Internship Report

🎓 **Submitted by: Shaiba Ali**

📅 **Duration: [14/07/2025] – [14/08/2025]**

🏢 **Organization: [Code alpha]**

📊 **Domain: Data Analysis**

Introduction

This report summarizes the work completed during my internship focused on Data Analytics and Sentiment Analysis. The main goal was to work on real-world datasets, understand data patterns, visualize meaningful insights, and perform sentiment classification on textual data.

Internship Objectives

- To perform structured Exploratory Data Analysis (EDA) on real-world datasets.
- To create meaningful and interactive data visualizations.
- To perform Sentiment Analysis on text data using NLP techniques.
- To gain practical experience in Python, Pandas, Seaborn, and TextBlob.

Tools and Technologies Used

Python, Pandas, NumPy, Matplotlib, Seaborn, TextBlob, Jupyter Notebook, Tableau (optional).

Task 2: Exploratory Data Analysis (EDA)

- Dataset Used: Superstore Sales (or similar).
- Cleaned the dataset, removed nulls/duplicates.
- Statistical summaries and visual exploration using Pandas and Seaborn.
- Key Insights: Technology category had the highest sales; Discounts negatively affected profit.

TASK 2: Exploratory Data Analysis (EDA)

◆ Objective:

To understand the structure, patterns, and relationships in the dataset using descriptive statistics and visual exploration.

◆ Dataset Used:

- **[Superstore Sales Dataset]** from Kaggle
- Contains records of orders with fields like Sales, Profit, Category, Region, Order Date, etc.

◆ Steps Performed:

1. **Loading and Inspecting the Data**
   - Used `pandas` to load and inspect rows and column types.
   - Verified the presence of null values and duplicates.
2. **Data Cleaning**
   - Removed duplicates using `df.drop_duplicates()`.
   - Checked for missing values and handled them using imputation or removal.
3. **Understanding Data Types**
   - Converted `Order Date` to `datetime` format.
   - Checked numerical and categorical columns.
4. **Statistical Summary**
   - Used `.describe()` to get mean, min, max, std, etc.
   - Identified outliers and skewed distributions.
5. **Feature Correlation**
   - Generated heatmaps to see correlation among features like Sales, Quantity, and Profit.
6. **Key Insights**
   - Discount and Profit are **negatively correlated**.
   - The **Technology** category contributes the highest sales.
   - **Central and West** regions showed top performance.

Task 3: Data Visualization

- Created bar charts, line plots, pie charts, and heatmaps.
- Used Seaborn, Matplotlib for visualizations.
- Insights: West and East regions contributed most to sales. Discount and Profit were inversely related.

Objective:

To convert insights from the dataset into easy-to-understand graphical representations to support decision-making.

◆ Tools Used:

- Python libraries: **Matplotlib, Seaborn**
- Optional: **Tableau** or **Power BI** (for dashboards)

◆ Visualizations Created:

1. **Bar Charts**
   - Sales by Category
   - Profit by Region
2. **Line Charts**
   - Monthly Sales Trends using `groupby(Order Date)`
3. **Pie Chart**
   - Sales Distribution by Shipping Mode
4. **Heatmap**
   - Correlation between features (e.g., Discount vs Profit)
5. **Countplots**
   - Customer Segment frequency
   - Product Category counts

◆ Insights Gained:

- Profits decline when discounts exceed 30%.
- Office Supplies have high sales volume but low profit margin.
- West region outperforms East in average sales per order.

Task 4: Sentiment Analysis

- Dataset: Amazon Product Reviews.
- Cleaned text and analyzed using TextBlob for polarity.
- Results: 65% Positive, 20% Neutral, 15% Negative.
- Visualized using countplot and WordCloud.

Objective:

To classify customer review text into Positive, Negative, or Neutral categories using Natural Language Processing (NLP).

◆ Dataset Used:

- **Twitter Reviews Dataset**
- Text data containing customer reviews and ratings

◆ Preprocessing Steps:

1. **Text Cleaning**
    - Removed punctuation, stop words, and converted text to lowercase.
    - Used regular expressions for basic normalization.
2. **Tokenization & Lemmatization**
    - Split sentences into words and reduced them to their base form.
3. **Sentiment Analysis**
    - Used **TextBlob** to calculate the polarity score.
    - Polarity > 0 = Positive, < 0 = Negative, = 0 = Neutral
4. **Visualization**
    - Countplot for Positive vs Negative vs Neutral
    - WordCloud to display most frequent words in Positive/Negative reviews

◆ Results:

- 65% reviews were **Positive**
- 20% were **Neutral**
- 15% were **Negative**

◆ Interpretation:

- Products had overall positive reception.
- Some negative reviews highlighted delivery delays and packaging issues.
- Useful for product improvement and marketing analysis.

Learnings

- Real-world dataset handling, data preprocessing.
- Visual analytics and NLP techniques.
- Hands-on experience with Python and its libraries.

Challenges

- Handling null/inconsistent data.
- Managing large datasets efficiently.
- Understanding mixed-sentiment texts.

Conclusion

This internship enhanced my practical knowledge in Data Analytics and NLP. I am now more confident in using tools like Pandas, Seaborn, and TextBlob for real-world problems.

References

- https://pandas.pydata.org
- https://seaborn.pydata.org
- https://textblob.readthedocs.io
- https://www.kaggle.com/datasets