**Relationship between transfer fees and team performance among Chelsea and Manchester City Football Club**

Registration Number: 240212137

Module Details: INF4000 Data Visualisation (AUTUMN 2024~25)

Number of Words:

|  |  |
| --- | --- |
|  |  |
|  |  |

**Figure 1: Composite Visualisation**

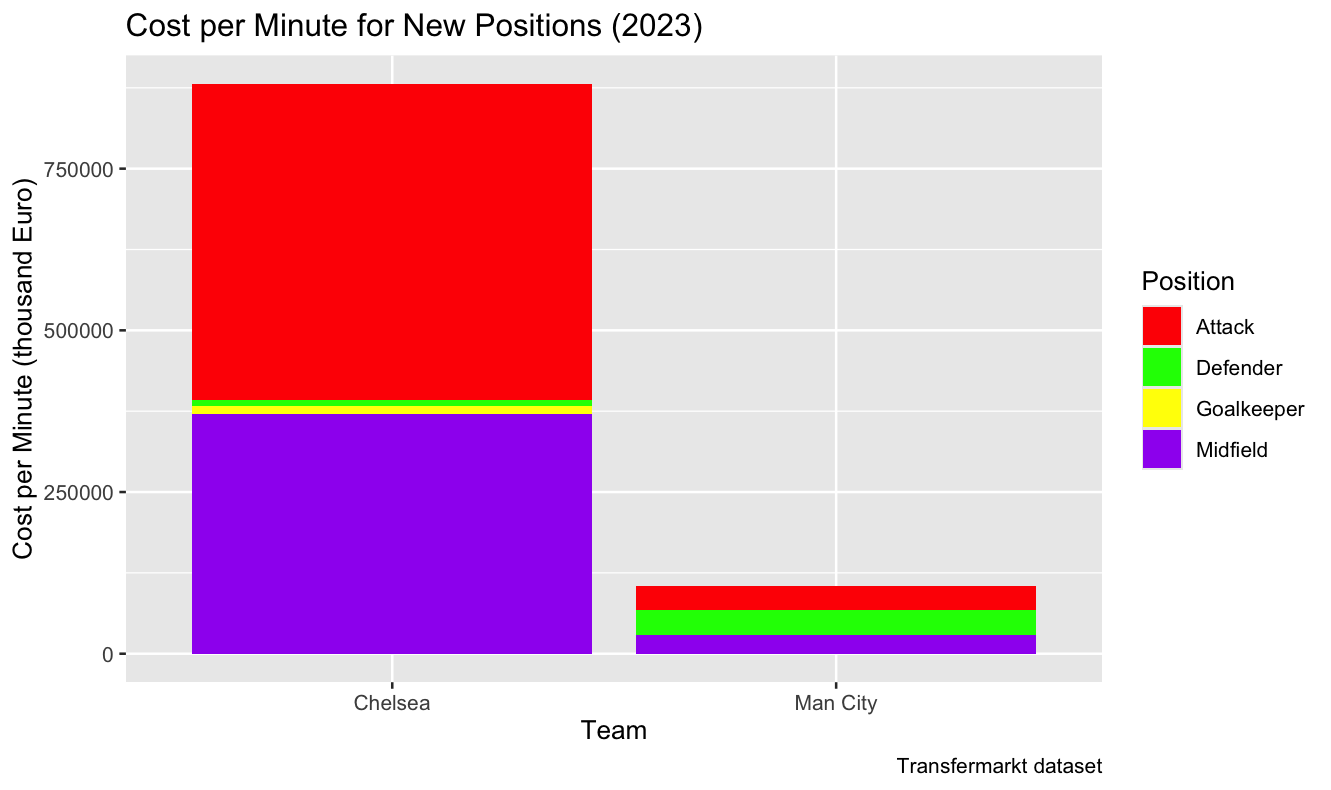
# Knowledge Building

The transfer fee of players is an exciting field for analysis from the perspective of team performance in football. Tomkins et al. (2010) state that they have found a significant connection between transfer fees and success. That can explain why the owners of clubs located in Saudi Arabia, Qatar, and America immediately think that spending money will lead to success (Brandon, 2024). That thought is not entirely unreasonable. Merten (2022) determines that believing in the traditional knowledge that increased transfer fees correlate to better performance is accurate. Moreover, clubs need to maintain a high transfer fee in the long term to influence future performance, and high transfer fees cannot influence performance in the short term. Therefore, they cannot be expected to influence performance in the long term.

