## 

## **Exploratory Data ANALYSIS INCLUDING VISUALIZATION**

**Sreelakshmi E U**,

SREE SANKARA COLLEGE KALADY

MSc STATISTICS WITH DATA SCIENCE

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 focuses on **Exploratory Data Analysis (EDA)** of a sales dataset to uncover business insights. The data was cleaned by handling missing values and duplicates, followed by descriptive statistics to understand its structure. Using visualizations, the analysis explored **monthly profit trends**, revealing seasonal fluctuations with notable peaks during late 2022–early 2023. Further, category-wise sales were examined to identify the **best-performing product categories**. The study provides actionable insights into sales performance, seasonal demand patterns, and product profitability, supporting data-driven decision-making.

1. **Introduction**

This project is centered on **Exploratory Data Analysis (EDA)** of a retail sales dataset with the objective of uncovering patterns, trends, and actionable business insights. In today’s data-driven world, businesses rely heavily on data analysis to understand customer behavior, optimize sales strategies, and improve profitability. By analyzing sales performance over time and across product categories, this project contributes to effective decision-making in retail management.

**Relevance**

Understanding sales trends, profit fluctuations, and product-wise performance is highly relevant to modern businesses that want to identify their best-selling products, forecast seasonal demand, and design targeted marketing strategies.

1. **Project Objective**

Elaborate the objective of doing the project, can include what is trying to illustrate using this project (in bullets, max 7 points). There should not be more than 5 objectives. If you are doing any hypothesis testing that also needs to be mentioned here. If any sample survey was done, please mention the target population for which the survey is intended.

### ****Technology Involved****

* **Python** (for data analysis and visualization)
* **Pandas & NumPy** (for data manipulation)
* **Matplotlib, Seaborn, and Plotly** (for visualization)
* **Jupyter Notebook** (for implementation and documentation)

### ****Background Material Survey****

Before starting the project, a review of standard practices in data analysis and visualization was carried out. This included:

* Descriptive statistics to summarize datasets
* Handling missing values and duplicates
* Time-series trend analysis
* Category-wise grouping and visualization

### ****Procedure Used****

1. **Data Cleaning** – Removing duplicates, handling missing values, and formatting date fields.
2. **Exploratory Data Analysis** – Performing descriptive statistics and generating insights.
3. **Visualization** – Using line charts, bar charts, and categorical plots to show trends.
4. **Insights & Interpretation** – Identifying seasonal profit fluctuations and best-selling categories.

