The Winning Formula in European Soccer

By

Nicholas Pappacena

August 26, 2021

MSDS 430: Python for Data Science

Introduction

Soccer has been my passion since the age of three. There is no greater feeling than getting one's feet under a soccer ball, running through the grass, dribbling by players, and scoring a goal that is then celebrated with teammates. In fact, data suggests that I'm not alone in this sentiment; with an estimated global fanbase of around 3.5 billon people, soccer is undoubtably the world's most popular game (Sourav 2021). As a former division 1 collegiate soccer athlete, I also understand the tactical side of the game at a high level—I like to believe at least. While my playing days are over, I still closely follow both the game's latest trends as well as the tactics implemented by coaches to establish a team's playing philosophy.

As can be expected in any competitive sport, there are winners and losers in every game. As such, the goal of my research project is not only to understand the key factors that determine which teams win and lose, but also why. Since some of the best soccer leagues in the world are located in Europe, I have focused my analysis on the following five European leagues during the period of 2017 to 2021: (1) the Premier League (England); (2) La Liga (Spain); (3) Serie A (Italy); (4) Ligue 1 (France); and (5) Bundesliga (Germany).

I was able to source all the necessary data for these leagues from FBref, which is a comprehensive website that contains soccer statistics from all over the world. Within this site, each individual league has its own landing page with multiple data tables available for analysis. This feature made the initial data collection process straightforward as all five leagues' data was acquired from a single source. Here, I focused on capturing in-game Key Performance Indicators (KPIs) that measure the following categories: (1) goalkeeping stats; (2) passing stats; (3) shooting stats; and (4) defending stats.

It goes without saying that the number of goals a team scores should have a direct correlation to such team's overall success. Similarly, teams that have the most goals scored against them will likely be regarded as having lower success rates. Excluding these highly obvious KPIs, I predict that teams that complete a high number of passes will be more successful than teams that complete a low number of passes. Moreover, I also assume that teams that demonstrate a higher number of shots saved by their goalkeeper will be more successful compared to teams that have a lower number of shots saved by their goalkeeper.

Data Preparation and Analysis

Before I dive into the data preparation process, I will provide some context as to how the European soccer leagues are organized. Every season, there are 20 teams in each of the following leagues: the Premier League (England), La Liga (Spain), Serie A (Italy), and Ligue 1 (France). The Bundesliga (Germany) is the only league that has 18 teams. In all the leagues, each team plays against all the other teams in its respective league twice, once at home and once away. This means that teams in the first four leagues will play 38 games each season (19 home, 19 away), while teams in the Bundesliga will only play 34 games each season (17 home, 17 away).

The team that wins the league is the team that accumulates the most points throughout a given season. Every win gives a team 3 points, every tie gives a team 1 point, and every loss gives a team 0 points. This is important to understand because my analysis measures team success through a derived metric called Points Per Game (PPG), which calculates the total number of points a team earned divided by the total number of games that team played. The reason I am not using total points as the measure for success is because this puts teams in the Bundesliga at a disadvantage since they play four less games a season than teams in the other leagues. Also, the

2020/21 Ligue 1 season in France was suspended early due to concerns over COVID-19. To have a consistent comparison, the PPG metric neutralizes these impacts.

To prepare my data, I first collected all relevant KPIs from FBref. Figure 1 shows all the metrics that were analyzed and provides a description for each metric:

Field Name	Description / Calculation
Squad	Name of team
League	The name of the league the squad plays in
Season	The calendar year the squad is playing each league
Points Per Game	Total Points / Total Games Played
Saves	Goalkeeper saves
Save %	Percentage of shots saved by a goalkeeper that are on target
Shots	Shots taken by squad
Shots on Target %	Percentage of shots that are on goal divided by total shots taken
Passes Dist_Total	Total distance, in yards, that completed passes have traveled
Passes Comp_Short	Passes completed within 15 yards
Passes Comp_Med	Passes completed between 16 and 30 yards
Passes Comp_Long	Passes completed over 30 yards
Key Passes	Passes that lead directly to a shot taken
Passes Comp Final 3rd	Completed passes that enter the final third of the soccer field closest to the opposing team's goal
Comp Crosses	Completed crosses into the opposing team's penalty box
Tackles in Def 3rd	Tackles made in the beginning third of the soccer field closest to the team's own goal
Tackles in Mid 3rd	Tackles made in the middle of the field
Tackles in Att 3rd	Tackles made in the final third of the soccer field closest to the opposing team's goal
Shots Blocked	Number of shots blocked by a field player (excluding the goalie)
Passes Blocked	Number of passes blocked by a field player (excluding the goalie)
Tackle Win %	Total tackles won divided by total tackles made

