**CHAPTER 1**

# INTRODUCTION

In a fast-evolving world where coffee has become an integral part of daily routines,

understanding consumer purchasing patterns and sales trends is essential for businesses to remain

competitive. This report undertakes a detailed analysis of coffee sales data, encompassing transaction records that include purchase dates, coffee types, payment methods, sales values, and customer preferences. By exploring these metrics, the analysis aims to uncover meaningful insights into sales performance and consumer behaviour, enabling businesses to refine their strategies and improve profitability. This data-driven approach provides a roadmap for responding to market demands and enhancing the customer experience.

The dataset provides a rich overview of 500 coffee sales transactions, highlighting variations in payment methods (e.g., cash and card), the popularity of specific coffee types such as lattes and americanos, and spending trends over time. By examining temporal patterns and categorical data, the analysis delves into correlations that reveal the dynamics of coffee sales, including peak purchasing hours, seasonal influences, and customer loyalty trends. The incorporation of advanced data visualization techniques, such as trend analysis and heatmaps, offers a clearer understanding of key drivers behind consumer choices and transaction values.

As coffee consumption continues to grow globally, fuelled by an increasing preference for premium and personalized options, insights from this analysis are invaluable for retailers, marketers, and supply chain managers. The findings will help optimize product offerings, streamline marketing efforts, and enhance operational efficiency. Moreover, by leveraging these insights, businesses can adapt to evolving customer expectations, ensuring a competitive edge in the marketplace. This report underscores the pivotal role of data science in shaping strategies that foster sustained growth and customer satisfaction in the thriving coffee industry.

In addition to understanding transaction trends and consumer preferences, this analysis also investigates the financial and operational implications of the sales data. By examining the distribution of revenue across different coffee types and payment methods, businesses can identify high-margin products and streamline pricing strategies. Moreover, the temporal data allows for the identification of peak sales periods, aiding in inventory planning and staff allocation. Such insights are critical for minimizing waste and optimizing resource utilization. This comprehensive approach ensures that the findings are not only descriptive but also actionable, enabling stakeholders to make informed decisions that drive efficiency and profitability across the coffee supply chain.

**CHAPTER 2**

# IMPLEMENTATION

## **2.1 Module Explanation**

**2.1.1 Objective:**

The objective of the coffee sales analysis is to explore and interpret trends within transactional

data, focusing on factors such as purchase timing, customer preferences, payment methods, and

revenue. By leveraging this analysis, coffee businesses aim to optimize their sales strategies, enhance customer experiences, and make data-driven decisions to improve overall profitability. The study also seeks to identify hidden opportunities and challenges, such as underperforming coffee variants or shifts in customer behaviour, to enable proactive business planning.

**2.1.2 Components:**

* **Data Collection**:
  + **Description**: Includes coffee sales records with transaction details like date, payment method, type of coffee, and revenue. A reliable and diverse dataset ensures the analysis is comprehensive and reflective of real-world operations.
  + **Usage**: Provides the foundation for analysis by capturing all relevant information about customer transactions.
* **Data Preprocessing**:
  + **Description**: Handling missing values, removing duplicates, and formatting data for analysis. This step ensures the data's accuracy and relevance for meaningful insights.
  + **Usage**: Essential for cleaning and structuring the dataset, ensuring it is error-free and

suitable for advanced analytics.

* **Exploratory Data Analysis (EDA):** 
  + **Description**: Uncovers patterns, customer behaviour, and key sales insights using summary statistics and visualizations. EDA provides a foundational understanding of the dataset and highlights trends.
  + **Usage**: Identifies correlations, trends, and anomalies to guide further analysis and

decision-making.

* **Statistical Analysis**:
  + **Description**: Applying techniques like trend analysis, hypothesis testing, and correlation studies to validate assumptions and uncover deeper insights.
  + **Usage**: Supports evidence-based conclusions by quantifying relationships and

verifying patterns in the data.

