Data Analysis Report – Ekstraklasa 2024/25

Author: Dawid Jasiński

Date: 13.02.2026

Tools: Python (Pandas, Matplotlib, Seaborn, Scikit-Learn), Excel

Data Source: FBref.com

Executive Summary

This report presents an analysis of the data for Ekstraklasa teams in the 2024/2025 season, focusing on identifying the factors that determine the number of points gained and on characterizing the statistical profiles of the teams. The analysis reveals a clear division between dominant teams, mid-table sides and teams struggling with structural problems. The key mechanism explaining the differences in points is attacking efficiency, in particular shot quality and the ability to generate real goal-scoring opportunities. Top teams such as Lech and Raków combine a high number of shots on target with above-average finishing efficiency and defensive stability. Teams in the relegation zone, on the other hand, are characterized by limited attacking intensity and a high number of defensive errors, which significantly hinders their ability to score points on a regular basis.

The exploratory analysis confirms that ball possession plays an auxiliary role and is not the decisive factor of success. Statistical relationships indicate that more important are those elements of play that directly translate into goals, such as the number of shots on target, goals per shot and goals per shot on target. High correlation coefficients between these variables and the number of points confirm that success in Ekstraklasa is primarily a function of efficiency rather than domination in possession. Defensive stability is equally important, which is particularly well illustrated by Raków, which, despite a moderate volume of attacking actions, achieved one of the highest positions in the table thanks to an outstandingly functioning defense.

The linear regression model used in the final part of the analysis provided additional quantitative confirmation of the observations from the exploratory stage. Variables related to shot quality and the number of shots on target per match turned out to be the strongest predictors of the number of points, while the number of goals conceded showed one of the strongest negative effects. The model achieved a high level of fit, which made it possible to estimate the expected number of points for each team. A comparative analysis showed that teams such as Korona and Motor performed above the level suggested by their statistics, while

Śląsk, Radomiak and Puszcza achieved results significantly lower than predicted. These differences indicate the influence of elements not captured by basic statistics, such as efficiency in key moments of the match, tactical organization or mental stability.

In summary, the analysis of the 2024/2025 Ekstraklasa season confirms that the final results of the teams are determined primarily by the ability to generate and convert high-quality goal-scoring opportunities and to limit defensive errors. The best teams combine attacking intensity, finishing efficiency and defensive stability, while teams from the bottom of the table struggle both with a lack of chance creation and defensive problems. The modeling results highlight the importance of efficiency and structural coherence and indicate that a significant part of the differences in points is the result of the quality of decisions made in the key moments of the game. These conclusions indicate directions in which teams can develop their strategies, and statistical analysis is a useful tool for assessing and optimizing their potential in subsequent stages of competition.

INTRODUCTION

The purpose of this report is to conduct a comprehensive statistical analysis of Ekstraklasa teams in the 2024/2025 season using data from the FBref.com service. The analysis aims to identify which aspects of play—offensive as well as defensive—have the greatest impact on the number of points earned and therefore on the final league table. The project covers the full scope of analytical work typical for data analyst roles: from data acquisition and cleansing, through exploration and visualization, to predictive modeling and the formulation of actionable insights.

The foundation of the report is a merged dataset containing classic match statistics (goals scored, goals conceded, goal difference), indicators describing style of play (possession, number of shots, number of shots on target per 90 minutes), and quality-oriented metrics such as goals per shot (G/Sh) and goals per shot on target (G/SoT). The data were sourced from multiple FBref sheets and required cleaning and structural unification. Team names were stripped of HTML fragments, text columns were converted to numerical types, and consistency checks were applied to ensure correctness.

The exploratory data analysis forms the central part of the study. It provides a multi-layered view of the league: first through basic statistics (possession, shots, goals), then through relationships between variables, and finally through detailed team profiles. Key visualizations—including “possession vs points”, “shot efficiency vs points”, and the correlation matrix—make it possible to identify which features genuinely influence results and which, despite being often highlighted in media, have limited explanatory power. The data clearly indicate that Ekstraklasa places more emphasis on shot quality and finishing efficiency than on possession as such.

A more detailed perspective is offered through the comparison of statistical profiles of top-four teams, mid-table sides, and bottom teams. Radar charts and detailed numerical summaries show that top teams achieve superiority not only through volume but also through quality: they create more chances, take more effective shots, and maintain significantly better goal differences. Bottom teams, in contrast, struggle with low offensive intensity and structural defensive issues leading to frequent concession of goals.

A linear regression model was applied to determine the extent to which selected variables explain the number of points. The model, built on standardized data, achieved a very high R^2 value, indicating that variables describing finishing efficiency and defensive performance strongly explain actual outcomes. Expected points were also calculated, allowing identification of teams performing above or below their statistical potential.

The report uses standard analytical tools widely applied in modern data analytics. Python (pandas, numpy, matplotlib, seaborn, scikit-learn) was used throughout the data preparation, exploratory analysis, and modeling process. The entire work was conducted in Jupyter Notebook, ensuring transparency, reproducibility, and a structured analytical workflow.

