Introduction to data visualization: Final Coursework

Shayan Bali*
K24084452

ACADEMIC HONESTY AND INTEGRITY

Students at King's are part of an academic community that values trust, fairness and respect and actively encourages students to act with honesty and integrity. It is a College policy that students take responsibility for their work and comply with the university's standards and requirements. Online proctoring / invigilation will not be used for our online assessments. By submitting their answers students will be confirming that the work submitted is completely their own. Misconduct regulations remain in place during this period and students can familiarise themselves with the procedures on the College website. ***I agree to abide by the expectations as to my conduct, as described in the academic honesty and integrity statement.***

1 PART 1: ANALYTICS

I have selected the Premier League dataset as my topic. This dataset consists of information about the Premier League from 1993 to 2024. This dataset includes the following entries:

- `season_end_year`
- `team`
- `position`
- `played`
- `won`
- `drawn`
- `lost`
- `gf`
- `ga`
- `gd`
- `points`
- `notes`

This dataset provides a great amount of information to gain a general overview of the performance of the Premier League from 1993 to 2024.

In this part, I propose my exploratory questions to investigate different aspects of this topic.

1.1 Question 1:

Analyze the development of team performance over time. Are there detectable trends in the performance of teams over the last decade? To answer this question, we can use different attributes of the provided dataset[1] by the tutor, like league position, points, goal difference (gd), and wins/losses. All of these attributes can be useful, but in my opinion, the best one among them is league position. The reason lies within the fact that the highest and lowest amounts for the other metrics can vary from one season to another, and their values are not consistent [10], but the league position has always the same value and is consistent [1]. The provided dataset has the league position by itself, so it will be enough for visualization. So I can plot league position, year by year, to find detectable trends. The only thing is doing a simple data pre-process to remove, for example, columns that are not mainly relevant, like "notes". Additionally, null and missing values should be considered.

1.2 Question 2:

Is there any relationship between the spending of teams and their performance? Is more spending resulting in better performance, necessarily? One of the interesting questions can be the investigation of the relationship between performance and spending. As I mentioned in the previous question, for performance, the "Position" attribute of the provided dataset will be enough. However, it doesn't have information about the spending of teams. Accordingly, an additional dataset will be required here. So I found a dataset [2] that consists of all player transfers to the Premier League from 2001 to 2025. Based on this dataset, I created another dataset that has the total spending of each team from 2001 to 2025. The mentioned dataset has all transfers in the Premier League from 2001 to 2025; however, this record is player by player, not team by team. So I aggregated the data by the field of "team" for each season [7]. After this data pre-process, now I can merge these two datasets, and then we can have another column "total_spending" to show the total spending of each team in each year. On that ground, the resulting dataset will be absolutely ideal for this question, because it has the required data for spending and performance [6].

1.3 Question 3:

Which county has the best performance in terms of football in England For visualizing and analyzing this question, I need the performance data of teams along with the records about the county and city of the teams presented in the Premier League during the past two decades. I searched a lot, but I couldn't find a related dataset, especially for the geographical details of teams. On that ground, I decided to create my own dataset about the teams in the Premier League from 2005 to 2025 with their city and county. With having this dataset, I can merge these two datasets and use them to visualize this question. To create this dataset, I crawled the Wikipedia website [3] for the Premier League, year by year, to get the teams in each season and information about their cities and counties. After that, I combined all these seasons to create the whole dataset and

[*] e-mail: shayan.bali@kcl.ac.uk

[1] https://www.kaggle.com/datasets/evangower/english-premier-league-standings

[2] https://www.kaggle.com/datasets/tugbatandogan/epl-summer-transfers-2000-2024-from-transfermarkt

[3] https://en.wikipedia.org/wiki/Premier_League

use it for this question. Ultimately, the combination of these two datasets will be ideal for this question.

2 Design and Discussion

For the design of the visualizations, there are some parameters that should be considered. The most important thing is the **type of chart** [9]. By choosing the correct type of visualization and chart, you can provide the highest amount of information while keeping it interpretable [3]. Moreover, the proper choice of parameters to plot is also very valuable, and placing them in their correct positions is important. Also, another thing that can be critical is visual encoding, which makes your visualization more understandable and more interpretable [9].

2.1 Question 1:

For the first question, to analyze the development of teams' performance over the last two decades, I decided to choose a **multi-line chart** to show the league position of teams over time. Regarding the type of chart, line charts are always a good and effective option for temporal data. Especially when we want to find the trends over time, a line chart is absolutely ideal, because it reveals these patterns over time, and you can see the progress and weaknesses of teams. For the **X-Axis**, I used the *season_end_year*, which shows chronological trends from 2005 to 2024. The **Y-Axis** of the chart is dedicated to the **performance of teams**. In my opinion, the position of teams in each season is the most reasonable parameter for evaluating the performance of teams. *Position* shows the rank achieved by a team each season. Since a lower number indicates a better outcome, the y-axis is inverted to emphasize improvement with upward movement, making it intuitive (e.g., 1st place is visually at the top) [8]. To make the plot clearer and more interpretable, we focused on 8 consistently strong teams (based on the most frequent top-6 finishes). Regarding visual encoding, distinct colors differentiate teams, and circle markers (o) emphasize each season's position. Moreover, gridlines and labeling are also added to enhance readability and allow precise tracking [5]. Unlike traditional tables or bar charts that only show one season, this chart allows users to compare historical trajectories across teams. The inversion of the y-axis aligns performance trends with intuitive visual metaphors (up = better). This design shows competition, like dominance, ups and downs, and changes. Additionally, this chart is also interactive, and when you hover over a point in the chart, it shows the position of that specific team in that season.

