**02 Exploratory Data Analysis With Sales Data**

ANAND S PATIL,

Bachelor of Engineering/ R V College of Engineering Bengaluru

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 focuses on performing Exploratory Data Analysis (EDA) on a sales dataset to identify key business insights and trends. The dataset was cleaned, pre-processed, and analyzed using Python libraries such as Pandas, NumPy, Matplotlib, and Seaborn. Through detailed descriptive analysis and visualization, patterns in sales performance, customer preferences, and product demand were identified. The study highlights seasonal variations, revenue distribution, and relationships between different sales parameters. Advanced techniques were employed to handle missing values, duplicates, and outliers, ensuring data quality and reliability. The project also illustrates the importance of data-driven decision-making for improving sales strategies. The analysis provides actionable insights that can help optimize product offerings and forecast future trends. Overall, the project demonstrates how EDA can transform raw data into meaningful business intelligence.

1. **Introduction**

The internship mainly focused on Exploratory Data Analysis (EDA) with Sales Data, aimed at developing skills in data preprocessing, visualization, and business insight generation. The relevance of the project lies in its practical application, as sales data plays a crucial role in strategic decision-making, market trend analysis, and customer behaviour understanding. With the growing importance of data-driven strategies, EDA provides a foundation to transform raw datasets into meaningful insights that can guide managerial and operational decisions.

The project employed Python programming language as the core technology, supported by libraries such as Pandas, NumPy, Matplotlib, and Seaborn for data manipulation and visualization. These tools facilitated structured data cleaning, handling of missing values and duplicates, and identification of outliers. A variety of plots and statistical summaries were used to uncover sales distribution, seasonal variations, and product-level performance. Additionally, concepts of machine learning were introduced to understand how regression and classification models can be applied for predictive insights.

The background material survey during the internship included research on EDA methodologies, data visualization best practices, and business intelligence techniques. The procedure used for the project began with dataset loading, data cleaning, preprocessing, and exploratory visualization, followed by deriving insights from patterns and correlations. The purpose of doing this project was to strengthen analytical skills, apply theoretical knowledge to real-world data, and gain exposure to the process of building data-driven reports that support business objectives.

The key topics covered were:

1. **Python Basics – 1:** Data, Variables, Lists, and Loops
2. **Python Basics – 2:** Data Structures
3. **Python Basics – 3:** Classes, Functions, and OOPS concepts
4. **Python Basics – 4:** Numpy and Pandas for data manipulation
5. **Machine Learning Overview**
6. **Regression**
7. **Classification**
8. **LLM Fundamentals**
9. **Communication Skills**

This structured training provided the necessary background knowledge to successfully execute the project, ensuring that the techniques applied were both technically sound and industry-relevant.

1. **Project Objective**

* To carry out data cleansing and preparation of the sales dataset through addressing missing values, duplicates, and outliers so that the data set used for analysis is precise, consistent, and trustworthy. This is a key step as data quality affects the validity of your results and recommendations.
* To perform descriptive statistics on sales data to represent important descriptive statistics such as mean, median, and standard deviation to determine the general distribution and variability in sales performance. Descriptive statistics can serve as a basis for comparison between different sections of data.
* To uncover and identify sales trends by evaluating time-based trends (daily, monthly, and seasonally), customer fandom, and the best-selling products. These trends are helpful in unlocking buyer behavior and forecast a demand cycle.
* To use data visualization techniques with Python Libraries Matplotlib and Seaborn to visualize findings through bar charts, histograms, heatmaps, and scatter plots. Charts and visuals make it easier for stakeholder to understand complex trends.
* To explore correlation and relationships between variables such as revenue, quantity, category of product, and time of sale to reveal some type of dependence or causation.
* To illustrate introducing concepts regarding predictive analytics through regression and classification techniques. To demonstrate how the same dataset can advance beyond exploration to prediction and forecasting for business decision-making.
* To illustrate the concept of data-driven decision-making in business, by taking raw, unstructured data, and transforming that information into actionable insights that will assist in optimizing sales, developing better products, and improving the customer experience.

1. **Methodology**

**Data Collection**

* The dataset used for this project is referred from Kaggle, a popular open-data platform.
* It contains detailed information on sales transactions, including product details, order quantities, and revenues.