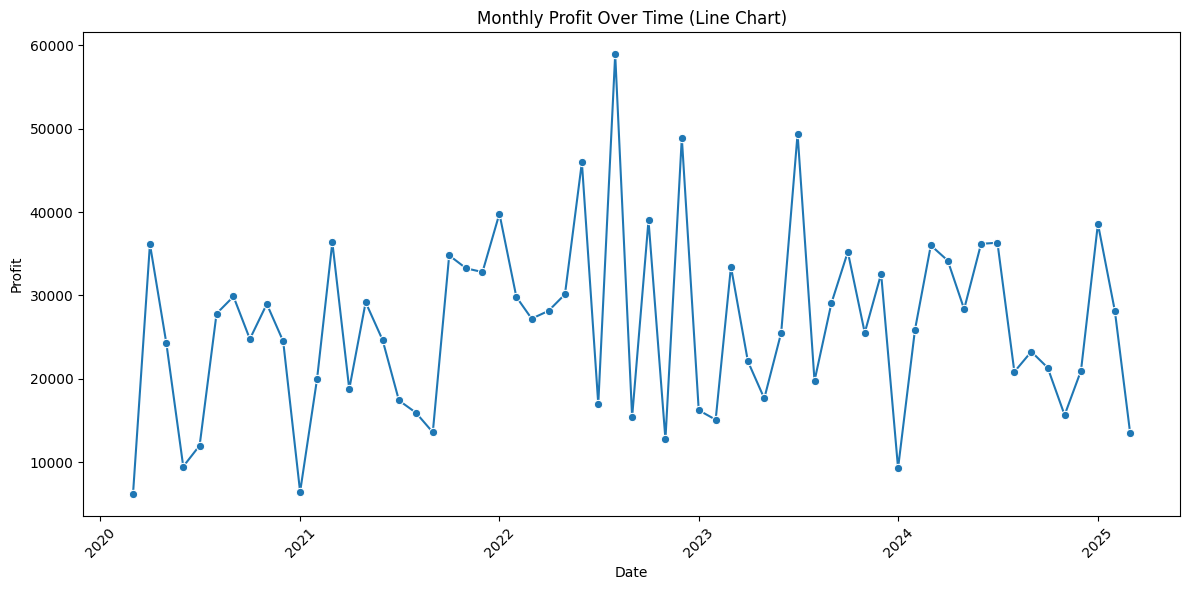
### ****Purpose of the Project****

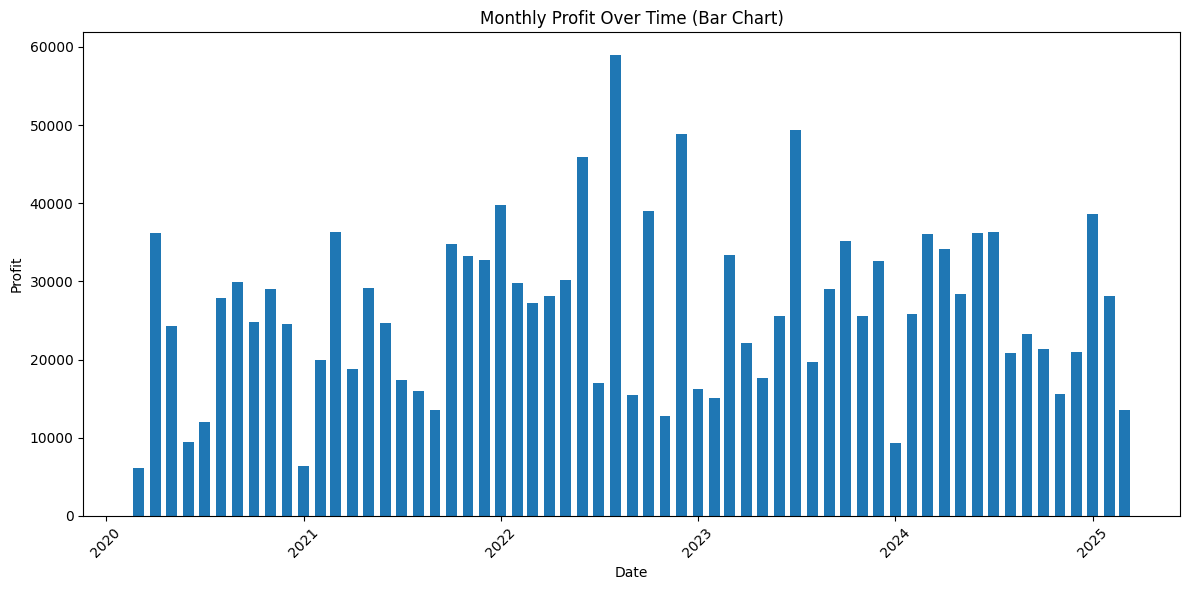
The purpose is to **gain hands-on experience in real-world data analysis**, apply theoretical knowledge of statistics and programming, and generate insights that could support business decisions in sales and marketing. It also helps in developing skills in data visualization, trend forecasting, and category analysis.

