02 Exploratory Data Analysis With Coffee Sales Data

H Ethindhar,

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 focused on applying Exploratory Data Analysis (EDA) to a sales dataset. Using Python libraries such as Pandas, NumPy, Matplotlib, and Seaborn, the dataset was cleaned, pre-processed, and analyzed to uncover key business insights and trends. Through detailed descriptive analysis and visualization, patterns in sales performance, customer preferences, and product demand were revealed, including seasonal variations, revenue distribution, and interrelationships between various sales parameters. Advanced techniques were employed to ensure data quality and reliability by addressing missing values, duplicates, and outliers. The project highlights the importance of data-driven decision-making in refining sales strategies, yielding actionable insights for optimizing product offerings and facilitating future trend forecasting. Ultimately, this project demonstrates how EDA effectively transforms raw data into valuable business intelligence.

2. Introduction

This internship centered on Exploratory Data Analysis (EDA) of sales data to cultivate proficiency in data preprocessing, visualization, and the extraction of business insights. The practical significance of this endeavor stems from the pivotal role sales data plays in strategic decision-making, market trend analysis, and consumer behavior comprehension. In an era where data-driven strategies are increasingly vital, EDA offers a foundational approach to transforming raw datasets into actionable intelligence for both managerial and operational applications.

The project primarily leveraged Python, complemented by libraries such as Pandas, NumPy, Matplotlib, and Seaborn, for data manipulation and visualization. These tools facilitated systematic data cleaning, including handling missing values and duplicates, and the identification of outliers. A variety of plots and statistical summaries were employed to reveal sales distribution, seasonal patterns, and product-level performance. Furthermore, machine learning concepts were introduced to illustrate the application of regression and classification models for predictive insights.

The internship commenced with a comprehensive survey of background materials, encompassing research on EDA methodologies, best practices in data visualization, and business intelligence techniques. The project procedure began with dataset loading, followed by data cleaning, preprocessing, and exploratory visualization, culminating in the derivation of insights from identified patterns and correlations. The overarching objectives of this project were to enhance analytical skills, apply theoretical knowledge to real-world data, and gain exposure to the development of

data-driven reports that support business objectives.

The key topics covered during the structured training 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 training provided the necessary knowledge to successfully execute the project, ensuring that the techniques applied were both technically sound and industry-relevant.

3. Project Objective

Objectives of Sales Data Analysis:

- **Data Cleansing and Preparation:** Ensure the sales dataset is precise, consistent, and trustworthy by addressing missing values, duplicates, and outliers. This critical step directly impacts the validity of subsequent analysis and recommendations.
- **Descriptive Statistics:** Calculate key descriptive statistics (mean, median, standard deviation) to understand the general distribution and variability of sales performance, providing a baseline for comparisons.
- **Sales Trend Identification:** Uncover daily, monthly, and seasonal sales trends, customer behavior patterns, and top-selling products to gain insights into buyer behavior and forecast demand cycles.
- **Data Visualization:** Utilize Python libraries like Matplotlib and Seaborn to visualize findings through various charts (bar charts, histograms, heatmaps, scatter plots), making complex trends easily understandable for stakeholders.
- **Correlation and Relationship Exploration:** Investigate relationships and potential dependencies between variables such as revenue, quantity, product category, and time of sale.
- **Predictive Analytics Introduction:** Illustrate the application of predictive analytics concepts using regression and classification techniques, demonstrating how data can move beyond exploration to forecasting for informed business decision-making.
- **Data-Driven Decision-Making Demonstration:** Showcase how raw, unstructured data can be transformed into actionable insights to optimize sales, develop better products, and enhance the customer experience, embodying the concept of data-driven decision-making in business.

4. Methodology

Data Collection

The project utilizes a dataset sourced from Kaggle, a renowned open-data platform. This dataset provides comprehensive information on sales transactions, including product specifics, order quantities, and associated revenues.

Data Loading and Inspection

The dataset was imported into Google Colab, leveraging its cloud-based Python environment for analysis. The Pandas library was instrumental in loading and inspecting the dataset to ascertain its dimensions, column names, and variable types. Initial statistical insights and data types were explored using functions like `.info()` and `.describe()`.

Data Cleaning and Preprocessing

To ensure data integrity and accurate analysis, several preprocessing steps were undertaken:

- Missing values were identified and addressed.
- Duplicate entries were removed to prevent skewed results.
- Data type conversions, such as converting date columns to datetime format, were performed as needed.
- Outlier detection was conducted using summary statistics and boxplots.