This project aims to expand upon the INF6027 finding, which suggests that maintaining high transfer fees over the long term is essential for influencing future performance, by exploring two examples: Manchester City Football Club, which retains its transfer fees over the long term, and Chelsea Football Club, which tends to invest in high transfers in the short term. Moreover, it will look at another important aspect that improves the team's performance: how well the club can invest wisely in the right players who enhance team performance directly.

First, Picture 1 illustrates Manchester City Football Club and Chelsea Football Club transfer fees from 2012 to 2022. During this period, Manchester City spent more than Chelsea in six out of eleven seasons, while Chelsea spent more in five seasons. This result indicates that Manchester City has held higher transfer fees in the long term than Chelsea, ultimately influencing their performance in 2023. Second, Picture 2 compares the number of wins for Manchester City and Chelsea Football Clubs in 2023. It shows that Manchester City achieved ten more wins than Chelsea, despite Chelsea spending almost twice as much on transfer fees. This outcome suggests that a high transfer fee in the short term does not affect short-term performance. Therefore, they cannot be expected to influence performance in the long term. Furthermore, another factor contributing to Manchester City's more tremendous success over Chelsea is their ability to invest wisely in the right players, directly enhancing team performance. Picture 3 depicts the distribution of spending on new positions in 2023, showing that while Chelsea invests in all positions, Manchester City chose not to spend on a goalkeeper. Both clubs invested the most in midfield players. Additionally, Manchester City allocated more funds to attack and defenders than Chelsea did. Picture 4 shows the costs per minute that Manchester City and Chelsea incur for players in various positions. The cost per minute reflects how much a club pays for each minute a player is on the field. A higher cost per minute may indicate that a player is either injured or not fitting into the manager's tactics, which can lead to less playing time and a reduced contribution to the team's performance. Conversely, a lower cost per minute suggests that a player fits well within the manager's strategy or remains injury-free, resulting in more playing time and a more effective contribution to the team's performance. Picture 4 reveals some key insights: Chelsea is the only club that invests in a goalkeeper, resulting in a higher cost per minute for that position compared to Manchester City. Chelsea also spends more on midfielders, leading to a higher cost per minute in that area. In contrast, Manchester City invests more in defenders, resulting in a higher cost per minute for that position than Chelsea. Interestingly, while Manchester City spends more on attackers, Chelsea's cost per minute for that position is greater than Manchester City's.

# Theoretical Frameworks



**Figure 2: Cost per Minute for New Positions**

## ASSERT Frameworks

### A (Asking)

Developing a good question to drive the design of interactive visualisation is critical. However, a question can become overly general and challenging, making it difficult or even impossible to respond significantly. Therefore, the project narrows the scope of the question by two factors, where it looks at the new positions that have come since 2023 rather than looking at the whole squad. Moreover, the project only looks at the cost per minute to evaluate how well Chelsea and Manchester City Football Club transfer decisions are rather than looking at cost per goal, cost per assist and cost per goal conceded, which are metrics that can also be used to evaluate how well the club’s transfer decision is. However, those metrics can only evaluate a specific position rather than the whole position, as the cost per minute does. Therefore, the project might need four different charts to evaluate the transfer decisions of Chelsea and Manchester City Football Club. Additionally, there are various methods for orderly constructing and evaluating the quality of the research question once the scope has been specified. The project chooses the three-part query that identifies a specific subject, the investigation to make regarding the information necessary for exploration, and the justification for why this exploration is essential (Ferster & Shneiderman, 2012). These three elements merged into a straightforward phrase, which is ‘We are looking at the cost per minute for new positions because we want to find out how new positions contribute to the team performance in order for my audience to evaluate how well the Chelsea and Manchester City Football Club transfer decision is’ that guides the visualisation process.

### S (Search for Information)

The project chooses to gather data using an inductive approach, using data from the Transfermarkt, a composite dataset consisting of multiple weekly updated CSV files that provide relevant attributes on competitions, matches, clubs, players, appearances, and transfers. The primary purpose of gathering data in an inductive way is to allow insights to arise naturally from the data itself. In contrast, the deductive approach used in experimental and hypothesis-testing approaches can introduce bias and hide essential insights (Thomas, 2006).

