{AUTUMN INTERNSHIP PROJECT REPORT FORMAT}

**Project Title –**

**Exploratory data-analysis with sales data on coffee\_sales dataset (Project 2)**

ARITRA BHATTACHARYA

AUTUMN INTERNSHIP, IDEAS-TIH

PRAFULLA CHANDRA SEN GOVERNMENT MEDICAL COLLEGE AND HOSPITAL (WBUHS)

Period of Internship: 25th August 2025 - 19th September 2025 (Do not change the dates)

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

1. **Abstract**

Exploratory Data Analysis (EDA) is a crucial component of Machine Learning. It includes Data Summarization, Data Visualization, Pattern Recognition, Anomaly Detection and Hypothesis Generation. In this era of big-data and huge influx of hospital record keeping in form of EHRs, a knowledge of ML becomes a dire need. Its application lies in the diverse field of healthcare and medicine where AI is becoming an inseparable attribute. This project is composed of a task of exploratory data analysis on a dataset known as coffee\_sales dataset using Python programming language. The analyzed data are represented in the form of descriptive statistics, bar plot and line plot and tables. This project proved with the knowledge regarding the Python programming and exploratory data analysis, thus forming the core of the application of ML and AI in healthcare and medicine.

1. **Introduction**

Exploratory Data Analysis (EDA) is an approach to analyzing data sets to summarize their main characteristics, often using visual methods. It is about discovering patterns, identifying anomalies, and formulating hypotheses. **[1]**

The [US National Institute of Standards and Technology define EDA](http://www.itl.nist.gov/div898/handbook/eda/section1/eda11.htm) as:

“*An approach/philosophy for data analysis that employs a variety of techniques (mostly graphical) to maximize insight into a data set, uncover underlying structure, extract important variables, detect outliers and anomalies, test underlying assumptions, develop parsimonious models and determine optimal factor settings.*” **[2]**

The key practices of EDA involve - **Data Summarization, Data Visualization, Pattern Recognition, Anomaly Detection** and **Hypothesis Generation. [3]**

Python is an interpreted, object-oriented programming language with dynamic semantics. Its high-level, built-in data structures, combined with dynamic typing and dynamic binding, make it very attractive for rapid application development, as well as for use as a scripting or glue language to connect existing components together. **[1]**

This project is directed towards the implementation of EDA on a sales dataset using the Python programming language. The importance of the project lies in the fact that EDA forms the basic and most vital step of a Machine Learning model development. Healthcare and medicine, like the most other fields, are rapidly evolving into taking AI and ML as inseparable attributes.

Also, this project is demands the application of the knowledge gained during the first two weeks of the internship on the topics of – **Basics of Python** (Data, variable, Lists, Loop, Data structures, Class, Functions, Object oriented programming, NumPy and Pandas) and **Machine Learning** (Regression, Classification and LLM fundamentals).

1. **Project Objective**

The objective of this project is to implement the knowledge gained regarding the introduction to Python programming language and Machine Learning to perform an exploratory data analysis on a given dataset. Hence the objectives of the study are –

1. Application of the knowledge regarding Python programming
2. Exploratory data analysis of the given dataset
3. Representation of the analyzed data in the form of descriptive statistics, bar plot and line plot and tables.
4. **Methodology**

This project was composed of a task of exploratory data analysis on a dataset known as coffee\_sales dataset. The dataset (<Coffe_sales.csv>) is a sales data which comprised of the following columns of entry – Hour of the day, cash type, money, coffee name, Time of the Day, Weekday, Month name, Weekday-sort, Month-sort, Date and Time. The dataset contains 3547 entries.

Exploratory data analysis was performed on the dataset using Python programming language.

The GitHub link containing the codes used in this project is given below.

<https://github.com/aritra01bach/IDEAS-TIH>

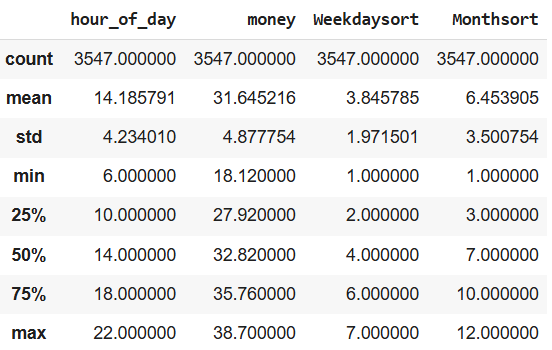
1. **Data Analysis and Results**

Both the data analysis and data presentation were performed using the Python programming language. The results obtained from analyzing the given dataset are shown below –