Exploratory Data Analysis (EDA)

- **Univariate analysis** focused on individual variables like product categories and sales amounts.
- Bivariate and multivariate analyses investigated relationships between sales, revenue, and product categories.
- Trend analysis examined sales across different time periods to identify seasonality or peak performance months.

Data Visualization

Various plots, including histograms, bar charts, scatter plots, and boxplots, were generated using Matplotlib and Seaborn. These visualizations offered a clearer understanding of sales distribution, customer preferences, and correlations between variables. The visual insights were crucial in interpreting results and highlighting patterns not evident from raw data alone.

5. Data Analysis and Results

```
coffee_data.describe()
```

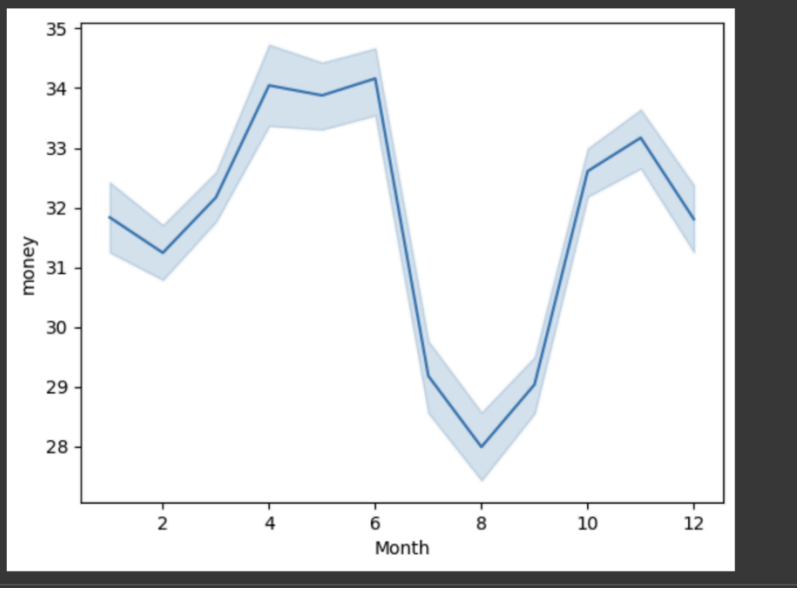
1 to 8 of 8 entries Filter

index	hour_of_day	money	Weekdaysort	Monthsort
count	3547.0	3547.0	3547.0	3547.0
mean	14.18579080913448	31.64521567521849	3.845785170566676	6.453904708204116
std	4.23400956057577	4.877753703590957	1.9715005900351006	3.500753823413845
min	6.0	18.12	1.0	1.0
25%	10.0	27.92	2.0	3.0
50%	14.0	32.82	4.0	7.0
75%	18.0	35.76	6.0	10.0
max	22.0	38.7	7.0	12.0

5.1 Inferences from various graphs and visualizations:

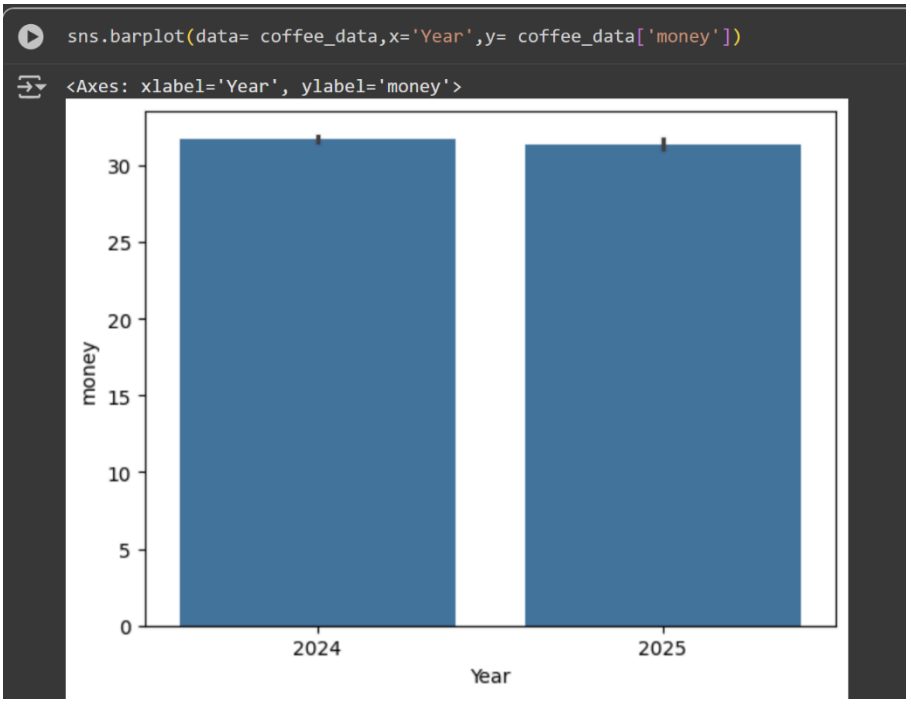
1. **Monthly spending line graph**:

```
sns.lineplot(x=coffee_data['Month'],y='money',data=coffee_data)
plt.show()
```