2.2 Question 2:

For the second question, to investigate the relationship between financial investment and team success, I designed a **scatter plot** where each point represents a team's seasonal performance, measured by total spending vs. final league position. The X-axis of the chart reflects the financial investments of teams, which is the core independent variable in our question. The Y-Axis represents *Position*, which is an ideal measure for the performance of teams. It is also inverted by placing 1st place at the top, aligning upward movement with success to make our chart more intuitive [8]. Regarding visual encoding, each color is associated with one team to make them distinguishable and also clearer [9]. A rate of alpha (transparency) can be used to avoid overplotting, since many teams cluster in low-spend / low-performance ranges. Gridlines and layout improve legibility and comparison. In this question, since our main focus is finding the relationship between spending and performance, a scatter plot is an ideal choice, especially since we can make it interactive by providing a filter option to just see the details about one specific team or also filter by the year to find the seasonal patterns. Moreover, when a user hovers on a specific point in the chart, the details about that specific part will be shown to provide additional information. This chart will help users to find hidden patterns between spending and performance.

2.3 Question 3:

For the third question, to investigate the performance of each county in the UK during the past two decades, a **choropleth map** is the most effective visualization [4]. The attached map visually communicates performance using color saturation mapped to each county's average performance rank (average position of teams in that county over the past two decades). Regarding the geographic encoding, each UK county is filled with a color based on its average *performance_rank*, which makes it easy to compare regions quickly. In terms of the legend and color, I use a divergent purple-to-pink palette that visually emphasizes extremes — darker purple represents better football performance (lower ranks), and a clear numeric legend (from 3.1 to 20.4) reinforces the meaning of the color scale [2]. Moreover, gray counties represent the areas that don't have any record, which helps users interpret gaps without confusion. Moreover, gray is a neutral color, which is reasonable for these areas. In addition, the map is interactive, and when you hover over a county, the average position of teams from that county will be shown; and when you click it, it will show the line chart of performance for the teams in that county, which is absolutely beneficial. On that ground, this design provides a great overview of teams' performance across the country.

References

[1] C. Barros and S. Leach. Performance evaluation of the english premier football league with data envelopment analysis. *Applied Economics*, 38:1449–1458, 02 2006. doi: 10.1080/00036840500396574

[2] C. A. Brewer. Color use guidelines for mapping and visualization. In A. M. MacEachren and D. Taylor, eds., *Visualization in Modern Cartography*, pp. 123–147. Pergamon, 1994.

[3] W. S. Cleveland and R. McGill. Graphical perception and graphical methods for analyzing scientific data. *Science*, 229(4716):828–833, 1985.

[4] B. D. Dent, J. S. Torguson, and T. W. Hodler. *Cartography: Thematic Map Design*. McGraw-Hill Education, 6th ed., 2008.

[5] S. Few. *Now You See It: Simple Visualization Techniques for Quantitative Analysis*. Analytics Press, 2009.

[6] E. Franck and M. Lang. A theoretical analysis of the influence of money injections on risk taking in football clubs. *Scottish Journal of Political Economy*, 61(4):430–454, 2014.

[7] J. Han, J. Pei, and M. Kamber. *Data Mining: Concepts and Techniques*. Morgan Kaufmann, 3rd ed., 2011.

[8] J. Heer, M. Bostock, and V. Ogievetsky. A tour through the visualization zoo. *Communications of the ACM*, 53(6):59–67, 2010.

[9] T. Munzner. *Visualization Analysis and Design*. CRC Press, 2014.

[10] A. Stafylidis, A. Mandroukas, Y. Michailidis, L. Vardakis, I. Metaxas, A. E. Kyranoudis, and T. I. Metaxas. Key performance indicators predictive of success in soccer: A comprehensive analysis of the greek soccer league. *Journal of Functional Morphology and Kinesiology*, 9(2):107, 2024.

Appendices

To create the initial designs, *Python* and *datawrapper* were used for question 1-2 and 3, respectively

Moreover you can have access to the dataset used for this coursework using following links

Dataset for question 1: EPL record from 1993-2024

Dataset for question 2 (transfer market): EPL Transfer Market

Dataset for question 3 (Teams' County): As I mentioned in the report, I have created my own dataset for county and city of EPL teams. To contribute future works in this field, I have made my dataset publicly available on Kaggle website, which can be used in future works. EPL teams from 2005-2025 with their county city