Figure 1: Data Dictionary

The metrics identified in Figure 1 have been gathered for every team across all leagues. As mentioned in the introduction, four calendar seasons worth of league data were analyzed: (1) the 2017/18 calendar season, (2) the 2018/19 calendar season, (3) the 2019/20 calendar season, and (4) the 2020/21 calendar season. A total of 98 combined teams played across all leagues for every calendar season. This implies that my final data set will consist of all the teams that played in the five leagues, across four different seasons, giving me a total population size of 392 teams.[1] With the data aggregated, cleansed, and prepared, the first part of my analysis focused on identifying existing correlations between all the metrics included in the data dictionary from Figure 1.

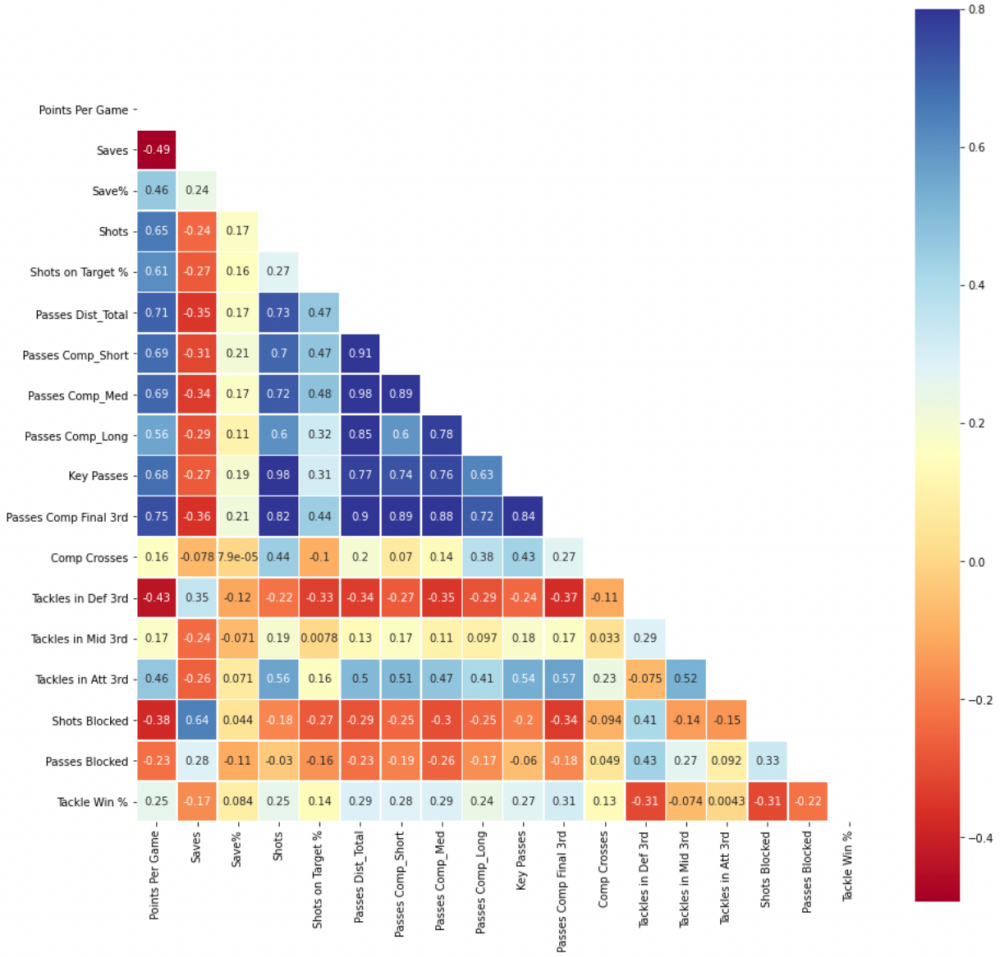

Figure 2: Correlation Half-Matrix

[1] Several of these teams will be repeated in the list because they will be recognized as a new team for every season played in the previously mentioned leagues.

