a public GitHub repository for verification.
GitHub URL : https://github.com/surai01072006-byte/Rittika_Surai

6. Conclusion

This project demonstrated how Python-based data analysis can reveal key sales patterns from coffee shop data. We found stable yearly revenue, balanced monthly demand, eight main coffee types, and higher night-time spending. The same workflow worked on a synthetic dataset, proving it is reproducible and adaptable. Future work could include predictive modeling and interactive dashboards for real-time business insights.

7. APPENDICES

1. References (https://www.syngendata.ai/)

2. Synthetic Data Link (C:\Users\surai\Downloads\Synthetic_Data_2025-09-17.csv