Exploratory Analysis of Coffee Sales Data

from

Dharani Krishna Sahithi,

Branch:AI&ML

BMS Institute of Technology,Bengaluru,Karnataka

PERIOD OF INTERNSHIP: 21ST JANUARY 2026 – 17TH FEBRUARY 2026

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

1. Abstract

This project presents an exploratory analysis of transactional coffee sales data to identify patterns in revenue generation across different products, weekdays, and time periods. The dataset consists of recorded sales transactions including coffee type, time of day, weekday, and revenue. Descriptive statistical analysis and visualization techniques were employed to understand business performance trends. The study identifies the highest revenue-generating coffee products, examines weekday and time-of-day variations in sales, and analyzes monthly revenue trends to detect potential seasonality. The findings provide insight into customer purchasing behavior and operational patterns. The analysis primarily focuses on descriptive and trend-based insights suitable for the dataset's structure.

2.Introduction

The retail and food service industry increasingly depends on data-driven decision-making to optimize operational efficiency and improve revenue performance. Transactional sales data provides structured information about product demand, time-based trends, and revenue patterns. Analyzing such data enables businesses to identify high-performing products, understand seasonal variations, and make informed strategic decisions.

This project focuses on the exploratory analysis of a transactional coffee sales dataset stored in CSV format. The dataset contains attributes such as hour of sale, coffee type, time of day, weekday, month, and revenue. By applying structured data analysis techniques, the project aims to uncover meaningful revenue patterns across products, weekdays, and time periods.

• Relevance of the Project

Understanding sales behavior is critical for retail analytics. Before implementing advanced predictive models, it is essential to perform exploratory data analysis (EDA) to understand:

Revenue distribution patterns
Product-wise contribution to total sales
Time-of-day impact on sales
Weekday vs weekend performance
Monthly trends and possible seasonality

This project demonstrates how foundational data analysis techniques can provide actionable business insights.

• Technology and Tools Used

The analysis was performed using:

• Python Programming Language
• **Google Colab** environment for execution
• **Pandas** for data manipulation and aggregation

• **NumPy** for numerical computations
• **Matplotlib** for data visualization
• **Scikit-learn** for implementing a baseline Linear Regression model

The dataset was processed using aggregation methods such as groupby(), descriptive statistical functions such as describe(), and visualization techniques including bar and line plots.

• Background and Material Survey

Exploratory Data Analysis (EDA) forms the first stage of any data science pipeline. Literature and standard data science workflows emphasize that understanding data structure, distributions, and trends is essential before applying predictive models. Time-series revenue analysis often involves:

Converting date columns into datetime format
Extracting temporal features (month, year)
Aggregating sales by time units
Identifying patterns such as seasonality or trend

This project follows these standard analytical procedures to derive insights from structured transactional data.

• Procedure Followed

The project was executed in a systematic manner:

- Imported and loaded the dataset in CSV format.
- Performed initial data inspection (shape, duplicates, missing values).
- Generated descriptive statistics for numerical columns.
- Converted the 'Date' column to datetime format and extracted Month and Year.
  Computed:
  1. Average sales per year
  2. Maximum sales by month
  3. Total sales by coffee type
  4. Average sales by time of day
- Conducted additional exploratory analysis:
  Best-selling coffee
  Weekly revenue analysis
  Daily and monthly revenue trends
- Implemented a baseline Linear Regression model for monthly revenue forecasting.
- Evaluated the model using MAE, RMSE, and R^2 metrics.

• Purpose of the Project

The primary purpose of this project was:

• To apply foundational Python and data analysis skills in a real-world dataset.
• To understand how structured sales data can be transformed into meaningful business insights.

- To demonstrate the complete data analysis workflow from inspection to visualization and basic predictive modeling.
- To gain practical experience in exploratory and trend-based analysis.

Topics Covered During the First Two Weeks of Internship Training

- Introduction to internship expectations
- Python Basics – Data types, Variables, Lists, Loops
- Data Structures in Python
- Functions and Object-Oriented Programming (OOPS)
- NumPy and Pandas fundamentals
- Machine Learning Overview
- Regression concepts
- Classification concepts
- LLM Fundamentals
- Communication Skills

These training sessions provided the foundational knowledge required to complete this exploratory data analysis project and implement a basic regression model.