Figure 2 shows that the strongest positive correlations exist among the team passing metrics, while the strongest negative correlations exist between the team tackling metrics and the team passing metrics. While informative, these correlations do not provide much value alone. The most significant correlation from Figure 2 that will be analyzed further is the relationship between the points per game metric and the other variables. Here, we start to learn about the individual in-game actions that impact team success across the European soccer leagues.

Passes Comp Final 3rd	0.746618
Passes Dist_Total	0.705325
Passes Comp_Med	0.694950
Passes Comp_Short	0.687334
Key Passes	0.680521

Figure 3a: Highest correlations to PPG

Comp Crosses	0.159660
Passes Blocked	-0.226673
Shots Blocked	-0.380940
Tackles in Def 3rd	-0.433981
Saves	-0.493437

Figure 3b: Lowest correlations to PPG

Figures 3a and 3b show the five highest correlations and five lowest correlations that exist between points per game and the other metrics. Figure 3a confirms one of my initial assumptions; the strongest positive correlations that contribute to team success, or points per game, are the in-game team passing metrics. More specifically, the more passes that a team completes in the opposition's final third of the field is the strongest indicator to determine the amount of points a team will earn in a game. Essentially, the higher up the soccer field a team tries to play, the better chance that team has to be successful.

However, Figure 3b rejects my second prediction that more shots saved by a team's goalkeeper positively correlates to team success. In fact, shots saved by a team's goalkeeper is the strongest negative correlation that impacts a team's ability to earn points in a game, which seems a bit counter intuitive. I find it surprising how defensive actions in general, such as saves, tackles, and blocks, all show negative correlations that impact a team's success.

Shocked that the data suggests that teams who make more defensive actions are less likely to be successful than teams with less defensive actions per game, I analyzed the 10 best and 10 worst teams—based on their average points per game earned over the past four seasons—to understand how the correlations identified in Figures 3a and 3b match up against real squad data.

Squad	League	Points Per Game	Comp Pass Final 3rd / Game	Pass Dist_Total / Game	Key Passes / Game	Saves / Game	Tackles in Def 3rd / Game	Shots Blocked / Game
Manchester City	Premier League	2.4	51.19	12,169.32	13.1	1.65	6.12	2.05
Paris S-G	Ligue 1	2.38	46.56	10,543.26	11.46	2.47	7.32	2.69
Bayern Munich	Bundesliga	2.37	50.13	11,711.71	13.49	2.1	6.12	2.43
Juventus	Serie A	2.28	38.49	10,161.77	11.93	2.32	7.16	3.26
Barcelona	La Liga	2.24	47.96	11,170.8	10.88	2.47	6.16	2.57
Liverpool	Premier League	2.24	45.01	10,918.31	11.65	2.02	6.66	2.12
Real Madrid	La Liga	2.07	41.67	10,187.09	12.32	2.47	7.3	2.37
Inter	Serie A	2.07	36.44	9,405.71	11.72	2.36	7.38	3.34
Atlético Madrid	La Liga	2.04	30.89	7,352.97	8.59	2.61	9.52	2.95
Napoli	Serie A	2.03	45.07	10,339.01	13.8	2.23	7.04	2.7

