A Rainy Day in Stoke: Comparing Soccer Players

Justin Chen, Morissa Chen, Mihir Dhoot, Ashwin Nallan

1 Introduction

Soccer is the world's most popular sport with an estimated fan base of around 3.5 billion people. However, this popularity is highly skewed, with a majority of viewers invested in few leagues and competitions. We aim to explore team and individual player performance statistics to compare athletes spanning across professional leagues. By building this comprehensive analysis of all football players, fans and enthusiasts can deepen their understanding of the sport worldwide thus reducing the popularity gap between leagues.

The primary objective is to identify and visualize player performance comparisons on a season-by-season basis across genders and leagues to improve fan engagement and knowledge level. This analysis will also enable us to explore interesting topics such as contract year surges, injury performance impact, national vs. club player performances and more.

2 Literature Review

2.1 How it's done today

Sports have seen a surge in analytical studies of player performance in the past decade. Examples include regression modeling to measure player performance change from new contracts in baseball [8] or clustering of basketball players to compare playing style across positions [3]. While soccer holds no exception to this trend, the existing research in this domain either evaluates performance longitudinally within teams or a cross-sectionally for a specific league [6]–rarely accounting for both. Furthermore, current comparisons of soccer players often hone in on a single metric of the game such as passing[7] or goal scoring [5]. No publicaclly available comprehensive analysis for multiple performance metrics across leagues was found.

The methodology observed in these studies include ANOVA testing [5], k-means and hierarchical clustering [12], as well as advanced but niche clustering methods [1]. These studies often use biometric and positional data from tracking devices and sensors that are not available to us [1]. While these studies use more intricate data, they lack bigger picture applications. For example, one clustering study explored the best way to map soccer data and define a distance metric between players [1]. While the research evaluates various clustering methods with detailed analysis on efficiency and validation, it did not interpret or contextualize the cluster results for broader applications.

2.2 What is new in our approach?

We believe that our holistic approach is novel for this sport. Studies have shown a difference in technical performance [10] as well as aggressiveness and refereeing [11] across football leagues. Differences in match characteristics are even more pronounced when comparing women's leagues [9] due to physiological differences in their anatomy [4]. While this research has identified such discrepancies and the negative perceptions it can cause, fans lack the ability to adjust for these differences and fairly compare player performance across leagues. We hope to fill this gap by evaluating the distribution for various performance metrics per season and normalizing the data across leagues. We will attempt to cluster the normalized data and identify similarities/differences between players across leagues. In taking this holistic view, our results can be readily contextualized in the broader "world soccer" picture and allow for greater user interactivity.

3 Methodology

Player performance and team performance statistics are scraped from 20 leagues across all available seasons using Selenium drivers. An SQLite3 database stores the data in a single database file that is integrated with Python very neatly. It required no servers or installation process and is easily portable. "SQLite Studio" is used for easy interaction and data readability. Since many advanced statistics are available only for recent years, an initial cleaning process identified the top 9 attributes that were present across the full dataset. This data was imported to Python Notebooks in Google Collab for the analysis and modeling.

Using this data we generate meaningful clusters of players using K-Means with the number of clusters chosen through the Elbow Method. Various iterations were tested using different subsets of attributes and methods of scaling/normalization, tune parameters, and initialize cluster centers. Model validity will be evaluated primarily through Silhouette plots and alignment of clusters with positional data. The effectiveness of the clustering then informs how well we are able to determine player similarity for any given individual player.

The second objective is to generate profiles for each league based on selected attributes to find significant differences between league playing styles. This information will also be used in our clustering model. We evaluate the distribution of key attributes for each league to identify where normalization and scaling can be used, both within a league and across all leagues. A variety of methods including z-score normalization, logarithmic scaling, and rank order scaling were explored.

We built a website to display the results from the data analytics along with various interactive tools. The first tool allows users to find players similar to a given player. A clustering algorithm along with a carefully designed similarity metric determines the edges between various players. Interactive visualisation provides for easy exploration of the player comparison data. The second tool allows users to explore the differences between various leagues across various important statistics. Users can select specific leagues and statistics to gain insight into the differences in playing style of leagues. The last tool allows users to explore clustering soccer players themselves with the freedom of choosing various normalization methods and number of clusters. The website is built as a Flask app with a Python backend that provides all the data and analysis required for the website. We used Javascript/HTML/CSS to render the front-end and d3 to visualize various infographics. The objective of this website is to help achieve our ultimate goal of improving fan engagement and expanding soccer knowledge level. We then tested the website on available users for feedback and evaluation.