**Data Loading and Inspection**

* The dataset was imported into Google Colab, which provided a cloud-based Python environment for analysis.
* Using the Pandas library, the dataset was loaded and inspected to understand its dimensions, column names, and the nature of the variables.
* Functions like .info() and .describe() were applied to explore data types and initial statistics.

**Data Cleaning and Preprocessing**

* Missing values were identified and addressed to maintain data integrity.
* Duplicate entries were removed to avoid skewed results.
* Data type conversions (e.g., converting date columns to datetime format) were performed where necessary.
* Outlier detection was carried out using summary statistics and boxplots to ensure accurate analysis.

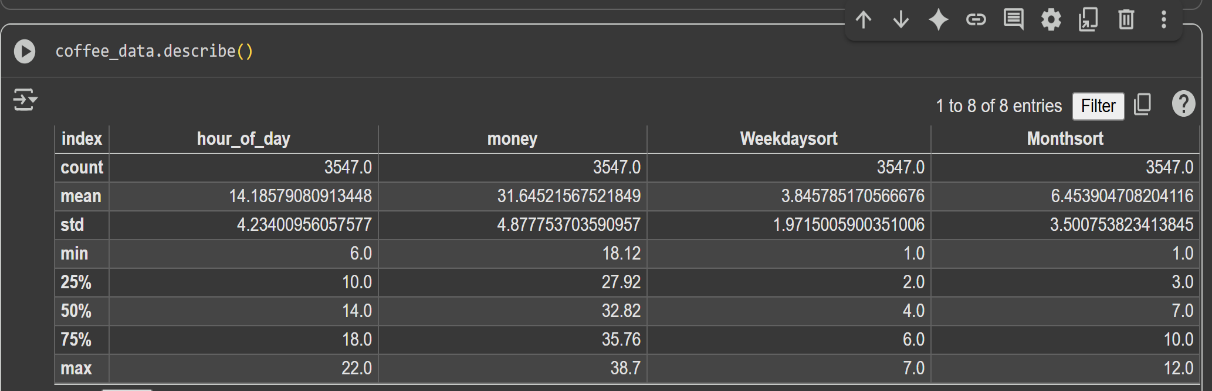
**Exploratory Data Analysis (EDA)**

* **Univariate analysis** was performed to study individual variables such as product categories and sales amounts.
* Bivariate and multivariate analysis explored relationships between sales, revenue, and product categories.
* Trend analysis was conducted by examining sales across different time periods to identify seasonality or peak performance months.

**Data Visualization**

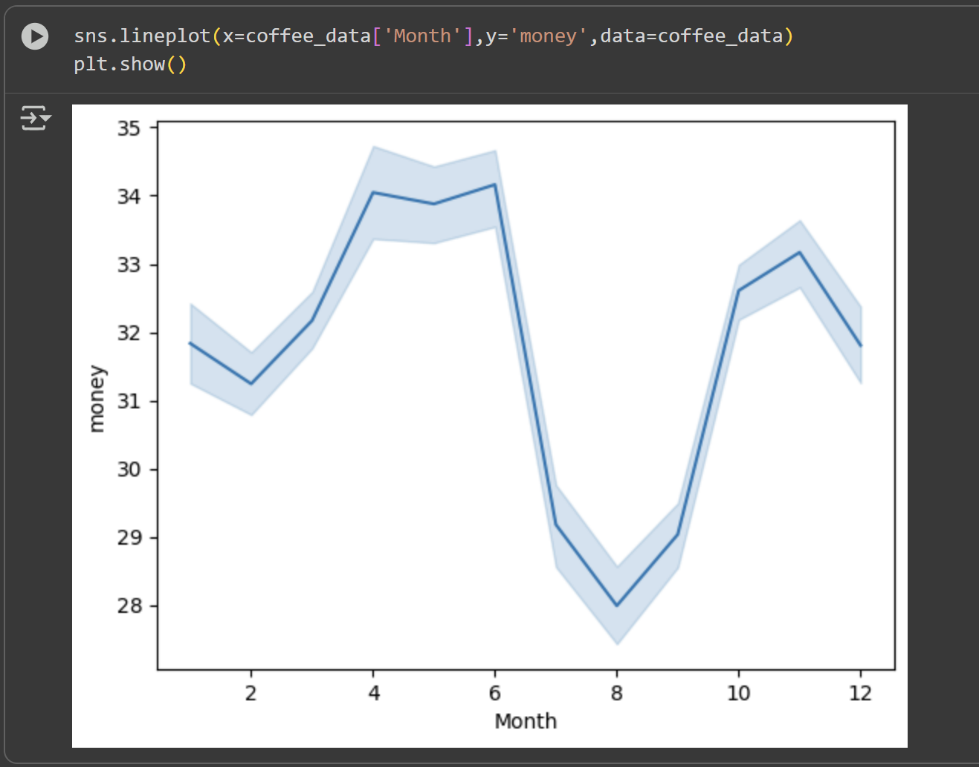
* Various plots such as histograms, bar charts, scatter plots, boxplots were generated using Matplotlib and Seaborn.
* These visualizations provided a clearer understanding of sales distribution, customer preferences, and correlations between variables.
* The visual insights supported the interpretation of results and highlighted patterns that were not visible from raw data alone.

