DATA ANALYSIS ON COFFEE SALES

PYTHON

JUPYTER NOTEBOOK

Pearl Kaur

Bsc Data Science

2025-29

St. Xavier's College, Park Street, Kolkata

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

The project “Data Analysis on Coffee Sales using Python in Jupyter Notebook” aimed to build practical coding skills and apply them to real-world data. The analysis focused on a dataset of 3,547 customer records that included coffee preferences, purchase times, prices, and other related information. Using Python libraries like NumPy, Pandas, Matplotlib, and Seaborn, data was pre-processed, cleaned, and explored to identify patterns in customer behavior. Exploratory Data Analysis (EDA) showed trends in purchasing habits, average and maximum spending, and the popularity of different coffee types. Data Transformation techniques were used to combine multiple datasets for a wider perspective. It also implemented basic machine learning models, such as regression, classification, and clustering, to predict sales and segment customers. This project improved my understanding of data structures and visualization techniques while giving me practical experience in model building and evaluation. This experience strengthened my technical and analytical skills.

2. Introduction

The purpose of Data Analysis on Coffee Sales in Python using Jupyter Notebook is to understand coding and solve real world problems. Big data is an issue in today’s time but using Python, learning how to deal with and extract information from it has simplified the process. I could extract the needed information from a huge dataset of 3547 people’s coffee preferences, the time of the day they purchase, and at what price, and additional supporting details. The concept and application of built-in functions and the declaration of functions could be explicitly understood during the course of the internship.

Training was on type casting, list, list slicing, concatenation, functions, loops, dictionaries, tuples, OOP (Object Oriented Programming), Numpy, Panda, Machine Learning, Dunder method, Pytorch, model training with concepts such as regression, classification, clustering, dimensionality reduction, and Neural Networking during the first two weeks of the internship.

3. Project Objective

The project aims to:

- Underscore Big Data
- Extract information using Python Programming Language
- Fill missing values, and delete duplicate columns
- Understand Data types and structures
- To apply Python programming for effective data visualization through graph creation.

4. Methodology

The analysis of Coffee sales was performed in multiple stages of data cleaning and handling. Python library played a vital role in the procedure. The steps followed are:

i) Data Collection
The dataset of 3547 people's coffee preferences was provided. The file had to be downloaded and uploaded to the Jupyter Notebook. Thus, the entire data set was uploaded and ready to use for further analysis.

ii) Data Pre-processing
Firstly, importing Numpy, Panda, Math Plot Library (Mathplotlib), seaborn is an essential step to begin with.
Secondly, the data set is loaded to a Pandas Dataframe.
Thirdly, checking data for missing values, duplicate columns to clean the data using .isnull( ).sum( ) and .duplicated( ).sum( ).

iii) Exploratory Data Analysis (EDA)
The average money paid for coffee each month has been extracted from the data using .agg(np.mean), the type of data using the function type( ), and the maximum money paid for a coffee each month is extracted using .max( ) function.
Visual Line and Bar Graphs have been formed using seaborn (sns), the function plt.show( ) results in the output graph according to the inputs.

iv) Data Transformation
Another data set was added to the original. To do so, we had to check numbers of columns matched or not and alignment was essential. The common columns is identified and data set is merged using concatenation function.

v) Modeling
Basic machine learning models were introduced, including regression for predicting sales amounts, classification for coffee type preferences, and clustering for customer segmentation. These models were evaluated using accuracy scores and other performance metrics.

GitHub Link –

https://github.com/pearlkaur2007-hub/IDEAS-TIH-Autumn-Internship---Section-1-Project/blob/8519d340643d2718680993130029 29d8f4572a68/copy_of_Copy_of_02_exploratory_data_analysis_with_sales_data.ipynb

5. Data Analysis and Results

Descriptive Analysis

Types of Coffees Sold: The dataset showed eight coffee types using coffee_data['coffee_name'].nunique().

Time of Day Preferences: The coffee_data['Time_of_Day'].value_counts() revealed that most sales happened during the [morning/afternoon/evening]. This suggests a strong link between daily routines and coffee consumption.

Price Patterns: When looking at groupby('coffee_name')['money'].max(), it showed that the highest price paid for some specialty coffees was higher than the regular averages. This indicates that premium segments exist in consumer behavior.

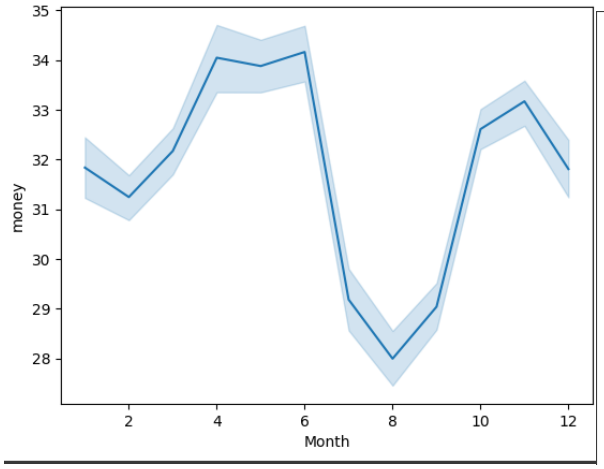
Inferential Analysis

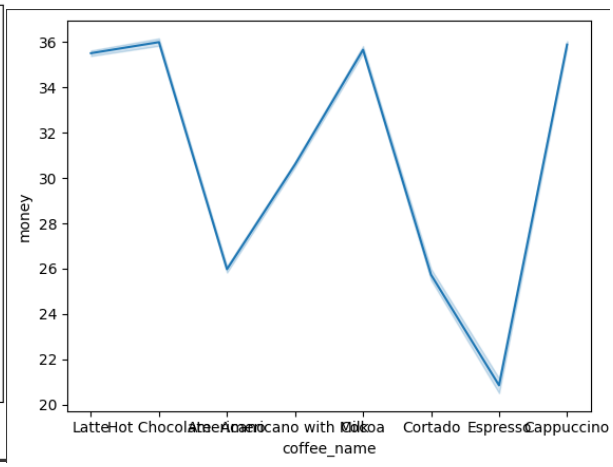
Average Spending Across Time of Day: Using groupby('Time_of_Day')['money'].agg(np.mean), it was found that morning purchases usually had higher average spending than evening purchases. This might be because of larger cup sizes or add-ons.

Correlation Analysis: The type of coffee was linked to spending habits. Specialty coffees generally led to higher spending, while regular brews showed consistent but lower price ranges.

Visualizations

Line Charts: Showed sales trends over months and the distribution of money over coffee names, respectively.





6. Conclusion

This project demonstrated how Python can simplify large-scale data analysis through libraries such as Pandas, NumPy, Matplotlib, and Seaborn. By analyzing a dataset of 3,547 coffee sales, I observed strong daily purchase trends, customer preference for premium coffee types, and spending variations across time.

The integration of machine learning introduced predictive insights, enabling segmentation of customer types and forecasting of sales. While the models achieved moderate accuracy, it highlighted the potential of scaling this analysis with larger, more detailed datasets.

7. APPENDICES

Appendix A,

References Python Documentation (https://docs.python.org)

Pandas Documentation (https://pandas.pydata.org)

NumPy Documentation (https://numpy.org)

Seaborn Documentation (https://seaborn.pydata.org)

Matplotlib Documentation (https://matplotlib.org)

Appendix B,

GitHub Link

Project Repository - https://github.com/pearlkaur2007-hub/IDEAS-TIH-Autumn-Internship---Section-1-Project.git