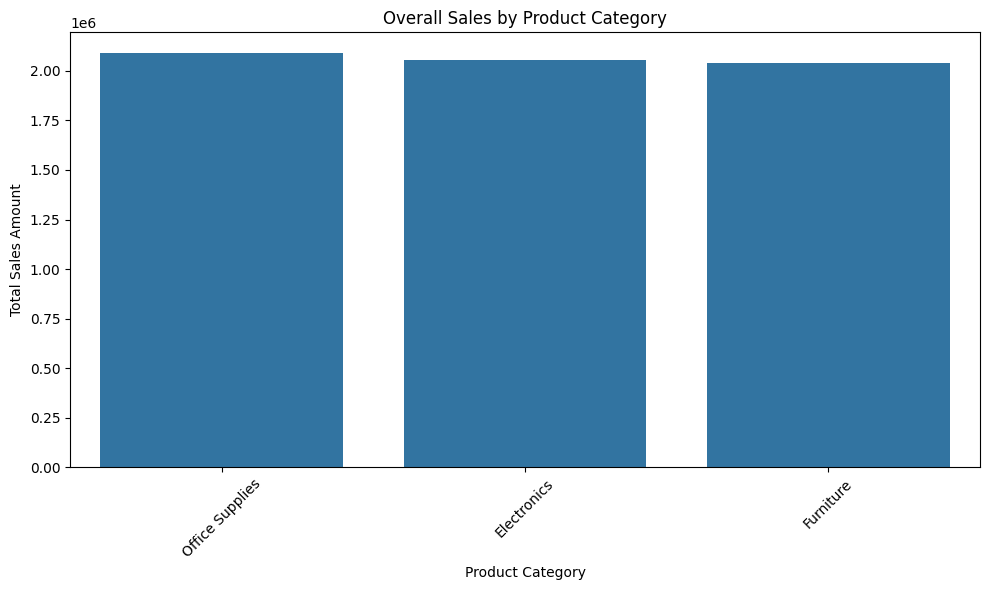
1. **Methodology**

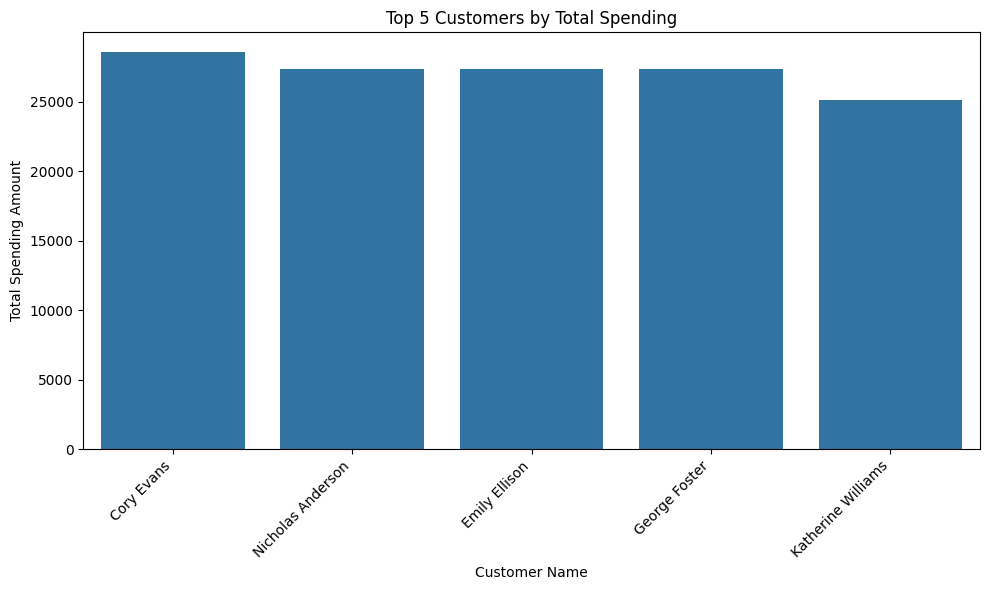
The project involved collecting and preprocessing retail sales data, cleaning missing and duplicate values, performing exploratory data analysis with Python libraries (Pandas, NumPy, Matplotlib, Seaborn, Plotly), visualizing sales and profit trends, analyzing category-wise performance, and optionally applying analytical models with train-test validation to generate actionable business insights.