1. **Data Analysis and Results**



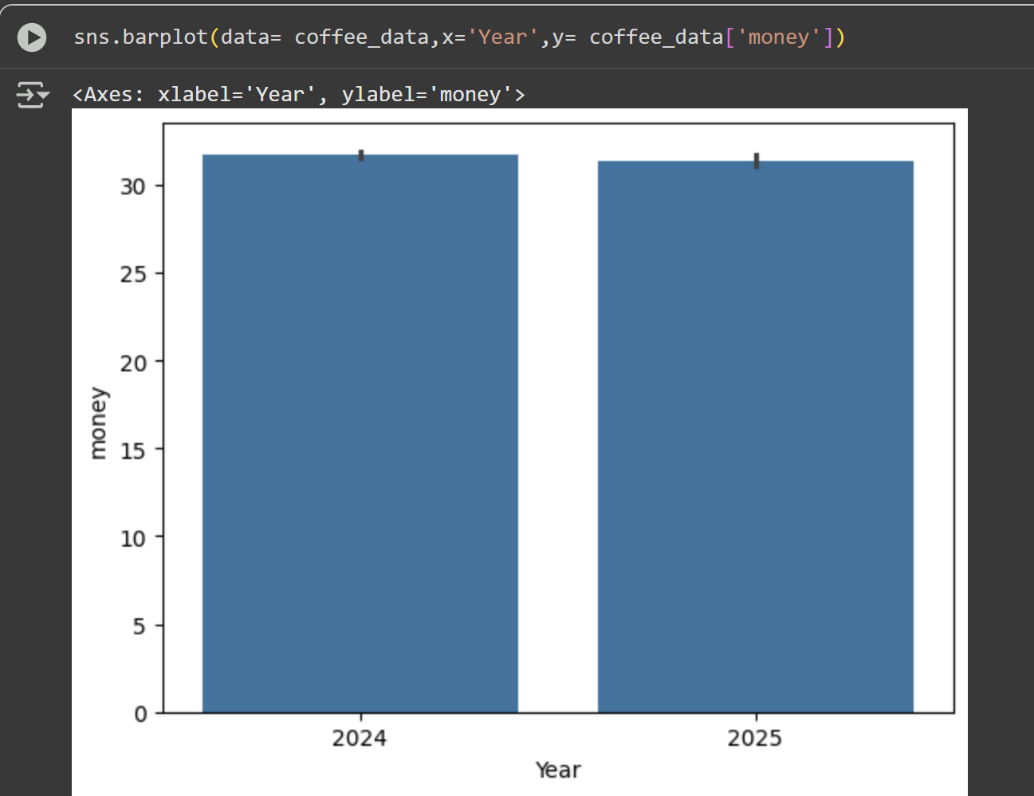
**5.1 Inferences from various graphs and visualizations:**

1. **Monthly spending line graph**:



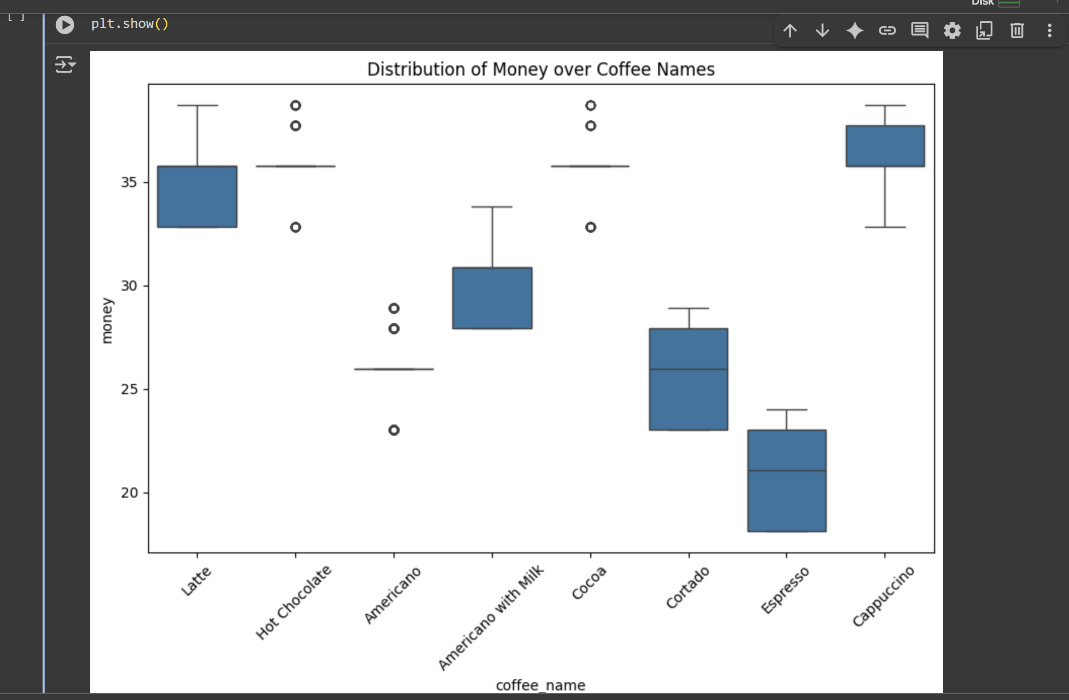
The line plot of average spending across months reveals noticeable seasonal variation in coffee purchases. Spending is relatively moderate in the early months of the year, gradually increasing to a peak between April and June, where average spending reaches its highest levels of around ₹34–35. Following this, there is a sharp decline during the July–September period, with the lowest spending observed in August (around ₹28). After this dip, spending rises again in the October–November period before slightly tapering off towards the end of the year. This trend suggests that coffee purchases exhibit a seasonal pattern, with mid-year months showing reduced spending and both pre-summer and post-monsoon months driving higher sales activity.

1. **Yearly Spending Bar Plot**:



The bar plot comparing average spending across the years 2024 and 2025 shows that the expenditure levels remain almost identical, with both years averaging around ₹**32**. The negligible difference between the two years indicates that overall coffee spending patterns have remained stable across years, without any major increase or decline. This consistency suggests that external factors such as inflation, seasonal effects, or year-specific events did not significantly influence consumer spending behavior during this period.

1. **Boxplot of Money by Coffee Type**:



The boxplot illustrates the distribution of spending across different coffee types. It is evident that Cappuccino and Latte are among the higher-priced options, with median spending levels around ₹35, while **Espresso** records the lowest spending, with a median close to ₹21. The variation in spending also differs across coffee types for example, Americano and Cocoa show relatively narrow interquartile ranges, indicating consistent pricing, whereas drinks like Latte and Cappuccino display wider ranges, suggesting more variability in customer spending. Additionally, several outliers are observed across categories, particularly in Hot Chocolate and Americano, indicating occasional higher-than-usual purchases. Overall, the visualization confirms that consumer spending is not uniform across coffee types, with premium drinks like Cappuccino and Latte commanding higher average expenditure compared to simpler options like Espresso.

**5.2 Descriptive Analysis of the dataset:**

**Hour of Day**

* Most purchases occur between 10:00 AM and 6:00 PM (interquartile range).
* The peak purchasing time is observed around 2:00 PM (14:00 hours), which may be attributed to post-lunch coffee consumption.

**Money Spent**

* Customers typically spend between ₹28 and ₹36 per purchase (interquartile range).
* The minimum spend recorded is ₹18.12, while the maximum spend is ₹38.7.