4 Data

Soccer data was scraped using Selenium on Python from FBref.com, a website that provides soccer statistics for a wide range of competitions. Statistics were scraped at the player level per season to compare individual players on a season by season basis. Additionally, statistics were scraped at the team level for additional context on individual performance.

On the player level, we scraped different categories of stats including:

- Standard stats: Position, team, age, goals, games played, etc.
- Goalkeeping: Saves, save percentage, etc.
- Shooting: Shots, shots on target, etc.
- Passing: Passes completed, passes attempted, pass types, etc.
- Defense: Tackles, interceptions, etc.
- Possession: Touches, dribbles, etc.

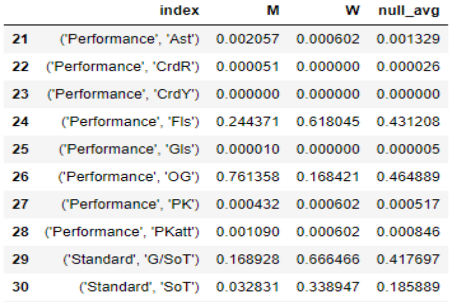
On the team level, we scraped standard stats including wins, losses, and draws, as well as team aggregates of the individual player stats mentioned above (shots, passes, tackles, etc).

We scraped data for a range of competitions including popular competitions (English Premier League & Men’s World Cup), lesser known leagues (South Korea K-League), and women’s competitions (NWSL & Women’s World Cup). We then filtered out all players who played less than 90 minutes or participated in less than 2 matches to remove small-sample skew.

The initial player dataset yields approximately 150000 rows and 160 attributes, with the removal of players with minimal playing time reducing the functional dataset to around 122000 rows. The team data yields around 7000 rows and 87(?) attributes.

4.1 Data Cleaning

Due to the variability in data availability across leagues and seasons we were not be able to use all the data scraped from FBRef. Furthermore, some stats were repeats or calculations based on other attributes. To narrow down our data to feed our clustering model, we first cleaned the data once again, removing seasons and attributes with more than 50% of their data missing. Below is a snapshot of our data cleaning process by attribute:



	index	M	W	null_avg
21	('Performance', 'Ast')	0.002057	0.000602	0.001329
22	('Performance', 'CrdR')	0.000051	0.000000	0.000026
23	('Performance', 'CrdY')	0.000000	0.000000	0.000000
24	('Performance', 'Fls')	0.244371	0.618045	0.431208
25	('Performance', 'Gls')	0.000010	0.000000	0.000005
26	('Performance', 'OG')	0.761358	0.168421	0.464889
27	('Performance', 'PK')	0.000432	0.000602	0.000517
28	('Performance', 'PKatt')	0.001090	0.000602	0.000846
29	('Standard', 'G/SoT')	0.168928	0.666466	0.417697
30	('Standard', 'SoT')	0.032831	0.338947	0.185889

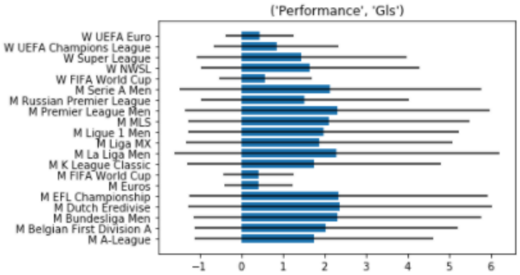
We ultimately found a common pool of 33 attributes that had more than 70% of its data available across a large subset of seasons, including the following:

- Age
- Matches Played
- Minutes Played
- Goals
- Shots on Target
- Assists
- Fouls
- Tackles + Interceptions
- Yellow Cards
- Red Cards

Through this second iteration of data cleaning for our player data, we were able to preserve 100,000 seasons across the 33 attributes. As will be discussed further in our results section, we were pleased with the range of attributes available for a forward or a defender's statistical profile, though we would have liked more detailed passing statistics to help distinguish a midfielder's impact. We then looked at potentially transforming and normalizing various attributes to yield more meaningful comparisons between players.