3.Project Objective

The primary objectives of this project are:

- To analyze the total revenue generated by different coffee products and identify the highest revenue-contributing items.
- To examine revenue variations across weekdays in order to understand weekly sales patterns.
- To study sales distribution across different time periods of the day (Morning, Afternoon, Night).
- To evaluate daily and monthly revenue trends to detect patterns, fluctuations, and potential seasonality effects.
- To derive meaningful business insights from exploratory data analysis using statistical aggregation and visualization techniques.
- To implement a baseline regression model for monthly revenue forecasting and evaluate its performance using standard metrics.

4.Methodology

4.1 Data Source and Collection

The dataset used in this project is a transactional coffee sales dataset provided as a CSV file (Coffe_sales.csv). The dataset contains 3,547 transaction records with 11 attributes including:

coffee_name
money (sales revenue)
Date
Time
Time_of_Day
Weekday
Month_name
and related temporal fields

The dataset was loaded into the Google Colab environment using the **pandas** library. No primary survey was conducted for this project. The analysis is entirely based on structured transactional data.

4.2 Tools and Technologies Used

The following tools and technologies were used:

Python (Google Colab environment)
Pandas – Data manipulation and aggregation
NumPy – Numerical operations
Matplotlib – Data visualization
Scikit-learn – Machine learning model implementation

4.3 Data Pre-processing and Cleaning

The following preprocessing steps were performed:

Loaded the dataset using **pd.read_csv().**
Checked dataset dimensions using **.shape.**
Verified duplicate records using **.duplicated().sum().**
Checked for missing values using **.isnull().sum().sum().**
Converted the Date column from object type to datetime using **pd.to_datetime().**
Extracted Month and Year from the datetime column using:
.dt.month
.dt.year
Created a Month_Index column for time-series regression modeling.

No missing values or duplicate records were found in the dataset.

4.4 Exploratory Data Analysis (EDA)

The analysis was performed using aggregation-based techniques:

○ Descriptive Statistical Analysis

Used **.describe()** to obtain:

- **Mean**
- **Standard deviation**
- **Minimum and maximum values**
- **Quartiles**

○ Aggregation-Based Analysis

Performed using **groupby():**

- **Total revenue by coffee type**
- **Maximum sales by month**
- **Average sales by year**
- **Average sales by time of day**
- **Revenue distribution by weekday**

○ Time-Based Trend Analysis

- Daily revenue trend visualization
- Monthly revenue trend visualization
- Weekly revenue pattern analysis

Visualizations were generated using **Matplotlib** to understand revenue fluctuations over time.

4.5 Machine Learning Model Development

Although the dataset was primarily intended for exploratory analysis, a basic time-series regression model was developed to forecast next month's revenue.

○ Model Used:

Linear Regression

○ Feature Engineering:

Created Month_Index as the independent variable.
Revenue was used as the dependent variable.

○ Train-Test Split:

Since the dataset consisted of monthly aggregated values:

Initial months were used as training data.
Final two months were used as test data.
A chronological split was used instead of random splitting to preserve time order.

○ Model Evaluation Metrics:

The model was evaluated using:

MAE (Mean Absolute Error)
RMSE (Root Mean Squared Error)
R^2 Score

Due to the small size of the test dataset, the R^2 value was unstable and negative, indicating that it is not a reliable metric for this dataset. However, MAE and RMSE showed that prediction deviation was approximately 2 units on average.
The model was then used to forecast the revenue for the next month.

4.6 GitHub Repository

All Python code used for analysis and modeling has been uploaded to a public GitHub repository:
https://github.com/Sahithi0406/ISI-IDEAS-PROJ.git

4.7 Flow of Analysis (Workflow)

The overall workflow of the project is summarized below:

Data Loading
Data Inspection
Data Cleaning and Type Conversion
Feature Engineering
Descriptive Statistical Analysis
Aggregation and Trend Analysis
Visualization
Regression Model Development
Model Evaluation
Forecasting Next Month Revenue

5. Data Analysis and Results

5.1 Descriptive Statistical Analysis

- The dataset consists of **3,547 transaction records with 11 attributes** including product name, time of day, weekday, and revenue.
- After performing initial inspection:

TOTAL DUPICATE ROWS	0
TOTAL MISSING VALUES	0

- Data types were verified and the Date column was converted to datetime format.
- New columns Month and Year were derived for temporal analysis.

Interpretation:
The average transaction revenue is approximately around the mean shown in the dataset. The standard deviation indicates moderate variation in transaction values. The absence of missing values confirms data consistency and suitability for analysis.