Although the analysis is extensive, its results should be understood as descriptive rather than predictive. The dataset covers only one season and includes just 18 teams, which limits the applicability of long-term forecasting models. Nonetheless, the findings provide a clear picture of the elements that drive success in Ekstraklasa and highlight which teams perform above or below expectations based on their statistical profile.

DATA CLEANING & PREPARATION

The data preparation process was a key stage of the entire project, as the raw data downloaded from FBref were spread across several sheets, had a non-uniform format and some of the variables required conversion and additional processing. The data covered six different sets of team statistics, such as the league table, standard statistics, shooting data, goalkeeping statistics, playing time and additional match parameters. To enable joint analysis, it was necessary to create a single, coherent dataset containing all key information.

The first step in preparing the data was to load all sheets and review their structure. Data from the league table required basic cleaning, including removal of empty records and conversion of columns from text types to numeric ones; this applied, among others, to values such as points, goals, matches or points per game. The most important element of the initial cleaning was unifying team names. In many sheets, team names included HTML links stored in square brackets, which prevented direct joins. Therefore, regular expressions were used to extract the proper name so that the same team name format would appear in all sheets.

The next step was to reduce unnecessary columns and select only those variables that were relevant for the later analysis. After an initial review of the data, key metrics were selected, such as the number of players, average age of the team, ball possession, number of goals, number of assists and total attacking output (G+A). Similarly, from the shooting statistics sheet, variables describing volume (shots per 90 minutes, shots on target per 90 minutes) as well as quality parameters, including goals per shot (G/Sh) and goals per shot on target (G/SoT), were selected. This helped reduce the size of the dataset and focus the analysis on those indicators that can actually explain team effectiveness.

All selected sheets were then merged into a single dataset using a join operation on the cleaned team name. As a result, a complete set was obtained, including 18 teams and 28 harmonized variables describing their statistical profiles. The following code snippet shows the process of merging the data:

```
df = df_table.merge(std_small, on='Squad', how='left', suffixes=('', '_std'))
df = df.merge(shoot_small, on='Squad', how='left', suffixes=('', '_shoot'))
```

After merging the sheets, a standard data quality validation was carried out, checking for missing values, distributions of key statistics and the consistency of basic measures such as the total number of goals scored and conceded. For example, the number of goals in the league table was compared with the number of goals recorded in the standard statistics sheets to ensure that the data covered the same set of matches. The results of these checks confirmed the consistency and correctness of the data, which made it possible to use them further without the need to fill gaps or correct values.

The prepared dataset became the basis for exploratory and modeling analyses presented in the subsequent parts of the report. Its structure allowed not only the analysis of individual indicators, but also the study of relationships between them, the comparison of team

profiles and the construction of a statistical model predicting the number of points. Thanks to the cleaning and transformation steps, the analysis could be conducted in a reliable and methodologically sound way.

EXPLORATORY DATA ANALYSIS

Exploratory Data Analysis (EDA) is the central part of the report. Its aim is to understand the structure of the league, identify the most important statistical relationships and assess how different aspects of play influenced the results of Ekstraklasa teams in the 2024/2025 season. Within EDA, both classic match metrics and more advanced efficiency indicators were analyzed, which made it possible to go beyond simple observations based on the league table. The analysis includes an overview of basic rankings, the study of relationships between key variables and the comparison of profiles of teams with different sporting outcomes.