### S (Structure the Information)

The project combines datasets of clubs, players, transfers, and appearances. Second, it extracts records from the combined dataset where the competition is the Premier League, and matches were played in 2023. Third, the project also extracts records of the transfers since 2023. Fourth, it aggregates the total minutes played by each player by summing these values. Moreover, it only extracts Chelsea and Manchester City Football Club records.  Furthermore, the project divides each player's transfer fee by the minute a player plays. Additionally, it groups the data by clubs and positions in each club. Then, it aggregates each position's mean cost per minute by calculating the mean of these values.

### E (Envision the Answer)

The project uses quantitative analysis to envision the answer, which includes the compare-and-contrast technique, where the project compares the average cost per minute of the same position in two different clubs.

### R (Represent the Visualisation)

There is no significant difference in perception between vertically or horizontally arranged stacked bar charts. The choice of layout will depend on which format best accommodates the value range and facilitates reading the item labels associated with each cluster (Kirk, 2019). Since the project only has two relatively short item labels, the project chooses vertical coordinates. Moreover, the project uses a warm (red, yellow) and cool colour (green, purple) to denote different positions, which increases the perceived contrast and alters the perception of both colours (Albers, 1963). Therefore, it helps preserve a clear distinction between the positions. Additionally, the project provides a label on the x-axis to denote the unit of the cost and a label on the y-axis to denote the team. In contrast, the label on the legend denotes which colour stands for which position. Furthermore, the title provides an overview of what the chart represents.

### T (Tell a Story)

Figure 2 highlights Chelsea's higher cost per minute for goalkeepers and midfielders than Manchester City Football Club's. Manchester City Football Club also has a higher cost per minute for defenders. Additionally, although Manchester City spends more on forwards, Chelsea's cost per minute for that position exceeds that of Manchester City.

## Grammar of Graphics

The project has four layered of grammar defining the components of a plot:

* Data and Aesthetics Mapping
* Geometric Object
* Scales
* Aesthetics

### Data and Aesthetics Mapping

The project uses the combined dataset, which contains information about the club’s name, positions in the club and the average cost per minute for each position. Then, it maps the club’s name to the x-axis, the average cost per minute to the y-axis and the positions to the fill aesthetic to colour the inside areas of geometric (Wickham, 2010).

### Geometric Object

The project uses a bar geometric with a statistical transformation equal to identity. This statistical transformation overrides the default count value and uses the value mapped to the y-axis, which is the average cost per minute.