Figure 4a: Top 10 teams PPG earned across Europe between 2017-2021

Squad	League	Points Per Game	Comp Pass Final 3rd / Game	Pass Dist_Total / Game	Key Passes / Game	Saves / Game	Tackles in Def 3rd / Game	Shots Blocked / Game
Málaga	La Liga	0.53	24.05	5,902.42	7.32	2.82	8.29	2.58
Norwich City	Premier League	0.55	23.18	7,112.82	7.87	3.39	11.03	5.18
Nürnberg	Bundesliga	0.56	22.44	6,494.06	6.74	3.15	10.06	3.59
Las Palmas	La Liga	0.58	27.21	7,507.53	7.66	3.82	7.55	3.05
Paderborn 07	Bundesliga	0.59	23.26	7,244.03	8.65	3.26	11.35	3.94
Brescia	Serie A	0.66	20.95	5,794.89	7.45	4.16	8.13	5.16
Frosinone	Serie A	0.66	20.74	6,085.37	8.16	3.84	8.84	4.11
Huddersfield	Premier League	0.7	24.84	6,642.47	6.91	2.7	9.8	3.51
Fulham	Premier League	0.71	26.86	7,801.04	8.79	3.45	8.63	3.64
Benevento	Serie A	0.71	22.25	6,743.26	8.3	3.13	9.14	3.78

Figure 4b: Bottom 10 teams PPG earned across Europe between 2017-2021

The teams that earned the most points per game across the European soccer leagues consistently recorded lower defensive actions per game and higher passing numbers per game compared to the teams with the lowest points earned per game. Analyzing the combined defensive actions from Figure 4a, the 10 best teams averaged 2.27 goalkeeper saves per game, 7.08 defensive tackles made in the defending third per game, and 2.65 blocked shots per game. Comparatively, Figure 4b shows that the 10 worst teams averaged 3.37 goalkeeper saves per game, 9.28 defensive tackles made in the defending third per game, and 3.85 shots blocked per game; all significantly higher than the numbers reported by the best teams.

From an offensive perspective, the top 10 teams averaged 43.34 completed passes in the opposition's final third per game compared to only 23.58 completed passes in the opposition's final third per game by the 10 worst teams. With an average difference of about 20 completed passes per game in the opposing team's final third, compounded at a full season's length, it is quantifiably evident why the worst teams in Europe struggle. When a soccer team is consistently forced to defend, it will generally lack possession of the ball. A team that's unable to maintain possession of the ball will have less opportunities to complete more passes, play higher up the pitch, and ultimately score more goals—all factors that contribute positively to team success.

Thus, when framed in the proper context, it has been proven that teams who regularly require their goalkeeper to save more shots will most likely not be dictating the rhythm and flow of the game. Playing defensive soccer might work for certain teams in situations where luck is on their side, but four years of data across the five best European soccer leagues suggest that building an offensive game plan that encourages both possession of the ball and increased passing combinations is the best strategy to sustain long-term success—a trait that all the best teams have.

Conclusion

My analysis sheds light into how soccer teams and coaches can plan for success in the future. Based on the results, my recommendation to any soccer coach would be to implement and build a soccer philosophy that rewards offensive playing styles over defensive structure. Figure 2 shows the relationships that exist between all metrics in this analysis. Prioritizing the metrics that have the highest correlation associated with team success is the best way to build a game plan.

While tactically and theoretically astute, my analysis excludes a key practical component that impacts any team's ability to be successful—the team's available financial resources. Within this premise, teams with high financial resources have the ability to buy some of the greatest players, which has traditionally been the most reliable way to improve team success. Yet, today, due to the economic impact of COVID-19, most teams are financially strapped. This financial strain, however, may offer soccer clubs' management and ownership the opportunity to think creatively to gain a competitive advantage over other clubs that still operate under traditional financial soccer models.

There is momentum for clubs to invest in analytical capabilities that will introduce new, innovative models for sustained financial and on-field performance success. By leveraging analytics as the key driver for management decision making, clubs would be able to hire coaches that encourage similar soccer philosophies to the one I recommend in my analysis. Moreover, analytics can be used to curate new scouting techniques that identify lesser-known players with incredible on-field potential who are not valued highly on the open market. In sum, achieving success in soccer is now becoming a science that will allow analytically driven teams to separate from clubs who still compete on intuition.

References

FBref. Accessed August 26, 2021. https://www.fbref.com/en/.

Sourav. “Top 10 Most Popular Sports in the World.” Last modified June 15, 2021. https://sportsshow.net/top-10-most-popular-sports-in-the-world/.