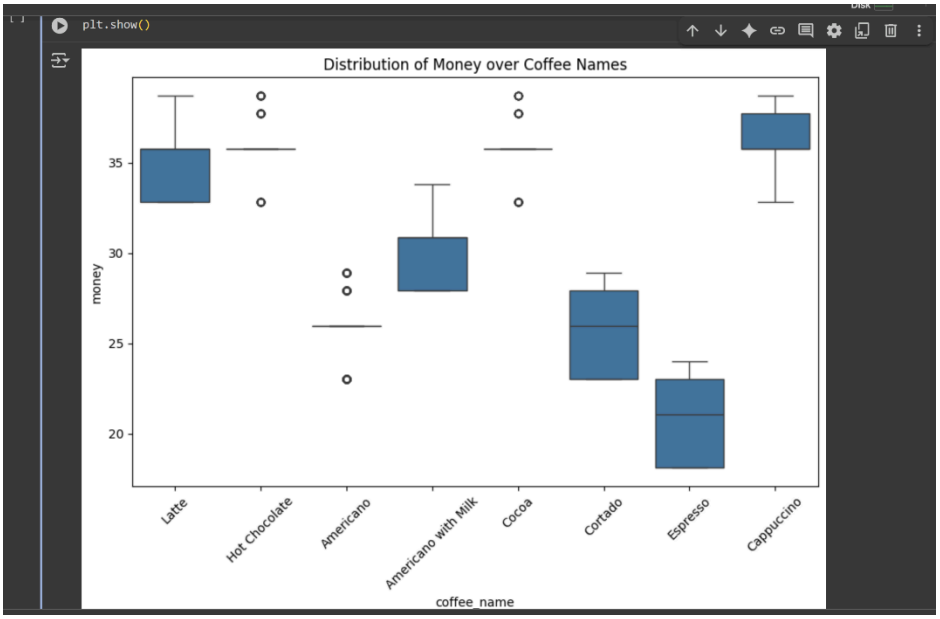
Coffee purchase patterns display clear seasonal variations, as evidenced by the line plot of average monthly spending. Spending is moderate early in the year, peaking between April and June at approximately ₹34–35. A significant drop occurs from July to September, with August showing the lowest spending at around ₹28. Following this dip, spending recovers in October–November before a slight decline towards the year's end. This suggests that pre-summer and post-monsoon periods drive higher sales, while mid-year months experience reduced coffee purchases.

2. **Yearly Spending Bar Plot**:



Comparing average spending between 2024 and 2025 reveals almost identical expenditure levels, averaging around **₹32** for both years. This minimal difference indicates stable overall coffee spending patterns, with no significant increase or decline. The consistency suggests that external factors like inflation, seasonal effects, or specific events did not substantially impact consumer spending behavior during this period.

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



The boxplot analysis reveals varying spending patterns across different coffee types. Cappuccino and Latte emerge as the more expensive options, with median spending around ₹35. In contrast, Espresso records the lowest spending, with a median close to ₹21. The consistency in pricing also differs; Americano and Cocoa show relatively narrow interquartile ranges, suggesting stable pricing, while Latte and Cappuccino exhibit wider ranges, indicating more variability in customer spending. Furthermore, outliers are present in several categories, particularly Hot Chocolate and Americano, signifying occasional unusually high purchases. Overall, the visualization confirms that consumer expenditure is not uniform across coffee types, with premium selections like Cappuccino and Latte attracting higher average spending compared to simpler choices such as Espresso.