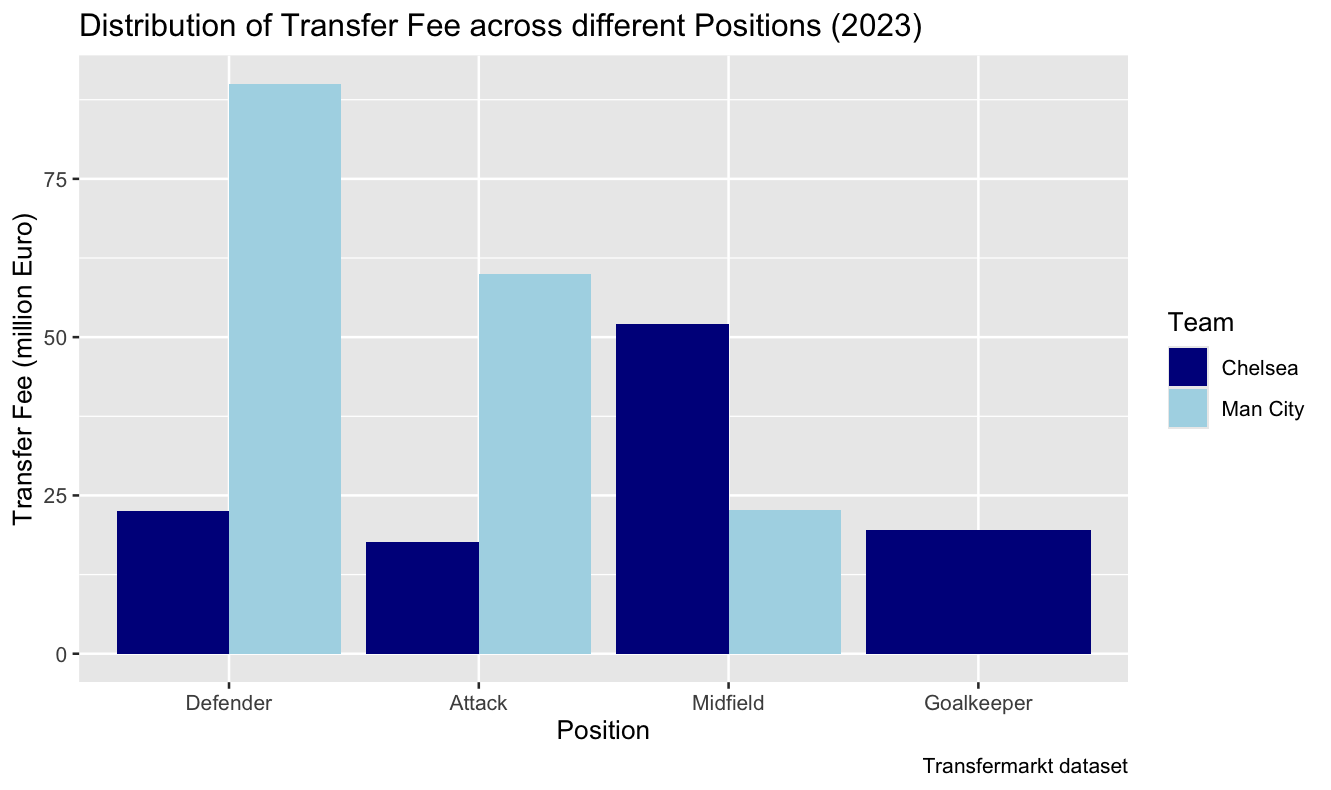
### Scales

The project manually scales the fill aesthetic by overriding the scale's default name, which is taken from the first mapping used for the fill aesthetic by renaming it as ‘Position’. Then it was used as a legend title. Moreover, the project maps a single aesthetic value to each position in the club in an alphabetical order, where red stands for attack, green for defender, yellow for goalkeeper and purple for midfield.

### Aesthetics

The project uses the x-axis to display the club’s name, the y-axis to display the average cost per minute in a thousand units, and legend labels to display the position’s name. Moreover, it displays the plot’s title as ‘Cost per Minute for new Positions (2023)’ to explain the main findings. Furthermore, the project uses the caption to provide information about the data source, which is Transfermarkt.

# Accessibility



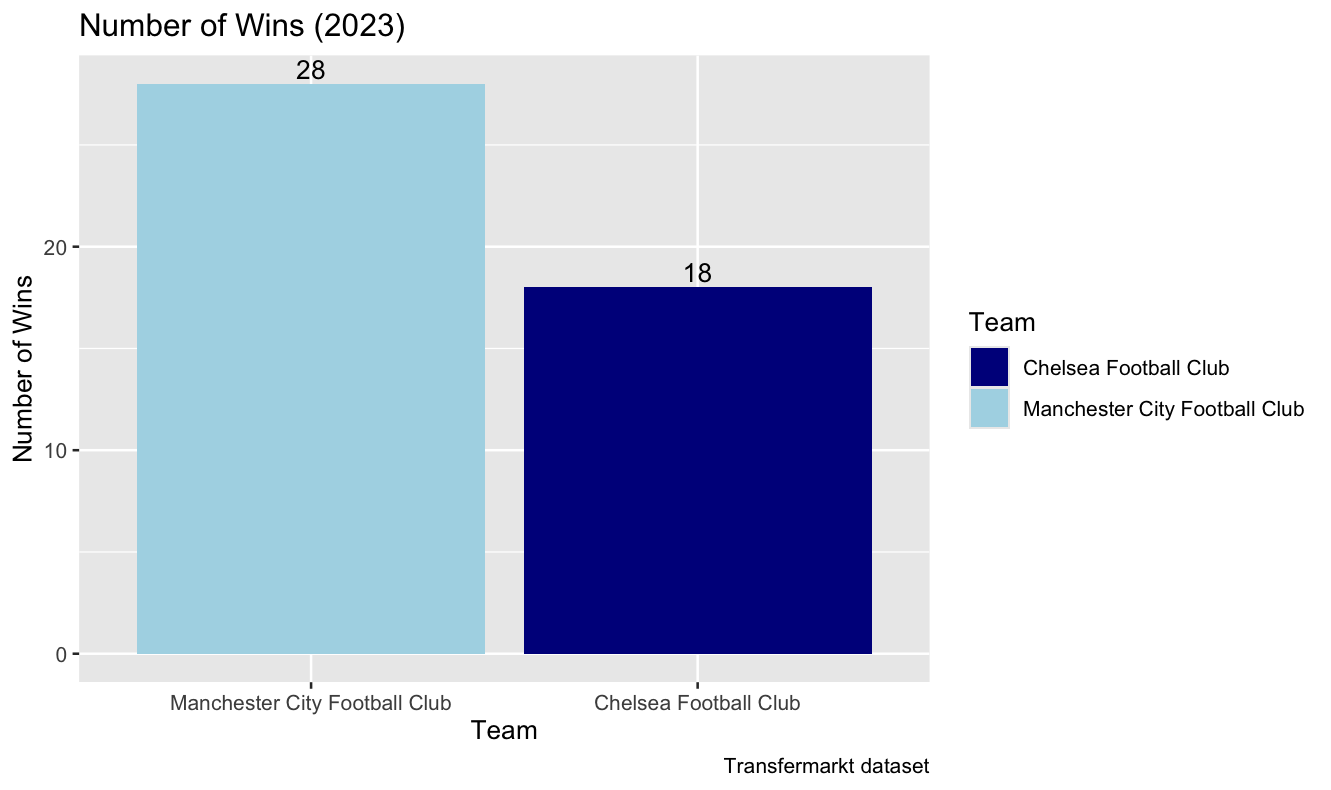
**Figure 3: Distribution of Transfer Fee across different Positions**

