Internship Final Report

Student Name: Ashish Mishra
University: University of Mumbai
Major: Electronics Computer Science
Internship Duration: April 1th, 2025 - May 5th, 2025
Company: ShadowFox
Domain: Data Science
Coordinator: Mr. Aakash

Objectives

1. To develop proficiency in data visualization using Python libraries such as Matplotlib and Seaborn for effective communication of insights.
2. To perform in-depth data analysis on real-world datasets, including Air quality and Sports data, to extract meaningful patterns and trends.
3. To enhance skills in data cleaning, preprocessing, and statistical analysis to address complex research questions.
4. To improve technical documentation skills by creating clear, user-friendly guides for technical and non-technical audiences.

Tasks and Responsibilities

1. **Visualization Library Documentation (Beginner Level)**

    - Developed a detailed guide on Matplotlib and Seaborn, two prominent Python visualization libraries.
    - Documented various plot types, including line plots, scatter plots, bar charts, histograms, pie charts, box plots, and heatmaps, with example code snippets and use cases.
    - Compared Matplotlib (highly customizable, ideal for complex plots) and Seaborn (user-friendly, aesthetically pleasing) based on ease of use, customization, interactivity, and performance with large datasets.
    - Compiled the guide into a well-structured PDF, ensuring accessibility for beginners.

2. **Air Quality Index (AQI) Analysis in Delhi (Intermediate Level)**

    - Analyzed a 2023 dataset of air quality metrics (CO, NO, NO2, O3, SO2, PM2.5, PM10, NH3) from Delhi.
    - Formulated research questions to explore pollutant distributions, correlations, and seasonal trends.
    - Conducted data cleaning and preprocessing using Pandas, followed by exploratory data analysis (EDA).

- Visualized findings using Matplotlib and Seaborn, creating histograms, box plots, scatter plots, heatmaps, and line plots to highlight pollutant trends and AQI fluctuations.
- Identified key insights, such as CO's high average concentration (3814.94 µg/m³) and strong correlations between PM2.5, PM10, and AQI.

3. **Cricket Fielding Analysis (Advanced Level)**
    - Analyzed fielding performance data for three players (Kuldeep Yadav, Lalit Yadav, Rilee Russouw) from a T20 match (IPL2367, Delhi Capitals).
    - Recorded fielding actions (clean picks, good throws, catches, dropped catches, runs saved) per ball, categorized by player, position, and outcome.
    - Calculated performance scores using a weighted formula: clean picks (1.5), good throws (1.2), catches (2.0), runs saved (1.0), and dropped catches (-1.0).
    - Visualized performance scores using a bar plot and organized data into a structured spreadsheet for strategic analysis.
    - Identified top performers and provided insights for optimizing fielding strategies.

Learning Outcomes

1. Mastery of Data Visualization: I gained expertise in creating and customizing visualizations with Matplotlib and Seaborn, learning to tailor plots to specific audiences and datasets.
2. Advanced Data Analysis Skills: The AQI analysis honed my ability to clean, preprocess, and analyze complex datasets, using statistical techniques like correlation analysis to uncover insights.
3. Sports Analytics Knowledge: The cricket fielding analysis introduced me to performance metrics in sports, teaching me to design and apply custom scoring systems.
4. Effective Technical Communication: Writing the visualization guide improved my ability to explain complex concepts in a clear, concise manner for diverse audiences.
5. Proficiency in Version Control: Managing a GitHub repository enhanced my understanding of collaborative workflows and the importance of organized codebases.
6. Time Management and Resilience: Balancing internship tasks with academic commitments sharpened my ability to prioritize and meet deadlines under pressure.

Challenges and Solutions

1. **Challenge: Navigating Complex Visualization Libraries**
    - **Issue:** Matplotlib's extensive customization options and Seaborn's dependency on Matplotlib were initially difficult to grasp.
    - **Solution:** I studied official documentation and practiced with small datasets, breaking down complex plots into manageable components. Online tutorials and forums like Stack Overflow provided additional clarity.
2. **Challenge: Managing Missing Data in AQI Analysis**
    - **Issue:** The AQI dataset contained potential missing values, which could skew analysis results.
    - **Solution:** I used Pandas functions (dropna(), fillna(method='ffill')) to address missing data. Visualizing missing values with heatmaps (sns.heatmap(df.isna())) ensured data integrity. I also validated data ranges to maintain consistency**.**
3. **Challenge: Structuring Unorganized Cricket Data**
    - **Issue:** The cricket dataset was unstructured, with inconsistent field names and missing values, complicating analysis.
    - **Solution:** I cleaned the dataset by standardizing column names, dropping irrelevant fields, and converting data types (Runs, BallCount) using Pandas. Referencing cricket terminology ensured accurate categorization of fielding actions.
4. **Challenge: Balancing Internship and Academic Commitments**
    - **Issue:** The internship coincided with academic exams, making it challenging to meet the original April 30th deadline.
    - **Solution:** I utilized the extended deadline (May 5th) to create a structured timeline, allocating specific hours to each task. I communicated with the coordinator on April 28th to clarify requirements, ensuring timely completion.

Conclusion

My internship at ShadowFox was an enriching experience that transformed my understanding of data science. Through tasks in visualization, air quality analysis, and cricket fielding analysis, I developed a robust skill set in Python, data analysis, and visualization. Overcoming challenges like complex libraries and unstructured data strengthened my problem-solving skills and confidence.

The internship highlighted the power of data in driving decisions and the importance of clear communication through visualizations. My GitHub repository stands as a testament to my efforts, showcasing my ability to tackle diverse projects. This experience has prepared me for future data science roles by equipping me with technical expertise, professional discipline, and a passion for innovation.

Acknowledgments

I am immensely grateful to ShadowFox for providing me with this opportunity to grow as a data scientist. My sincere thanks to my mentor, Mr. Hariharan, for his invaluable guidance and encouragement. I am also thankful to Mr. Aakash, the coordinator, for his prompt support and clear communication, particularly regarding task requirements and deadlines.