Final Project Report

2023 – 2024 Soccer Player Statistics and Their FC25 Ratings

By: Jake Thomas

1. Introduction

EA Sports FC (formerly FIFA) is a soccer video game franchise who's ratings of players hold a large weight amongst fans and football communities around the world. Statistics are a key indicator in a soccer players performance, and have a great deal to do with their ratings in EAFC. In EAFC, each player's rating is based on their performance in the past few seasons, but mainly takes into account the season directly prior to the new game. This is why I have decided to analyze the players last season's statistics in order to see their impact on the players FC25 ratings.

In this project, I have scraped FBref[1] statistics and performance metrics to see if there is a correlation between the players 2023-24 statistics and the players FC25 ratings that I obtained from Kaggle[2], as well as find out which clubs and leagues have the highest average FC25 rating.

2. Data

This project uses two data sources: FBref's[1] 2023-24 player statistics, and Kaggle's[2] FC25 player rating data.

2.1 FBref Player Statistics

The first data source I used in this project is FBref[1], a website that tracks in-depth statistics about every player that plays in the top 5 leagues in the world (Premier League, La Liga, Serie A, Bundesliga, Ligue 1). The dataset I used contains 2852 rows of match statistics of each player for the 2023-2024 season. The actual dataset contains around 40 columns, but I only used 10 of the most relevant features, seen in the data dictionary. I scraped this data from FBref[1] using the Python package Selenium in the notebook 00_Fbref_scraping.ipynb.

The dataset needed a good amount of cleaning, including imputing missing values like a handful of missing ages and nationalities. I did research on these players and imputed the data manually. There were many duplicate players in the dataset because of mid-season transfers where players switched clubs during the season. These players had their statistics for the 2023-2024 season divided between two teams, so I needed to add their totals for these statistics and consolidate them into one row each. I did this by creating a data frame of the duplicates (duplicate names **and** ages to ensure players with the same name were not affected), copying that data frame, and then summing each players goals, assists,

[1] https://fbref.com/en/comps/Big5/2023-2024/stats/players/2023-2024-Big-5-European-Leagues-Stats
[2] https://www.kaggle.com/datasets/nyagami/ea-sports-fc-25-database-ratings-and-stats?select=male_players.csv

matched played, and expected goals and grouping them by the player. Finally, I dropped the duplicates from the original data frame and remerged the cleaned/summed rows back into the original data frame. This cleaning was done in 02_FBref _cleaning.ipynb.

2.2 FC25 Player Ratings

The second data source I used is from Kaggle[2]. It is a downloadable CSV file that contains the EA Sports FC25 overall ratings and physical/player attributes for every player in the world. This was scraped from the EA Sports Website by Davis Nyagami. The CSV file contains 16,159 rows of these ratings and statistics for each player, based on the 2023-2024 season. The file has 58 columns, but again I only used 10 of the most relevant features.

This data set did not need much cleaning, as it was already well-established on Kaggle. However, I did need to drop most of the irrelevant columns so I was left with the 10 columns that I used for my analysis. I also needed to replace many of the players' names with names that matched the other FBref data set. For example, a player named Vinicius Junior, was referred to as "Vini Jr." in this data set. I needed to rename this example, as well as many others, so I would be able to merge the two data sets using the player names as the primary key. This cleaning was done in 01_FC25_import_cleaning.ipynb.

2.3 Merging Player Statistics and Ratings

Next, I merged the two datasets using horizontal integration on the primary key of Player Name. I **intersected** the datasets using an inner join, so that only players that are in both datasets (play in the top 5 leagues) will be included in the final dataset. As a result, I was left with 2109 players from the FBref[1] dataset, with their FC25 ratings and attributes added on. The data still needed a bit of cleaning after the merge. There were still a few duplicates because of some occasions of players having the exact same name, so I dealt with those by dropping the duplicates but leaving the first instance. Then I renamed all of the columns to more interpretable names, and replaced each league name from its abbreviation to the full league name. Finally, I removed the "cm" and "kg" strings from the Height and Weight columns, and transformed them to numeric types. This cleaning was done in 03_Fbref_FC25_merge.ipynb.

The final edit I made to the dataset was in 04_visualizations.ipynb, where I created the Goal Contributions column. This is simply the players Goals + Assists, and is used to give us a better idea of how the player impacts games without always scoring.