Accessibility in visualisation design focuses on providing the audience with meaningful insights. These insights should relate to the topic and be associated with their specific requirements. They should be accomplished in a manner that does not demand excessive effort to notice, interpret, or understand (Kirk, 2019).

The first element of accessibility relates to how relevant the visualisation is to the subject (Kirk, 2019). The subject of Figure 3 is how the transfer fees are distributed across different positions. To clarify this subject, the project places the various positions along the x-axis, the corresponding transfer fees along the y-axis and use the length of vertical bars to represent values. This setup helps the visualisation answer the subject.

The second element of accessibility in visualisation design is the suitability of design, which looks at the usefulness of the visualisation from different viewpoints, which are understanding the subject, representation design, time, attitude and emotion, and presentation design. This element highlights the understanding of how to use visualisation rather than the consequences of visualisation use (Kirk, 2019). However, the project only mentions understanding the subject and representation design. The first viewpoint is understanding the subject, which indicates what audiences know and do not know about the subject, significantly impacting how they decide whether the visualisation is accessible (Kirk, 2019). However, the project does not require the audience to know football to understand the subject; when someone mentions fees, it means money paid for a particular piece of work or a specific right or service. In these circumstances, the amount of money paid for a transfer and what things the clubs can transfer, such as players. However, the project does not view players in particular. Instead, it looks at the player’s role from a broader perspective. Therefore, the subject can be understood as money paid within a role. The second viewpoint is representation design, which indicates the perceived complexity of uncommon or unfamiliar chart types. Therefore, it is understandable how this exposes a risk of confusion (Kirk, 2019). In order to choose which type of chart will be used, it is important to decide what the visualisation’s subject is first. Thereby, choose the appropriate chart type. The project chose the clustered bar chart, one of the most widely used and recognisable bar charts, and it is easily understood thanks to its familiarity and versatile use (Inforiver, n.d.).