4.2 Data Normalization

4.2.1 Key Differences By League: We started by looking at each statistic's distribution by league, and identified 2 hidden factors influencing the distributions of these statistics by league. The first is that each league plays a different number of matches. This directly inflates pure counting stats for leagues with more games, as shown below.



In addition, each league's play style varied slightly, leading to differences in statistical profiles of each league even when taking the number of matches or minutes played into account.

4.2.2 Key Differences By Gender: In concordance with previous research, we found that Women had fewer yellow and red cards per game. Another distinguishing attribute was goals. Women's soccer competitions generally featured more goals per 90 minutes compared to male soccer competitions. We thought most of this variation could be accounted for, however, through normalizing on a league-by-league basis, depending on statistic.

4.2.3 Final Normalization Ultimately, we used a variety of methods to normalize the data for different attributes. The adjustments made are as follows:

- Playing time stats (Matches Played, Minutes Played, and Starts) were all adjusted over the total number of matches in each league.
- Goals, Assists, and Shots on Target, were adjusted over 90 minutes following common football reporting standards.
- Fouls and Tackles for every player were divided by the max values within a players league.
- We took the log of Yellow cards and Red cards due to its highly skewed distribution.
- We used the z-score of a players age within its individual league because age followed a relatively normal distribution within each league with varying means in different leagues.

We also obtained the rank order of Matches Played, Goals, Assists, Penalty Kicks, Shots on Target, Yellow cards, and Red cards within each league due to the high variation of both the mean and standard deviation of these attributes in different leagues. Once all our data was prepared for use, we created a Google Collab for our team, and added an SQLite database containing the final set of player data.

4.2.4 Team Data Our team-wide data helped validate some of our findings from the player data. We found a similar trend where goals varied most directly with matches played per season. This variation on a team-data level was a key foundation to our League Comparator tool, so we chose to avoid any further normalization on the team-wide data.

5 Results

Our final clustering model uses 24 total attributes and computed similarity using euclidean distances. Using the elbow method we picked a cluster number of 6. The cluster sizes ranged from 3235 to 39120. We were not able to obtain a high silhouette score (0.24) for this clustering model due to the lack of sufficient performance statistics. However in exploring the clusters we were able to distinguish distinct characteristics of players in the 6 clusters.

- Cluster 0 was predominantly defenders, midfielders, and defender/midfielder hybrid players. These were low performing players that did not accumulate significant field time and subsequently had low numbers of Tackles/interceptions, fouls, and red/yellow cards. They also tended to be younger.
- Cluster 1 was mid tier forwards and midfielders that had high playing times.
- Cluster 2 was a small subset of aggressive defenders and midfielders that had very high numbers of Tackles and interceptions.
- Cluster 3 was our forwards that did not have much playing time. Although their Shots on Target and Goals/Assists per 90 minutes were not significantly lower than cluster 1 they had significantly lower metrics for defensive statistics. They also tended to be younger players.
- Cluster 4 was our high scoring forwards and midfielders. They also had equal or higher numbers of tackles/interceptions, fouls, and yellow/red cards as some defensive players.
- Cluster 5 were defenders with the highest playing times.

These cluster results then informed finding similar players to each existing player. Because our clusters revealed groups of players with distinct characteristics, we can then be more confident in utilizing the same similarity metric (euclidean distance) to identify players with similar characteristics to that of a specific individual player. Then, for each player and each season played, pairwise distances were calculated, and the most similar player to each season a given player played was then classified as "similar". For example, Lionel Messi has 15 seasons worth of statistics, and thus has 15 similar players, though some were duplicates since he was similar to the same player in multiple seasons.

Overall, we think the clustering and the resulting similarity measure were largely in line with what we hoped to accomplish. The two main points for further improvement we noticed were both related to data. The first, as discussed briefly in the data section, was the lack of ability for the model to handle midfielders due to the lack of passing statistics. Strong defensive or scoring midfielders were clustered with defenders or forwards, respectively. Midfielders that we know anecdotally mainly focus on passing, however, were often clustered with players that had very little impact on the game in Cluster 1.

The second was that women players often had other women players considered "most similar" to them, even though women made up a small minority of the overall dataset. We think that further data normalization could help make cleaner comparisons between men and women.

