Fundamentals Of Information Visualisation

Introduction

“The greatest value of a picture is when it forces us to notice what we never expected to see". [1] Words from John Turkey which encouraged me into creating visualisations on a topic I am fervent about, football. I hoped to encounter patterns/ trends I did not expect, coming from someone who believes they have great knowledge about the sport.The data being used for this project are player statistics from the 2021-2022 season across Europe's top 5 leagues.

Data Description

The data is heavily detailed consisting of 143 columns regarding player details and statistics, the column description can be seen in Appendix[A].

There are 2921 records within the dataframe, only 1 row contains a null value, due to such a small number I removed the value.This was useful as it reduced the amount of data wrangling needed (mean/median imputation .Figure 1 is a table which shows the sum of rows that do/don’t contain null values.I preferred tabular format to graphical representation as I think viewers are able to process this information quickly without a potential cognitive load, it is also easier to compare the values through a table.

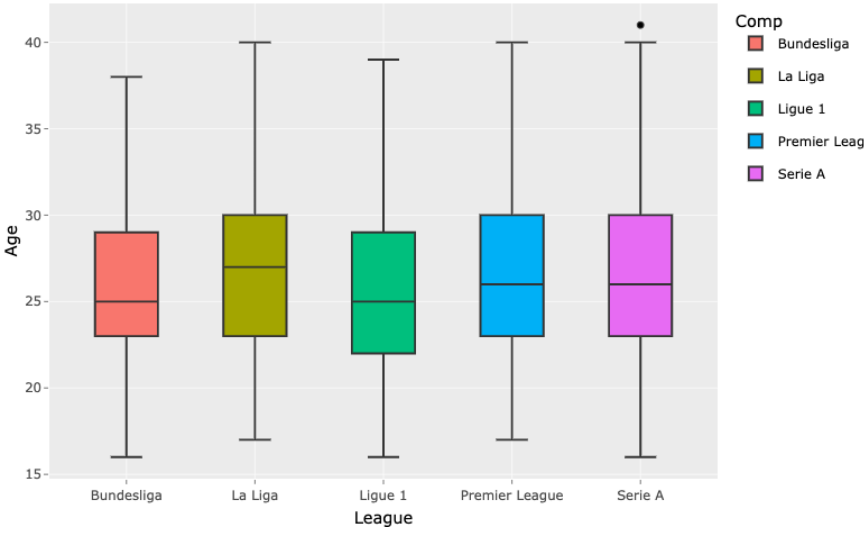
valid <chr>
null
not null

2 rows

(Figure 1)

As age is a major factor in my research questions I thought it’d be best to have an overview. I decided to use a box plot as it allows comparison of different categories

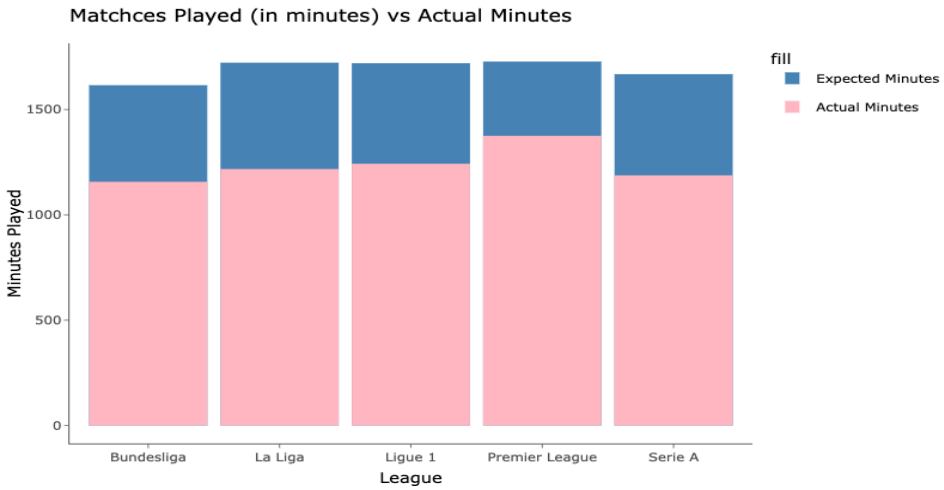
(Figure 2)



easier and in the same graph. In figure 2 you can see the average footballer age ranges from 23-30, footballers in these leagues tend to start at 16 and retire in their late 30’s.