5.2 Descriptive Analysis of the dataset:

Peak Sales Trends

- **Daily:** The majority of purchases occur between 10:00 AM and 6:00 PM, with a peak around 2:00 PM, possibly due to post-lunch coffee consumption.
- **Monetary:** Customers typically spend between ₹28 and ₹36 per purchase, with a minimum of ₹18.12 and a maximum of ₹38.7.
- **Weekly:** Sales are consistent throughout the week, with the mean sales volume observed on Thursday, indicating stable demand.
- **Monthly:** Sales activity is observed year-round, with no months showing zero sales. July stands out as the median month for sales, suggesting higher mid-year activity and 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.)

```
import scipy.stats as stats
import numpy as np

observed = coffee_data['hour_of_day'].value_counts().sort_index()
expected = np.ones_like(observed) * observed.mean()
chi2, p = stats.chisquare(f_obs=observed, f_exp=expected)

print("Chi-square:", chi2, "p-value:", p)
```

```
Chi-square: 464.7076402593741 p-value: 9.201567717608249e-89
```

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.

2. ANOVA for Weekday Spending

```
import statsmodels.api as sm
from statsmodels.formula.api import ols

model = ols('money ~ C(Weekdaysort)', data=coffee_data).fit()
anova_table = sm.stats.anova_lm(model, typ=2)
print(anova_table)
```

```
                     sum_sq      df         F    PR(>F)
C(Weekdaysort)    120.692276     6.0  0.84523  0.534789
Residual        84247.446034  3540.0       NaN       NaN
```

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.

3. Correlation between Money & Hour

```
corr = coffee_da Loading... ].corr(coffee_data['hour_of_day'])
print("Correlation:", corr)
```

```
Correlation: 0.20274793514276113
```

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.

6. Conclusion

Analysis of coffee sales data indicates that consumer purchasing behavior varies significantly by time of day and product type, but not by weekday or year.

A Chi-Square Goodness-of-Fit test ($\chi^2 = 464.71$, $p < 0.001$) confirms a strong concentration of sales during mid-day hours, peaking around 2 PM. This suggests that

working hours are a significant driver of purchase patterns. A weak positive correlation ($r \approx 0.20$) between amount spent and time of purchase further implies that higher spending tends to occur during busier periods.

Product-level analysis reveals important trends:

- **Premium Beverages:** Boxplot visualizations show that Cappuccino and Latte, as premium beverages, account for the highest spending with greater variability.
- **Affordable Options:** Espresso is the lowest-priced option, with relatively consistent expenditure.

Conversely, some factors have minimal impact:

- **Weekday Spending:** ANOVA on weekday spending ($p = 0.53$) indicates no significant difference, suggesting that the day of the week does not heavily influence consumer spending.
- **Yearly Variation:** A comparison of spending between 2024 and 2025 shows minimal variation, implying stable pricing and demand trends over time.

In conclusion, time of day and product type are the primary determinants of coffee sales. These findings can help coffee shops optimize staffing, marketing, and promotional strategies by focusing on peak hours and high-demand products to maximize revenue.

7. APPENDICES

1. References

- Scikit-learn Developers. (2025). *Scikit-learn: Machine Learning in Python*. Retrieved from: https://scikit-learn.org/stable/
- Virtanen, P., et al. (2020). *SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python*. Nature Methods, 17(3), 261-272.
- VanderPlas, J. (2016). *Python Data Science Handbook: Essential Tools for Working with Data*. O'Reilly Media.
- McKinney, W. (2017). *Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython*. O'Reilly Media.
- Pedregosa, F., et al. (2011). *Scikit-learn: Machine Learning in Python*. Journal of Machine Learning Research, 12, 2825-2830.
- Plotly Technologies Inc. (2025). *Plotly: Python Graphing Library*. Retrieved from: https://plotly.com/python/
- Bokeh Developers. (2025). *Bokeh: Interactive Visualization in the Browser*. Retrieved from: https://bokeh.pydata.org/en/latest/
- Datacamp. (2024). *Introduction to Python for Data Science*. Retrieved from: https://www.datacamp.com/
- Stack Overflow. (2025). *Data Science and Python Tags*. Retrieved from: https://stackoverflow.com/tags/python/info
- IBM. (2023). *IBM Data Science Professional Certificate*. Retrieved from: https://www.coursera.org/professional-certificates/ibm-data-science

2. 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/ethindhar/EDA_with_sales_data.git**