* **Visualization Tools**:
  + **Description**: Creating charts and plots using libraries such as Matplotlib and Seaborn to present findings effectively and make the data accessible to all stakeholders.
  + **Usage**: Transforms complex data into clear, actionable insights through visuals like bar charts, scatter plots, and heatmaps.
* **Predictive Modelling**:
  + **Description**: Employing machine learning tools to forecast future trends based on historical data, enabling proactive decision-making.
  + **Usage**: Helps businesses anticipate demand fluctuations, optimize inventory, and plan marketing campaigns efficiently.

**2.1.3 Workflow:**

* **Data Collection and Loading:** Importing the dataset into a structured format using

tools like Pandas, ensuring compatibility with analysis workflows.

* **Data Cleaning**: Addressing inconsistencies, such as missing or duplicate data, ensuring

dataset integrity and reliability.

* **Exploratory Data Analysis**: Analysing trends in sales volume, popular coffee types,

payment methods, and seasonal variations using visual and statistical tools.

* **Visualization**: Generating visual summaries like bar charts, scatter plots, and heatmaps

to highlight key insights, making patterns easily interpretable.

* **Advanced Analytics**: Incorporating statistical and predictive techniques to understand

customer preferences and forecast sales trends.

* **Reporting**: Summarizing findings with actionable recommendations for stakeholders,

presented in an easily digestible format for strategic decision-making.

**2.1.4 Benefits:**

* **Improved Decision-Making:**  
  By identifying top-selling coffee types, peak sales periods, and customer payment

preferences, businesses can make informed decisions on inventory management, pricing

strategies, and promotional campaigns. This ensures resources are allocated efficiently

and aligns operational efforts with customer demand, reducing waste and enhancing

overall profitability.

* **Enhanced Customer Satisfaction:**  
  Understanding customer preferences and behaviour allows businesses to tailor their

offerings, such as introducing popular coffee variants or creating personalized

promotions. By meeting customer expectations and delivering a superior experience,

businesses can foster loyalty and encourage repeat visits, contributing to long-term

growth.

* **Revenue Optimization:**  
  Insights derived from sales analysis enable businesses to identify and capitalize on

untapped opportunities, such as upselling complementary products or leveraging high-

demand periods. Predictive analytics can further guide proactive actions to maximize

revenue, ensuring the business remains competitive in the market.

## **2.2 Explanation of each module code**

## The data preparation phase focuses on ensuring the coffee sales dataset is clean and ready for analysis:

**2.2.1 Cleaning and Preprocessing**

**Data Cleaning and Preparation :**

Data cleaning and preparation are essential steps to ensure the dataset is accurate and ready for

analysis. The following processes were undertaken:

* **Handling Missing Values**:
  + Checked for missing values in critical columns such as money and coffee name.
  + Rows with missing values in these columns were removed to maintain data integrity.
* **Correcting Data Types**:
  + Converted the date column to a datetime format for proper chronological analysis.
  + Ensured that the money column was in a numeric format to facilitate calculations.
* **Removing Duplicates**:
  + Identified and removed duplicate entries to prevent skewed results. This was done by checking for identical rows across all columns.
* **Standardizing Text Data**:
  + Normalized entries in the coffee\_name column to ensure consistency (e.g., all lowercase, removing extra spaces).
  + Coded categorical variables like cash\_type into numerical values for easier analysis.
* **Outlier Detection and Treatment**:
  + Analyzed the money column for outliers that could indicate errors (e.g., unusually high or low transaction amounts).
  + Outliers were reviewed individually to determine whether they should be retained

or removed.

* **Feature Engineering**:
  + Created additional features such as extracting the month from the date column to analyze monthly sales trends.
  + Generated customer segmentation features based on total spending per customer.
* **Sorting and Indexing**:
  + Sorted the dataset by date and time to facilitate chronological analysis.
  + Set an index based on transaction IDs for improved data retrieval efficiency.
* **Data Validation Checks**:
  + Conducted consistency checks to ensure all entries conformed to expected formats.
  + Verified that total sales calculated from individual transactions matched expected totals from summary reports.
* **Final Review and Documentation**:
  + Conducted a final review of the cleaned dataset to ensure all cleaning steps were applied correctly.
  + Documented all changes made during the cleaning process for transparency and reproducibility.