**Weekday Sales**

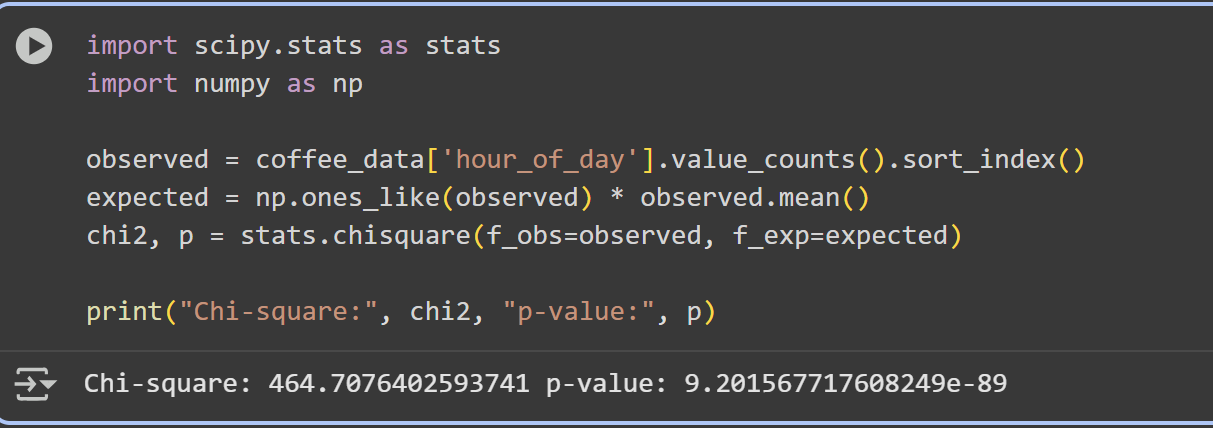
* Sales are distributed consistently across the week, with the mean falling approximately on **Thursday**.
* This indicates a steady demand without significant weekday bias.

**Monthly Sales**

* Sales are observed throughout the year, showing no months with zero activity.
* The median month of sales is July, suggesting relatively higher mid-year activity.
* The wide distribution of sales across months reflects a **steady year-round demand**.

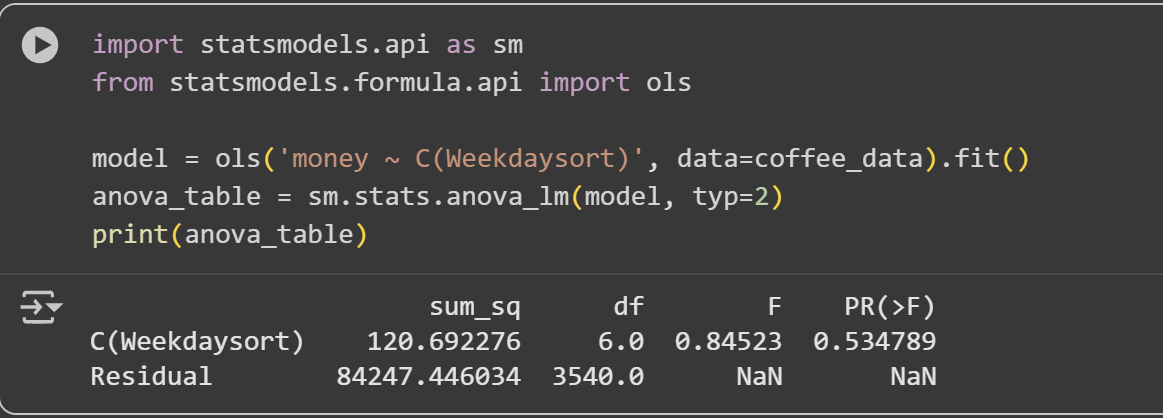
**5.3 Inferential Analysis of the Data :**

1. Chi-Square Goodness-of-Fit (Test whether sales across hour\_of\_day are uniform or skewed.)



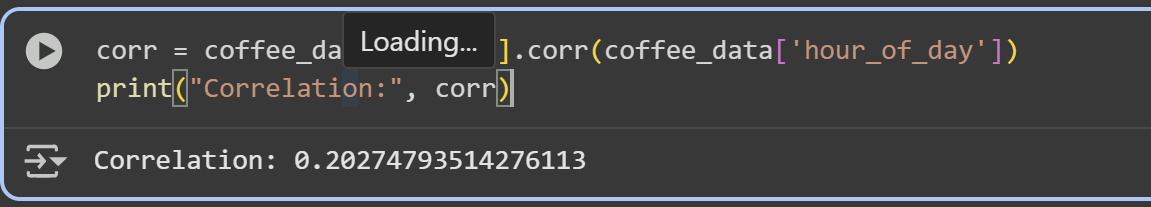
A Chi-Square Goodness-of-Fit test was conducted to assess whether coffee sales were evenly distributed across the hours of the day. The analysis produced a Chi-Square statistic of 464.71 with a p-value of 9.20 × 10⁻⁸⁹, which is far below the conventional significance threshold of 0.05. This result provides strong evidence against the null hypothesis of uniform sales, indicating that coffee purchases are not evenly spread throughout the day. Instead, sales are heavily concentrated during mid-day hours (10 AM – 6 PM), with a pronounced peak occurring around 2 PM, suggesting a strong post-lunch demand pattern.