A final thing that came to mind was the matches played and minutes played column.I wanted viewers of my visualisation to understand that matches played is not the most accurate factor to use as a player can have played 30 matches yet only have a total of 30 minutes played. Figure 3 shows in each league the comparison between the average matches played (in total minutes for easier comparison) and average minutes played.The graph solidifies my thought as on average the players

play around 70% of the matches they're said to have played.



(Figure 3)

Initial Questions

Using this statistical data I aim to with focus on how factors like age, discipline and statistics affect and reflect player performance.Specifically I will aim to answer these questions :

1. Does player age determine the frequency in which players dribble with the ball?
2. Do younger players tend to be offside more?
3. What player in the Arsenal Squad had the worst player misconduct?
4. Which Arsenal Midfielder was best in possession?

My reason for choosing question 1 was because there I could think of logical arguments as to how age does or does not affect a player's need to dribble. This reminded me of a quote '*The first question you should always ask yourself is whether visualizing your data is really necessary. Would it make the story easier to understand?* '[2] I believe this question and the latter all fit into this description.
I chose question 2 as there is no obvious answer, common thought would suggest younger players are more naive and wonder offside more, however it can be argued older players may be less mobile and are caught offside more .
I chose question 3 in order to focus on one team and create a multivariate data visualisation. Similar to question 3 I wanted to do the same but on a larger scale with more variables..

Visualisation Strategies

"Each set of data has particular display needs, and the purpose for which you're using the data set has just as much of an effect on those needs as the data itself."[3] This quote resonates well with me as during the visualisation process I found that for each question required various different techniques in order to obtain meaningful results from the broad set of data I initially started with.

From Appendix B I first loaded the dataset from the project directory.I came to notice some of the character encoding was inaccurate, the accents on player names was not being recognised by R due to the incorrect formatting being used, later on this would lead to problems trying to create graphs when displaying player names therefore to fix this I applied **iconv()** function which manually converted the character encoding from "ISO-8859-1", to = "UTF-8".
I searched for null values within the data frame so I looked for entries with null values using !(not) on **complete.cases(df)**, this returns a FALSE if at least one value on a row is missing, with this I was able to find 1 record that had null values. I just removed this value as it was less than 1% of the dataset and would not skew my results.

Question 1

Firstly I used a transformation technique which is part of the dplyr package in R studio called **select**(). This trimmed the data frame to

consist of relevant values to help produce a visualisation regarding player age and player dribbling attempts per game.I performed a data transformation technique from the same package called **Cut()**. I decided to use it on the Player age variable in order to split the player ages into groups, I used groups of 3 as I felt players within 3 years of each other tend to play similar as they have similar amounts of experience. I used the **seq**() method in order to perform the group splitting with **cut**().I made a new column called age group rather than overwriting the age one so I could double check results. I further transformed the data using the **group_by**() and **summarise**() methods concurrently with pipeline operators in order to pass in my original data frame to produce a new data frame which consists of all the age groups and their mean average dribble attempts. This was obtained using the **mean**() method on the dribble attempts.Finally I then plot the averages of each age group as a bar chart using **ggplot**() and **ggplotly**() to make it interactive.

Question 2

Select() was used to narrow down the data frame. I used a new transformation technique **subset()** in order to filter the data, I wanted to only select players who had at least played 5 matches, reason being I didn't want bias in the results I achieved, some football teams play extreme “high lines” which cause players to be offside more frequent and the other extreme is a “low block” where there is less space behind a defence for a player to be offside therefore I wanted a vast amount of teams for each player to have faced before taking results into consideration. I then decided to use the subset() method again, reason number one being I wanted to reduce the number of data entries so my visualisation can be clearer and in order to target players where receiving balls near an oppositions backline is required, this will quickly remove players who play in positions like Goalkeeper or defence.With a more specific and accurate dataset I then plotted my interactive scatter plot.