**2.2.2 Visual Representation ,Exploratory Data Analysis(EDA) and Statistical Inference**

We conducted various visualizations and statistical analyses to explore data distributions and relationships between key variables. This included Exploratory Data Analysis (EDA) for visual insights and Statistical Inference using Confidence Intervals (CI) and Hypothesis Testing.Visual representation plays a crucial role in analyzing coffee sales data, as it allows for the effective communication of insights through graphical formats. In this analysis, we utilize three key visualizations:

* **Pie Chart**: This chart is employed to identify the best-selling coffees by displaying the
* proportion of total sales attributed to each coffee type. It provides a clear visual indication
* of which products are most popular among customers.
* **Histogram:** Used to analyze the total profit by day of the week, histograms illustrate the
* frequency distribution of profits across different days. This helps in observing how profits
* fluctuate and identifying trends that may inform operational decisions.
* **Box Plot:** This visualization is utilized to depict the profit distribution by day for a week.

Box plots show the median, quartiles, and potential outliers in profit amounts for each

day, allowing us to identify variations in profits throughout the week and understand

which days yield higher or lower profits.

Together, these visual representations enhance our understanding of sales patterns and customer

preferences, enabling data-driven decision-making.

**Statistical Inference**

Following EDA, **statistical inference** was used to provide quantitative insights. This involved **Confidence Intervals (CI)** and **Hypothesis Testing** as follows:

1. **Confidence Intervals (CI** ):Confidence intervals (CIs) are a statistical method used to estimate the range within which a population parameter, such as the mean, is likely to fall based on sample data. In your analysis of coffee sales, we calculated the mean transaction amount as 34.14 with a standard deviation of 4.75.We computed a 95% confidence interval for the average transaction amount, resulting in an interval of approximately (33.72, 34.56). This indicates that you can be 95% confident that the true average amount spent by customers lies within this range.This statistical insight provides valuable information for making informed decisions regarding pricing strategies and revenue forecasting in your coffee business.
2. **Hypothesis Testing**:

Hypothesis testing is a statistical method used to make inferences about population parameters based on sample data. In our coffee sales analysis, we may want to test specific assumptions or claims about customer behavior or sales performance.

* Setting Hypotheses: For example, we might set up a hypothesis test to determine whether there is a significant difference in average transaction amounts between cash and card payments:
  + Null Hypothesis (H0*H*0​): There is no difference in average transaction amounts between cash and card payments.
  + Alternative Hypothesis (Ha*Ha*​): There is a significant difference in average transaction amounts between cash and card payments.
* Conducting the Test: Using statistical tests such as t-tests or ANOVA, we can analyze the data and calculate p-values to determine whether to reject or fail to reject the null hypothesis. The choice of test depends on whether we are comparing two groups (t-test) or more than two groups (ANOVA).
* Interpreting Results: If we find a p-value less than our significance level (commonly set at 0.05), we would reject the null hypothesis, suggesting that payment method does indeed affect transaction amounts. This insight could inform marketing strategies or promotions tailored to specific payment methods.