5.2 Revenue Analysis by Coffee Type

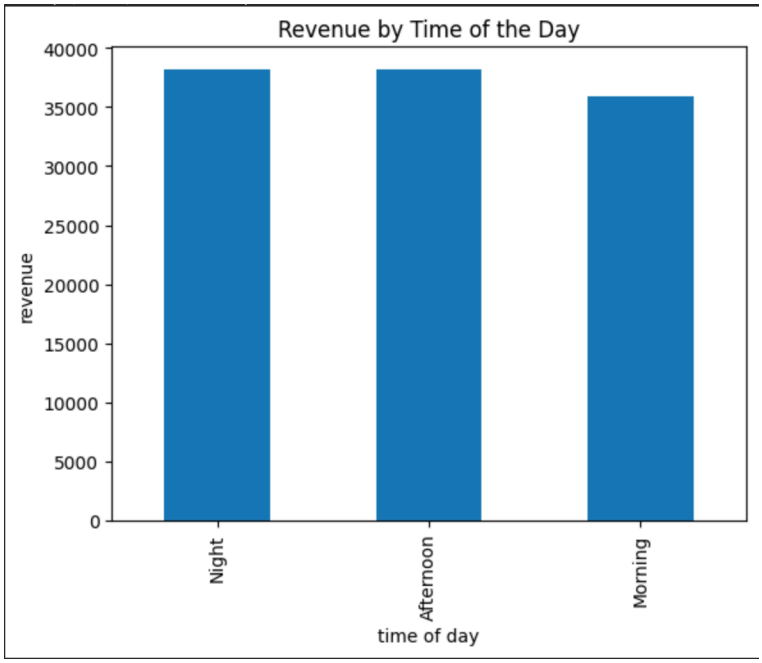
coffee_name	
Americano	14650.26
Americano with Milk	24751.12
Cappuccino	17439.14
Cocoa	8521.16
Cortado	7384.86
Espresso	2690.28
Hot Chocolate	9933.46
Latte	26875.30

The total revenue generated by each coffee product was computed using **groupby** aggregation.

Interpretation:
Among all products, [LATTE] generated the highest revenue.
This suggests strong customer preference for this product. Lower revenue products may require promotional strategies or pricing review.

5.3 Revenue Analysis by Time of Day

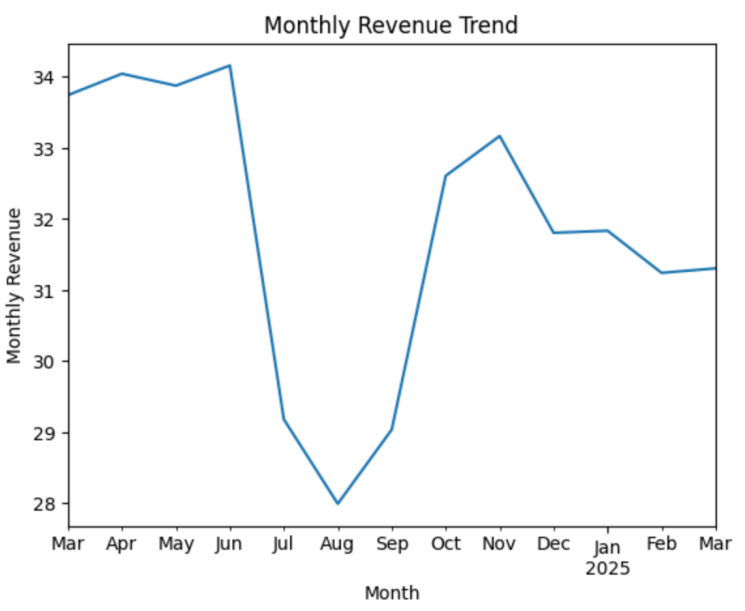
Average revenue was analyzed across Morning, Afternoon, and Night.



Interpretation:
The [NIGHT] period shows the highest average revenue.
This indicates peak purchasing hours and can help optimize staffing and inventory planning.

5.4 Monthly Revenue Trend Analysis

Monthly revenue was aggregated and visualized to identify trends and potential seasonality.



Interpretation:
Revenue shows noticeable fluctuations across months. A significant dip is observed during mid-year, followed by gradual recovery. This may indicate seasonal demand variation or external operational factors.

5.5 Predictive Modeling (Linear Regression)

A basic Linear Regression model was implemented to forecast next month revenue.