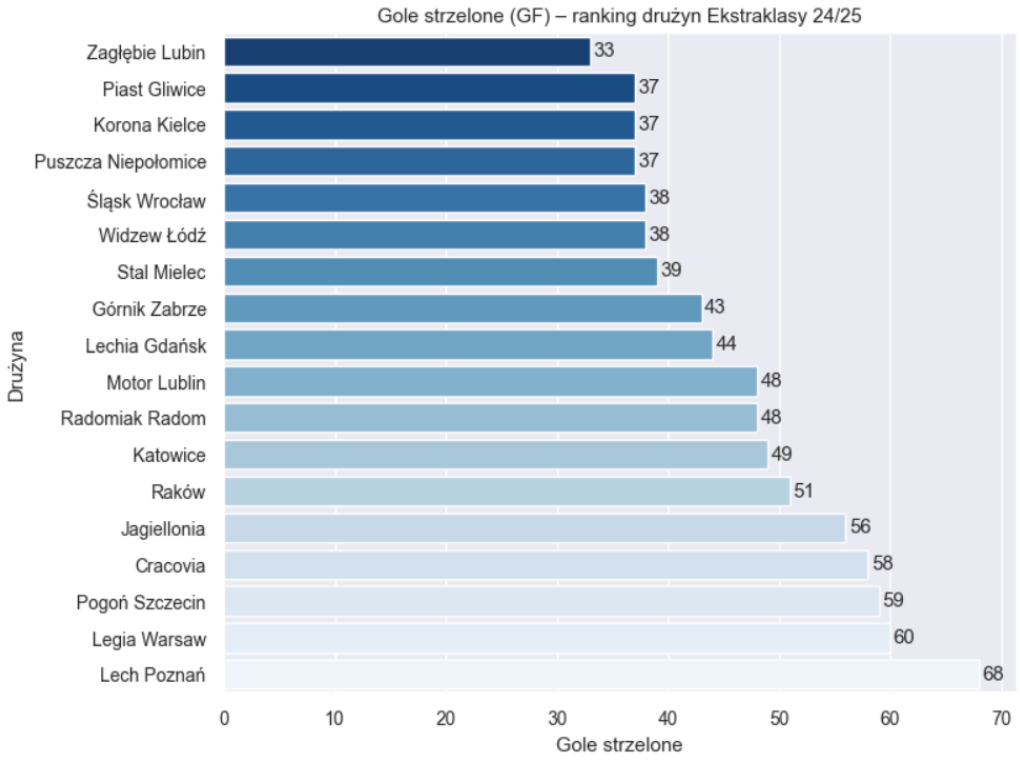
League Overview

The first step was to analyze basic attacking and defensive statistics such as the number of goals scored and conceded, goal difference and average ball possession. Already at this stage, clear differences appear between dominant teams and teams struggling to maintain stability in their play. Lech Poznań stands out as the most complete team, combining high possession with a large number of shots and above-average finishing efficiency, which translates into the highest number of goals scored in the league and one of the best goal differences. Raków, on the other hand, represents a different way of achieving results – despite lower possession than Lech, it has the best-functioning defense in the league, confirmed by the smallest number of goals conceded. Its effectiveness stems more from stability and organization of play, especially in the defensive line.

At the other end of the spectrum are teams that struggled throughout the season. Puszcza, Stal Mielec and Zagłębie Lubin are characterized not only by a low number of goals scored, but also by poor offensive efficiency and very unstable defensive play. These teams show a combination of weak chance creation and high susceptibility to conceding goals, which prevents them from competing on equal footing with better organized opponents. Already at the level of basic statistics it is clear that the differences between top teams and those at the bottom of the table go beyond goal difference alone and extend to the overall quality of play, both in terms of maintaining possession and building attacks as well as reacting to threats in their own defensive third.

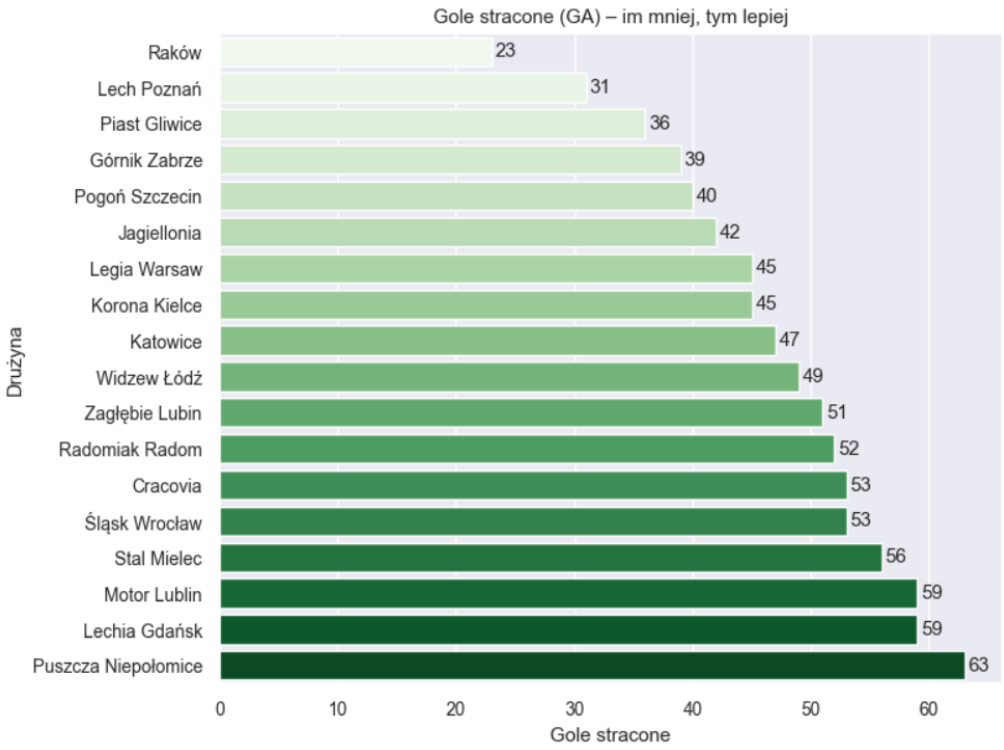
This initial analysis helps to understand that the 2024/2025 Ekstraklasa season is characterized by a clear dichotomy between dominant teams and those fighting for survival. Differences in efficiency, attacking intensity and defensive organization are visible even in the simplest metrics. The subsequent parts of the analysis expand these observations using more detailed statistical relationships, shedding light on why some teams were able to consistently build an advantage while others faced problems on many fronts.

Figure 1. Goals scored (GF)



Source: Own elaboration based on FBref data.

