IDEAS TIH AUTUMN INTERNSHIP 2025 PROJECT REPORT

Project Title:

02-Exploratory Data Analysis of Sales Data

CHANDRANI DUTTA

4 yrs. BSc. Mathematics Honours, Scottish Church College

Period of Internship: **25th August 2025 - 19th September 2025**

Report submitted to: **IDEAS – Institute of Data Engineering, Analytics and Science Foundation, ISI Kolkata**

1. Abstracts

This project presents a comprehensive study on the preprocessing and exploratory analysis of two sales datasets, one pertaining to coffee transactions and the other to mobile phone sales. Both datasets were subjected to systematic data quality checks, including the identification of duplicate entries, assessment of missing values, and verification of data type consistency. Preprocessing procedures were carried out to ensure structural integrity and prepare the datasets for meaningful analysis.

Exploratory Data Analysis (EDA) was then performed to examine sales trends across temporal dimensions such as years, months, weeks, and times of the day, as well as across product categories and brands. Visualizations were employed to highlight seasonal patterns, peak revenue periods, and product-level performance. Correlation studies further revealed relationships between key variables, including unit cost, unit price, and total sales. To extend the scope of analysis, synthetic data was generated and integrated into the original datasets, enabling the evaluation of underrepresented scenarios and testing the robustness of findings.

Overall, the project demonstrates how rigorous preprocessing, systematic visualization, and synthetic augmentation together enhance data quality, reveal hidden trends, and support data-driven decision-making. The methodologies applied in this study illustrate the versatility of data analytics across industries, from retail coffee sales to mobile phone markets, thereby laying the foundation for predictive modeling and business intelligence applications.

2. Introduction

In today’s world, data plays a very important role in understanding how businesses work and how customers behave. Companies collect huge amounts of sales data, and analyzing this data can help them make better decisions. In this project, I worked on two sales datasets — one about **coffee sales** and another about **mobile phone sales**. By comparing and studying these datasets, I was able to practice data cleaning, preprocessing, visualization, and even generating synthetic data to test new scenarios.

The project started with checking the quality of the data, such as finding duplicate values, missing values, and making sure the structure of the dataset was correct. After cleaning the data, I carried out descriptive and time-based analysis to look at sales by years, months, weeks, weekdays, and times of the day. I also studied product and brand-level information to see which products performed the best and during which time periods sales were the highest.

The technologies I used were **Python** for coding and libraries like **NumPy, Pandas, Matplotlib, and Seaborn** for data analysis and visualization. All the work was done in **Google Colab**, which gave me an interactive way to write and run my code step by step. I used different types of visualizations such as bar charts, histograms, and time series plots to make the results easy to understand.

Finally, I also generated synthetic data and added it to the original datasets. This helped me test how the analysis changes when we add new rows and whether the findings remain consistent.

The main purpose of this project is to show how preprocessing and visualization can help businesses find useful insights from data. By working on both coffee and mobile sales datasets, I learned how data science techniques can be applied in different industries to identify customer behavior, improve sales performance, and support data-driven strategies.

Training Topics Covered in First Two Weeks of Internship:

- Data and Variables
- Lists and Loops
- Data Structures
- Classes and Functions
- Object-Oriented Programming (OOPs)
- NumPy and Pandas
- Machine Learning Overview
- Regression Lab
- Classification Lab
- LLM Fundamentals
- Communication Skills

3. Project Objective

The main objectives of this project are:

- To pre-process the sales datasets by identifying and handling missing values, duplicate entries, and inconsistencies to ensure reliable analysis.
- To analyze sales distribution across different time periods, pricing ranges, and product categories (coffee types and mobile brands) in order to identify key market trends.
- To visualize the data using appropriate charts and plots so that patterns and insights can be easily interpreted and communicated.
- To generate and analyze synthetic data, extending the scope of study where the original datasets were insufficient or lacked certain scenarios.
- To illustrate how systematic preprocessing and visualization can support business intelligence, informed decision-making, and future predictive modeling.

4. Methodology

The methodology adopted for this project involved several structured steps to ensure the sales datasets were properly prepared, analysed, and visualized for meaningful insights.

Data Collection

- The Coffee Sales Dataset was provided by the instructor.
- The Mobile Phone Sales Dataset was collected from an open-source repository on GitHub.
- The Excel file was converted into a **CSV file** for easier handling.
- The project was executed in a **Jupyter Notebook** (.ipynb) for interactive coding and analysis, and the working environment was set up in **Google Colab**.

Data Quality Checks

- Checked the number of rows and columns.
- Verified if there were duplicate columns or duplicate rows.
- Checked for missing values in rows and columns.
- Verified data types of each column to ensure correctness.

Data Preprocessing

- Loaded the datasets into Pandas DataFrame.
- Checked datasets dimensions, column types, and basic statistics.
- Prepared the datasets for analysis by standardizing structures.

Exploratory Data Analysis (EDA)

- Descriptive analysis was performed to compute basic statistics such as mean, median, standard deviation, and range.
- Time-based analysis included studying sales by year, month, quarter, week, weekday, and time of the day.
- Product and brand-level analysis included identifying top-performing products, profit margins, and unit sales.
- Visualization methods such as histograms, bar plots, line charts, and correlation heatmaps were used to make findings easier to interpret.

Synthetic Data Generation

- 100 synthetic rows were generated for each dataset to simulate additional records.
- Synthetic data was merged with the original datasets to create extended versions.
- The same analyses (EDA and visualizations) were re-run on the combined datasets to compare with original results.

Documentation and Reporting

- A set of questions was designed to support targeted analysis (attached in Appendix).
- All results, tables, and visualizations were compiled into a structured report.
- A flowchart of the process was created to summarize the methodology.
- A GitHub repository was prepared to store the notebook, datasets, flowchart, and report files for easy access.

