My Data Science Internship at Charlton Athletic FC: A Journey Through Football Analytics

I recently had the incredible opportunity to join Charlton Athletic Football Club as a Data Science Intern for the 2024/25 season. This opportunity came after completing a challenging entry task as the final stage of the interview process, where I was required to analyze a large dataset of football players and make recommendations for the club's first-team squad.

The Entry Task

To secure the internship, I had to successfully complete an analysis project as the final task of the selection process. The project involved a comprehensive dataset containing per-90 performance data of football players from 13 different leagues. This dataset included key metrics such as "Play Duration," which represents the total number of seconds a player has played, and "Match Share," the proportion of matches that correspond to the player's playing time. My goal was to clean, process, and analyze the data to identify and recommend three players suitable for Charlton Athletics' first-team squad.

Tools and Technologies

To accomplish this task, I utilized a range of tools and technologies, including:

- **Python: For data manipulation and analysis.**
- **NumPy and Pandas: To efficiently handle and process the large dataset.**
- **Scikit-learn: For statistical modeling and deeper data analysis.**
- **Matplotlib: To create visual representations of the data.**
- **Tableau: To enhance and make the visualizations more interactive and accessible.**
- **GitHub and Git: For version control and managing the project.**
- **Open Sea: For initial data exploration.**

The Process: A Six-Day Challenge

I completed the entire project in six days, carefully planning each day to tackle different phases of the task:

- **Day 1: Data Exploration, Cleaning, and Transformation**

I started by exploring the dataset to understand its structure, identifying missing or anomalous data points. I used Python's Pandas library to clean the data by handling missing values, removing duplicates, and transforming columns for consistency.

- **Day 2: Data Analysis and Visualization with Matplotlib**

Next, I analyzed the cleaned data, focusing on key performance metrics such as passing accuracy, defensive contributions, and goal-scoring opportunities. I created visualizations using Matplotlib, including heatmaps and scatter plots, to reveal patterns and insights.

- **Day 3: Visualization Enhancement Using Tableau**

To improve the clarity and impact of the visualizations, I transitioned from Matplotlib to Tableau. This allowed me to create more dynamic, interactive dashboards that effectively illustrated player performance metrics, making the data more accessible to the recruitment team.

- **Day 4: Documentation and Presentation Creation**

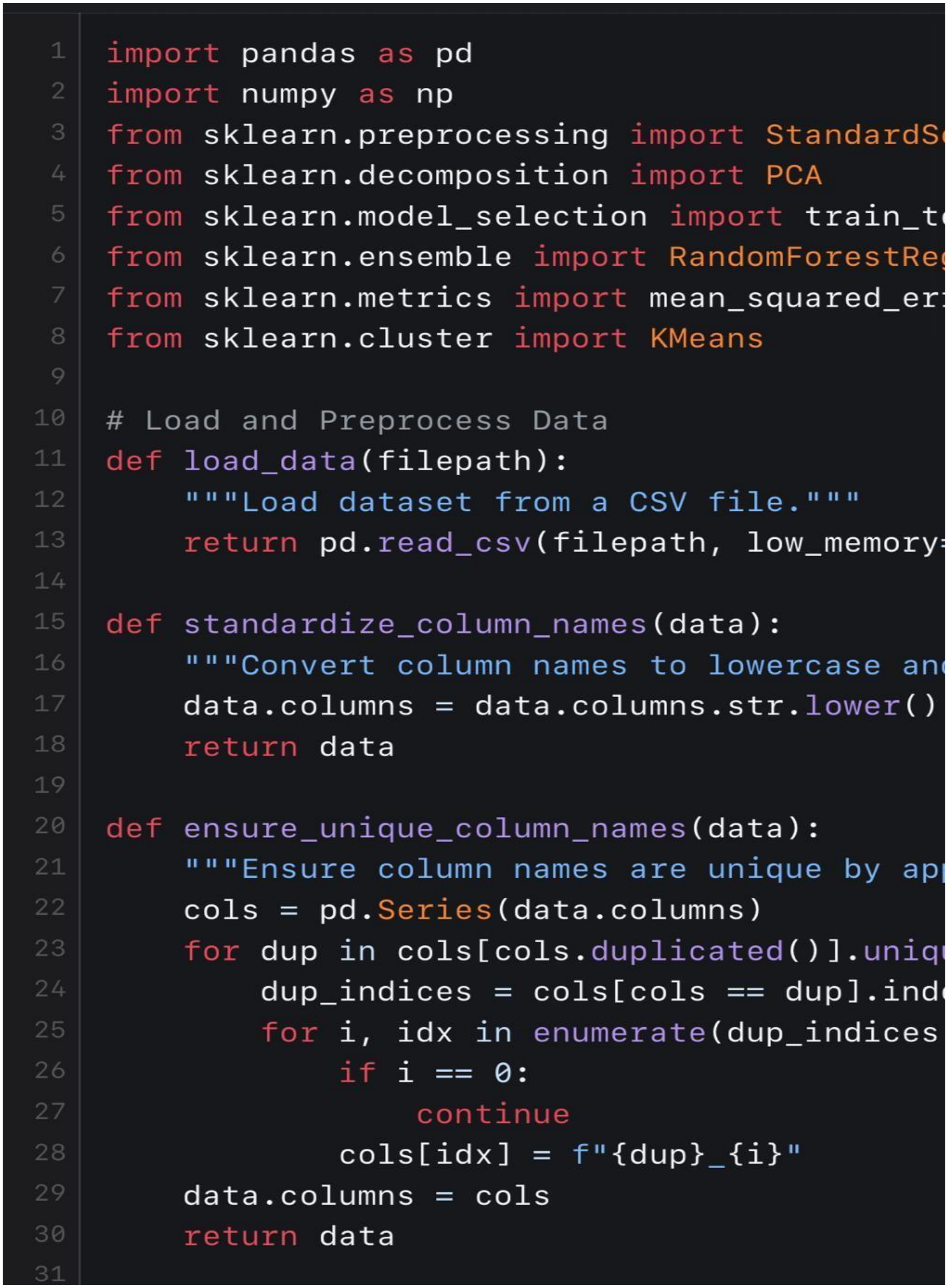
With the analysis completed, I documented my findings and developed a detailed presentation to summarize my recommendations. I made sure the presentation was clear, concise, and supported by robust analytics and visualizations.

