**GitHub Portfolio Link:** [My Portfolio (keaveneys.github.io)](https://keaveneys.github.io/)

**GitHub Repository Link (use if portfolio link is not clear/available):** [**BP0289088**](https://github.com/Keaveneys/BP0289088/tree/main)

# BPP Coursework Cover Sheet

Please use the table below as your cover sheet for the 1st page of the submission. The sheet should be before the cover/title page of your submission.

|  |  |
| --- | --- |
| **Programme** | BSc (Hons) Data Scientist |
| **Module name** | Data Science Professional Practice |
| **Schedule Term** | May 2024 |
| **Student Reference Number (SRN)** | BP0289088 |
| **Report/Assignment Title** | Public Dataset |
| **Date of Submission**  ***(Please attach the confirmation of any extension received)*** | 28/08/2024 |
|  | |
|  | |
| Declaration of Original Work:  I hereby declare that I have read and understood BPP’s regulations on plagiarism and that this is my original work, researched, undertaken, completed and submitted in accordance with the requirements of BPP School of Technology.  The word count, excluding contents table, bibliography and appendices, is 1438 words.  Student Reference Number: BP0289088 Date:28/08/2024 | |

[BPP Coursework Cover Sheet 1](#_Toc175650146)

[Glossary 3](#_Toc175650147)

[Executive Summary 4](#_Toc175650148)

[Introduction 4](#_Toc175650149)

[Data Engineering 5](#_Toc175650150)

[Analysis 8](#_Toc175650151)

[Visualisations 13](#_Toc175650152)

[Discussion 15](#_Toc175650153)

[Future Iterations 15](#_Toc175650154)

[References 16](#_Toc175650155)

[Appendix 17](#_Toc175650156)

[Appendix A – Clustering Outputs 18](#_Toc175650157)

## Glossary

|  |  |
| --- | --- |
| Term | Definition |
| Match\_count | The number of matches each team has participated in. This is largely a validation exercise that we have captured the expected number of matches. |
| Shots | The number of shots recorded by a team. |
| shotsOnTarget | The number of shots which were either converted or would have been converted as a goal without interruption from the goalkeeper or a defender. |
| xG | The Expected Goal value accumulated by the team. (StatsBomb, n.d.) |
| Goals | The number of goals scored by the team. |
| Own\_goals | The number of own goals scored by the team. |
| Passes | The number of successful passes recorded. |
| Players | The number of unique players used in a match, including substitutes. |
| Pts/MP | The number of points recorded by a team, on average per match. |
| Rank | The teams final league position. |
| SHAP | Shapley Additive Explanations. |
| VIF | Variance Inflation Factor. |
| WCSS | Within Cluster Sum of Squares. |

## Executive Summary

This report examines the relationship on-pitch performance indicators has on a teams performance, recognising that data adoption within elite football is still relatively nascent and the ability for decision makers within the club to action produced insights is low. By examining these indicators, we can begin to understand what teams do well and support decision makers in making informed decisions.

A clear output of this analysis details that a possession-based style of play is important to achieving results and teams are recommended to reflect on the decisions made both in and out of possession to maximise their threat whilst disrupting their oppositions.

## Introduction

Research question: How can statistical analysis be used to identify the differences between high-performing and low-performing top-tier football teams, and how could teams use these insights to enhance their on-field performance?

Null hypothesis (H₀): There is no apparent relationship between the identified variables and a teams success.

Alternate hypothesis (H1​): A relationship between the identified variables and team success is recognised and can be actioned by practitioners.

Whilst StatsPerform, n.d highlight the significant evolution of football data availability in the previous 25 years, Van Haaren notes that one of the key challenges facing data analysts in football currently is an inability to ‘translate the data into answers to the questions that practitioners have’ (SciSports, 2021). This project was designed to deliberately tackle a key question all football clubs are likely to be reflective of; what distinguishes my teams play-style from others, and is that indicative of on-field performance? Thunberg Kalt, 2024 concludes that there are ‘two more or less distinct clusters’ present within their analysis of domestic football data but the analytical outputs are challenging to consume and are likely to fall into the challenge Van Haaren is critical of. The approach detailed in this report will take influence from Thunberg Kalt and other sports analytics reports, whilst attempting to produce an output and accompanying narrative that can be understood by practitioners.

## Data Engineering

To tackle the research question, and improve the reliability of the outputs, we will perform analysis on all matches within a league to minimise the potential influence of extreme variability within individual football matches. StatsBomb have released all matches within the 2015/2016 season for the top 5 leagues in Europe. This relates to 1823 total competitive, domestic top-flight matches being targeted.