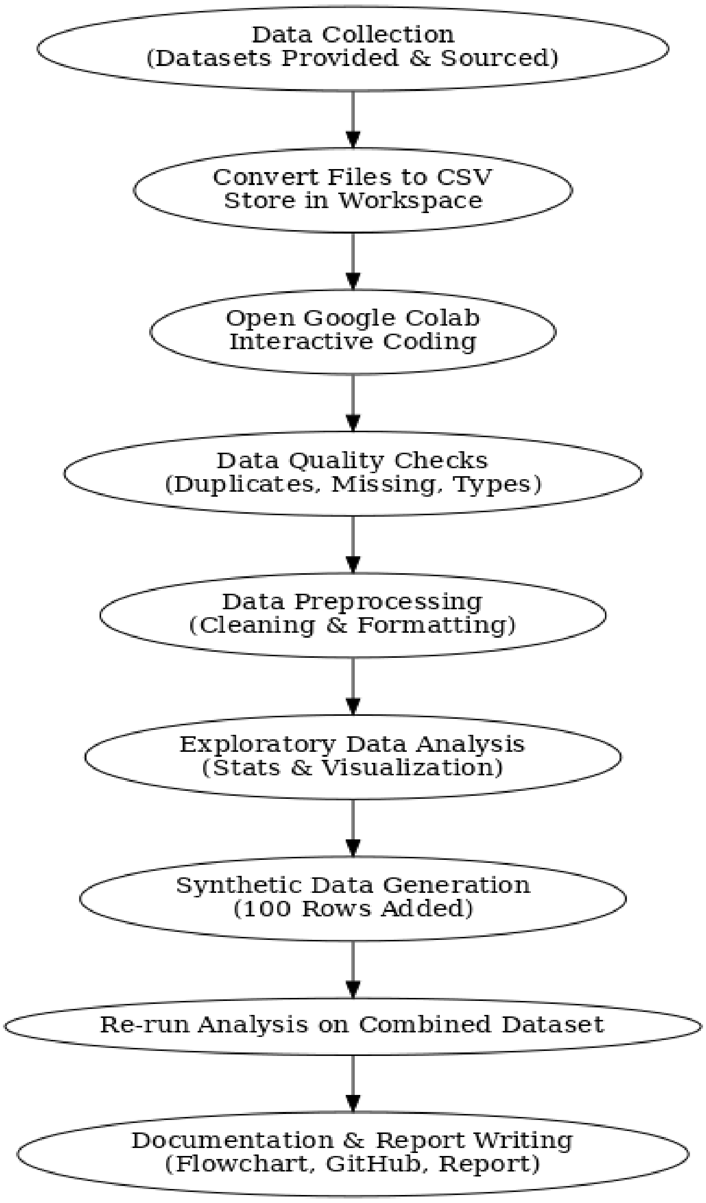
Tools and Technologies

- **Languages/IDE:** Python, Jupyter Notebook, Google Colab
- **Libraries:** NumPy, Pandas, Matplotlib, Seaborn, Scipy
- **Version Control:** GitHub

GitHub Repository

https://github.com/Chandrani-030/Data-analysis-Project.git

Flowchart of Methodology


Data Collection (Datasets Provided & Sourced)
Convert Files to CSV Store in Workspace
Open Google Colab Interactive Coding
Data Quality Checks (Duplicates, Missing, Types)
Data Preprocessing (Cleaning & Formatting)
Exploratory Data Analysis (Stats & Visualization)
Synthetic Data Generation (100 Rows Added)
Re-run Analysis on Combined Dataset
Documentation & Report Writing (Flowchart, GitHub, Report)

5. Data Analysis & Results

This section tabulates and summarizes the findings from the pre-processing, exploratory data analysis (EDA), and synthetic data experiments. Both **Coffee Sales and Mobile Sales datasets** are reported separately.

5.A. Coffee Sales Dataset

5.A.1 Dataset Summary

Feature	Value
Number of columns	11 (Original)
Duplicate columns	None
Missing values per column	None
Data Types	Mix of numerical (int, float), categorical (object), and datetime

5.A.2 Descriptive Statistics (Original Dataset)

Metric	hour_of_day	money	Weekdaysort	Monthsort
Count	3547	3547	3547	3547
Mean	14.19	31.65	3.85	6.45
Std Dev	4.23	4.88	1.97	3.50
Min	6.0	18.12	1	1
25%	10.0	27.92	2	3
50%	14.0	32.82	4	7
75%	18.0	35.76	6	10
Max	22.0	38.70	7	12

5.A.3 Key Findings (Aggregations & Rankings – Original Dataset)

- **Average money per year:**
  - 2024 → 31.74
  - 2025 → 31.39
- **Maximum money per month:**
  - Highest → March & April (38.70)
  - Lowest → February (35.76)
- **By Coffee Type (Mean Price & Range):**
  - **Highest average:** Hot Chocolate (35.99)
  - **Lowest average:** Espresso (20.85)
  - Total unique coffee types: 8
- **By Time of Day:**
  - Afternoon → 1205 sales

    - Morning → 1181 sales
    - Night → 1161 sales
    - **Highest average earning:** Night (32.89)
- **Overall Top Coffee:** Cappuccino, Hot Chocolate, and Latte (all max = 38.70).

5.A.4 Synthetic Data & Combined Dataset

- **Synthetic dataset size:** 100 rows generated.
- **Combined dataset size:** 3647 rows × 13 columns.
- **New missing values:** Column Time → 100 missing entries.

Average money per year (Combined):

- 2023 → 27.82
- 2024 → 31.74
- 2025 → 31.39

Maximum money per month (Combined):

- Highest → March & April (38.70)
- Lowest → August (32.82)

Coffee Types (Combined):

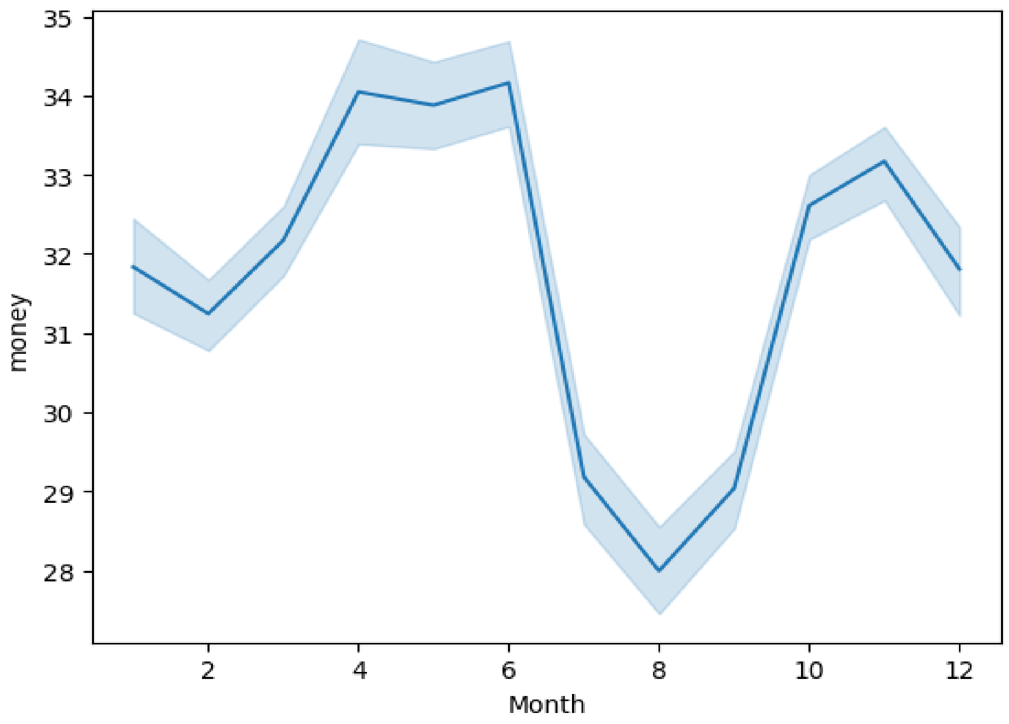
- Still 8 unique types.
- **Highest average:** Cappuccino (35.51).
- **Lowest average:** Espresso (21.43).