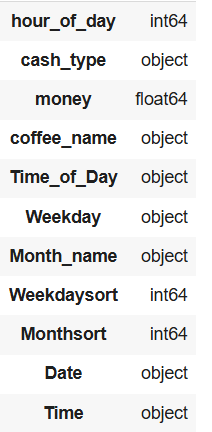
* **Descriptive analysis –**

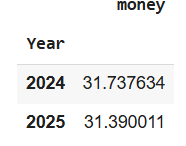
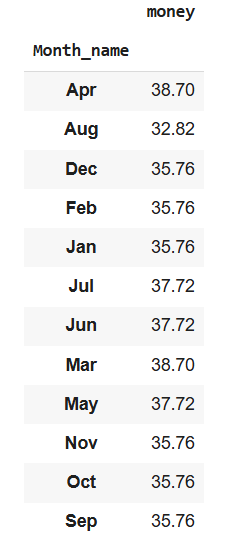
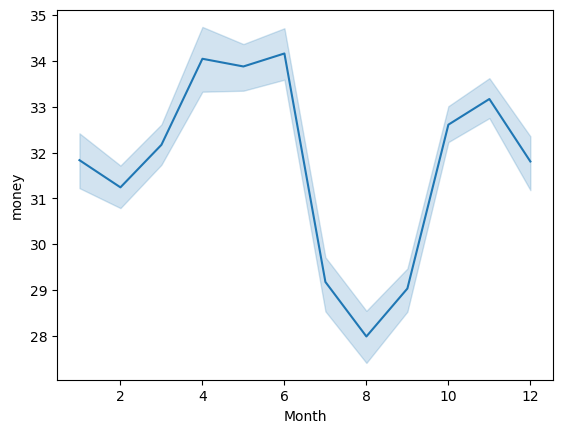
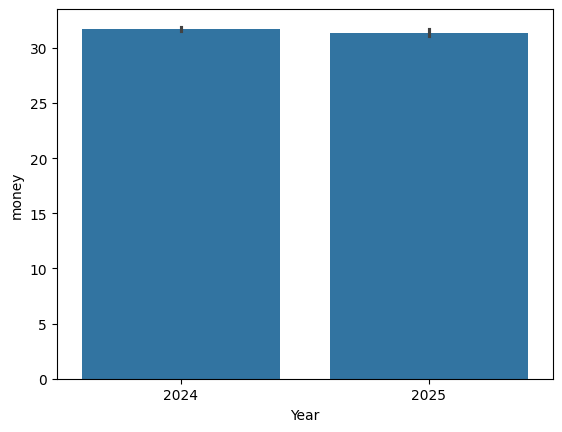
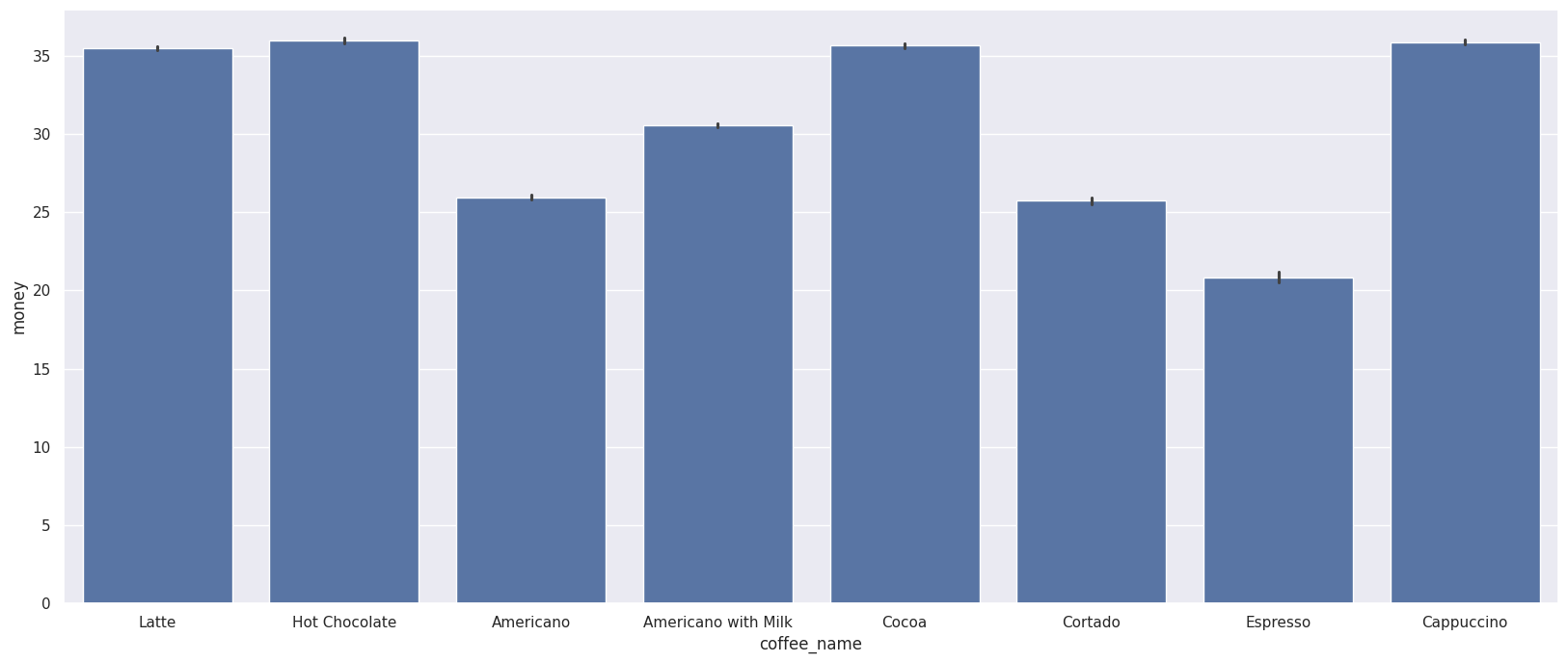
|  |  |
| --- | --- |
| 1. Number of columns in the dataset | 13 |
| 2. Number of duplicate columns | **0** |
| 3. Number of missing values in each column | **0** |
| 4. Total number of missing values | **0** |