6 User Interface

6.1 Potential Users

Our project primarily targets fans to increase engagement and interest. We hope that our product will serve as a go-to point for soccer enthusiasts to explore data from all leagues and better understand their favorite players or rising stars.

This UI component is especially valuable when considering new soccer fans that will benefit from this tool. The complex structure of professional soccer involves far more leagues and international competitions than other sports. This can be confusing and intimidating for new fans, and we hope to remove this barrier of entry by enabling easier comparisons across leagues.

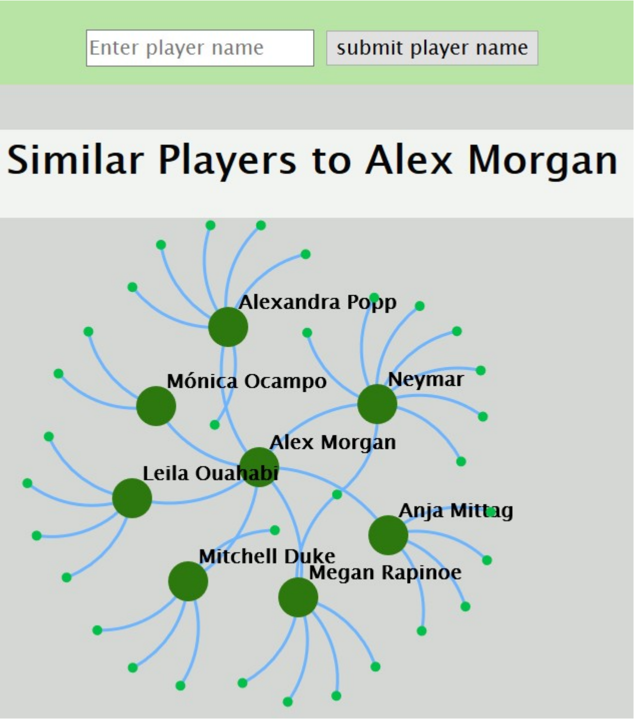
This project could also be used by coaches, players, and the rising sports betting market. Sports coaches and executives can take advantage of our results for club-level decisions in marketing, playing time, training, and new player acquisitions. For the sports betting market, our analysis will quantify the variation in competition levels across leagues which has historically caused mispricing and miscalculations of profitable odds [2].

6.2 UI Tools

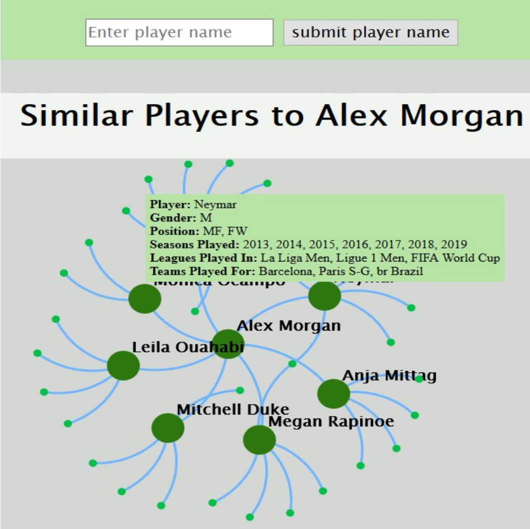
We built an interactive website with tools to allow users to compare players and teams in the world of soccer. The three tools are as follows:

6.2.1 Player Comparator : Users can input the name of a certain player (i.e. their favorite). A graph is then generated that displays nodes as players and edges connecting nodes which indicate that players are similar. Players displayed include the queried player q, the $S = [s_1, ..., s_n]$ players similar to q, and $M = [m_1, ..., m_n]$ players similar to S. Users can then hover over a node to view player information such as name, position, or teams played for, as well as double click to search for similar players to that node.

