**Project Progress Report**

**PROJECT IDEA: Predicting the winners of the 2024 Champions League Football Tournament**

**Data Source:**

Our chosen data source for this project is the Kaggle dataset *Champions League Era Stats.* This data source contains different datasets. However, since the championship is club-based and mostly dependent on past performances by the different clubs and players, we will most likely be using the *All-Time Rankings by Club* data set and the *Player Goal Totals* or *Player Goal Details* datasets. The *All-Time Rankings by Club* dataset contains information about each club’s participation record, total number of games played, number of championships won, games drawn, and games lost. The other data sets contain information about goal rankings of the different players.

This data set is very trustworthy as it is sourced from the UEFA.com official website. UEFA is the official body that organizes the Champions League tournament, so we can be certain about the authority of the source of this data set.

**Source:** Champions League Era Stats <https://www.kaggle.com/datasets/basharalkuwaiti/champions-league-era-stats?rvi=1&select=AllTimeRankingByClub.csv>

This is an instance of the *Rankings by Club* dataset:

**A screenshot of a computer

Description automatically generated**

**Data Cleaning Strategies**

Although the data provided in these data sets seem to be relatively clean, we might need to perform some cleaning strategies just to double-check and ensure that everything works as we expect. Some of these strategies include the following:

* **Standardizing Team Names:** Some team names are inconsistent or abbreviated (e.g. "Man United" vs "Manchester United"). We will need to standardize all names to a consistent format.
* **Fixing invalid or inconsistent values:** We will need to scan for impossible values like negative goal numbers, ensure totals matches played is the accurate sum of its components (wins, draws and losses), etc. There might be need to fix or remove invalid records.
* **Filtering unnecessary columns:** We could remove any redundant, unused, or irrelevant columns that may not be essential in solving the problem, to further simplify the data.
* **Normalize encoding:** Our dataset is currently encoded in a UTF-16 format. There might be need to convert the encoding to UTF-8 to allow for compatibility. This might not be necessary considering the possibility of loss of some characters not available in the UTF-8 encoding format.

**Initial Data Exploration:**

Three very important factors for determining success in this championship will be past performance considering previous titles won , overall points secured, and record number of goals in the tournament. It only makes sense to explore the distribution of titles and goals among the top contending teams for this championship. Examples below:

**Distribution of Previous titles won:**

A screenshot of a computer

Description automatically generated

**Distribution of Points:**

**A graph of points won by club

Description automatically generated**

**Distribution of Goals Scored**

**A graph of goals and goals

Description automatically generated**