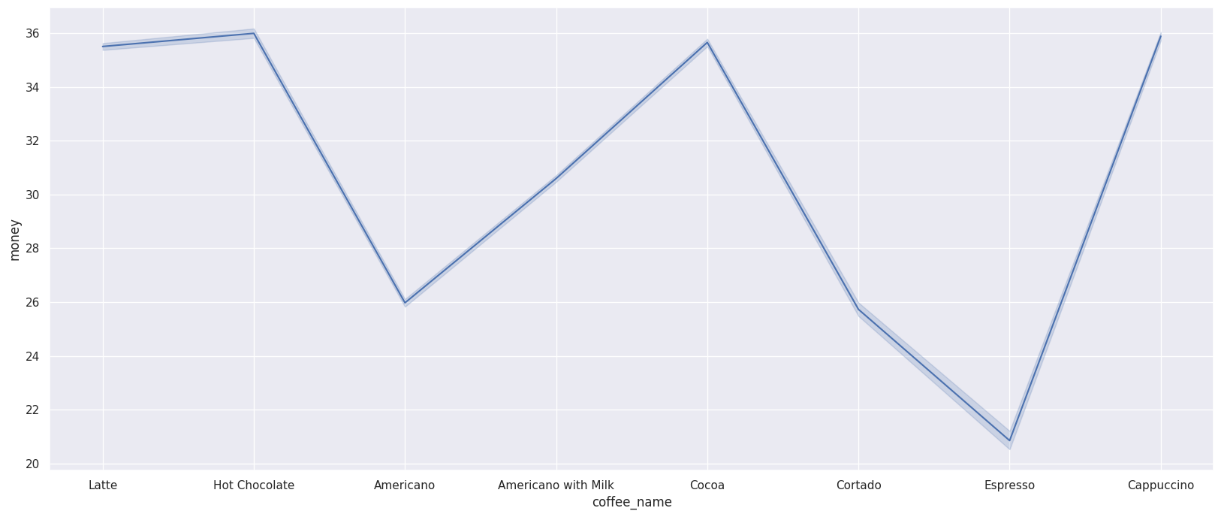
* **Basic statistics of the data –**

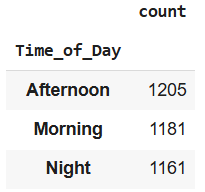
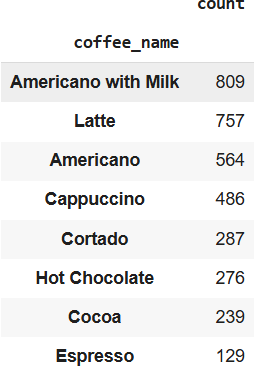


* **Data types of the dataset –**



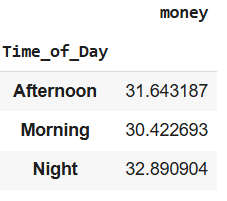
* **Average money for each year –**
* **Maximum money for each month –**
* **Line plot showing distribution of money over months –**
* **Bar plot showing density of money over years –**
* **Bar plot and line plot showing distribution of money over coffee names –**



* **Times of the day in the dataset –**
* **Types of coffee in the dataset –**
* **Maximum money over coffee names –**



* **Average money over the times of the day –**



1. **Conclusion**

This project provided the pathway to the introduction of exploratory data analysis which is an integral and crucial step while dealing with a dataset. Along with a detailed introductory knowledge on Python programming, the project also introduced the colorful and rapidly evolving world of Machine Learning. In this era of big-data and huge influx of hospital record keeping in form of EHRs, a knowledge of ML becomes a dire need. The knowledge gathered during this project finds its application in the diverse field of healthcare and medicine where AI is becoming an inseparable attribute. Python programming and exploratory data analysis is thus forming the core of the application of ML and AI in healthcare and medicine.

1. **APPENDICES**

* **References –**

1. <https://www.ibm.com/think/topics/exploratory-data-analysis>
2. <https://graphworkflow.com/>
3. Camizuli, E. and Carranza, E.J. (2025). Exploratory Data Analysis (EDA). In The Encyclopedia of Archaeological Sciences, S.L. López Varela (Ed.). <https://doi.org/10.1002/9781119188230.saseas0271>

* **GitHub link –**

<https://github.com/aritra01bach/IDEAS-TIH>

* **Dataset link –**

[Coffe\_sales.csv](file:///C:\Users\ADMIN\Downloads\Coffe_sales.csv)