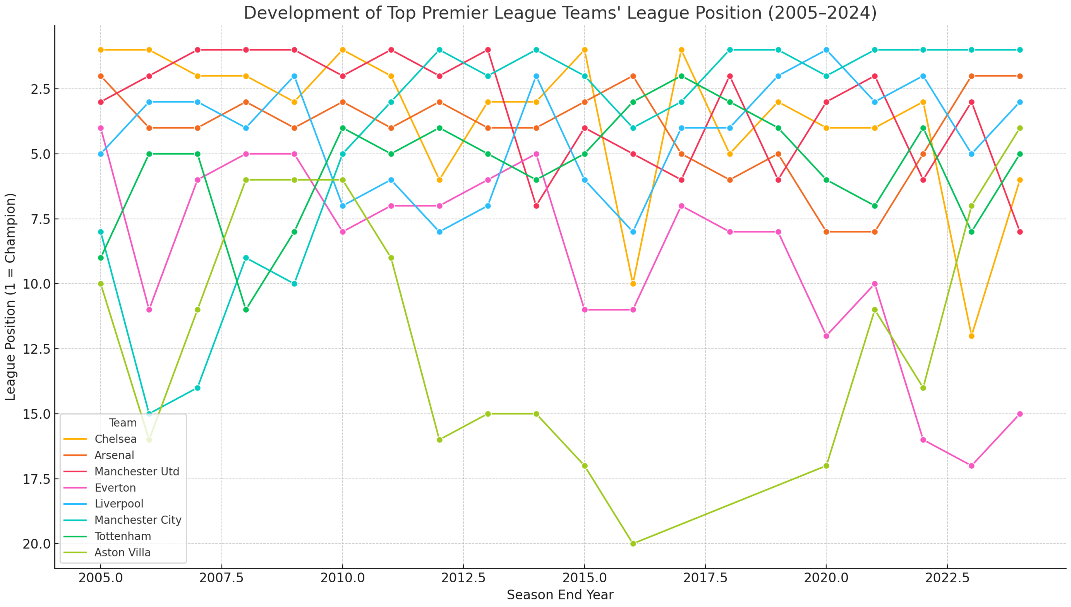

Figure 1: (Question 1, Main design): League Position Trends of Top Premier League Teams (2005–2024). This line graph shows the rankings for eight major clubs. Lower positions on the y-axis indicate better performance (1 = Champion) which is highlighting trends like Manchester City's rise and fluctuations among traditional top teams. Designed by Python.