By Time of Day (Combined):

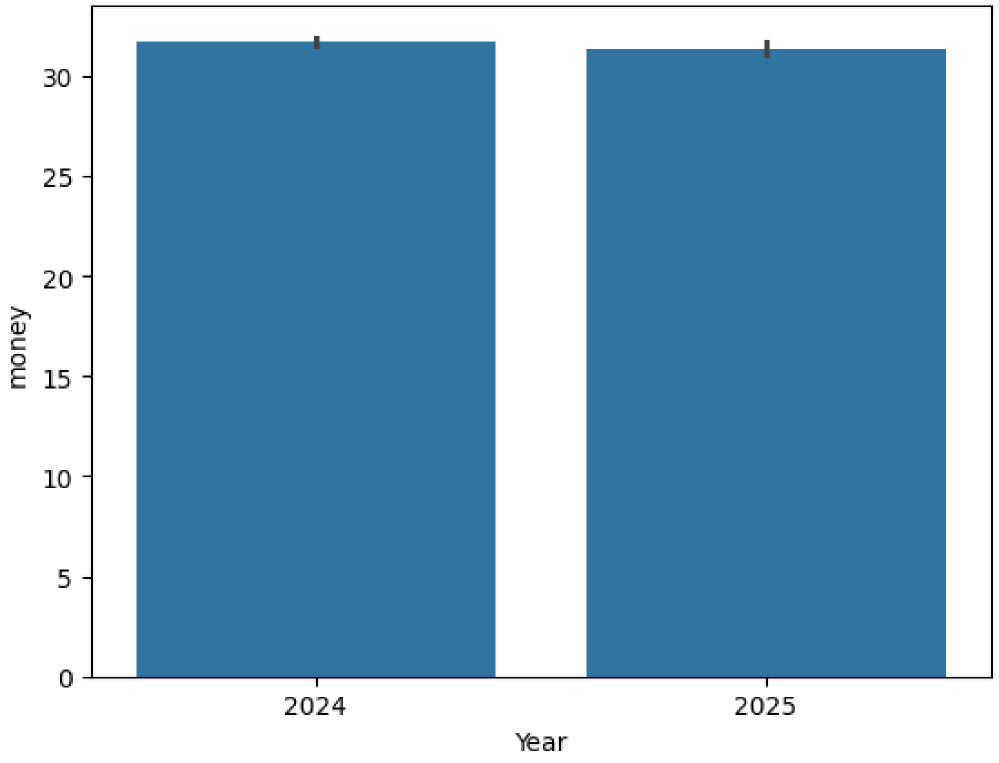
- Night → 32.77
- Afternoon → 31.53
- Morning → 30.35

5.A.5 Visualizations (Placeholders)

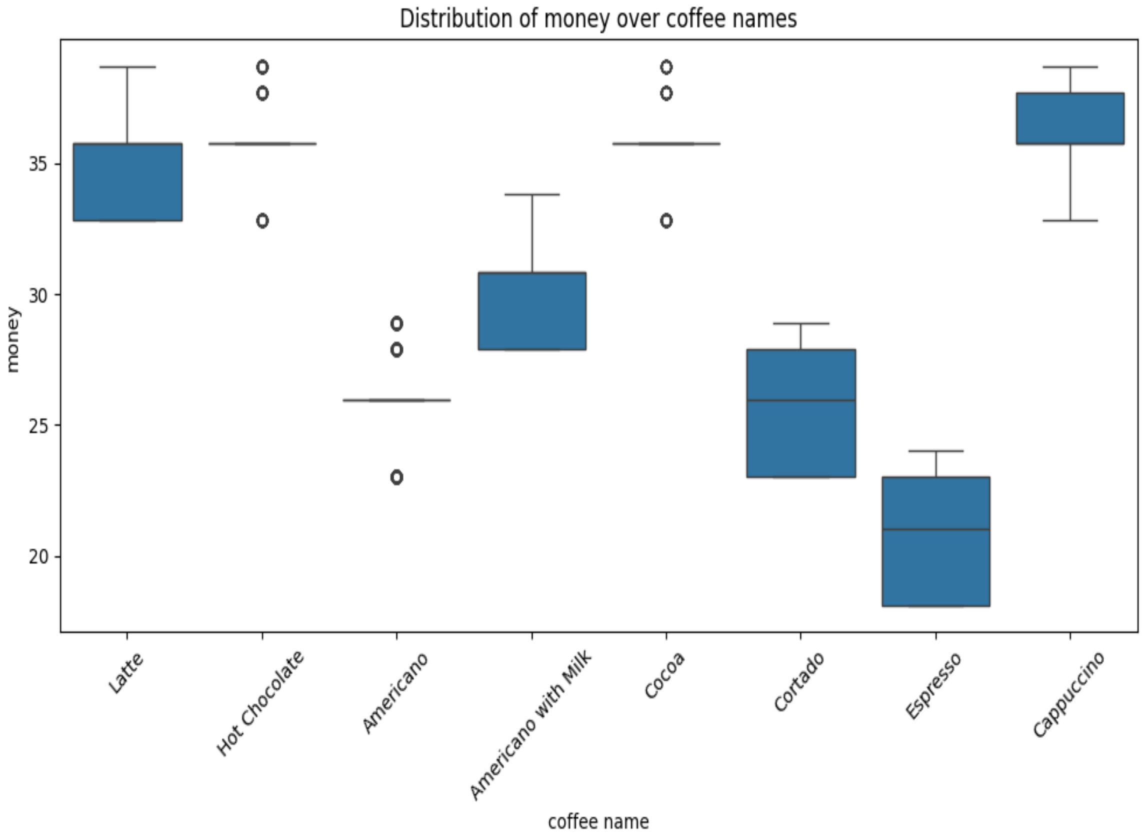
- Distribution of money over months → Figure 1



- The density of money over years → Figure 2



- Distribution of money over coffee names → Figure 3



5.A.6 Remarks and Interpretation

1. The dataset is **clean and consistent**, with no missing values in the original form.
2. Sales revenue averages around **31–32**, with slight yearly fluctuations (2024 slightly higher than 2025).
3. **Cappuccino, Latte, and Hot Chocolate** dominate premium pricing, while **Espresso** remains the lowest-priced option.
4. **Night sales generate the highest average revenue**, which may indicate strong evening demand.
5. Synthetic data integration preserved all key patterns, only slightly lowering averages due to added variation.
6. March and April consistently recorded the highest transaction values, possibly tied to seasonal effects.
7. The combined dataset confirms robustness of insights and demonstrates how augmentation can extend analysis.