Given the size of the data engineering request, and the availability of the statsbombpy package within Python, Python has been selected as the appropriate tool to conduct data ingestion and the subsequent data transformation steps to deliver prepared data for the analysis task. There is a manual exercise to select the appropriate league/season to target. A depiction of the ETL package can be found in Figure 1, with supporting notes in Figure 2:

A diagram of a process

Description automatically generated

Figure 1

A close-up of a text

Description automatically generated

Figure 2

Figure 3

Similar to Thunberg Kalt, 2024, the selected variables were largely selected on subject matter knowledge about the types of variables which could be used to understand team performance. A limitation to the current list of variables is that it only captures actions taken by the team, and not actions taken by the opposition so variables like number of shots faced are not presented. There is also little manipulation undertaken to understand the context the action takes place in, so the distance of passes, the number of passes made in the final third or the average position of the most defensive player is never considered, for example. A list of the target variables is available in the glossary.

A decision was made during data engineering to reduce the value outputs from their absolute counts to their “per-match” count by dividing the total number of actions by the number of matches the team participated in, in a similar approach as adopted by Akhanli and Hennig, 2022.

Manual manipulation of the team names is required to enable a merge between the two datasets; it is recognised that this is not an ideal column to perform a merge with given it cannot be seen as unique, a key component of good data quality (Askham et al., 2013).

## Analysis

Beginning with a correlation matrix revealed strong correlation between features in the set. This told me that there was a risk of multi-collinearity.

Whilst having reduced the data to just “per-match” values should mitigate some of the risk for collinearity to influence the product outputs, I decided to conduct VIF analysis to determine the extent to which multi-collinearity could be present in the variables, this is presented in Figure 3.

A screenshot of a computer

Description automatically generated

Figure 3 – some incredibly high VIF values and even some infinity values show there is multi-collinearity present.

It is unsurprising the values are so high at this stage as we can expect features like ‘shots/90’ and ‘SOT/90’ to share a ‘high level of correlation’ (Sambandam, 2022).

I implemented a random forest regression model to display the contributions of each features output to the model predictions, using SHAP values. The idea was to identify which features have a high influence on the model so those that do not could be removed and re-assessing the VIF values at this stage. The outputs can be seen in Figures 4 and 5, the features ‘Pts/MP’ has been plotted as the y-variable.

A graph with blue squares

Description automatically generated with medium confidence

Figure 4

A graph of a game

Description automatically generated with medium confidence

Figure 5 – the feature ‘Rk’ clearly has a strong predictive power (0.91 on the correlation matrix) but is diluting the influence that other features can have. The variables not at match level were dropped for a similar reason.

Figure 6 displays the output of the revised SHAP analysis.

A screenshot of a graph

Description automatically generated

Figure 6 shows that goals/90, passes/90, xG/90 play more important roles than others at predicting the ‘Pts/MP’ feature.

Having identified features with an impact on the ‘Pts/MP’ feature, I re-ran the VIF test with an improvement in results but still exceptionally high, as shown in Figure 7. It is recommended that future iterations of this product consider a dimensionality reduction technique, such as principal component analysis to mitigate the prevalence of multi-collinearity.

A screenshot of a computer screen

Description automatically generated

Figure 7

I immediately recognise that a clustering technique is required and have elected to use k-means as it is more computationally efficient than alternative methods and the steps taken to mitigate the prevalence of extreme outliers should prevent potential sensitivities, as identified by Banoula, 2023 as a disadvantage of k-means, if not appropriately handled.

Having conducted WCSS, average silhouette scores and Calinszki-Harabasz analysis, I elected that four appeared to be the optimal number of clusters.

## Visualisations

To present this information in a clear manner, I plotted each of the features I identified as having an influence on the dependent variables with the clusters as legends, as depicted in Figure 8.

I deliberately opted not to plot the centroid centres as felt this could add additional confusion to the report, with no benefit to the end-user as cluster allocation can be clearly ascertained through the legend. This legend uses the ‘colorblind’ palette within seaborn visualisations to support users who have potential difficulties at distinguishing between colours.

