Final Capstone Project Proposal

I. Introduction

The project focuses on performing an Exploratory Data Analysis (EDA) of financial and performance metrics of football teams across 15 European leagues, including England, Spain, Italy, Germany, France, Portugal, Netherlands, Türkiye, Belgium, Russia, Greece, Austria, Ukraine, Denmark, and Switzerland. The goal is to analyze the relationship between financial spending and team performance, identify patterns and trends over time, and provide an interactive dashboard for users to explore these insights. The chosen project type is an EDA because it allows for in-depth analysis and visualization of the data, making it easier to uncover trends and correlations.

Given my background as a striker for 12 years in my city club, where I climbed all categories, combined with my passion for football and permanent following of international and European leagues, I bring a deep understanding of the sport to this project. Additionally, as a dedicated fan of Real Madrid, a club that consistently dominates global football both in sports achievements and financial prowess, I am well-equipped to analyze and interpret the data effectively.

II. Project Objective

The primary objective of this project is to analyze the financial efficiency and performance metrics of football teams across various European leagues. Specifically, the project aims to answer the following questions:

- How does financial spending correlate with team performance across different leagues?
- What patterns and trends can be observed in financial efficiency over time?
- How do different leagues compare in terms of financial and performance metrics?

By the end of the project, the goal is to provide a comprehensive analysis that highlights key insights and trends, presented through an interactive dashboard that allows users to explore the data dynamically.

III. Data Description

The dataset to be used in this project includes financial and performance data for football teams across 15 major European leagues. The key features of the dataset include:

- **Columns**: league, team, season, revenue, spent, net, goals_for, goals_against, wins, ties, losses, and other financial metrics.
- **Leagues Included**: England, Spain, Italy, Germany, France, Portugal, Netherlands, Türkiye, Belgium, Russia, Greece, Austria, Ukraine, Denmark, Switzerland.
- **Source**: The dataset has been gathered from the website https://www.transfermarkt.com/.
- **Structure**: The dataset contains 4,342 entries and 21 columns, with a mix of numerical and categorical data. Some columns, such as performance metrics (goals, wins, etc.), have missing values that will be handled during the data cleaning phase.

IV. Methodology

The approach to accomplish the project objectives includes the following steps:

1. **Data Cleaning and Preprocessing**:
   - Handle missing data by using appropriate strategies such as removal, imputation, or interpolation.
   - Transform categorical variables into numerical formats suitable for analysis.
   - Create new features, if necessary, to enhance the analysis.

2. **Exploratory Data Analysis (EDA)**:
   - **Descriptive Statistics**: Compute and visualize summary statistics for key numerical and categorical columns.
   - **Correlation Analysis**: Generate a correlation matrix and visualize it with a heatmap to identify significant relationships between financial and performance metrics.
   - **League-Wise Analysis**: Aggregate data by league and analyze trends over time using grouped bar charts, line plots, and stacked area charts.
3. **Interactive Dashboard Development**:
   - Develop an interactive dashboard using Streamlit or Plotly Dash (I didn't yet decide), incorporating all visualizations and allowing users to explore the data by filtering by league, team, or season.

V. Expected Deliverables

The final deliverables for this project include:

- **GitHub Repository**: A public repository containing all relevant notebooks, scripts, data, and documentation.
- **Interactive Dashboard**: A fully functional and deployed dashboard that allows users to explore the data interactively. The dashboard will feature various visualizations, such as scatter plots, bar charts, and correlation heatmaps.
- **PowerPoint Presentation**: A concise presentation summarizing the project objectives, methodology, key findings, and a demonstration of the interactive dashboard.

VI. Timeline and Tasks

The project will be completed according to the following timeline:

Week 1:

- Set up the GitHub repository and organize the project structure.
- Conduct initial data exploration and document the dataset.
- Perform data cleaning and preprocessing.
- Begin the EDA by calculating descriptive statistics and creating initial visualizations.
- Continue with correlation analysis and league-wise analysis.

Week 2:

- Start developing the interactive dashboard.
- Finalize the interactive dashboard and ensure all visualizations are correctly integrated.
- Prepare the PowerPoint presentation and finalize the README.md file.
- Review and refine the project deliverables.
- Submit the project by creating a PR and ensure all files are correctly linked and organized.

VII. Potential Challenges

Some potential challenges that might arise during the project include:

- Difficulties in finding an appropriate dataset.
- The dataset contains missing values in key columns like performance metrics, which may complicate the analysis.
- Integrating multiple visualizations into a single dashboard may lead to performance issues. This will be mitigated by optimizing the code and using efficient data structures.
- Understanding the context behind certain financial and performance metrics might require domain-specific knowledge.
- Present the project in English for the first time.