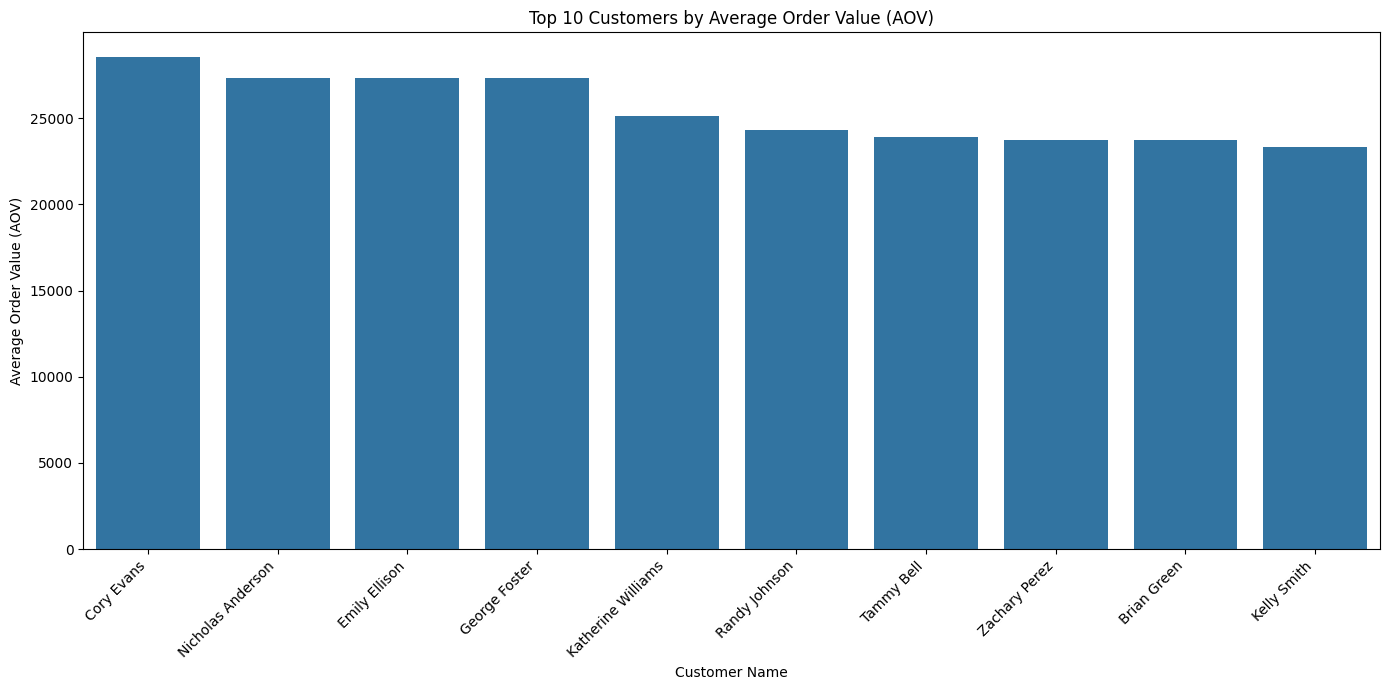
1. **Data Analysis and Results**

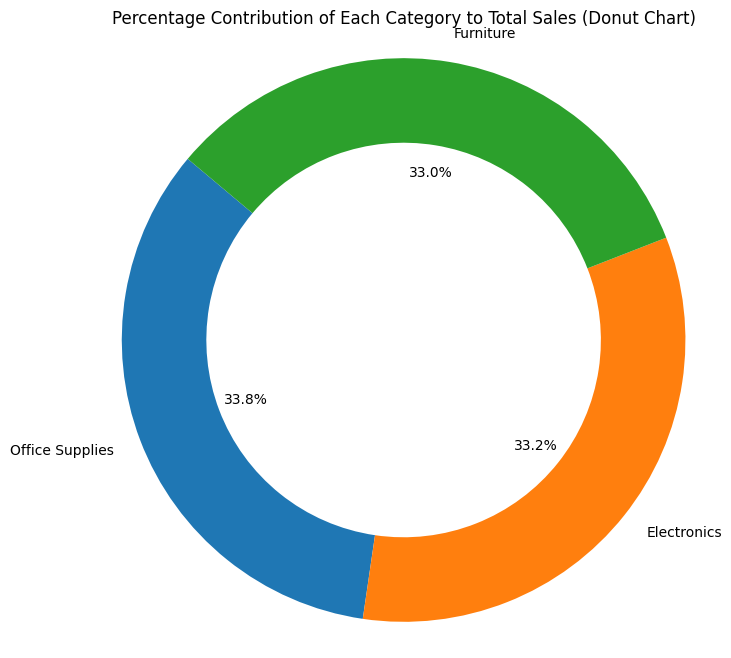
****

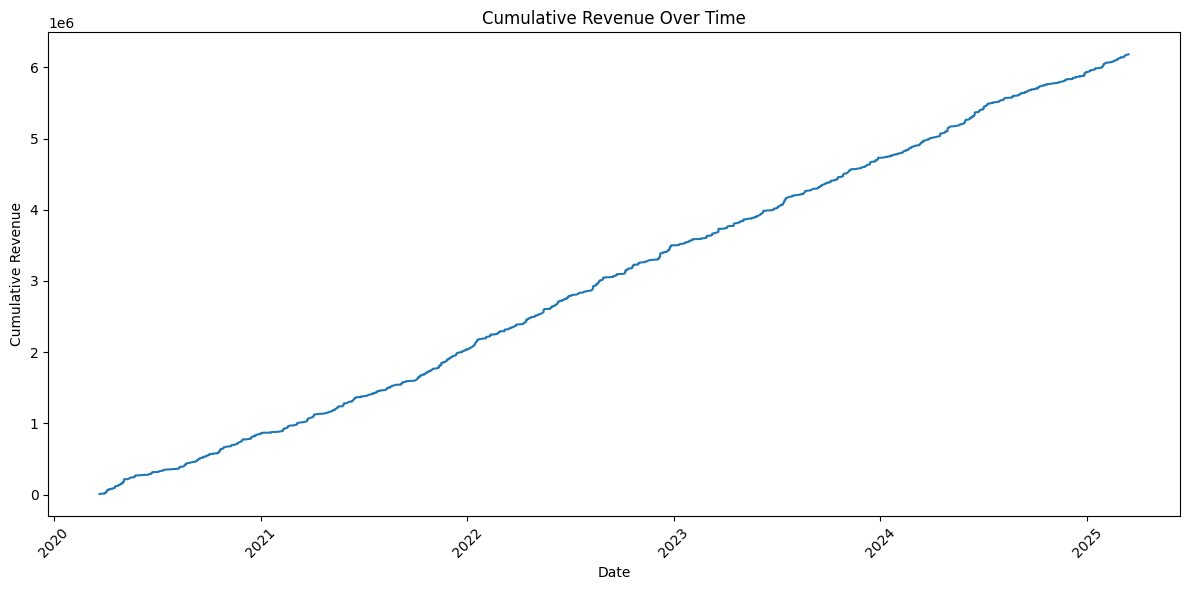
****

****

****

****

****

****

1. **Conclusion**

The analysis showed that profits follow seasonal patterns with peaks during late 2022–early 2023, and **Electronics emerged as the best-performing category**. Despite fluctuations, overall sales and profit trends indicated business growth.

1. **APPENDICES**

### ****References****

1. Pratap Dangeti, Statistics for Machine Learning, Packt Publishing.
2. Tom M. Mitchell, Machine Learning, McGraw Hill.
3. Towards Data Science (https://towardsdatascience.com/)
4. Kaggle – Retail Sales Datasets (https://www.kaggle.com/)
5. Documentation of Python libraries: Pandas, NumPy, Matplotlib, Seaborn, Plotly.

### ****Survey Questionnaire****

### How often do you purchase products from our store?

* + Weekly / Monthly / Occasionally

1. Which product categories do you prefer?
   * Electronics / Furniture / Clothing / Others
2. What is the usual budget range of your purchases?
   * < ₹1,000 / ₹1,000–₹5,000 / > ₹5,000
3. How satisfied are you with product quality?
   * Very Satisfied / Neutral / Dissatisfied
4. Would you recommend our store to others?
   * Yes / No

**Document Links**

**Colab Notebook link** :

<https://colab.research.google.com/drive/1myk-dzczTFjYQQmPzjmuE3GDUDzO31ls?usp=sharing>

<https://colab.research.google.com/drive/1wc-KZQW6YdcFAjpeIHJ80cbI246-ymd3?usp=sharing>