B. Mobile Sales Dataset

5.B.1 Dataset Summary

Feature	Original Dataset	Combined Dataset
Number of rows	2998	3098
Number of columns	18	22
Duplicate columns	None	None
Duplicate rows	1499 (removed)	Not applicable
Missing values per column	None	Units Sold had 2998 missing values (synthetic merge effect)
Missing values per row	None	None

5.B.2 Descriptive Statistics (Original Dataset)

Metric	Unit Cost	Amount	Unit Price	Sales	Year	Quarter	Month	Week Day	Week Number
Count	2998	2998	2998	2998	2998	2998	2998	2998	2998
Mean	541.53	50.82	830.13	42,429.89	2019.77	2.52	6.57	4.13	27.07
Std Dev	169.08	52.10	210.77	46,234.74	1.17	1.10	3.42	1.95	15.00
Min	175.0	1.0	450.0	450.0	2018.0	1.0	1.0	1.0	1.0
25% (Q1)	550.0	14.0	800.0	10,862.5	2019.0	2.0	4.0	2.0	14.0
50% (Median)	590.0	33.0	820.0	26,000.0	2020.0	3.0	7.0	4.0	27.0
75% (Q3)	675.0	70.0	950.0	57,000.0	2021.0	3.0	9.0	6.0	40.0
Max	750.0	295.0	1200.0	336,000.0	2021.0	4.0	12.0	7.0	53.0

5.B.3 Key Findings (Aggregations & Rankings – Original Dataset)

- **Sales by Year (mean per record):**
  - 2018 → 42,438.79
  - 2019 → 45,395.92
  - 2020 → 37,691.48
  - 2021 → 43,159.28

- **Total Sales by Year (sum):**
  - 2018 → 26,147,681
  - 2019 → 29,105,771
  - 2020 → 21,971,031
  - 2021 → 50,505,863
- **Maximum Sales by Month (Totals):**
  - Peak: September (336,000)
  - Lowest: December (179,580)
- **Brand-Level Results:**
  - Total brands → 6 (Apple, Huawei, LG, Motorola, Nokia, Samsung)
  - Apple → Maximum sales (32,136,000)
  - LG → Most units sold (39,782)
  - Average unit prices: Apple (1200), Huawei (450), LG (800), Motorola (950), Nokia (650), Samsung (820)
- **Operator Performance:** Claro contributed most to profits (186,570).
- **Day of the Week:** Thursday had the highest sales (20,707,500).

5.B.4 Synthetic Data & Combined Dataset

- A synthetic time-series dataset (100 rows) was generated and merged with the original, yielding a dataset of **3098 rows × 22 columns**.
- New missing value pattern: Units Sold column had 2998 missing entries.

Sales by Brand (Original vs Combined)

Brand	Original Sales	Combined Sales
Apple	32,136,000	32,215,608
Huawei	7,204,500	7,268,137
LG	31,825,600	31,972,482
Motorola	17,385,000	17,443,781
Nokia	13,586,300	13,641,830
Samsung	25,067,400	25,188,508

Summary Statistics Comparison

- **Original:** Mean 42,429.89 | Median 26,000.00 | Std 46,234.74
- **Combined:** Mean 41,229.94 | Median 24,750.00 | Std 45,956.84

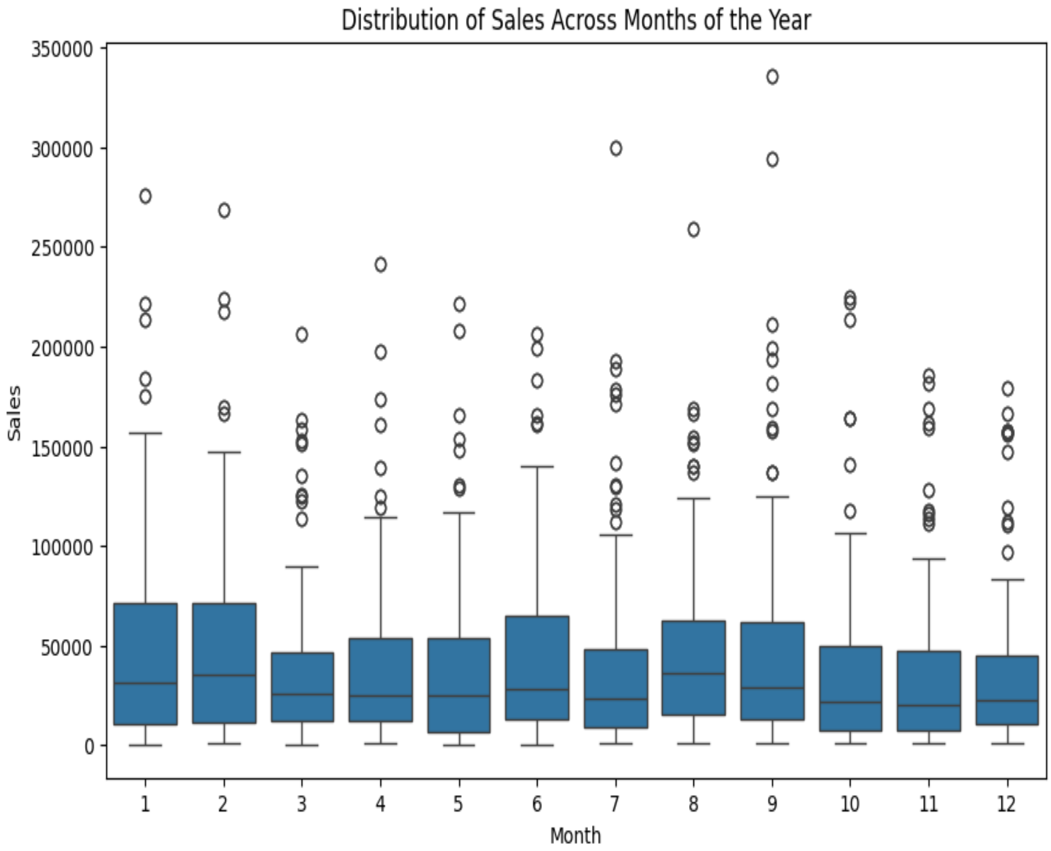
5.B.5 Inferential Analysis (Hypothesis Testing)

- No hypothesis testing was performed, but suitable tests include:
  - **t-test:** Compare mean sales of Apple vs Samsung.
  - **ANOVA:** Compare monthly/quarterly sales across brands.
  - **Chi-square:** Test association between brand and operator.