Figure 2. Goals conceded (GA)

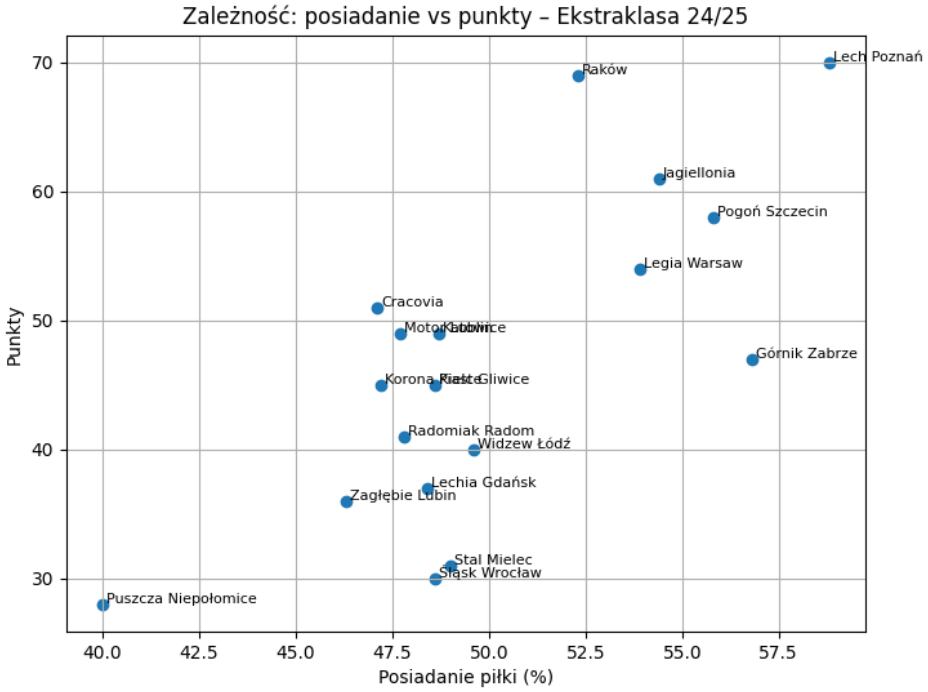


Source: Own elaboration based on FBref data.

Key Statistical Relationships

The next step was to examine the relationship between ball possession and the number of points earned. The scatter plot showed that although top-table teams tend to maintain higher levels of possession, this relationship is not strongly deterministic. Lech and Raków, located in the upper-right part of the chart, combine high possession with a high number of points, but Górnik—despite one of the highest possession levels—finishes much lower. On the other hand, Puszcza, which had the lowest possession in the league, earned by far the fewest points. This indicates that possession can play a supporting role but is not the key factor determining results.

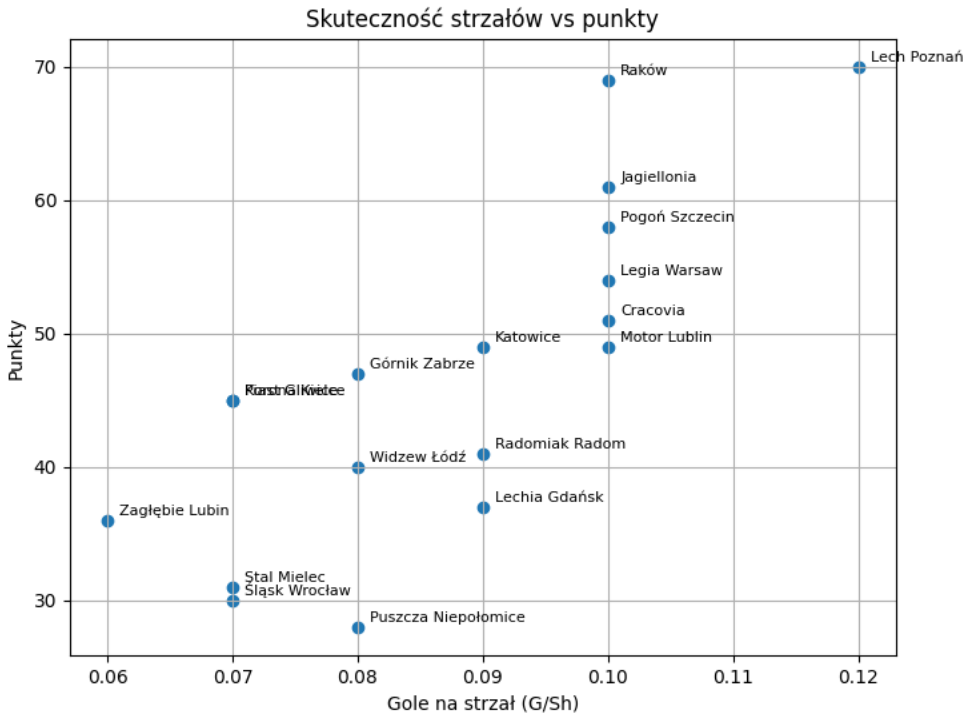
Figure 3. Relationship: possession vs points



Source: Own elaboration based on FBref data.