**CHAPTER 3**

# RESULTS AND DISCUSSION

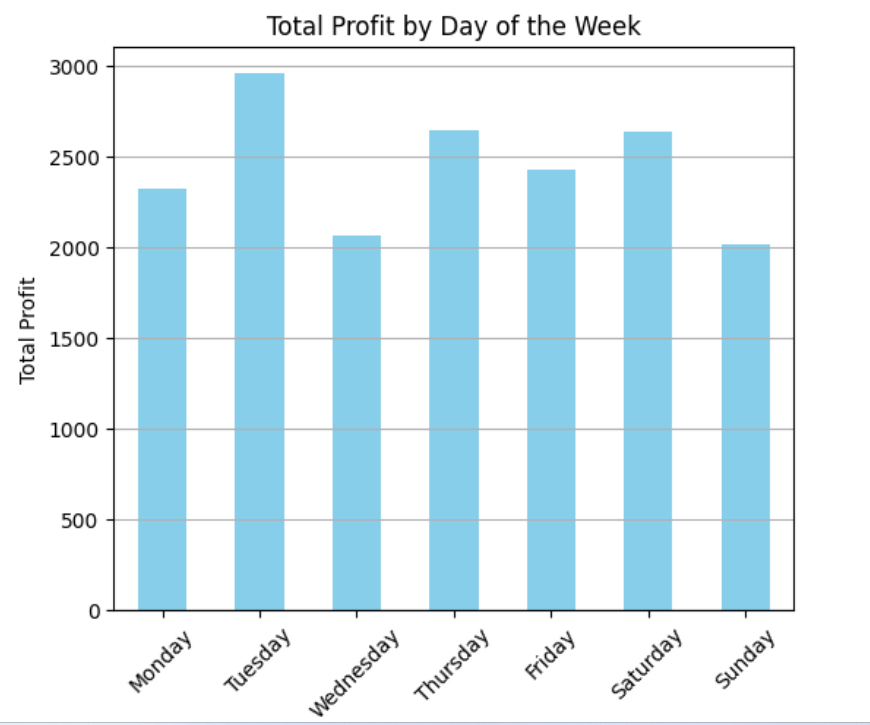
This chapter presents the results and visualizations through detailed graphs.

A graph with blue rectangular objects

Description automatically generated

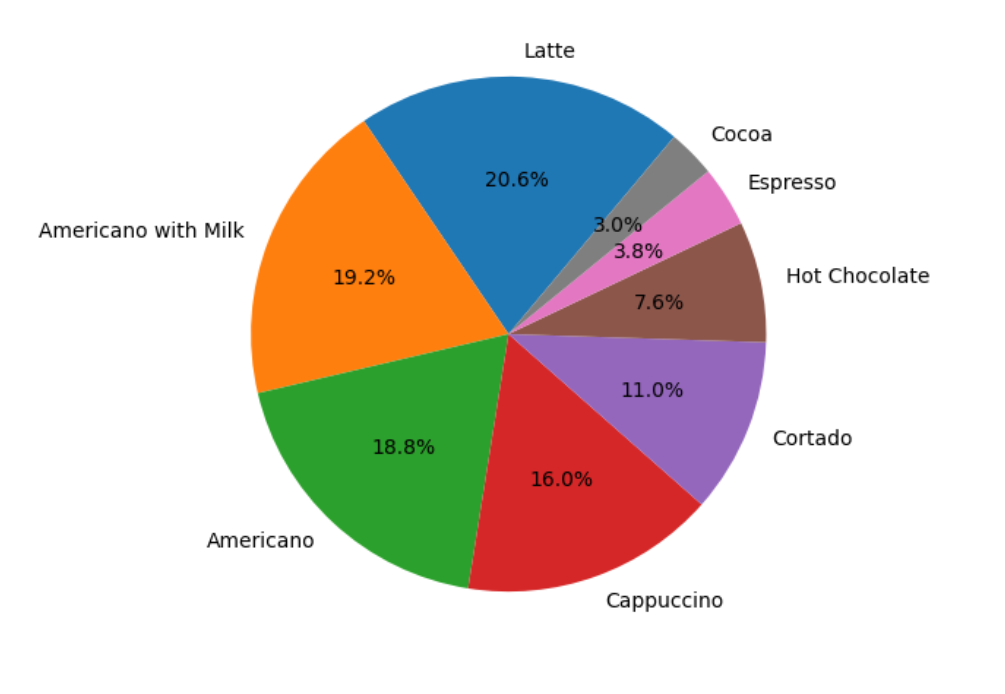
## Fig .3.1. Box Plots

In Fig. 3.1 the box plot graph for profit distribution by day of the week effectively illustrates the variations in profits across different days. It highlights key statistics such as the median, quartiles, and potential outliers, providing insights into which days generate higher or lower profits. This analysis is essential for optimizing operational strategies and maximizing profitability throughout the week.



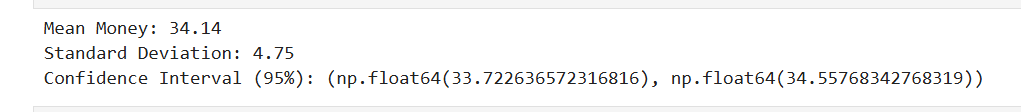
## Fig.3.2. Histogram

Fig.3.2 shows total profit by day of the week , highlighting the total profit by day .Here ,the highest profit is on Tuesday.