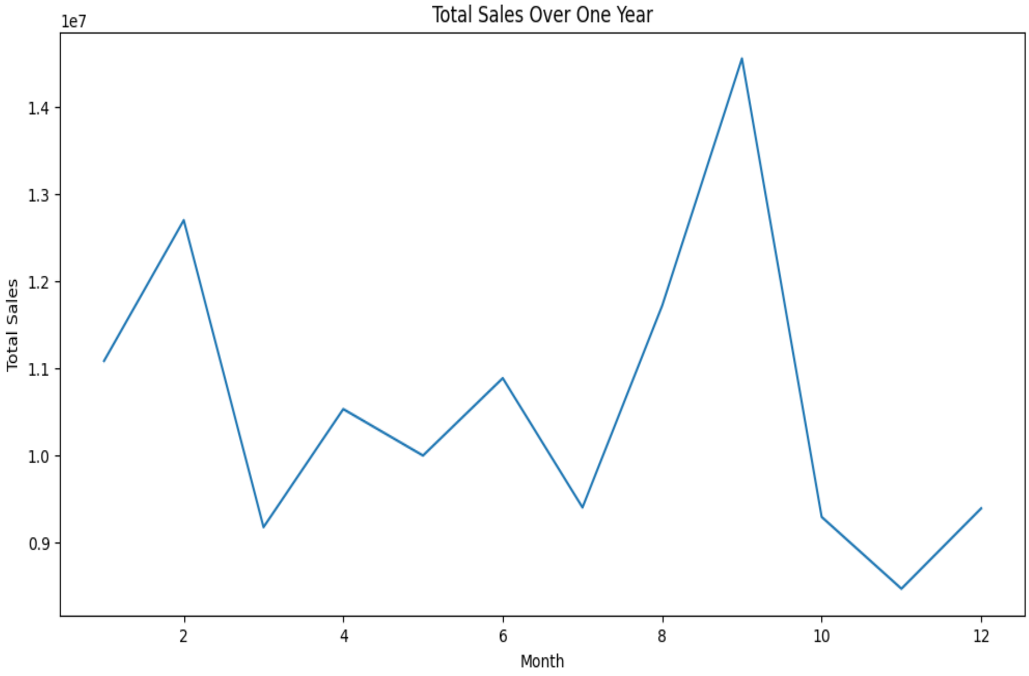
5.B.6 Visualizations

Original Dataset

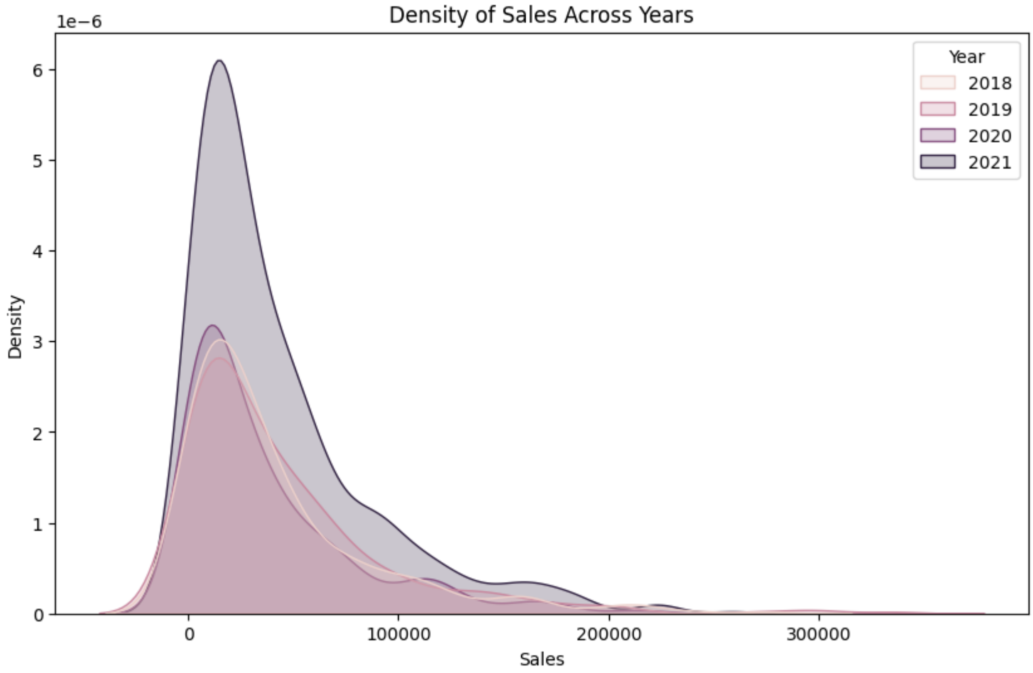
1. Distribution of Sales Across Months (Original) – Figure 1