This allows users to expand their soccer knowledge by finding new players similar to ones they already know of. For example, new soccer fans resulting from the success of the US Women's National Team (USWNT) may wish to delve deeper into the world of soccer. They could then search for a prominent USWNT player such as Alex Morgan:



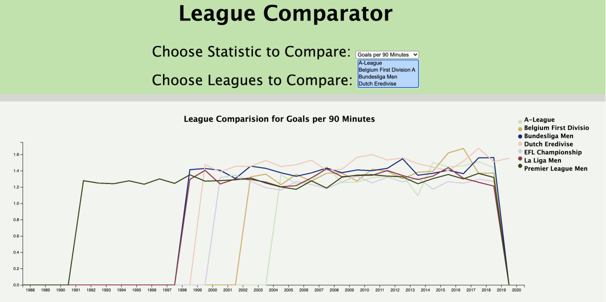
They could then find other prominent women players like Alexandra Popp or Anja Mittag who play in Europe, or one of the best male players in the world like Neymar. Hovering over Neymar's node then brings up player information that then guides further education:



Player: Neymar
Gender: M
Position: MF, FW
Seasons Played: 2013, 2014, 2015, 2016, 2017, 2018, 2019
Leagues Played In: La Liga Men, Ligue 1 Men, FIFA World Cup
Teams Played For: Barcelona, Paris S-G, br Brazil

Users could then search for Neymar's highlights at the FIFA World Cup with Brazil, or double-click on Neymar's node to see which players were similar to Neymar.

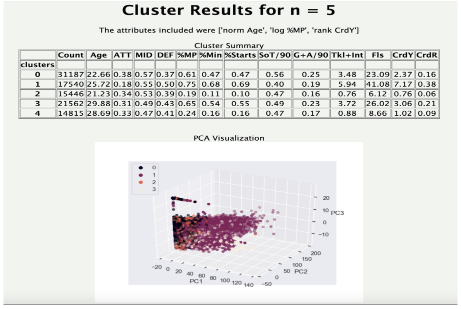
6.2.2 League Comparator : Our second tool can be used to compare various leagues across multiple statistics like Goals and Assists per 90 min, Fouls and Yellow Cards. This will assist users in understanding the contexts of different leagues, which could stimulate interest in new leagues or develop a better understanding of how players will perform in different leagues. This tool also serves as the basis for normalization of player performances based on league style while computing player similarity.


League Comparator

Choose Statistic to Compare: Goals per 90 Minutes

Choose Leagues to Compare: A-League, Belgium First Division A, Bundesliga Men, Dutch Eredivise

6.2.3 Clustering Tool : Finally, we built a tool to allow users to re-cluster the player data based on the attributes they are interested in. This is partly an interesting way for fans to group and classify players in ways that are interesting to them, and view the cluster-wide statistical profiles. A major usage for this tool, however, would be for coaches and general managers for soccer teams to group players on the attributes the team considers most important, and identify potential acquisitions accordingly. For example, adding filters for age and position, and then clustering by goals, shots on target, and assists, a general manager could identify a subset of young, promising players that score goals at a high rate.



Cluster Results for n = 5

The attributes included were ['norm Age', 'log %MP', 'rank CrdY']

Cluster Summary

	Count	Age	ATT	MID	DEF	%MP	%Min	%Starts	SoT/90	G+A/90	Tkl+Int	Fls	CrdY	CrdR
clusters														
0	31187	22.66	0.38	0.57	0.37	0.61	0.47	0.47	0.56	0.25	3.48	23.09	2.37	0.16
1	17540	25.72	0.18	0.55	0.50	0.75	0.68	0.69	0.40	0.19	5.94	41.08	7.17	0.38
2	15446	21.23	0.34	0.53	0.39	0.19	0.11	0.10	0.47	0.16	0.76	6.12	0.76	0.06
3	21562	29.88	0.31	0.49	0.43	0.65	0.54	0.55	0.49	0.23	3.72	26.02	3.06	0.21
4	14815	28.69	0.33	0.47	0.41	0.24	0.16	0.16	0.47	0.17	0.88	8.66	1.02	0.09

PCA Visualization

6.2.4 User Testing : Due to COVID-19, we were unable to conduct widespread user testing to refine our project. However, from the select user evaluations we were able to conduct, the following takeaways were most important, especially in guiding future iterations of this analysis:

- The Player Comparator and League Comparator tabs were very intuitive and easy to use, and the visual scheme across the whole tool was very appealing.
- When hovering over a node in the Player Comparator tab, it would have been cool to see the player's statistical profile in addition to basic biographic information.
- The clustering tab output was very interesting. It was a little difficult to choose which attributes to include (and what normalization to use), so including a default value might help.