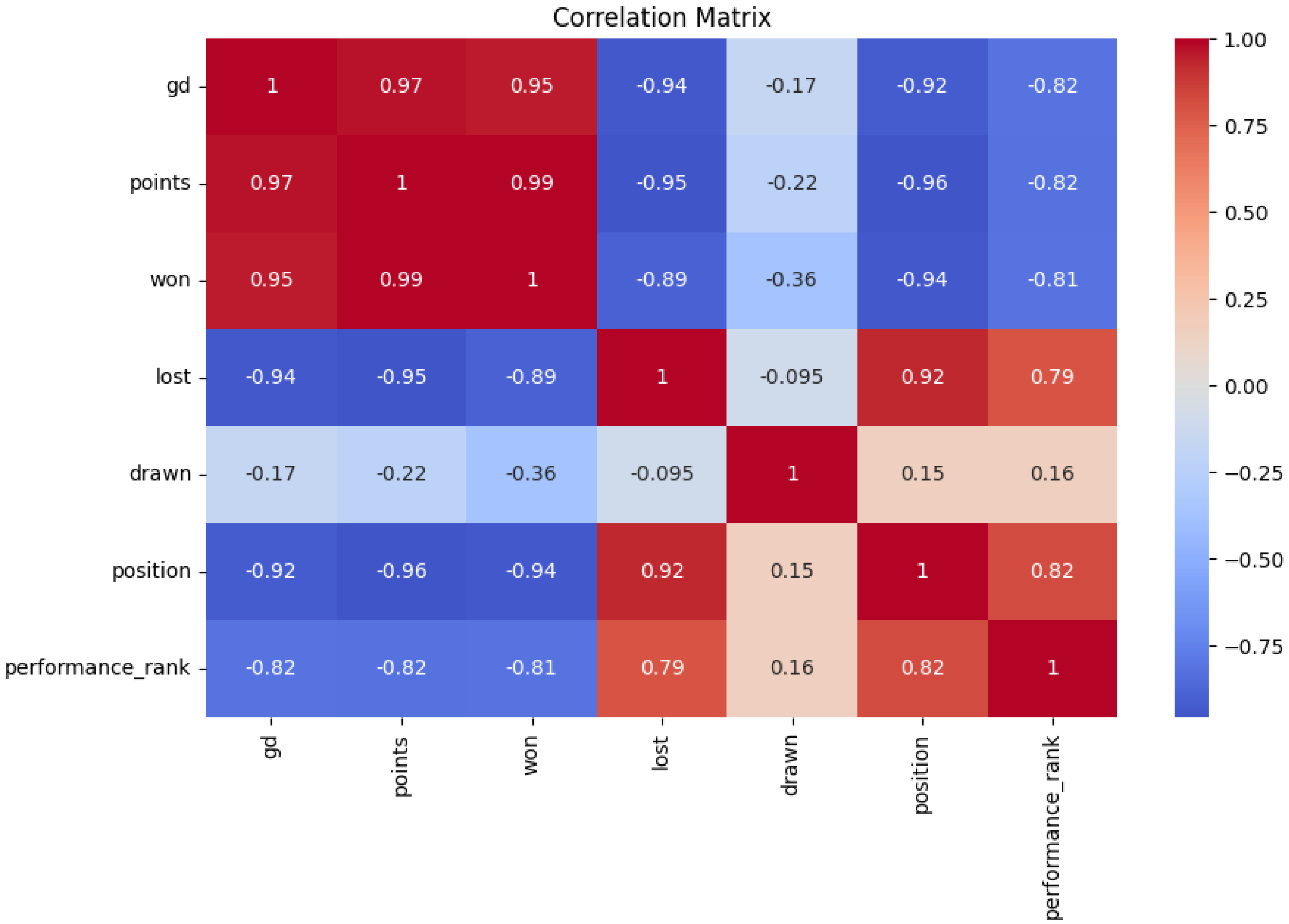

Figure 2: (Question 1 Genenral overview of provided dataset): Correlation Matrix of Key Performance Metrics. This heatmap shows the correlations between various Premier League performance metrics. Great positive correlations exist among points, goals difference, and wins. losses and league position are negatively correlated with team success. Designed by Python.

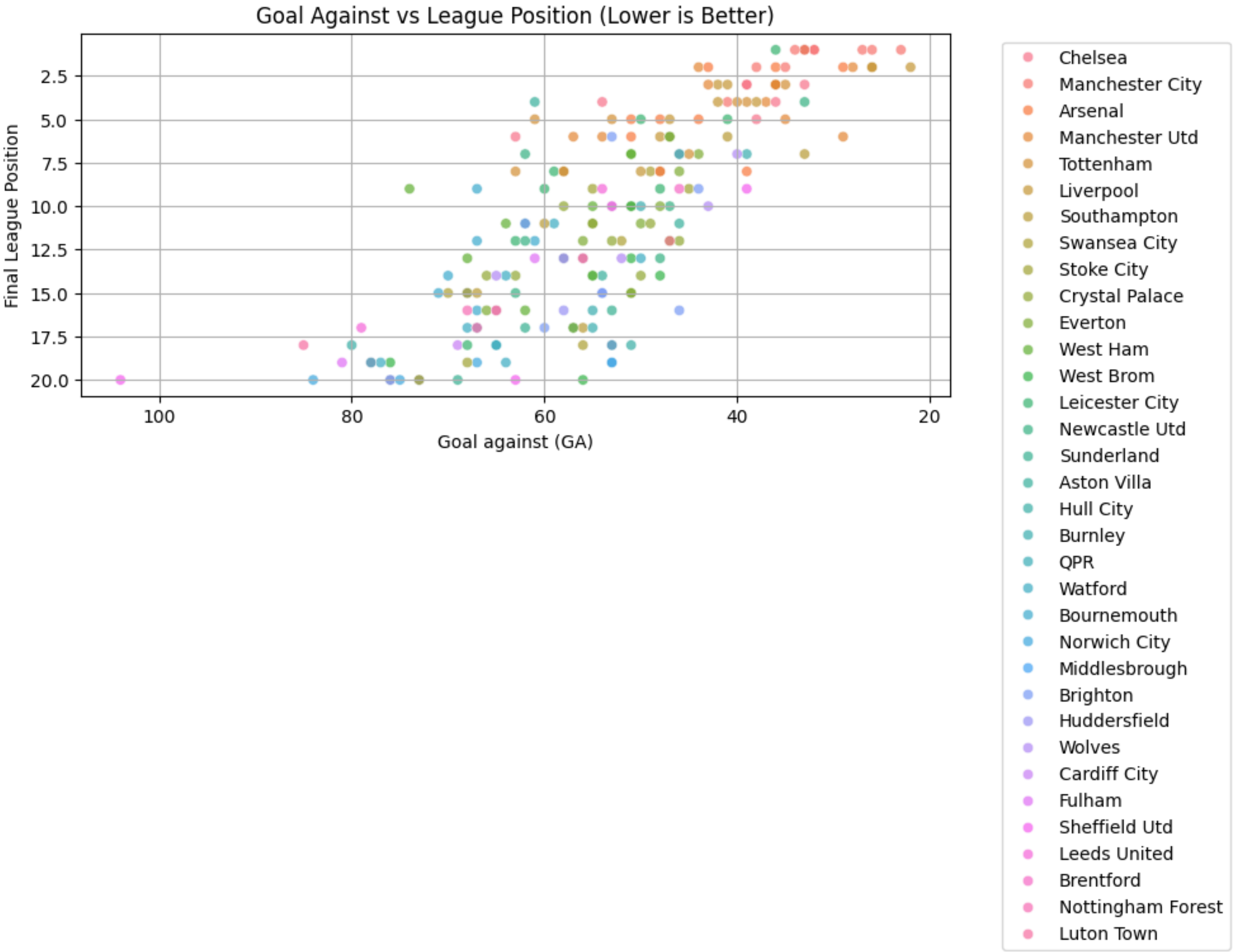

Figure 3: (Question 1 – Alternative Design): Goal Against vs Final League Position. This scatter plot shows the relationship between goals against (GA) and final league position for Premier League teams. A clear trend shows that teams with fewer goals against are more likely to achieve better (lower) league positions which highlights the importance of defense strength. Designed by Python.