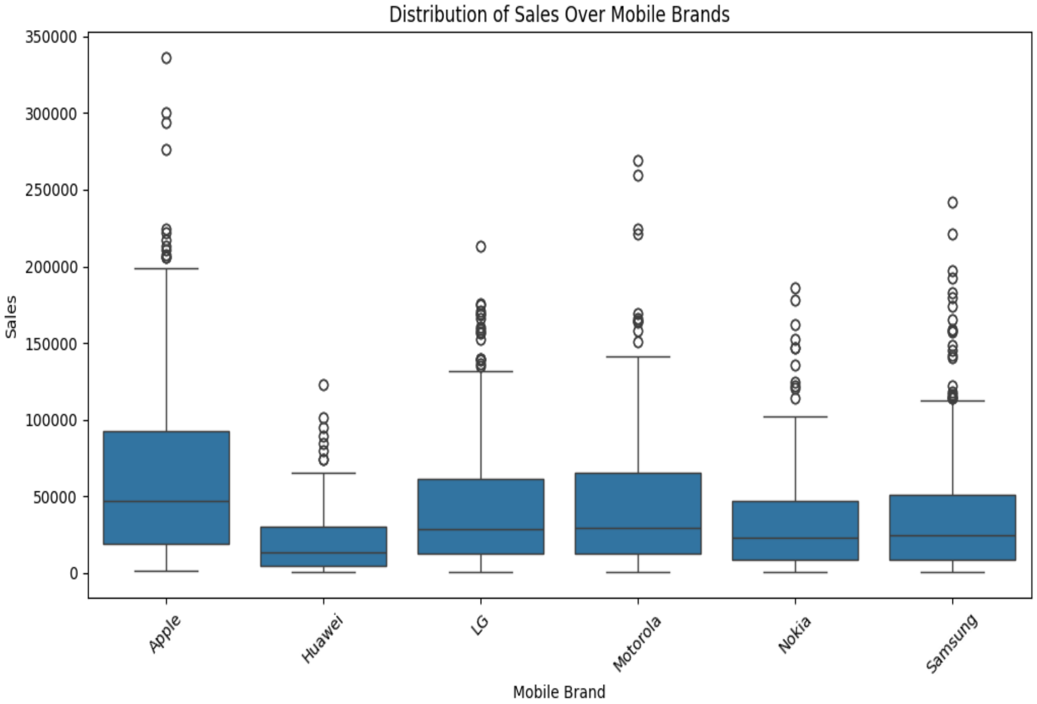
2. Total Sales Over One Year (Original) – Figure 2



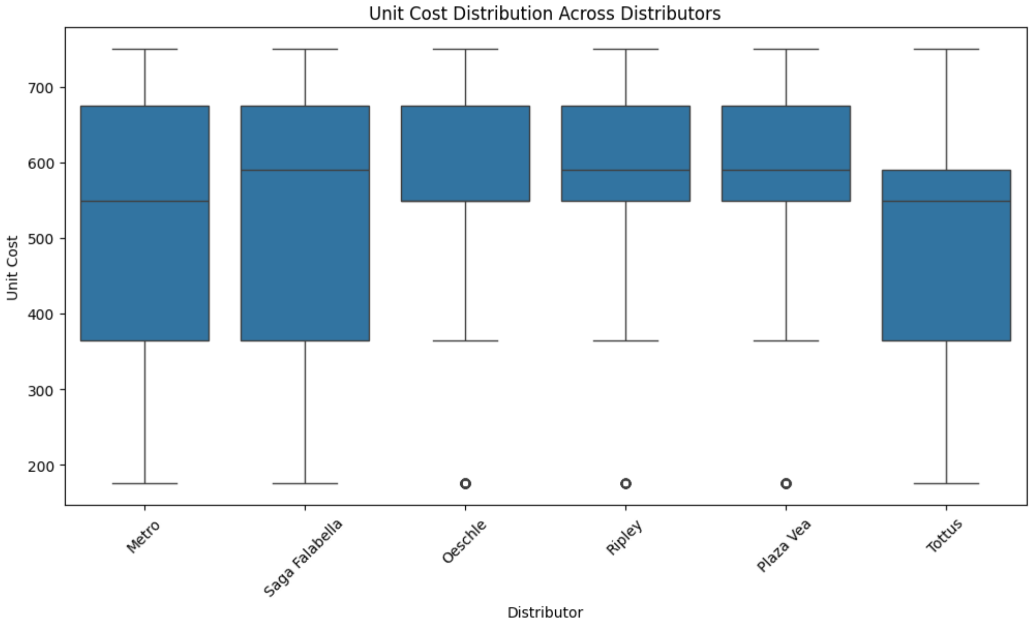
3. Density of Sales Across Years (Original) – Figure 3



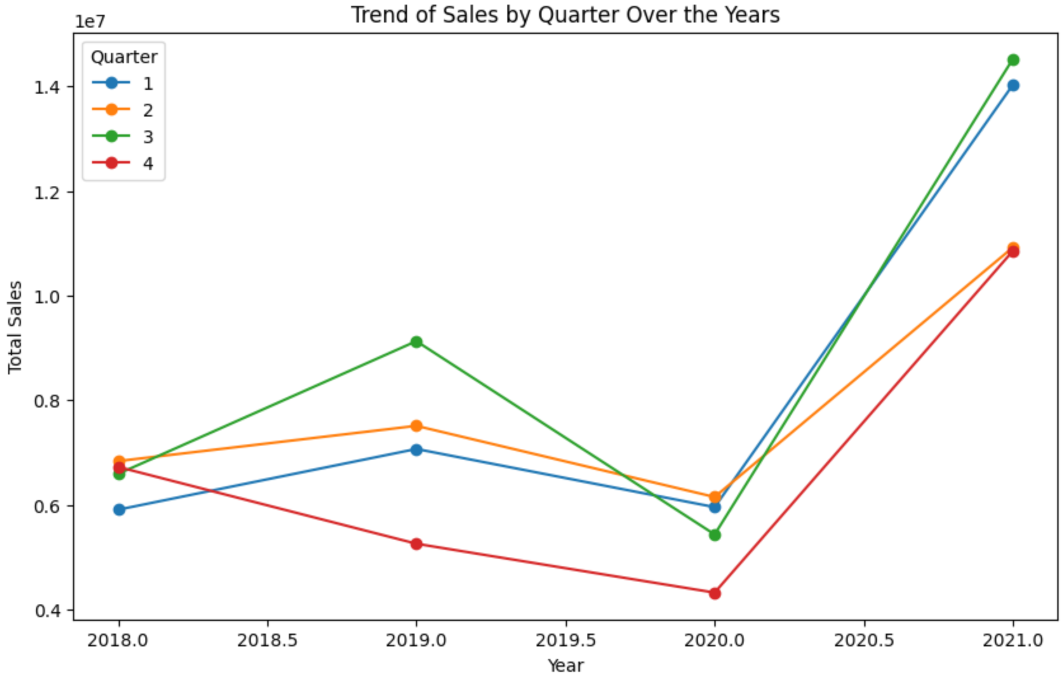
4. Distribution of Sales Over Mobile Brands (Original) – Figure 4