7 Conclusion

Through our interactions with the statistical profiles of soccer players on a season-by-season basis, we were able to draw clear comparisons between players not currently thought of as similar. Our emphasis on pairwise-distance similarity allowed us to compare, for example, a women's soccer star striker like Alex Morgan to a men's soccer star winger like Neymar, or players in lesser-know leagues like Korea to famous stars in Europe.

There were some areas of improvement for future analyses, however:

(1) We were unable to use many of the attributes we scraped, such as Expected Goals, as these "newer" statistics have just recently begun to be tracked.
(2) Data on passing was also sparse, and likely affected our clustering model and similarity graph's ability to accurately compare midfielder "passing maestros".
(3) We noticed that our similarity graph still largely paired women together, despite womens' seasons representing only 4% of the data; it would be interesting to investigate why that was.
(4) We would like to add a way for users to view and ideally interact with a player's statistical profile on the Player Comparator tool, thus providing more information and context for the player comparisons.

Ultimately, we think we were largely successful in creating a way for soccer fans to invest in new leagues and players. We hope to build on this work to raise interest in women's soccer, and make soccer in general more accessible to all.

8 Work Division

All team members contributed similar effort.

9 Bibliography

References

[1] Serhat Emre Akhanli. 2019. *Distance construction and clustering of football player performance data.* Ph.D. Dissertation. University College London.
[2] Alistair C.Bruce. Johnnie E.V.Johnson Anastasios Oikonomidis. 2015. Does transparency imply efficiency? The case of the European soccer betting market. *Economics Letters* 128 (March 2015), 59–61. https://doi.org/10.1016/j.econlet.2015.01.015
[3] Michael Armanious. 2019. *Men's U-Sports Basketball Analysis.* https://bookdown.org/michael_arman7/Basketball-Analysis/
[4] Ingvild Merete Aksdal Arve Vorland Pedersen and Ragna Stalsberg. 2019. Scaling Demands of Soccer According to Anthropometric and Physiological Sex Differences: A Fairer Comparison of Men's and Women's Soccer. *Frontiers in Psychology* 10 (April 2019). https://doi.org/10.3389/fpsyg.2019.00762
[5] Pedro Henrique Moreira da Silva, Júlio Garganta, José Maia, and Pedro M. Santos. 2012. Tracking Performance in Football – An Example Using Goal Scoring Data. *The Open Sports Sciences Journal* 5 (2012), 181–187.
[6] J. Duch, Joshua S. Waitzman, and L. Amaral. 2010. Quantifying the Performance of Individual Players in a Team Activity. *PLoS ONE* 5 (2010).

[7] J. Li, Y. Deng, Z. Zeng, J. Liu, and Q. Hu. 2012. Re-evaluation of Individual Performance in the 2012 UEFA Euro Cup Tournament. *2012 Eighth International Conference on Semantics, Knowledge and Grids* (2012), 283–284.

[8] Heather M. O'Neill. 2014. Do Hitters Boost Their Performance During Their Contract Years? *Fall 2014 Baseball Research Journal* (2014).

[9] Magni Mohr Julen Castellano Anna Wilkie Paul S Bradley, Alexandre Dellal. 2014. Gender differences in match performance characteristics of soccer players competing in the UEFA Champions League. *Human Movement Science* 33 (Feb 2014). https://doi.org/10.1016/j.humov.2013.07.024

[10] Chen Dai Hongyou Liu Qing Yi, Ryan Groom and Miguel Ángel Gómez Ruano. 2019. Differences in Technical Performance of Players From 'The Big Five' European Football Leagues in the UEFA Champions League. *Frontiers in Psychology* 02 (Dec 2019). https://doi.org/10.3389/fpsyg.2019.02738

[11] Ryan M. Sapp, Espen E. Spangenburg, and James M. Hagberg. 2018. Trends in aggressive play and refereeing among the top five European soccer leagues. *Journal of Sports Sciences* 36, 12 (2018), 1346–1354. https://doi.org/10.1080/02640414.2017.1377911

[12] Silvia Chiusano Yingying LI and Ing. Vincenzo D'Elia . 2010. Modeling Athlete Performance Using Clustering Techniques. *International Symposium on Electronic Commerce and Security* (2010), 169–171. https://doi.org/10.1.1.404.1235