Table 1 Data Dictionary

Column	Type	Source	Description
Player	Text	FBref	The name of the player
Nation	Text	Kaggle	The nation the player represents in international competitions
Position	Text	Kaggle	The position(s) the player plays
Club Team	Text	Kaggle	The club team the player plays for
League	Text	FBref	The league the club is in
Age	Numeric	Kaggle	The age of the player
Matches Played	Numeric	FBref	The number of matches the player played in the 2023-2024 season
Goals	Numeric	FBref	The number of goals the player scored in the 2023-2024 season
Assists	Numeric	FBref	The number of assists the player got in the 2023-2024 season
xG	Numeric	FBref	The number of goals the player was expected to score in the 2023-2024 season
Overall	Numeric	Kaggle	The overall rating of the player in the FC25 game
Height	Numeric	Kaggle	The height of the player in centimeters
Weight	Numeric	Kaggle	The weight of the player in kilograms
Pace	Numeric	Kaggle	The pace rating of the player in the FC25 game
Shooting	Numeric	Kaggle	The shooting rating of the player in the FC25 game
Passing	Numeric	Kaggle	The passing rating of the player in the FC25 game
Dribbling	Numeric	Kaggle	The dribbling rating of the player in the FC25 game
Defending	Numeric	Kaggle	The defending rating of the player in the FC25 game
Physicality	Numeric	Kaggle	The physicality rating of the player in the FC25 game

3. **Analysis**

This project aims to analyze the effects of players statistics from the 2023-24 season on the players EA FC25 ratings. I have also explored some broader ideas about the clubs and leagues, as well as the relationship of a players age and their overall FC25 rating.

3.1 Player Summary Statistics

Before I began some of my deeper analysis, I wanted to get a better understanding of the data by exploring some summary statistics. I found that the average overall rating in my data set was 74.61 and the median overall rating was 75. These are very close to each other, which indicated to me that the data distribution is likely symmetrical or is not skewed. Here are a couple plots illustrating this point:

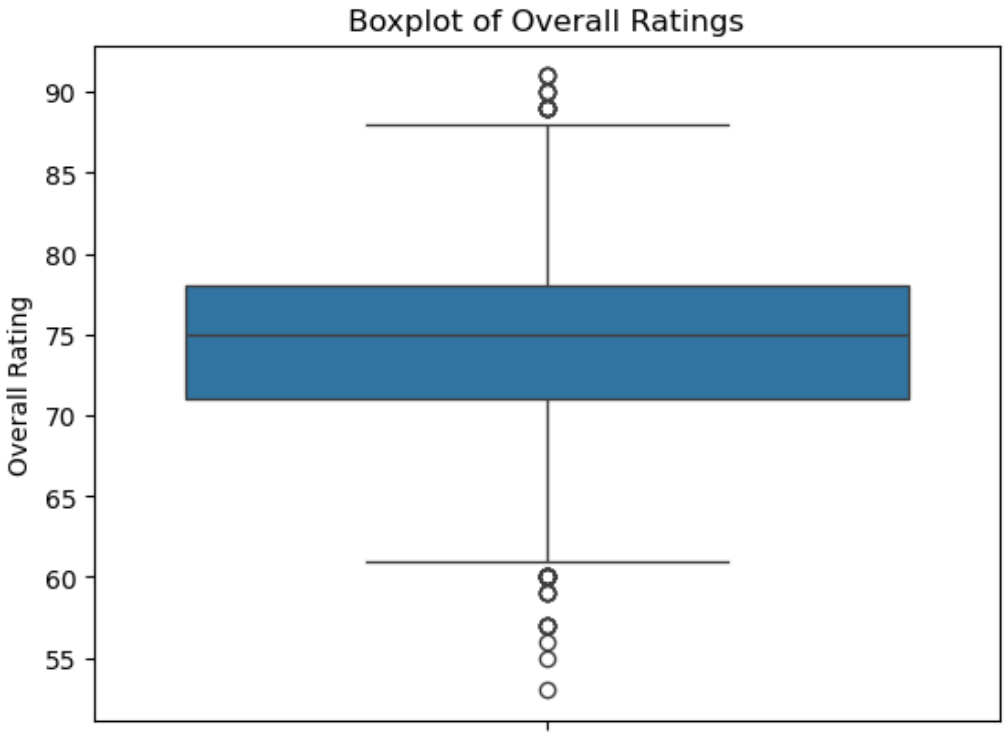

Figure 1: Overall Rating Boxplot

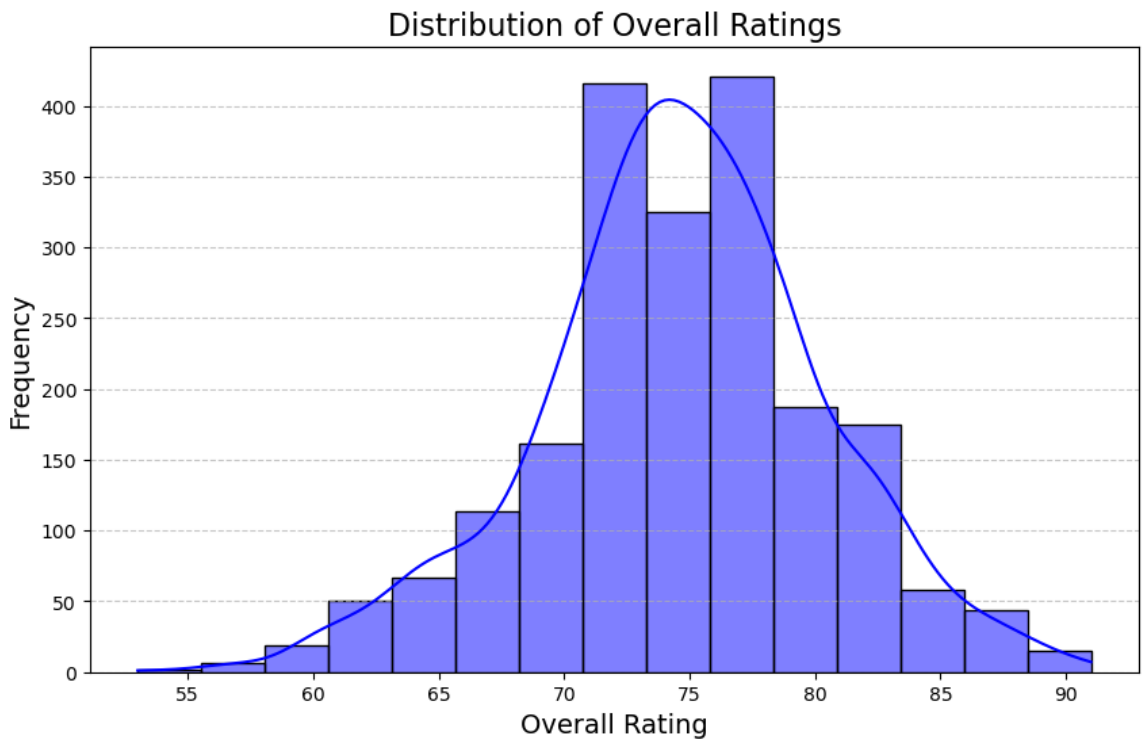

Figure 2: Overall Rating Histogram

I repeated this descriptive statistics process for the Goal Contributions column. These statistics were extremely right skewed, with the average number of goal contributions being 3.47, and the median being 2. Some plots that illustrate this as well the outliers are included below:

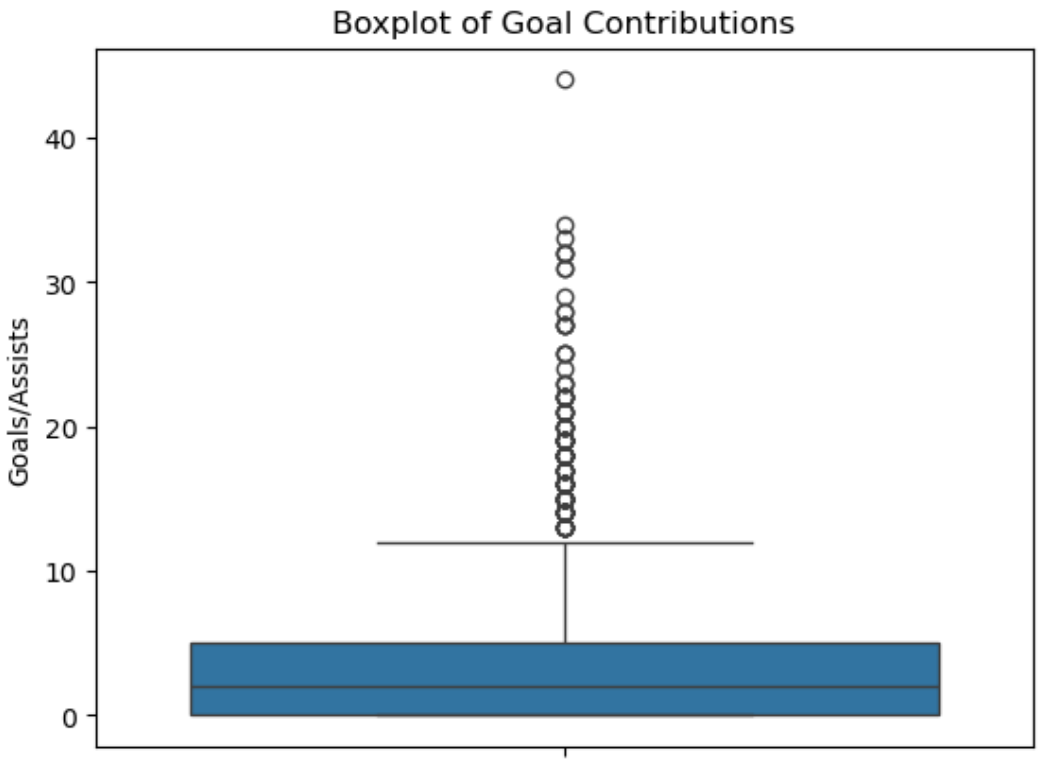

Figure 3: Goal Contributions Boxplot

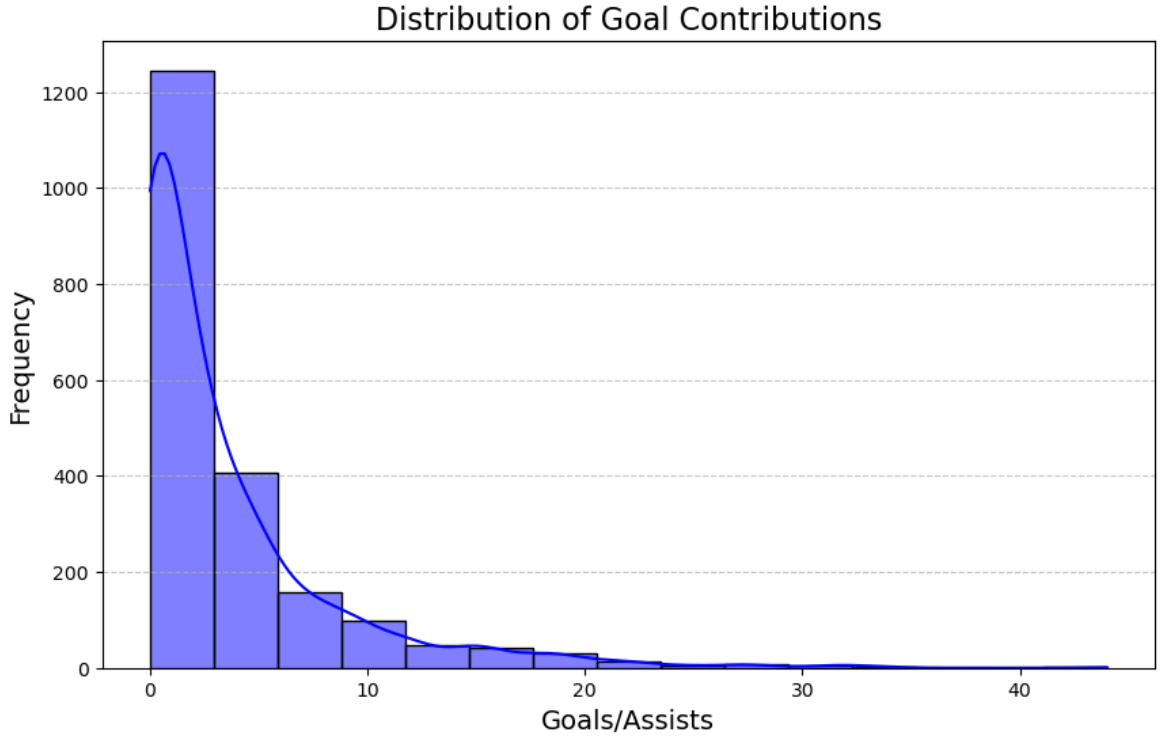

Figure 4: Goal Contributions Histogram

My final descriptive analysis was on the Age column. I found that the average age for a player in this data set was 26.24, while the median was 26. Similar to Overall Ratings, this indicated that the Age data was pretty evenly distributed, as you can see in the plots below:

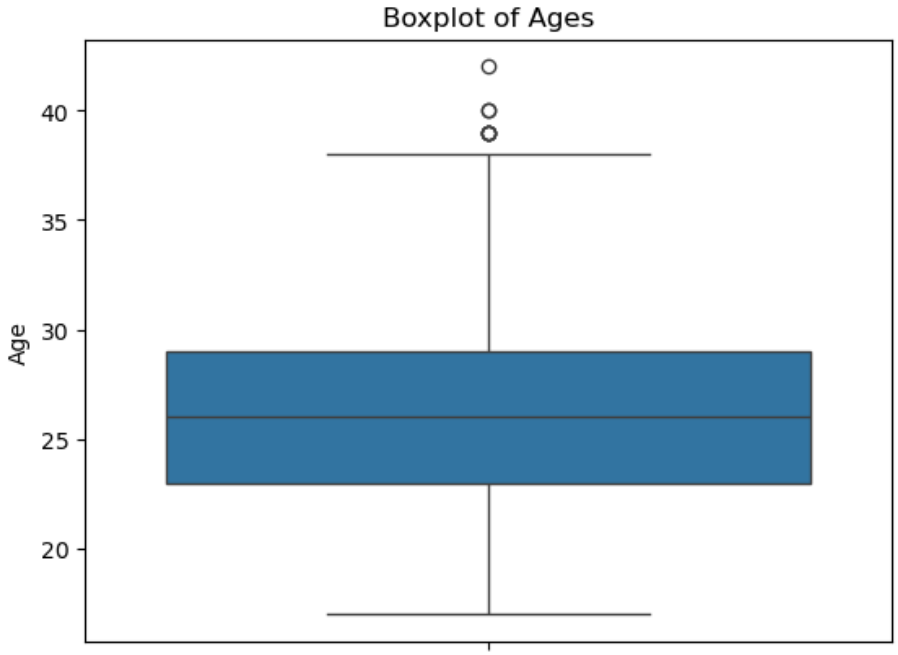

Figure 5: Age Boxplot

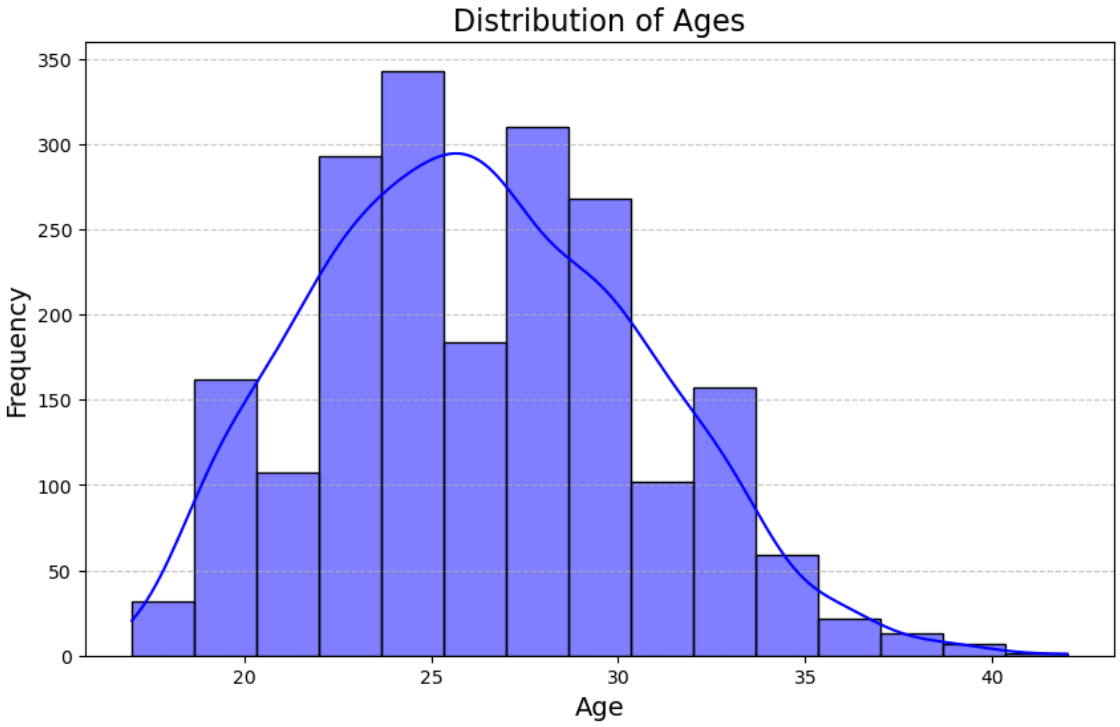

Figure 6: Age Histogram

3.2 Goal Contributions and Overall Rating Correlation

Next, I examined the correlation between players Goal Contributions (Goals + Assists) and their Overall Rating in FC25. The Pearson Correlation Coefficient was 0.4875, which I interpreted as a moderate linear correlation. I did a p-value test, and found that the p-value was 0.00, which indicated that we reject the null hypothesis and therefore significant correlation between the two columns exists.

My theory as to why the correlation was not strong as I predicted, is that there are many highly rated defenders and goal keepers in the data set that do not score goals or provide assists very often. The goal contributions metric is much more applicable to midfielders and forwards, despite there being plenty of very good players in other positions, which makes sense as to why the correlation was only moderate. Here is a scatterplot displaying the correlation of these two features:

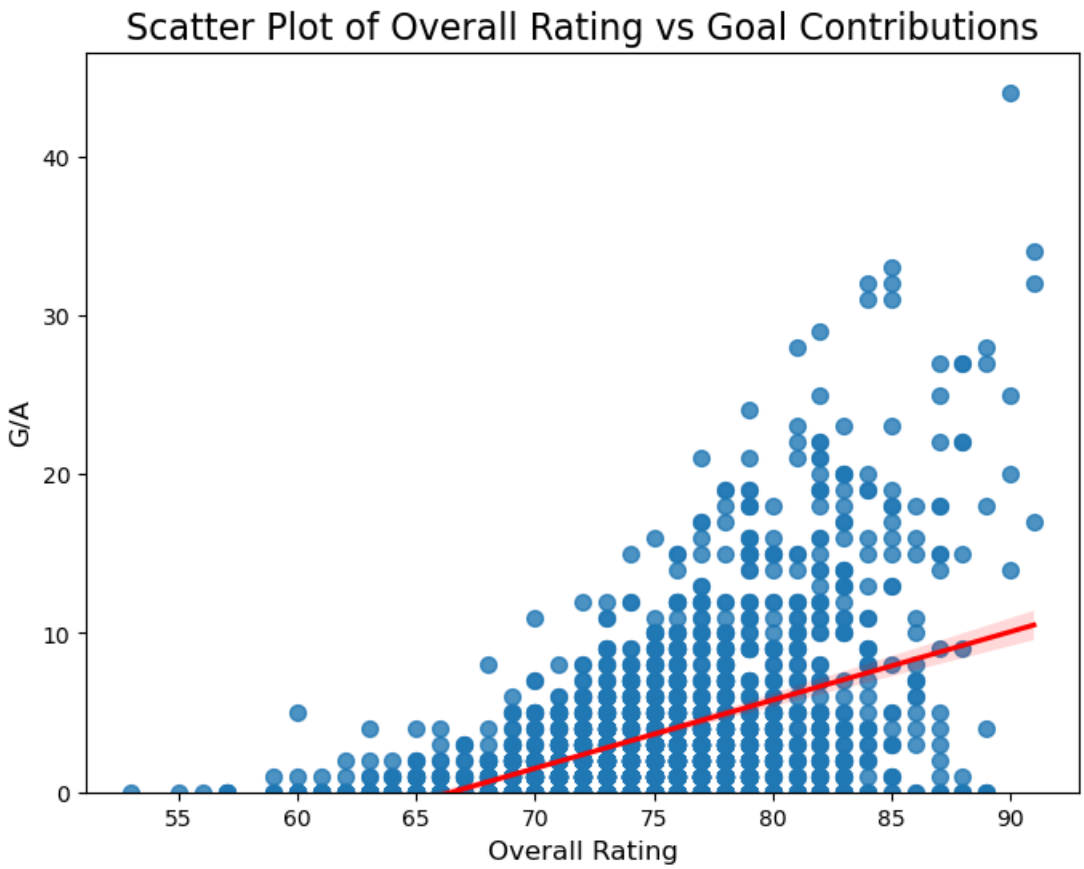

Figure 7: Overall Rating vs Goal Contributions Scatterplot

3.3 Age and Overall Rating Correlation

Next, I examined the correlation between players Age and their Overall Rating in FC25. The Pearson Correlation Coefficient was 0.3899, which I interpreted as a moderate linear correlation, however less than the Goal Contributions correlation. I did a p-value test, and found that the p-value was 0.00, which indicated that we reject the null hypothesis and therefore significant correlation between the two columns exists. Here is a scatterplot that displays these findings:

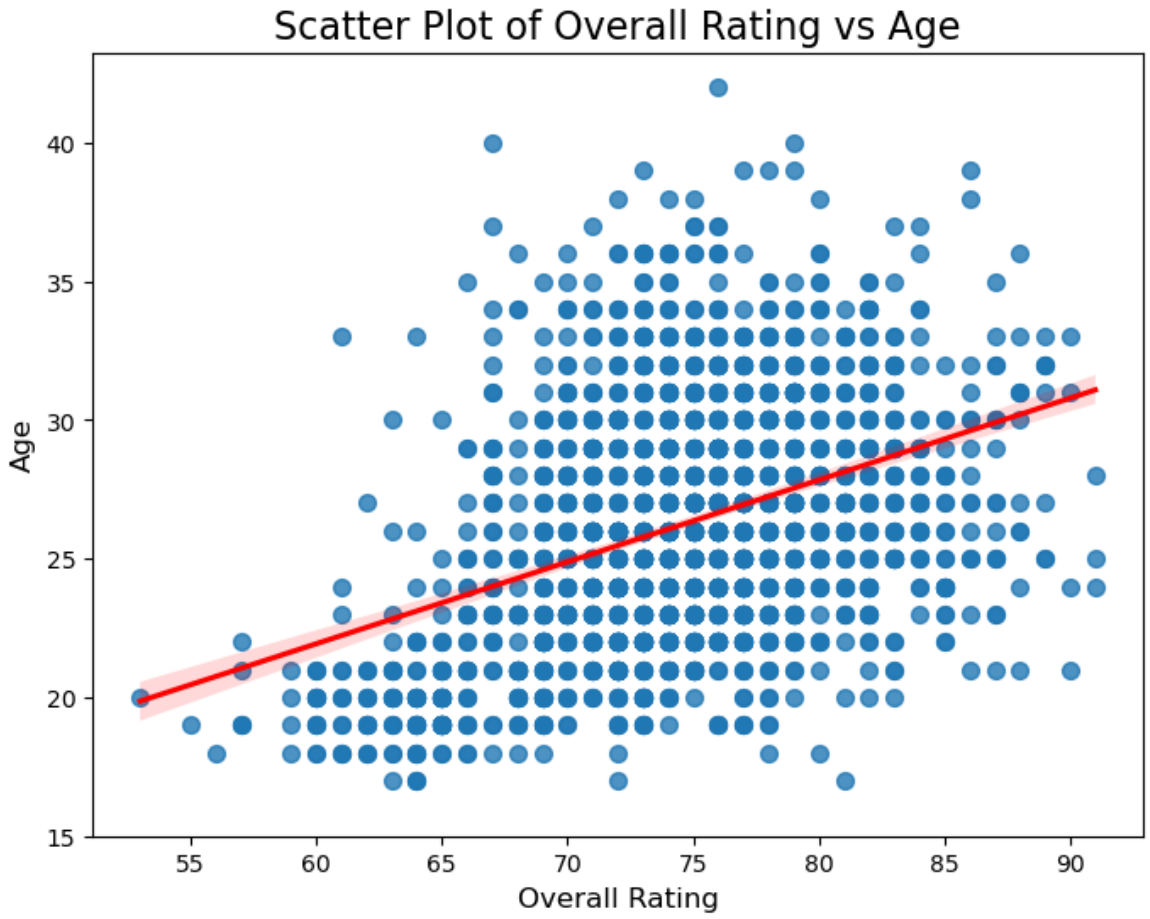

Figure 8: Overall Rating vs Age Scatterplot

3.4 Average Ratings per Club

I wanted to understand which clubs had the highest average overall ratings, and see if it aligned with my expectations based on which teams are widely regarded as the top in Europe. My analysis proved to be accurate, as all of the top ten teams had very good seasons last year as a result of having very good players. Here is a visualization that shows the top 10 teams with the highest average player ratings:

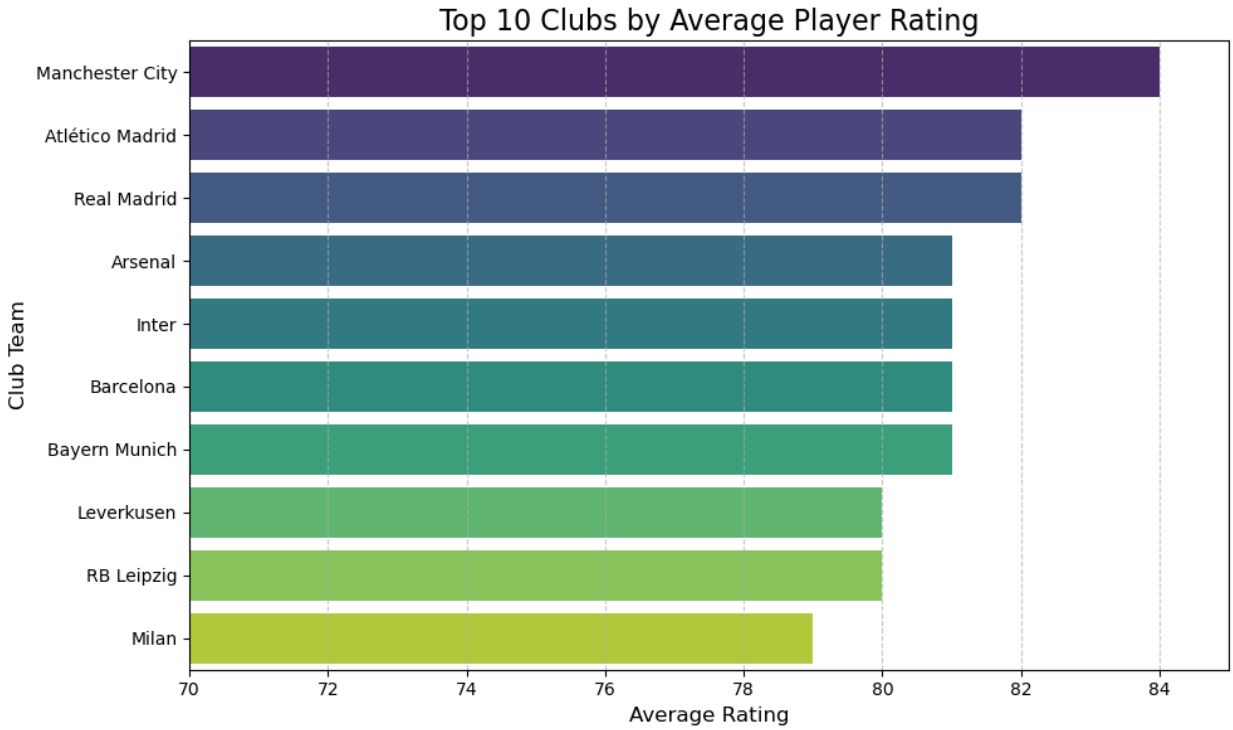

Figure 9: Top 10 Clubs by Average Player Rating Bar Chart

3.5 Average Ratings per League

I repeated this analysis for the 5 European Leagues that I included in my research, and found that the Premier League (English league) had the highest average overall rating, but by a small margin. La Liga (Spanish League) was right behind, before a somewhat significant drop off between it and Bundesliga (German League). Finally, Serie A (Italian League) and Ligue 1 (French League) rounded out the bottom of the average ratings. This bar chart depicts these findings:

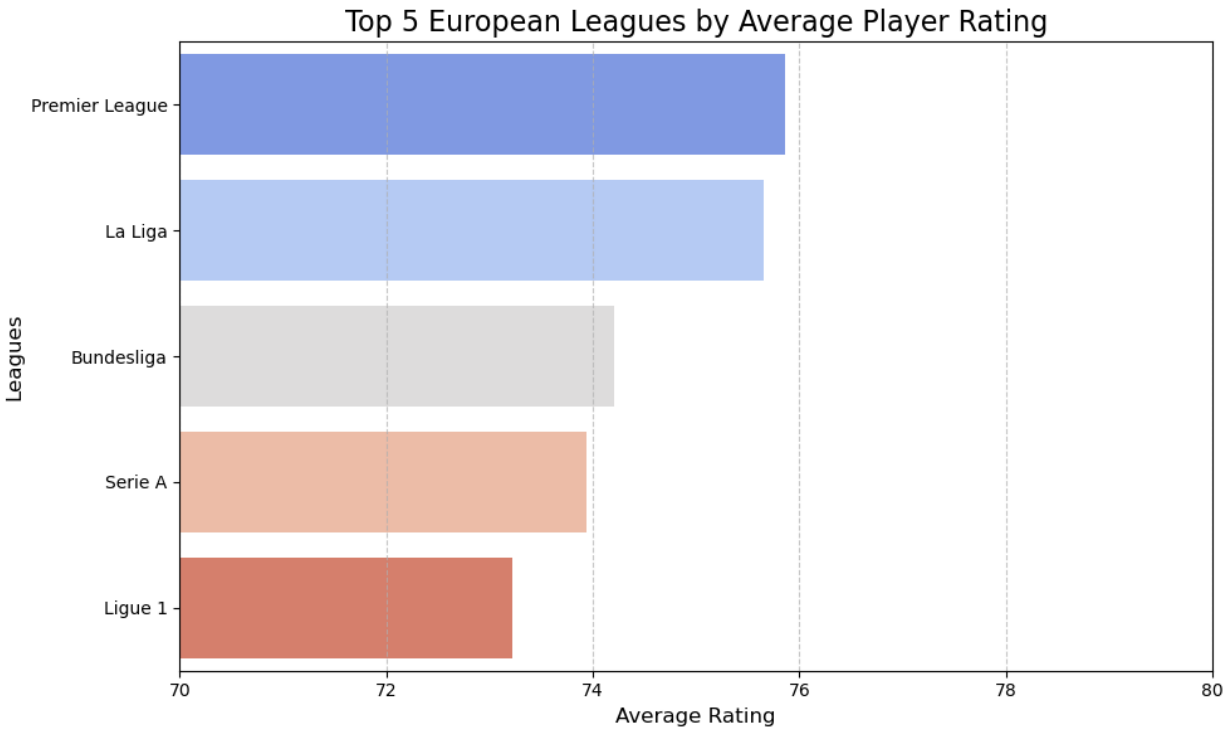

Figure 10: Top 5 European Leagues by Average Player Rating Bar Chart

4. Conclusion

In this project, I analyzed many different 2023-2024 seasons statistics of players in the top 5 European Leagues, along with their FC25 ratings and attributes. In summary, from the analysis questions presented in my proposal, I found the following results.

1. *Is there a strong correlation between a players goal and assist tally (Goal Contributions) and their FC25 overall rating?*

   There is a moderate correlation between a players goal contributions and their FC25 overall rating. This was somewhat surprising to me, until I realized that the goal contributions metric does not do a great job accounting defenders and goal keepers who may have high ratings.

2. *Is there a strong correlation between a players age and their FC25 overall rating?*

   There is a moderate correlation between a players age and their FC25 overall rating. I found this interesting because I wasn't sure what to expect, but the fact that there is any correlation at all is intriguing.

3. *Which club has the highest average FC25 player rating?*

   The club with the highest average FC25 player rating is Manchester City with an average player rating of 84. This is not surprising as they are currently widely regarded as the best team in the world. The rest of the teams in the top 10 are also very prestigious and successful clubs so they make sense as well.

4. *Which league has the highest average FC25 player rating?*

   The league with the highest average FC25 player rating is the Premier League. It is extremely close however, with La Liga only behind by 0.2. This checks out because I believe most football fans around the world would agree that the English and Spanish leagues have the highest quality players.

This project has many limitations, including discrepancies in the names of the players in each dataset, players with the same name, and a lack of impactful defensive statistics. These limitations resulted in many players being excluded from the analysis, and a future implementation could include imputing these players back into the data set. The biggest limitation of this analysis is that I am only examining offensive statistics (Goals, assists, and expected goals) which leaves defenders and goalkeepers as possible outliers when it comes to their FC25 ratings compared to their goal contributions. In the future, I would like to expand this analysis to include defensive/goalkeeping statistics such as clean sheets, tackles, and saves.