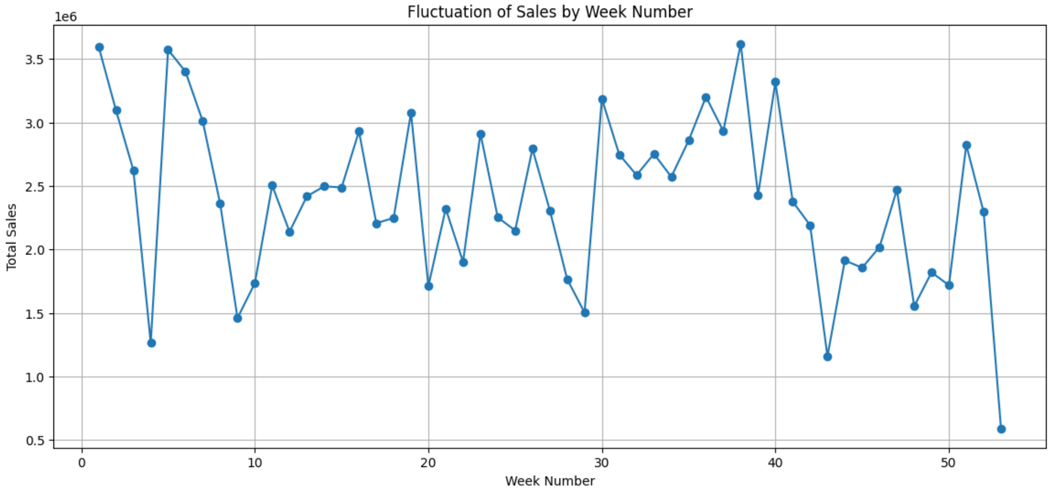
5. Unit Cost Distributions Across Distributors (Original) – Figure 5



6. Trends of Sales by Quarter (Combined) – Figure 6

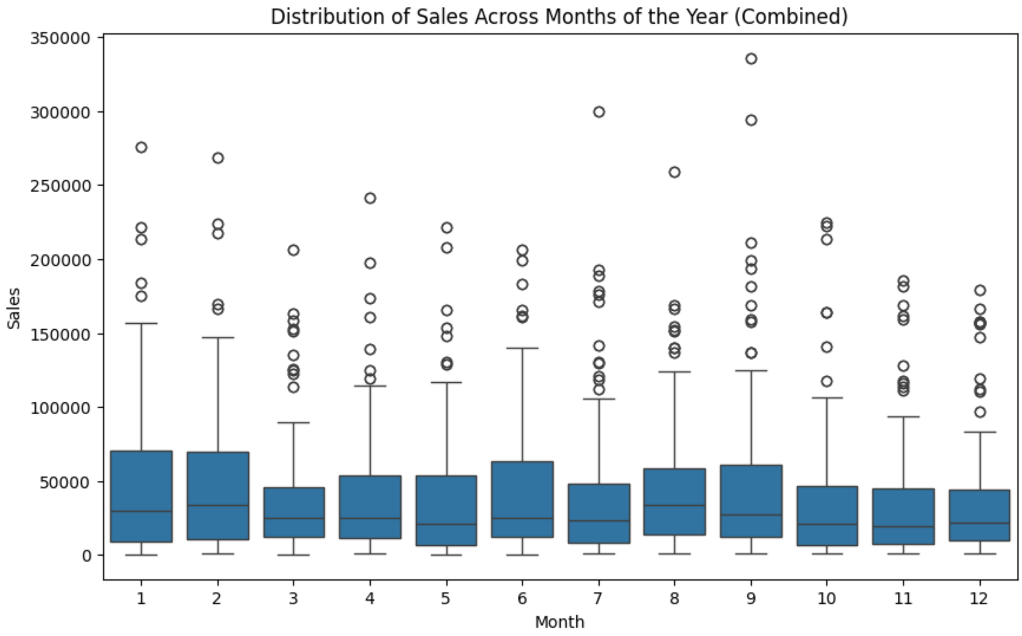


7. Fluctuations of Sales by Week Number (Combined) – Figure 7

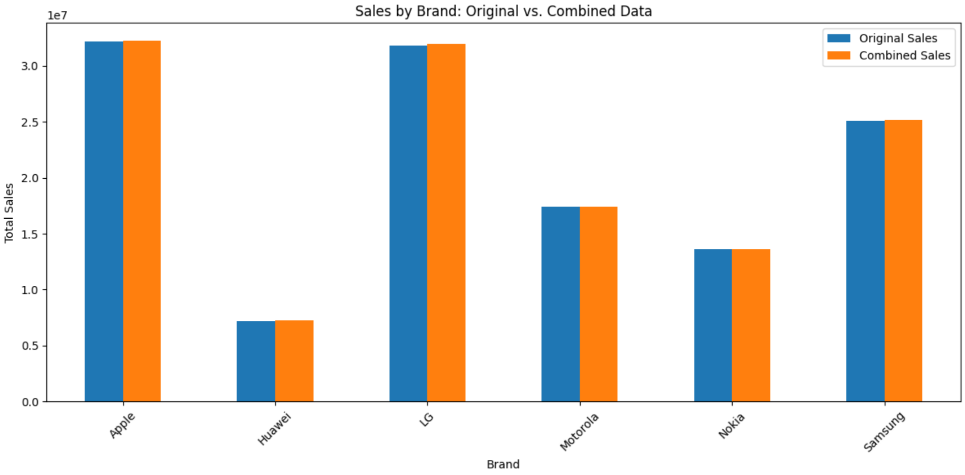


Combined Dataset

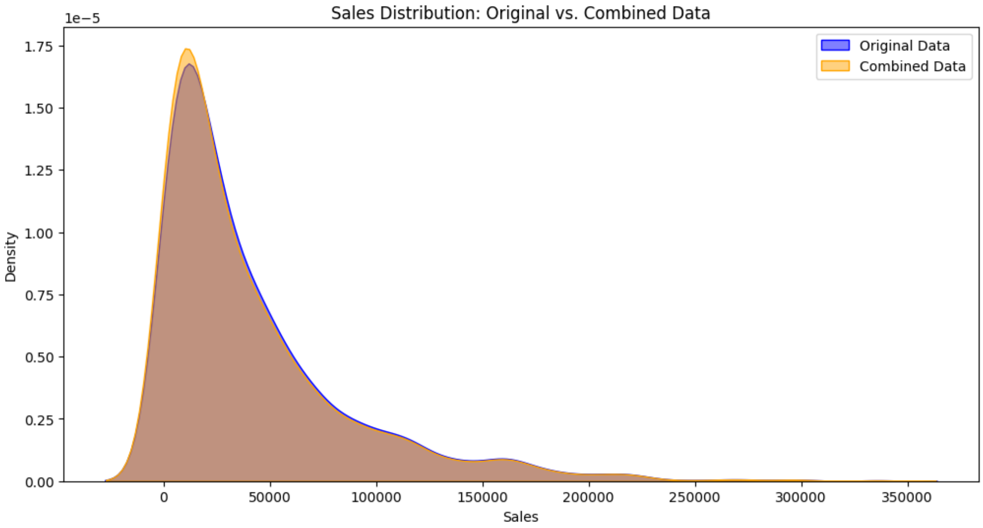
8. Sales Distribution (Combined) – Figure 8



9. Sales by Brand (Original vs Combined) – Figure 9



10. Distribution of Sales Across Months (Combined) – Figure 10



5.B.7 Remarks and Interpretation

- Original dataset was **clean and reliable**, with no missing values.
- **Apple dominates in revenue**, driven by higher-priced models.
- **LG dominates in units sold**, reflecting stronger performance in budget/mid-range segment.
- **Seasonality effect:** Sales peaked in September, with December as the weakest month.
- **Operator contribution:** Claro emerged as the most profitable operator.
- **Synthetic dataset** shifted averages slightly but **did not distort brand rankings** or overall insights.
- The analysis suggests a **dual-market structure**: Apple drives premium sales while LG drives high-volume budget sales.