Far stronger relationships emerge when analyzing shot efficiency. The relationship between goals per shot (G/Sh) and points is almost linear: teams that need fewer shots to score a goal tend to occupy higher positions in the table. Lech, the only team with G/Sh around 0.12, is clearly ahead of the rest in both efficiency and points. At the opposite extreme is Zagłębie, with the lowest shot efficiency in the league, which helps explain its poor final position. These charts show that in Ekstraklasa success is driven primarily by attacking efficiency rather than sheer shot volume.

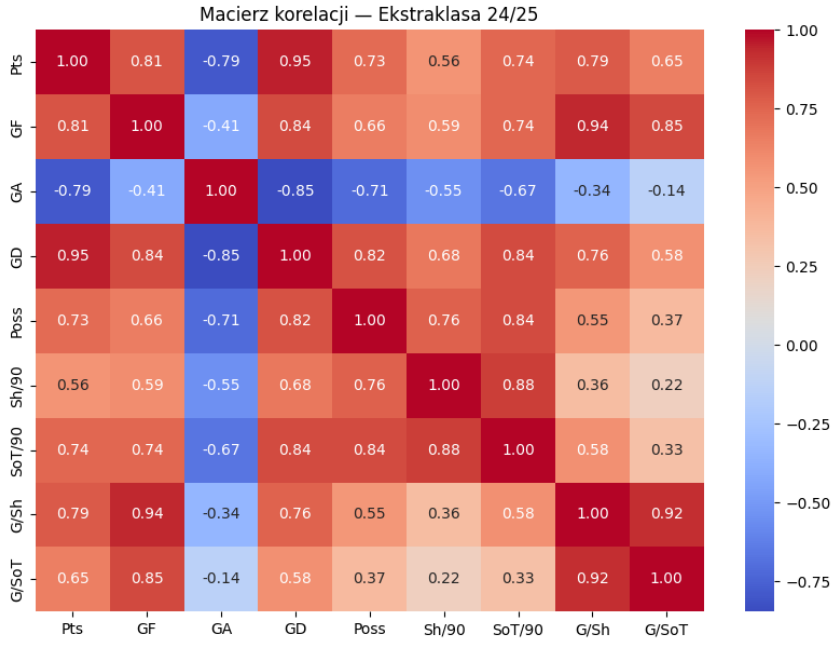
Figure 4. Shot efficiency vs points



Source: Own elaboration based on FBref data.

Additional confirmation comes from the correlation analysis. The correlation matrix shows that the strongest "statistical explanation" of points is provided by goal difference (r = 0.95), the number of goals scored (r = 0.81) and shot efficiency indicators (G/Sh = 0.79 and G/SoT = 0.65). Ball possession, although often highlighted in media analyses, shows only moderate correlation (r ≈ 0.73). In practice, this means that teams do not win matches by simply keeping the ball longer, but by being able to create and convert clear goal-scoring opportunities. Defense is equally important: the number of goals conceded is strongly negatively correlated with points (r = –0.79), confirming that limiting goals against is a fundamental success factor.

Figure 5. Correlation matrix



Source: Own elaboration based on FBref data.

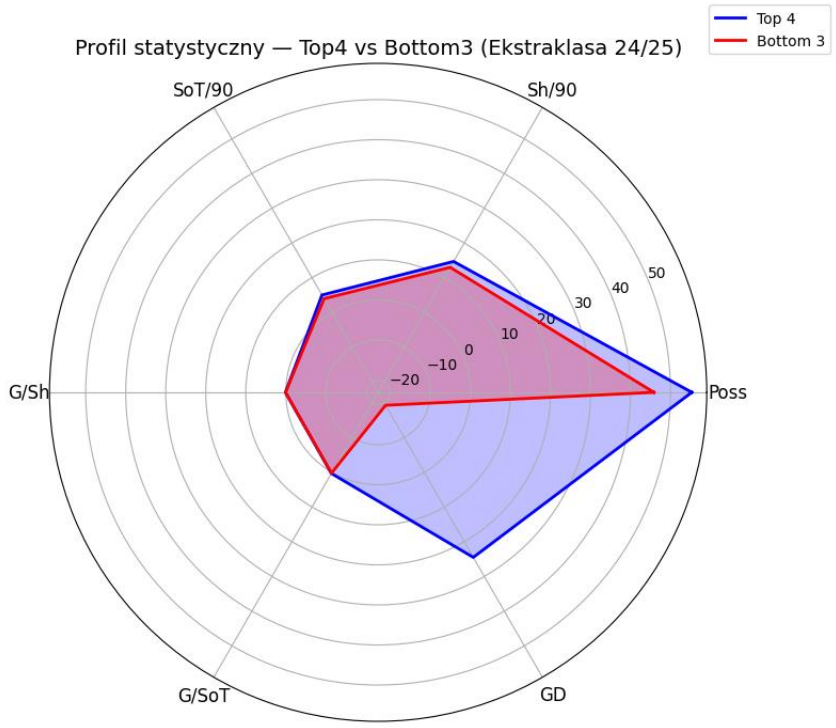
Team Profiles

A comparison of the statistical profiles of top-four teams and those in the relegation zone clearly shows differences in style and quality of play, both in attack and defense. Top teams display much greater attacking intensity: they keep the ball more often, control the flow of the game for longer periods and are frequently present in the opponent's half. This translates into a higher number of shots, more shots on target and more goal-scoring situations, which in turn lead to more goals scored and better goal differences.