- **Day 5: Code Review and Enhancement**

I reviewed my code to ensure it was efficient and accurate, making necessary improvements. I also ensured that my code was well-documented, enabling reproducibility of the results.

- **Day 6: Submission and Presentation**

On the final day, I submitted my analysis and recommendations to the recruitment and analysis team at Charlton Athletic FC. I then presented my findings via Microsoft Teams, followed by a Q&A session where I addressed questions about my methodology, data sources, and conclusions.



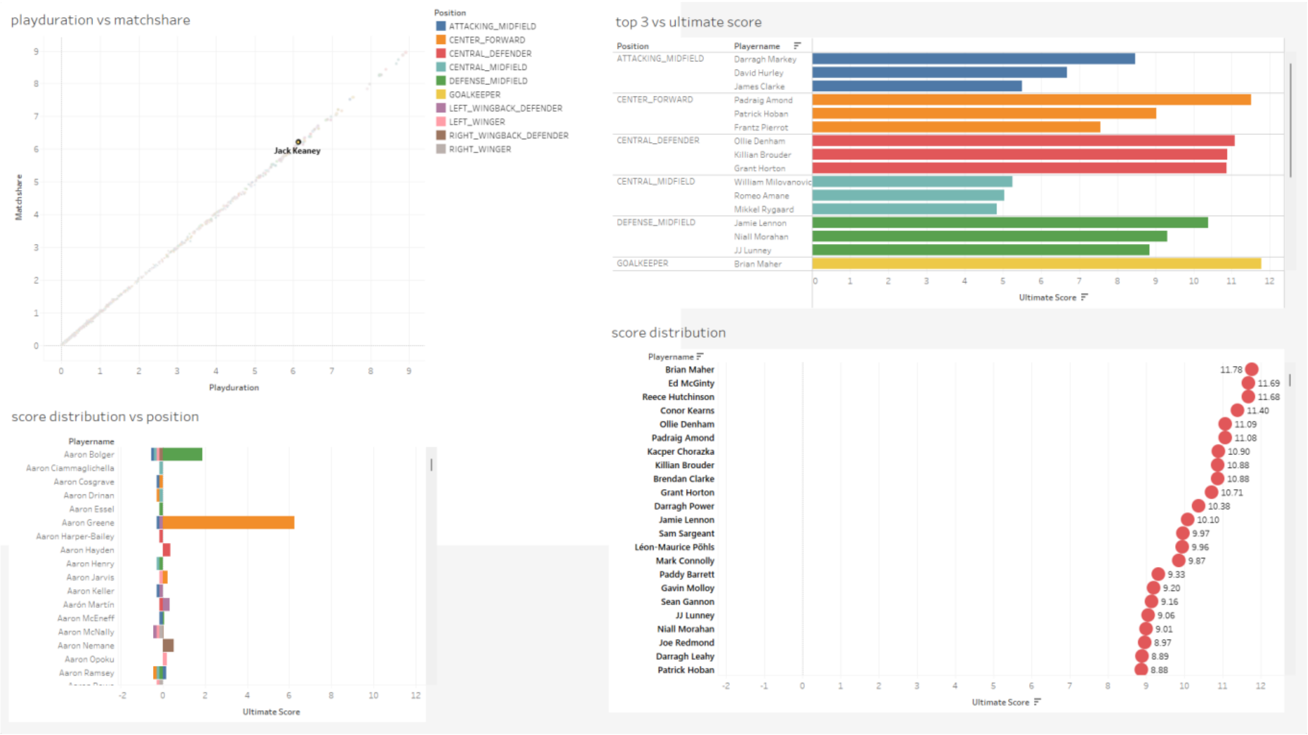
```python
import pandas as pd
import numpy as np
from sklearn.preprocessing import StandardSc
from sklearn.decomposition import PCA
from sklearn.model_selection import train_te
from sklearn.ensemble import RandomForestReg
from sklearn.metrics import mean_squared_er
from sklearn.cluster import KMeans

# Load and Preprocess Data
def load_data(filepath):
    """Load dataset from a CSV file."""
    return pd.read_csv(filepath, low_memory=

def standardize_column_names(data):
    """Convert column names to lowercase and
    data.columns = data.columns.str.lower()
    return data

def ensure_unique_column_names(data):
    """Ensure column names are unique by app
    cols = pd.Series(data.columns)
    for dup in cols[cols.duplicated()].uniqu
        dup_indices = cols[cols == dup].inde
        for i, idx in enumerate(dup_indices
            if i == 0:
                continue
            cols[idx] = f"{dup}_{i}"
    data.columns = cols
    return data

```



Outcome and Reflection

Successfully completing this entry task was a pivotal moment in securing my internship. The project was a challenging yet rewarding experience that enabled me to apply my data science skills to a real-world scenario in football recruitment. It underscored the importance of data cleaning, visualization, and clear communication when presenting complex analyses.

I am excited to continue contributing to Charlton Athletic FC's success by using data analytics to make informed decisions in player recruitment and beyond.