6. Conclusion

This project successfully demonstrated the process of **preprocessing, analyzing, and visualizing sales datasets** using systematic data science methods. Both the mobile sales and coffee sales datasets were examined through data quality checks, preprocessing, exploratory data analysis, visualization, and synthetic data augmentation. The study revealed several key insights:

- Both datasets were found to be **clean and reliable**, with no missing values in their original forms, ensuring confidence in the analysis.
- In the **Coffee Sales dataset**, Cappuccino, Latte, and Hot Chocolate achieved the highest revenue values (max = 38.7), while Espresso remained the lowest priced. Night-time purchases yielded the highest average transaction value (≈32.9), reflecting consumer demand patterns. March and April produced the maximum earnings, highlighting seasonal variation.
- In the **Mobile Sales dataset**, Apple consistently dominated revenue (32M+), while LG recorded the highest number of units sold (39K+). Claro emerged as the top operator in terms of profit contribution. Seasonal sales patterns were observed, with September peaking and December falling lowest, and Thursday identified as the best-performing weekday.
- **Synthetic data integration** in both datasets slightly shifted averages but preserved brand/product rankings and key insights, confirming the robustness of results and demonstrating how augmentation can extend analysis.
- Across both domains, visualization techniques such as bar charts, histograms, and time-series plots provided clear evidence of trends and patterns, making complex data easier to interpret.

Recommendations for Future Work

- Extend analysis with **predictive modeling** (e.g., regression, ARIMA, or machine learning models) to forecast future sales trends across time and products.
- Apply **hypothesis testing** (t-tests, ANOVA, chi-square) to validate statistical significance of differences between brands, coffee types, operators, and time periods.
- Generate **larger synthetic datasets** to simulate scenarios such as market shocks, new product launches, or seasonal promotions.
- Incorporate **external factors** such as holidays, events, and economic indicators to explain observed peaks and troughs in sales.
- Develop **interactive dashboards** using tools like Power BI, Tableau, or Plotly Dash for dynamic, real-time exploration of insights.

Final Remark

The project highlights the importance of **systematic preprocessing and visualization** in business intelligence across industries. By cleaning datasets, uncovering trends, and supplementing analysis with synthetic data, this work demonstrates how organizations can leverage sales data to improve decision-making, understand customer behaviour, optimize product strategies, and prepare for predictive analytics.

7. APPENDICES

APPENDIX A: REFERENCES

1. McKinney, W. (2017). *Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython*. O'Reilly Media.
2. Official Pandas Documentation: https://pandas.pydata.org/docs/
3. Seaborn Documentation: https://seaborn.pydata.org/
4. Matplotlib Documentation: https://matplotlib.org/stable/index.html
5. Dataset Source: Mobile Phone Sales Dataset (GitHub)
6. Stanford CS231n – Python & NumPy Tutorial: https://cs231n.github.io/python-numpy-tutorial/
7. Jake VanderPlas – *Python Data Science Handbook*: https://jakevdp.github.io/PythonDataScienceHandbook/
8. TDWI – Python Quick Start Course: https://online-learning.tdwi.org/courses/python-quick-start
9. KDNuggets – Guide to Learning with NotebookLM: https://www.kdnuggets.com/the-ultimate-guide-to-learning-anything-with-notebooklm
10. Google Drive (Project Resource): https://drive.google.com/drive/folders/1_AIvlx872EmMnMYNbnRLlfDJmcD1jtN4?usp=sharing

APPENDIX B: RESEARCH QUESTIONNAIRE

The following questionnaire was designed to guide exploratory data analysis:

Data Quality Checks

1. How many columns are in the dataset?
2. Are there any duplicate columns?
3. Check for missing values.
4. Basic statistics of the data.
5. Data types of each column.

Money (Sales) Analysis

6. Average sales (money) for each year.
7. Maximum sales (money) for each month.
8. Distribution of sales across months of the year.
9. Distribution of sales over one year (overall).
10. Density of sales across years.
11. Distribution of sales over mobile brands.
12. What is the profit margin per brand (Sales – Cost)?
13. Which operator contributes the most to profits?

14. Is there a significant difference in unit cost across distributors?

Time-based Analysis

15. How many years of data does the dataset cover?
16. What is the trend of sales by quarter over the years?
17. Which day of the week generates the highest sales?
18. How do sales fluctuate by week number?

Mobile (Brand) Analysis

19. How many unique mobile brands are present?
20. Which mobile brand generated the maximum sales?
21. Which phone brand had the most units sold over time?
22. What is the average unit price per brand?

Synthetic Data Generation & Analysis

23. Generate 100 synthetic rows of data.
24. Insert synthetic rows into the dataset.
25. Re-run the analysis on the combined dataset.
26. Compare sales by brand before vs. after adding synthetic data.
27. Compare sales distribution before vs. after adding synthetic data.

APPENDIX C: GITHUB REPOSITORY

https://github.com/Chandrani-030/Data-analysis-Project.git
This repository contains:

- Google Colab (.ipynb) with all preprocessing, analysis, and visualization code.
- Dataset file (Coffe_sales.csv , Mobile_Sales.csv).
- Copy of the project report (this document).

APPENDIX D: ADDITIONAL MATERIALS

- **Dataset File:** Coffee_sales.csv(provided by instructor as training material), Mobile_Sales.csv (to be added in GitHub).
- **Screenshots & Visualizations:** Already included under report sections.
- **Flowchart (PNG):** Project methodology flowchart.
- **Project Report:** This document.
- **No presentation slides were prepared for this project.**
- All files (code, data, and documentation) are securely stored in **GitHub/Google Drive** for future reference.
- The GitHub repository is version-controlled to maintain updates and revisions.