Teams in the relegation zone lag behind not only in the number of chances created but also in defensive stability. Conceding many goals and frequent defensive errors mean that these teams rarely manage to impose their own style of play, and many of their attacking moves break down before they can generate real danger. Importantly, the differences between the groups concern not only the volume of actions but also their consequences: top-four teams can maintain structure both in attack and defense, while bottom teams concede goals after simple mistakes and struggle to build sustained advantages over the course of a match.

Interestingly, pure shot-conversion efficiency—measured as goals per shot or goals per shot on target—does not differ as dramatically between the groups as one might expect. This suggests that the key problem for teams in the relegation zone lies less in their finishing skills and more in the low number of chances created and the high number of defensive errors. Ultimately, these structural differences—insufficient attacking intensity and unstable defending—lead to the points gap visible at the end of the season and constitute the main factor separating title contenders from relegation candidates.

Figure 6. Statistical profile – Top4 vs Bottom3



Source: Own elaboration based on FBref data.

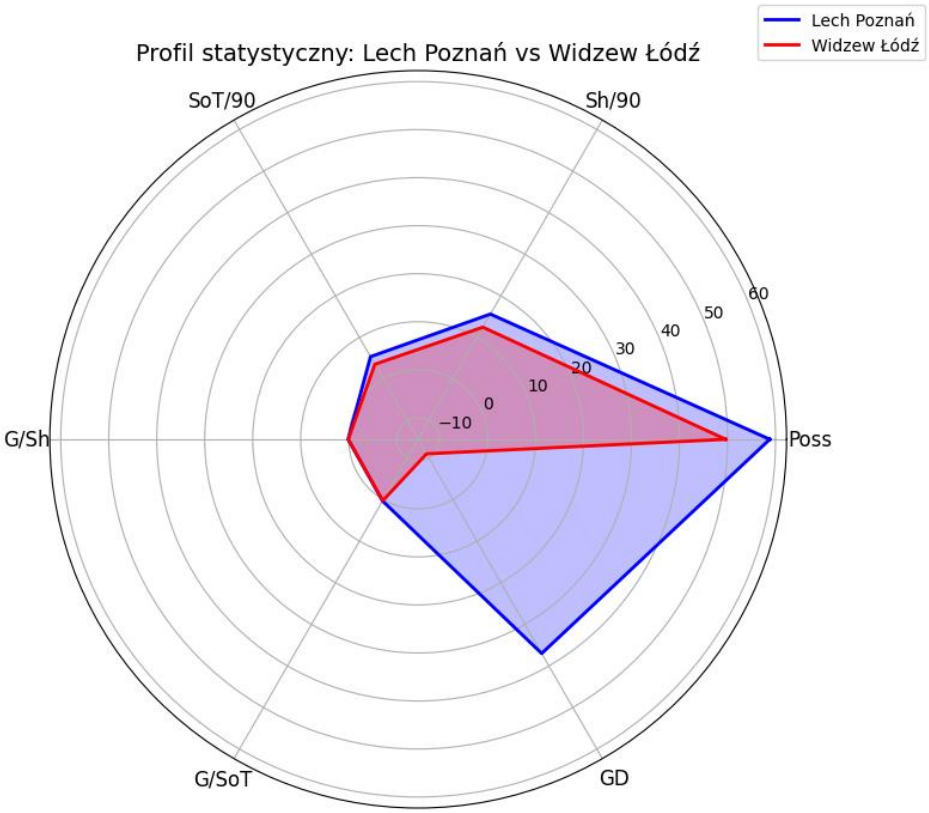
A further complement to the analysis is a more detailed case study of Lech Poznań and Widzew Łódź, which clearly illustrates the differences between a team fighting for the highest goals and one occupying a mid-table position. Lech presents the profile of a dominant team both in terms of playing style and effectiveness. It maintained one of the highest possession levels in the league, which translated into more attacking actions and a significantly higher volume of shots and shots on target. As a result, Lech was able to control the tempo of matches, and the high quality of the chances created, combined with very good finishing efficiency, led to a strong goal advantage and, ultimately, a high points total.

Widzew's profile differs from Lech's in almost every aspect. The team generated fewer shots, reached clean shooting positions less often, and its possession remained at a level typical of mid-table teams. Although Widzew's shot efficiency was not drastically below the league average, the low attacking intensity meant that the team had limited ability to build a goal advantage. Defensive performance was also a significant problem: Widzew conceded goals much more often than top teams, which further complicated point accumulation over the season.

Comparing these two teams on a radar chart reveals a clear asymmetry: Lech outperforms Widzew in almost every dimension—from possession level, through number and quality of shots, to goal difference. Widzew, by contrast, presents a more conservative and considerably less effective profile in key aspects of play. This case study clearly shows that the difference in points between the two teams does not stem from a single metric but from the cumulative

effect of multiple areas in which Lech consistently outperforms its rival. The comparison effectively summarizes the broader EDA findings: top teams combine high volume of chances, good shot quality and stable defending, whereas mid-table sides typically lack one or several key components.