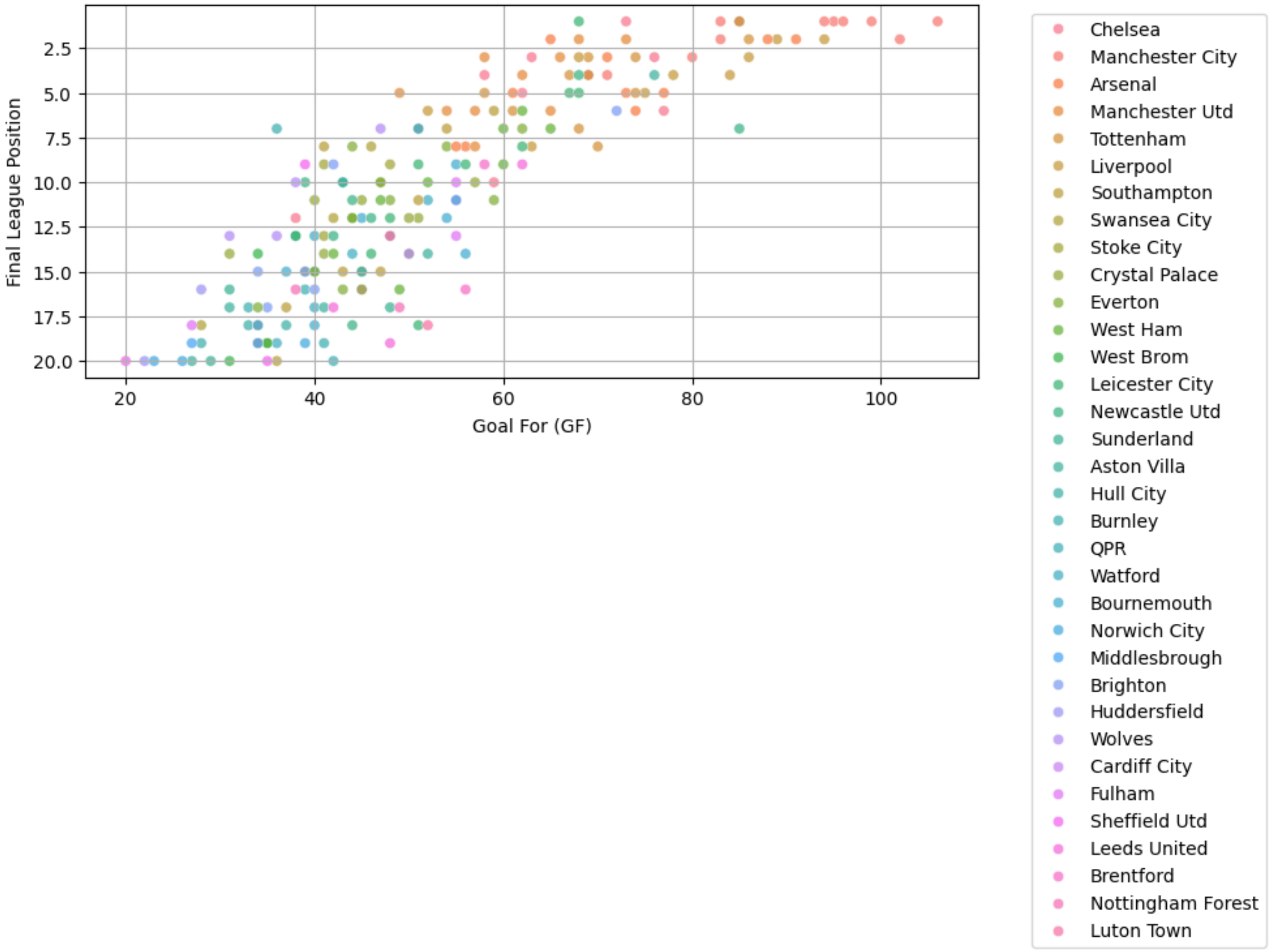

Figure 4: (Question 1 – Alternative Design): Goal For vs Final League Position. This scatter plot shows the relationship between goals scored (GF) and final league position. Teams with higher goal are generally achieving better rankings, that shows strong attacking performance is closely linked to league success. Designed by Python.