Data Split:

Training set: 9 months
Validation set: 2 months
Test set: 2 months

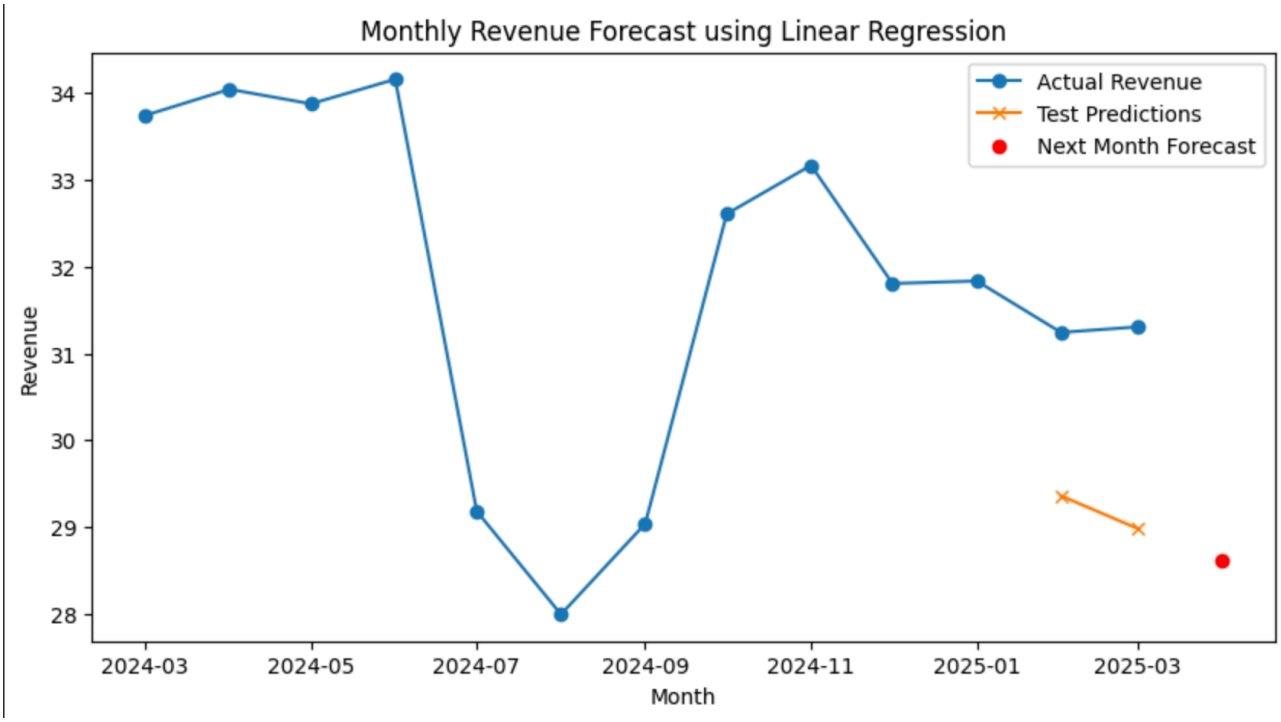
Evaluation Metrics:

```
print("validation mae : ",mae_val)
print("validation RMSE : ",rmse_val)
print("validaation R2 : ",r2_val)
```

```
validation mae :  1.9003981048289909
validation RMSE :  1.9110803004756425
validaation R2 :  -17158.07838554977
```

R^2 score negative (due to small test size)

Interpretation:
The MAE of approximately 2 indicates that the model's prediction deviates from actual revenue by around 2 units on average. Considering the revenue range (approximately 28–34), this corresponds to a moderate prediction error.
The R^2 score is unstable due to the extremely small test dataset and should not be considered a reliable performance indicator.
The predicted revenue for the next month is approximately 28.61 units, suggesting a slight downward trend.



6. Conclusion

This project performed an exploratory analysis of transactional coffee sales data to identify revenue patterns across products, weekdays, and time periods. The analysis revealed clear differences in revenue contribution among coffee types, with certain products generating significantly higher total sales. Weekly analysis showed stronger performance during weekdays, while time-of-day analysis indicated variation in average revenue across different periods.

Monthly revenue trends displayed noticeable fluctuations, including a mid-year decline followed by recovery, suggesting possible seasonal influences. These findings highlight the importance of structured exploratory analysis in understanding business performance.

A baseline Linear Regression model was implemented to forecast next month's revenue. The model produced a projected value of approximately 28.61 units. While the MAE and RMSE indicated moderate prediction error, the limited number of monthly observations restricted model reliability. Therefore, the forecast should be interpreted as a trend-based estimate rather than a precise prediction.

Overall, the project demonstrates how descriptive analysis combined with basic predictive modeling can provide meaningful business insights from structured transactional data.

7.APPENDIX

GITHUB LINK:
https://github.com/Sahithi0406/ISI-IDEAS-PROJ.git