# Visualisation choice

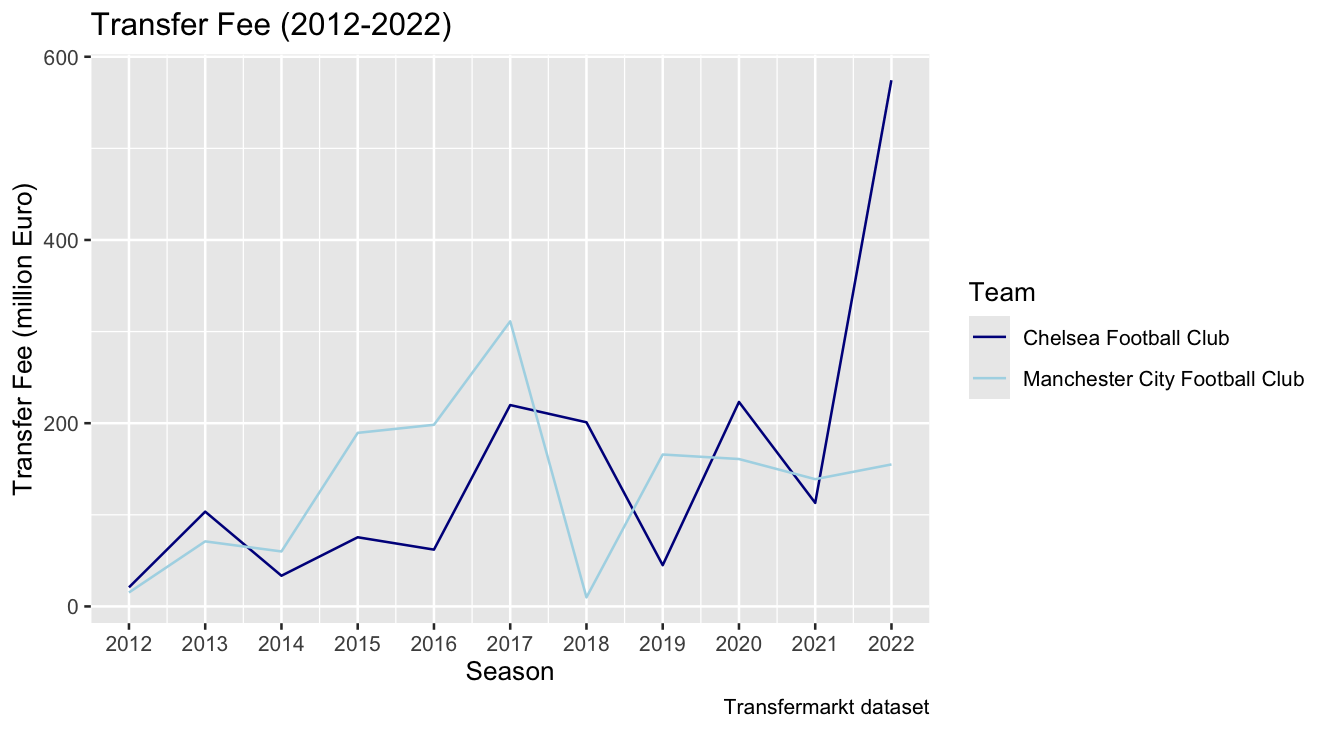


**Figure 4: Number of Wins (2023)**

The goal of the visualisation is to determine whether Chelsea or Manchester City Football Club has more wins by comparing the number of wins between Chelsea and Manchester City Football Club. Consequently, a bar chart illustrates numerical values for various category items. The chart features line marks (bars) where the size characteristics (length or height) are utilised to represent the quantitative value for each item. Moreover, it enhances both broad understanding and detailed reading, which improves the viewer's perceptual accuracy (Kirk, 2019). Nevertheless, it does not elicit happiness, as Kirk (2019) mentions that a perspective on data visualization suggests that anything can be represented as a bar chart. This potential is genuine, yet it remains an idea that lacks joy. In this circumstance, the number of wins represents the quantitative value and Chelsea and Manchester City Football Club represent the category items. Then, it compares the length of the vertical bars to conclude whether Chelsea or Manchester City Football Club has more wins than the other.

Additionally, several alternatives to the bar chart can enhance visual appeal. The project opts for a Treemap and a Donut Chart. The first alternative, the Treemap, illustrates category items by assigning larger sizes to those with greater values and smaller sizes to those with lesser values. This design allows the audience to perceive significant differences in quantitative values easily. It is particularly effective with a small number of category items, making all items visible, which is ideal for this project since there are only two categories. However, the Treemap is unsuitable for datasets with many category items or time series information. While it can display multiple category items, it becomes challenging for the audience to discern significant differences among the values. Additionally, it cannot represent information from a single category item over time within a box (Liu, 2022). The second alternative is the Donut Chart, which functions similarly to the Treemap but uses arcs instead of rectangles. This chart also works best with a limited number of category items and effectively illustrates the distribution of total quantitative values as a percentage (full circle). In this case, where the project has only two category items, the audience can easily understand the distribution of quantitative values in an arc. However, like the Treemap, the Donut Chart can accommodate only up to four category items, as it becomes difficult to read the labels and interpret the quantitative values of each group (Liu, 2022).