Figure 7. Statistical profile: Lech Poznań vs Widzew Łódź



Source: Own elaboration based on FBref data.

Overall, the exploratory analysis demonstrates that Ekstraklasa 2024/2025 is a league in which success is determined primarily by attacking efficiency and defensive stability rather than the number of shots taken or sheer dominance in possession. These findings form the foundation for the next part of the report, which applies regression modeling to determine the exact weight of individual variables in explaining the number of points.

Predictive Modeling

In the final stage of the analysis, a linear regression model was applied to determine the extent to which selected statistical variables explain the number of points gained in the 2024/2025 season. The model was not intended to serve as a forecasting tool but as an interpretive one—its task was to indicate which aspects of play, among the previously analyzed parameters, are statistically the most important in the context of final results. The use of regression makes it possible to look at Ekstraklasa from a quantitative perspective: not only to notice differences between teams but also to understand how strong the impact of individual variables is on points when other factors are taken into account simultaneously.

Before building the model, the data were standardized to ensure comparability of all variables regardless of their units and scales. Standardization also makes it possible to interpret regression coefficients as a relative measure of the impact of each factor on the dependent variable, which in this case is crucial for drawing meaningful conclusions. The model included variables describing both the volume of attacking actions (such as shots per 90 minutes and shots on target per 90 minutes) and quality parameters (overall shot efficiency and shot-on-target efficiency), as well as ball possession and basic goal-related metrics.

The regression model demonstrated a very high degree of fit, with an R^2 value of around 0.95. This means that the selected variables explain approximately 95% of the variance in points between Ekstraklasa teams, which is exceptionally high for sports data, typically characterized by substantial variability and modeling difficulties. At the same time, such a high R^2 should be interpreted with caution, as the number of observations is small and the model describes a completed season, which naturally increases the degree of fit. Nevertheless, the results allow for accurate interpretation of the relationships between statistics and table positions.

The strongest predictor of the number of points turned out to be goals per shot on target (G/SoT). A high value of this parameter means that a team not only takes shots but does so in a deliberate and efficient manner, converting a large share of its shots on target into goals. The next most important variables were the number of shots on target per 90 minutes and overall shot efficiency (G/Sh). Their significance supports the conclusion that success in Ekstraklasa depends on the quality of attacking actions and the ability to finish chances. It is not the sheer number of shots that matters most, but their accuracy and effectiveness. Interestingly, traditional metrics such as total goals scored received lower coefficients because their impact is largely captured by higher-order variables that better describe the quality of the attacking process rather than its raw outcome.

Table 1. Linear regression results

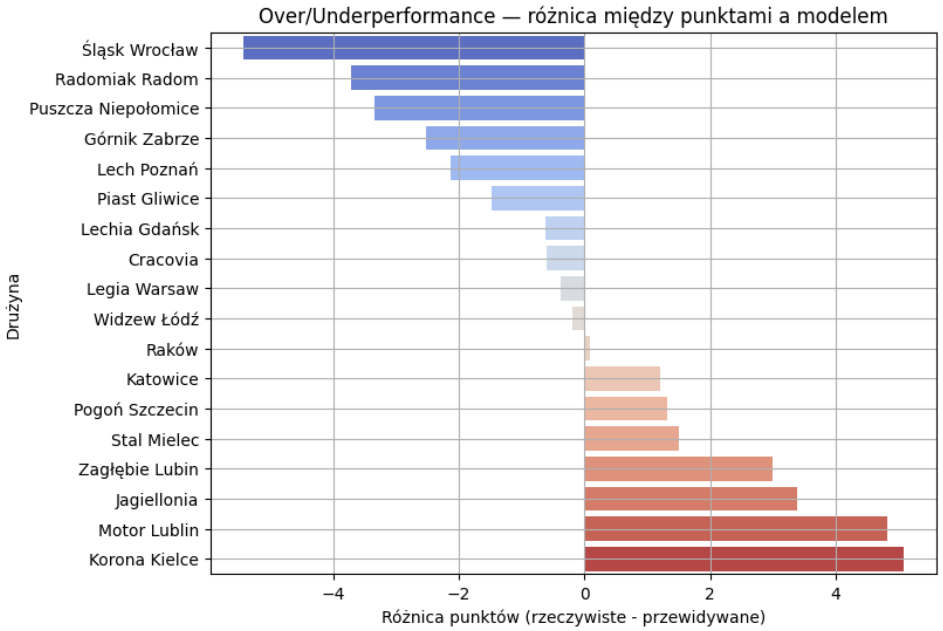
Feature	Coefficient
G/SoT	5.561082
SoT/90	3.439537
G/Sh	1.573352
Sh/90	-0.668302
GF	-0.869304
Poss	-1.316131
GA	-7.632773

Source: Own elaboration based on FBref data.