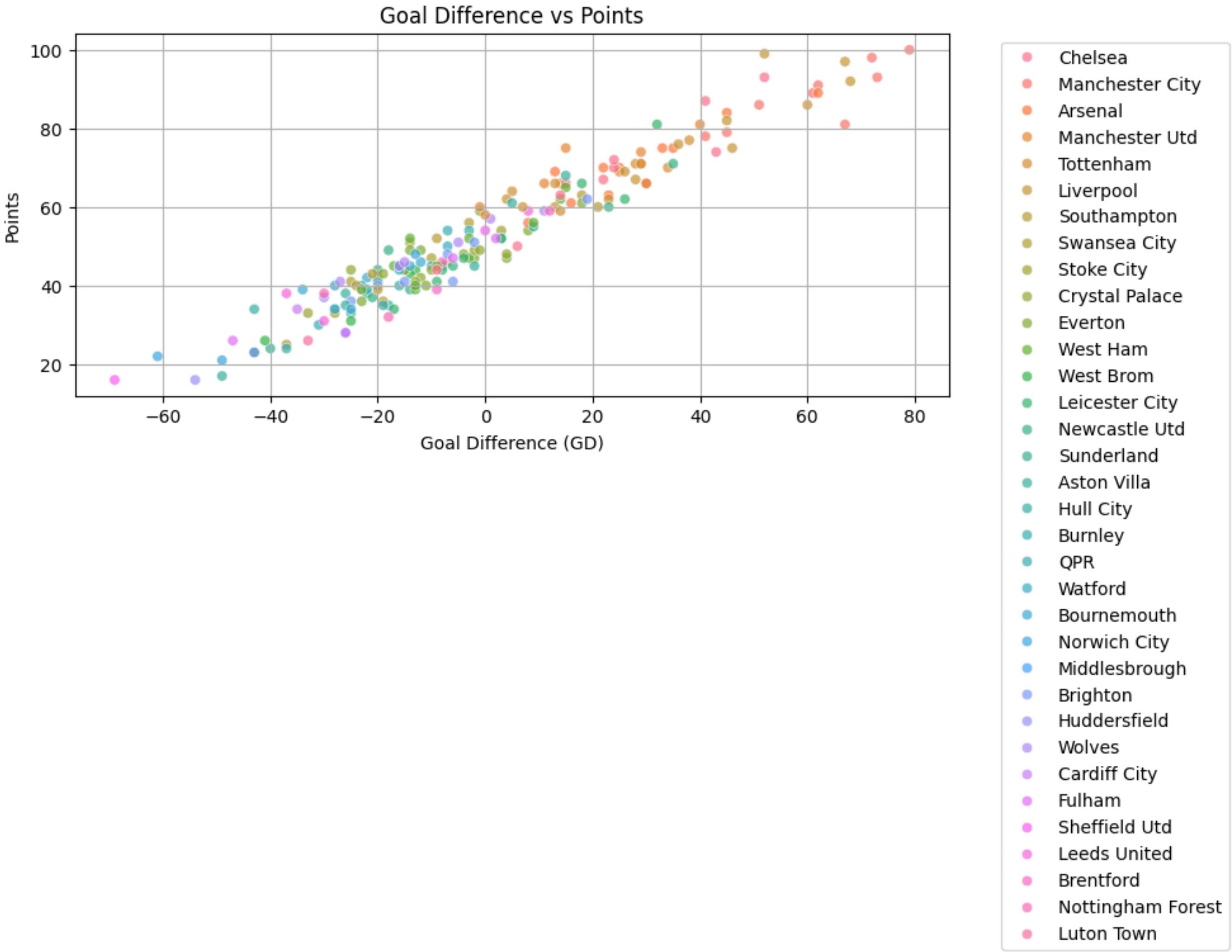

Figure 5: (Question 1 – Alternative Design): Goal Difference vs Points. This scatter plot shows the positive relationship between goal difference (GD) and total points earned. Teams with higher goal differences are more likely to get more points. it shows that both attack and defense are effective on overall success. Designed by Python.