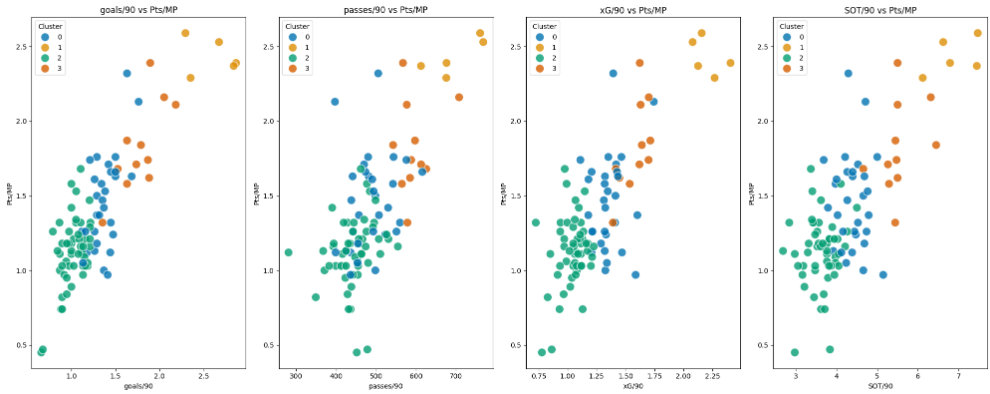


Figure 8 – an enlarged copy is available in the appendix.

I also decided to present the mean average of the feature values associated with each cluster to support practitioners in understanding the type of team performance within each cluster, as shown in Figure 9. This is accompanied by a summary text which guides users in understanding the narrative accompanying the text, this can be found in the discussion section of this report.

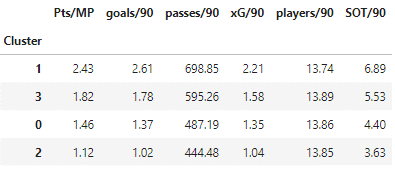


Figure 9

## Discussion

Returning to the research question, teams who can retain possession of the ball and amass a higher number of passes are more likely to produce more shots on target throughout the game and subsequently generate a higher expected goal (xG) value.

There is no suggestion that having a 'deeper' squad, and utilising your bench more frequently is correlated to an increased performance on the pitch as the cluster with the highest number of players used per 90 produce the 3rd best, out of 4, points per match.

The recommendation to clubs is to review their season results and determine which cluster their performance is likely to result in. Recommended steps to improve their ability to increase the number of passes include a review of how a team approaches set-piece situations, the risk appetite towards possession retention and the defensive shape they deploy and the impact that has on an opposition teams ability to retain control.

It should be recognised that as a new data scientist, my limited expertise might have influenced the project’s outcomes, and more experience could have led to a more robust model. Analysing football data will always present challenges, and analysis of the 2015/16 season is no exception; Leicester won the Premier League when the betting odds were famously set at 5000/1 at the start of the season.

## Future Iterations

Future iterations of this product should investigate:

* Identifying a more exhaustive suite of variables to support team performance analysis.
* Further review of the influence feature removal has on SHAP values, VIF values to improve the feature selection process.
* Utilising PCA or other dimensionality reduction techniques to potentially reduce the implications that multi-collinearity might have on outputs.
* Investigating if alternative clustering algorithms, like DBSCAN, could improve the clustering output.

## References

Akhanli, S.E. and Hennig, C. (2022). Clustering of Football Players Based on Performance Data and Aggregated Clustering Validity Indexes. [PDF] p.5. Available at: https://arxiv.org/pdf/2204.09793 [Accessed 22 Aug. 2024].

Banoula, M. (2023). K-Means Clustering Algorithm. [online] Simplilearn.com. Available at: https://www.simplilearn.com/tutorials/machine-learning-tutorial/k-means-clustering-algorithm.

Sambandam, R. (2022). TRC Insights. [online] TRC Insights. Available at: https://trcmarketresearch.com/whitepaper/cluster-analysis-gets-complicated/#:~:text=Collinearity%20is%20a%20problem%20in [Accessed 22 Aug. 2024].

SciSports (2021). State of the Football Analytics Industry in 2021. [online] SciSports. Available at: https://www.scisports.com/state-of-the-football-analytics-industry-in-2021/.

StatsBomb (n.d.). What Is xG? How Is It calculated? [online] StatsBomb | Data Champions. Available at: https://statsbomb.com/soccer-metrics/expected-goals-xg-explained/.

StatsPerform (n.d.). From Descriptive to Predictive: Unlocking the Potential of Football Data to Go beyond Tracking. [online] Stats Perform. Available at: https://www.statsperform.com/resource/from-descriptive-to-predictive-unlocking-the-potential-of-football-data-to-go-beyond-tracking/.

Thunberg Kalt, C. (2024). Cluster Analysis on Football Teams Performance Data. [PDF] Available at: https://uu.diva-portal.org/smash/get/diva2:1833447/FULLTEXT01.pdf [Accessed 22 Aug. 2024].

## Appendix

### Appendix A – Clustering Outputs