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:

```
import pandas as pd
```

```
clubs = pd.read_csv("AllTimeRankingByClub.csv", encoding='utf-16')
clubs
```

	Position	Club	Country	Participated	Titles	Played	Win	Draw	Loss	Goals For	Goals Against	Pts	Goal Diff
0	1	Real Madrid CF	ESP	53	14	476	285	81	110	1047	521.0	651.0	526.0
1	2	FC Bayern München	GER	39	6	382	229	76	77	804	373.0	534.0	431.0
2	3	FC Barcelona	ESP	33	5	339	197	76	66	667	343.0	470.0	324.0
3	4	Manchester United	ENG	30	3	293	160	69	64	533	284.0	389.0	249.0
4	5	Juventus	ITA	37	2	301	153	70	78	479	301.0	376.0	178.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...
531	532	CS Stade Dudelange	LUX	1	0	2	0	0	2	0	18.0	0.0	-18.0
532	533	Rabat Ajax FC	MLT	2	0	4	0	0	4	0	20.0	0.0	-20.0
533	534	Keflavík	ISL	4	0	8	0	0	8	5	35.0	0.0	-30.0
534	535	US Luxembourg	LUX	5	0	10	0	0	10	3	43.0	0.0	-40.0
535	536	FC Avenir Beggen	LUX	6	0	12	0	0	12	1	56.0	0.0	-55.0

536 rows × 13 columns

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.

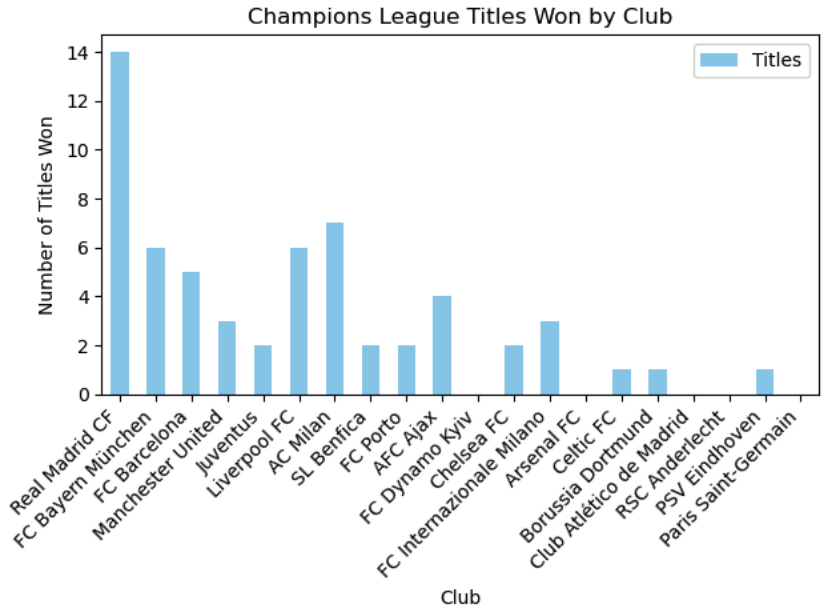
Initial Data Exploration:

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:

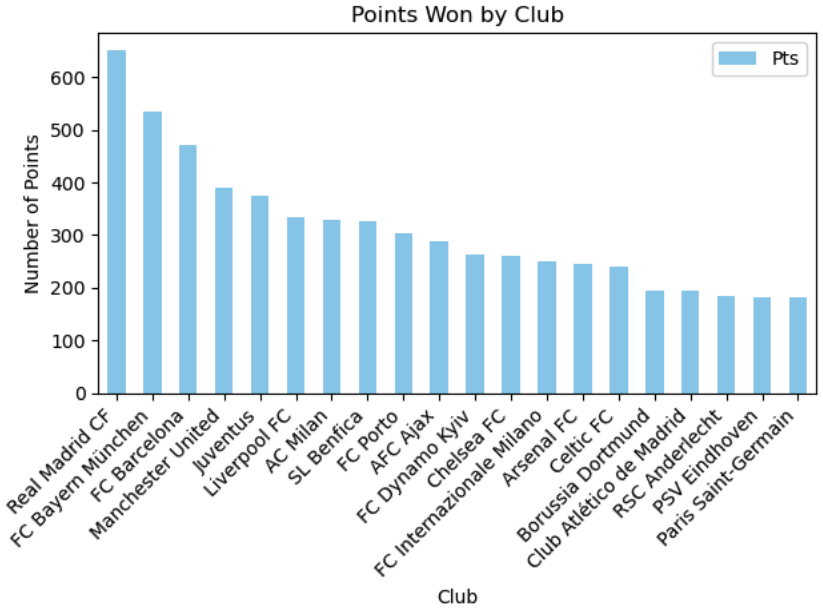
Distribution of Previous titles won:

```
plt.figure(figsize=(12, 6))
top_20.plot(x='Club', y='Titles', kind='bar', color='skyblue')
plt.xlabel('Club')
plt.ylabel('Number of Titles Won')
plt.title('Champions League Titles Won by Club')
plt.xticks(rotation=45, ha='right')
plt.tight_layout()
plt.show()
```

```
<Figure size 1200x600 with 0 Axes>
```



Distribution of Points:



Distribution of Goals Scored