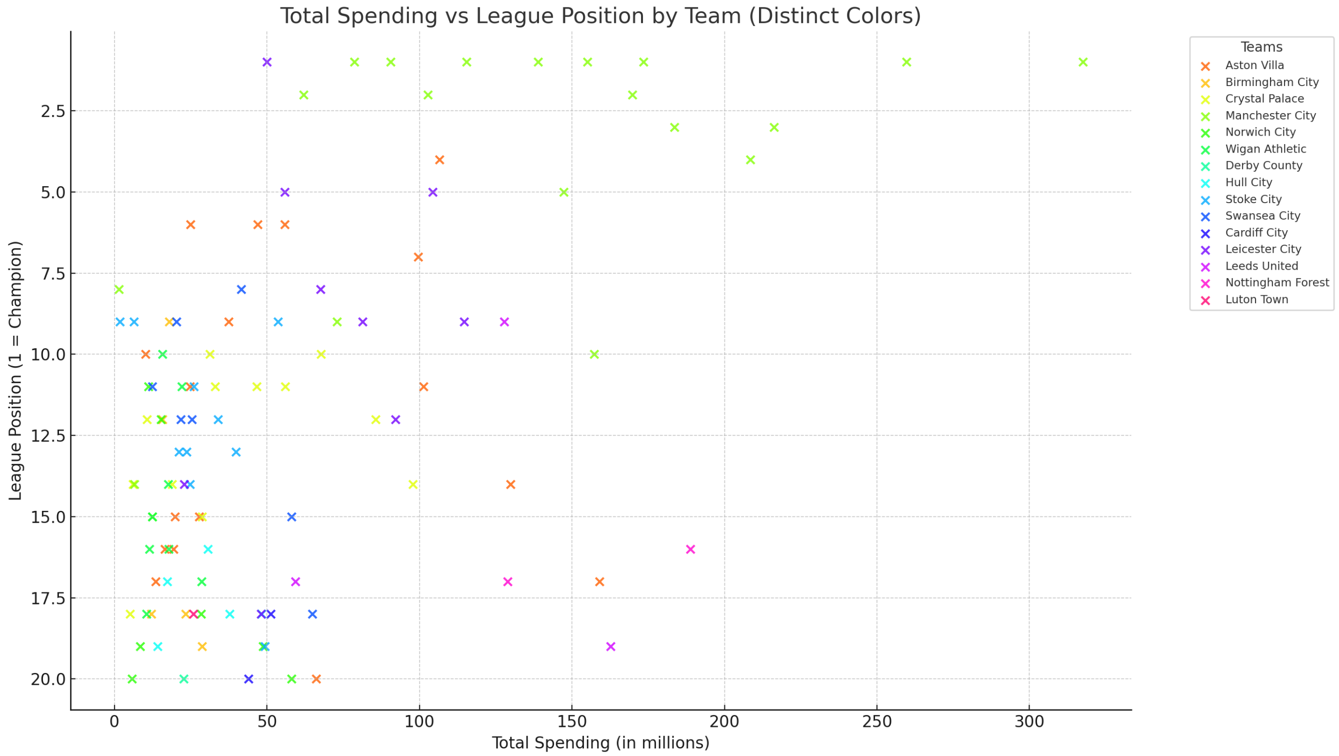

Figure 6: (Question 2): Total Spending vs Final League Position by Team. This scatter plot shows the relationship between total team spending and league position, with distinct colors for each club. Higher spending is mostly associated with better (lower) league positions, but consider that financial investment alone doesn't guarantee top performance. Designed by Python.

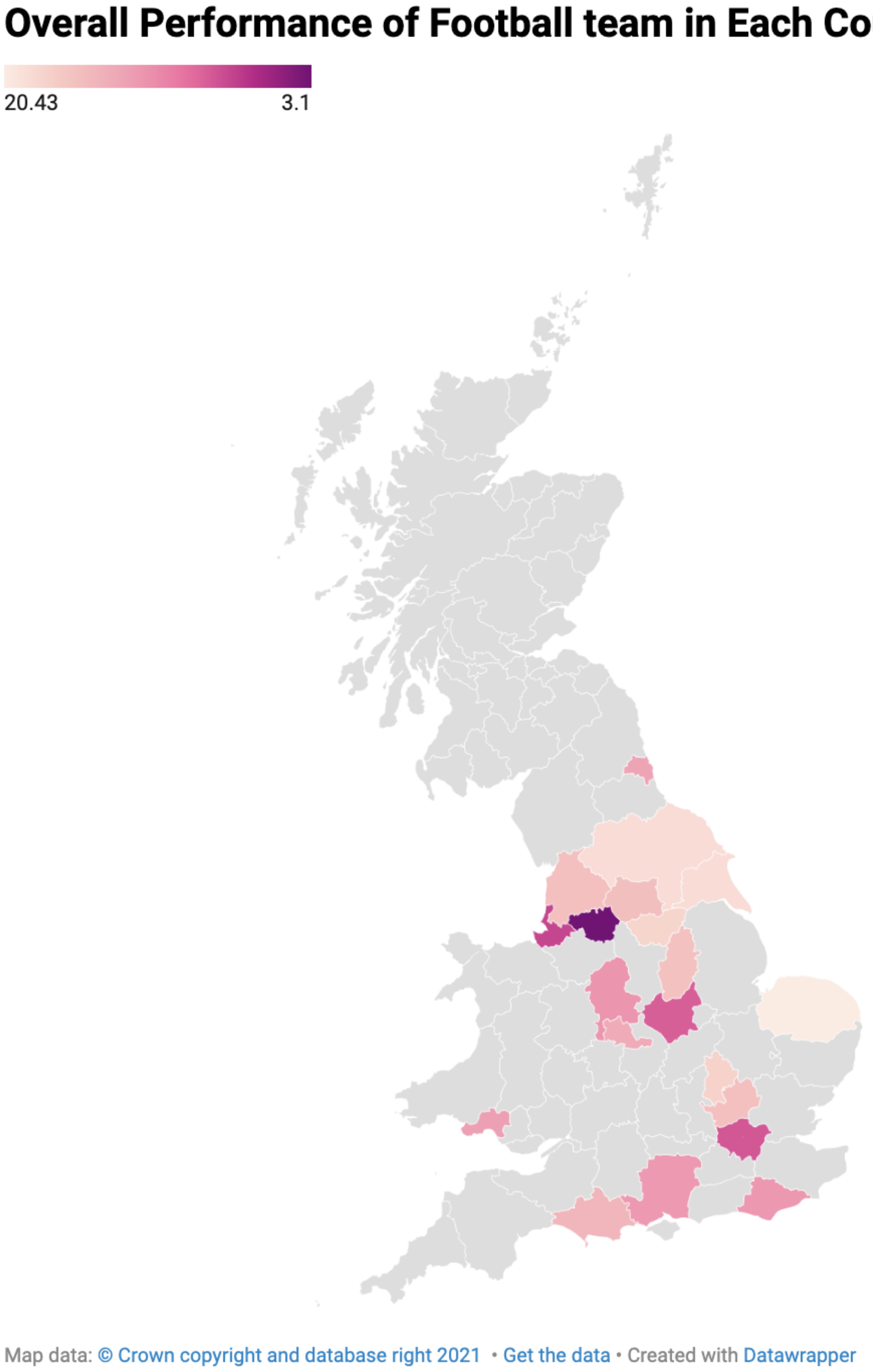
Overall Performance of Football team in Each County
20.43
3.1
Map data: © Crown copyright and database right 2021 • Get the data • Created with Datawrapper