The number of goals conceded also showed a strong negative impact on points. A high value of this variable reduces a team's total points in a predictable and consistent way, confirming that a solid defense is a cornerstone of success. In the context of regression, this means that even teams with strong attacking organization and efficiency have limited ability to compensate for defensive weaknesses—if they concede many goals, the model immediately reflects this with a lower predicted points value.

After building the model, expected points were calculated for each team, which made it possible to identify sides performing above or below their statistical potential. The results showed that teams such as Korona and Motor Lublin gained more points than the model predicted based on their statistics, which may indicate an ability to take advantage of key match moments, individual brilliance of certain players or well-executed set pieces. Conversely, Śląsk, Radomiak and Puszcza achieved results significantly below expectations, suggesting issues with efficiency in phases decisive for points—such as match endings, one-on-one situations or defending a lead.

Figure 8. Over/Underperformance



Source: Own elaboration based on FBref data.

The difference between actual and expected points is a particularly interesting element of the analysis. It shows which teams were more effective than their statistical profile suggests and which underperformed relative to their potential. Teams with positive differences demonstrated higher efficiency in the key moments of the season, while those with negative differences may be struggling with tactical, mental or organizational issues not directly reflected in raw statistics.

The regression model and the expected-points analysis are an important part of the report, as they close the analytical section in a numerical and precise way. They form a logical complement to the observations from EDA, confirming that the foundations of success in Ekstraklasa are attacking efficiency, the intensity of chance creation and defensive stability, rather than the sheer number of shots or possession dominance. This made it possible to prepare a clear picture of which teams built their positions on solid statistical foundations and which deviated from the results suggested by their data, illustrating how valuable predictive analysis can be in assessing football performance.

Key Findings

The analysis of Ekstraklasa 2024/2025 data allows several key conclusions to be drawn regarding the mechanisms that determine league outcomes. The most important of these is the dominant role of attacking efficiency. Teams finishing high in the table were characterized not only by a larger number of shots but, above all, by higher quality and efficiency of those attempts. Relationships revealed in the correlation analysis and scatter plots clearly indicate that goals per shot and goals per shot on target are among the variables most strongly correlated with the number of points earned. This implies that success depends less on the sheer intensity of attacking play and more on the ability to convert that intensity into effective actions. For teams at the bottom of the table, there was a noticeable deficit in both the volume of attacking actions and defensive stability, resulting in greater vulnerability to conceding goals and difficulty building an advantage over the course of the season.

A second key finding concerns the clear role of defensive organization, with Raków as the best example. Its high position in the table was primarily the result of an exemplary defense that reduced the number of goals conceded and allowed the team to collect points even in matches with relatively few attacking opportunities. The regression model confirmed the importance of this factor by identifying goals conceded as one of the variables most strongly and negatively correlated with points. Finally, the expected-points analysis showed that some teams achieved results above or below the level suggested by their statistics. Teams such as Korona and Motor Lublin distinguished themselves through high efficiency in decisive moments of matches, whereas Śląsk, Radomiak and Puszcza delivered weaker results than their parameters would suggest. Taken together, these observations indicate that final league outcomes are the result of a combination of attacking efficiency, defensive organization and the ability to exploit key moments in matches, rather than being a simple function of the number of shots taken or ball-possession levels.

Recommendations

The conclusions from the analysis make it possible to formulate several practical recommendations regarding the development directions of Ekstraklasa teams, particularly those in the lower part of the table. The most important area requiring improvement is the quality of chance creation. Teams in the relegation zone do not differ dramatically in terms of pure shot-conversion efficiency, but generate far too few opportunities to score points consistently. It is therefore advisable for these teams to increase the intensity of their attacking play, both by speeding up transitions and by improving the quality of passing in the final third. In practice, this may mean adjusting the tempo of play, introducing more dynamic wide players or placing greater emphasis on patterns that more frequently lead to high-quality shooting positions.

A second crucial area is defensive organization. The data clearly show that the number of goals conceded has a very strong impact on points, and in many cases teams' problems did not stem from the number of shots faced but from a lack of stability and concentration in critical phases of matches. Improving defensive structure and increasing tactical discipline can help teams both limit goals conceded and better control match flow. Teams with defensive issues should analyze their matches in terms of organization during transitions and positioning in their own penalty area, as many goals conceded resulted from individual errors or a lack of cover.

Another recommendation is a broader use of analytics in match preparation and squad-management decisions. The expected-points analysis shows that some teams performed significantly below their statistical potential, which may indicate poor management of key phases of games, a lack of mental stability or problems with finishing in high-pressure situations. Teams with a negative balance between actual and expected points should focus on detailed analysis of the phases in which they lose control, and on working on mentality and precision in decisive moments, so as to minimize the cost of individual mistakes.

In summary, to compete effectively in Ekstraklasa, teams should aim to increase attacking intensity while maintaining defensive stability and optimizing the process of converting chances. The key to improving results is not merely increasing the number of shots or dominating possession, but achieving coherence between all phases of play and the ability to make high-quality decisions in moments that determine the outcome of matches. Implementing these recommendations can help teams improve their points efficiency and make better use of their potential in future seasons.