# Implications and Improvements



**Figure 5: Transfer Fee (2012-2022)**

First, the project examines the ethical considerations surrounding data visualisations. It is essential to recognise six common types of misleading visuals associated with unethical data practices: scale distortion, data not adding up, arbitrary dual y-axes, manipulated y-axis, 3D distortion, and omitted or concealed data (eisquare, n.d.). However, the project focuses on omitted or concealed data, which occurs when only data supporting a specific narrative is included. In this project's circumstances, the visualisation consists of the transfer fees of Manchester City and Chelsea from 2012 to 2022. It informs that Manchester City spent more than Chelsea in six out of eleven seasons, while Chelsea spent more in five seasons. This result leads to the conclusion that, over the long term, Manchester City has maintained higher transfer fees than Chelsea, which may have influenced their performance in 2023. Thus, it reinforces the narrative that high transfer fees can positively impact long-term performance. However, when the project expands its analysis to include transfer fees from 2012 to 2023, it informs that Manchester City and Chelsea each spent more in six out of eleven seasons. This result challenges the conclusion that Manchester City has consistently maintained higher transfer fees than Chelsea. As a result, it cannot support the narrative that high transfer fees directly influence long-term performance, nor can it explain why Manchester City has more wins than Chelsea in 2023.

Second, the project explores ways to improve the visualisation. While several aspects could be addressed, the focus here is on enhancing the visual design specific to this topic. One alternative design to consider is a clustered bar chart. In this chart, the seasons would be plotted on the x-axis, and the transfer fees would be represented on the y-axis. Each season would feature two bars, with colours defined in the legend. The dark blue bar would denote Chelsea Football Club's transfer fees, while the light blue bar would represent Manchester City Football Club's transfer fees.

# R Code, GitHub Pages

The source code in SQL, used to integrate and transform data, and R, used to visualise the insights for this project, can be found in the SQL and R folders. The cleansed data used to analyse can be found in the Data folder, while the report of words can be found in the Report folder. For more information, please refer to the Readme.md file. Here is the GitHub link:

# References

* Tomkins, P., Riley, G., & Fulcher, G. (2010). Pay as you play: The true price of success in the Premier League era. GPRF Publishing
* Brandon, B. (2024). Does money really buy success? Empirical analysis of money spent effect on English Premier League team’s performances/success: A panel data analysis. Applied Economics. Bryant University. Retrieved from <https://digitalcommons.bryant.edu/cgi/viewcontent.cgi?article=1233&context=eeb>
* Ferster, B., & Shneiderman, B. (2012). Interactive visualization: Insight through inquiry. The MIT Press.
* Thomas, D. R. (2006). A general inductive approach for analysing qualitative evaluation data*.* American Journal of Evaluation, 27(2), 237–246. <https://doi.org/10.1177/1098214005283748>
* Kirk, A. (2019). Data visualization: A handbook for data-driven design (2nd ed.). SAGE Publications.
* Albers, J. (1963). Interaction of color (3rd ed.). Yale University Press.
* Wickham, H. (2010). A layered grammar of graphics. Journal of Computational and Graphical Statistics, 19(1), 3–28. <https://www.jstor.org/stable/25651297>
* Inforiver. (n.d.). Guide to bar charts: When to use them & how to design. Inforiver Insights. Retrieved December 7, 2024, from <https://inforiver.com/insights/guide-to-bar-charts-when-use-them-how-design/#:~:text=Bar%20charts%20are%20some%20of,and%20are%20versatile%20in%20use>.
* Liu, X. (2022, December 5). 5 alternatives to the bar chart. Dev Genius. <https://blog.devgenius.io/5-alternatives-to-the-bar-chart-f53660c0e57f>
* **Eisquare. (n.d.).** Ethical data visualisation: Key considerations [Blog post]. Eisquare. Retrieved December 7, 2024, from <https://www.eisquare.co.uk/blogs/ethical-data-visualisation#:~:text=The%20main%20ethical%20considerations%20during,accurately%20show%20the%20underlying%20data>.