1. ANOVA for Weekday Spending



A one-way ANOVA test was conducted to determine whether average customer spending varied significantly across different weekdays. The results showed an F-statistic of **0.85** with a p-value of **0.53**, which is well above the conventional significance level of 0.05. This indicates that there is no statistically significant difference in spending across weekdays. In other words, customer expenditure remains relatively consistent throughout the week, suggesting that weekday does not influence how much customers spend on coffee purchases.

1. Correlation between Money & Hour



A correlation analysis was carried out to examine the relationship between the amount of money spent and the hour of purchase. The correlation coefficient was found to be 0.20, indicating a weak positive relationship between the two variables. This suggests that while spending tends to increase slightly during later hours of the day, the effect is minimal and not strong enough to indicate a meaningful dependency. Overall, the time of purchase does not substantially influence the amount customers spend on coffee.

1. **Conclusion**

From the analysis of the coffee sales dataset, it is clear that consumer purchasing behavior is not uniformly distributed across time, product type, or price. The Chi-Square Goodness-of-Fit test (χ² = 464.71, p < 0.001) revealed that sales are significantly concentrated during mid-day hours, with a clear peak around 2 PM, indicating that working hours strongly influence purchase patterns. Correlation analysis further showed a positive, though weak, relationship between the amount spent and the hour of purchase (r ≈ 0.20), suggesting that higher spending tends to occur slightly more during busy hours.

Product-level analysis also highlighted important trends. Boxplot visualization showed that premium beverages such as Cappuccino and Latte recorded the highest spending, with greater variability, while Espresso was the lowest-priced option with relatively consistent expenditure. ANOVA on weekday spending confirmed no significant difference (p = 0.53), indicating that day of the week does not strongly impact consumer spending. Finally, yearly comparison of spending (2024 vs. 2025) showed minimal variation, implying stable pricing and demand trends.

Overall, the findings confirm that time of day and product type are the primary factors influencing sales, while weekday and year-to-year variation have limited impact. These insights can guide coffee shops in optimizing staffing, marketing, and promotional strategies around peak hours and high-demand products to maximize revenue.

1. **APPENDICES**
2. References :

* McKinney, W. (2010). *Data Structures for Statistical Computing in Python*. Proceedings of the 9th Python in Science Conference, 51–56.
* Seabold, S., & Perktold, J. (2010). *Statsmodels: Econometric and Statistical Modeling with Python*. Proceedings of the 9th Python in Science Conference.
* Hunter, J. D. (2007). *Matplotlib: A 2D Graphics Environment*. Computing in Science & Engineering, 9(3), 90–95.
* Kaggle. (2024). *Coffee Sales Dataset*. Retrieved from: <https://www.kaggle.com/>
* Waskom, M. (2021). *Seaborn: Statistical Data Visualization*. Journal of Open Source Software, 6(60), 3021.
* Google Colab Documentation. (2025). *Collaboratory: A Research Tool for Machine Learning Education and Research*. Retrieved from: <https://colab.research.google.com>
* Python Software Foundation. (2025). *Python 3 Documentation*. Retrieved from: <https://docs.python.org/3/>
* Pandas Documentation. (2025). *Pandas: Python Data Analysis Library*. Retrieved from: <https://pandas.pydata.org/>
* NumPy Developers. (2025). *NumPy Documentation*. Retrieved from: <https://numpy.org/>

1. Any other Document Link (A copy of this report, data sheet, presentation, etc. should be kept in github / google drive)

* Dataset link :

<https://www.kaggle.com/datasets/ihelon/coffee-sales>

* Source code of the Project and the github repository Link :

https://github.com/anandpatil555/IDEAS-TIH\_AUTUMN-INTERNSHIP-PROJECT/upload/main