Figure 7: (Question 3): Overall Performance of Football Teams by County in England. This choropleth map visualizes the average league performance of football teams in UK counties. Darker color indicate better-performing counties. Designed by Datawrapper.

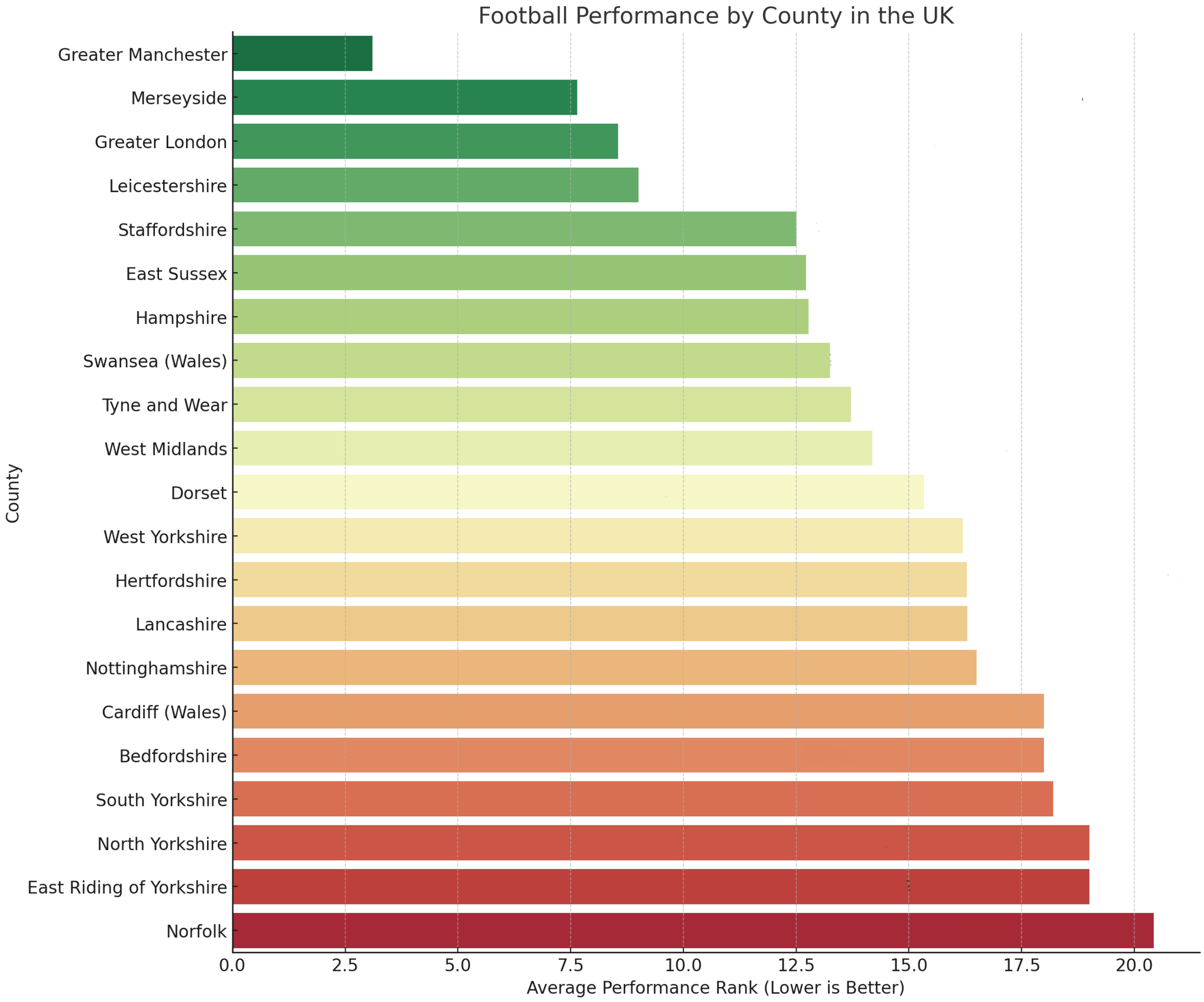

Figure 8: (Question 3 – Alternative Design): Average Football Team Performance by County. This horizontal bar chart ranks UK counties based on the average performance of their football teams. Lower values indicating better performance. Greater Manchester, Merseyside, and Greater London lead the rankings. it shows the counties with consistently strong football clubs. Designed by Python