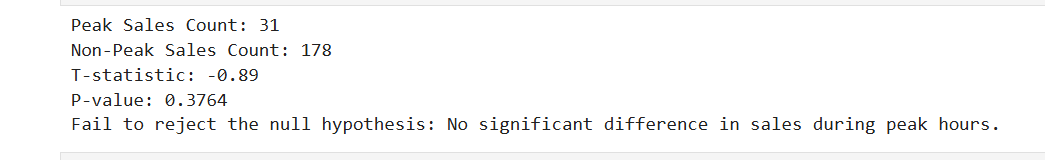
## Fig.3.3. Pie Chart

In Fig.3.3 the pie chart illustrates the distribution of coffee sales by type, showcasing the best-selling varieties. Each segment represents the proportion of total sales for each coffee type, allowing for quick identification of popular products.



## Fig.3.4. Confidence Interval

In Fig.3.4 the confidence interval (CI) analysis performed on the transaction amounts revealed a mean money value of **34.14** with a standard deviation of **4.75**. The calculated 95% confidence interval for the average transaction amount is approximately **(33.72, 34.56)**, indicating that we can be 95% confident that the true average amount spent by customers falls within this range. This statistical insight is valuable for understanding customer spending behavior and informing pricing strategies.



## Fig.3.5. Hypothesis Testing

In Fig.3.5 the hypothesis testing on peak working hours focused on determining whether there are significant differences in sales during different times of the day. By analyzing transaction data, we aimed to identify specific hours that yield higher sales volumes. The results of this analysis can inform staffing and operational strategies to optimize service during peak hours.

**CHAPTER 4**

# CONCLUSION

The coffee sales data analysis project provided an in-depth exploration of customer

behavior and sales performance over a specified period, utilizing the dataset from coffeee1.csv. This dataset includes transaction records that detail various aspects such as date, time, payment method, transaction amounts, and types of coffee sold, covering the period from March 1 to April 12, 2024.

Through visualizations like pie charts, we identified the best-selling coffee types, which highlighted popular products and informed inventory management strategies. The box plot showcased profit distribution by day of the week, revealing variations in daily earnings and helping to pinpoint which days are most profitable. Additionally, a histogram analyzed total profit by day, providing insights into patterns in daily earnings and guiding strategic decisions for future sales initiatives.

Statistical analyses were also conducted to deepen our understanding of customer spending behavior. A confidence interval analysis on transaction amounts revealed a mean value of **34.14** with a 95% confidence interval of approximately **(33.72, 34.56)**. This indicates that we can be 95% confident that the true average spending falls within this range. Furthermore, hypothesis testing was performed to analyze peak working hours, determining whether significant differences exist in sales during various times of the day. The results provided insights into specific hours that yield higher sales volumes, which is crucial for optimizing staffing and operational efficiency during busy periods.

In conclusion, the project successfully integrated visualizations and statistical methods to derive actionable insights that can drive business decisions in the coffee industry. By identifying best-selling products through pie charts, analyzing profit distributions with box plots and histograms, and applying confidence intervals and hypothesis testing to understand customer behavior and peak sales times, this project provides a holistic view of operations. The insights gained aim to enhance profitability while improving customer satisfaction through informed decision-making. Ultimately, this comprehensive approach can guide strategic planning for future operations in the coffee business, ensuring that resources are allocated effectively to meet customer demand and maximize revenue potential.

# REFERENCES

1. Python Software Foundation. (n.d.). *Jupyter Notebook and JupyterLab Documentation.* <https://jupyter.org/>
2. Kaggle. (n.d.). Mobile Device Usage and User Behavior Dataset. [Coffee Sales](https://www.kaggle.com/datasets/ihelon/coffee-sales)
3. Python Software Foundation. (n.d.). Python Documentation. <https://docs.python.org/3/>
4. OpenAI. (n.d.). ChatGPT for coding assistance and Python exploration. <https://chatgpt.com/>