Question 3

For question 3 I narrowed down using **select**(). I then used %>% with the **subset**() method to filter through only arsenal player who have at least played 5 games in order to give more accurate results, I then wanted to arrange the dataframe by the red card column in order to see the actual values of the red cards so i know how to visualise my data, this was key because when plotting a multivariate data chart you need to make sure the values align with the factor that represents it in the chart (e.g shape or colour).I then used the **names**() method to further transform the data frame, this was to make visualisations more clearer.The data was then ready to be plotted.

Question 4

I decided to carry out majority of transformations in one go using %>%, I used subset() method in order to select only Arsenal players, in the same method I used boolean operators (& |) in order to select players whose primary position is midfield.From this subset I then used the **select**() method to narrow down the dataset and get rid of redundant data, I then performed a data transformation using the **mutate**() method. This was specifically used because the PasTotCmp.column was in percentage format (a whole number). In order to fit the format for visualisation better and make it less confusing, I needed the data in decimal format, hence I divided all the results in that column by 100. I used the **names**() method to rename the attributes.

Exploration Of New Questions And Visualisations

To my surprise, for Q1, there was no correlation between players being offside and

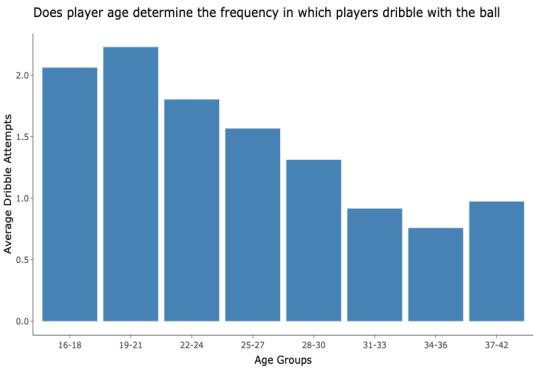
their age, this led me to think it may have nothing to do with the players physical ability to stay onside, maybe it's a mentality/ workrate relation. In football being offside can be seen as being lazy as you're too sluggish coming from an offside position. I looked along the data frame attribute list and came across presses per game, players who work hard press a lot during a game, it is not possible to be lazy and press a lot therefore I posed the question **(Q5) - Do Players Who Tend To Be Offside More Press Less?** The process to moulding the data into shape was fairly simple, using the piping operator I first transformed the data using **select**() to pick out relevant attributes (Offsides and Presses), I then used a transformation to filter the data twice, firstly to select players who have played 5>= matches (again for that variable of reliability) and then selected players nearing 1+ offsides a game to make sure the chart was not too clustered.

Q3's visualisation surprised me by how regular players foul in the Arsenal squad. It made me think about the possible causes as to why they would. **(Q6) - Do players who get dribbled past more often tend to foul more ?** To prepare the data for this scatterplot visualisation I transformed the data using the select() function to select relevant data, this then led to me filtering the data using subset, I only chose players with 15+ games and more than 2 fouls per game in order to select players that foul the most and reduce clustering on the chart. I then used **names**() again to rename the data attributes for clarity.

My favourite visualisation was the parallel plot, I felt it was a good multivariate data visual as it was clear and allows a big group to be plotted, context wise it does well at identifying weaknesses in players games. It was great being able to see how Arsenal players compare against each other, however this led me to **(Q7) - Where do Arsenal's Midfielders Rank against Europe's top 5 leagues?** I used a range of transformation methods with the pipeline operator, i used the **subset**() method to pick all players that do not play for arsenal, then narrowed down using select().I altered the attribute for total pass completion, I divided all data entries for that column by 100 in order to turn a percentage to a decimal so that the visual encoding variable fits better.To finally prep for plotting the parallel plot the **group_by**() and **summarise**() methods are used together to produce a data frame containing all 5 leagues and their averages for the relevant midfielder stats.

Analysis of Visualization Design

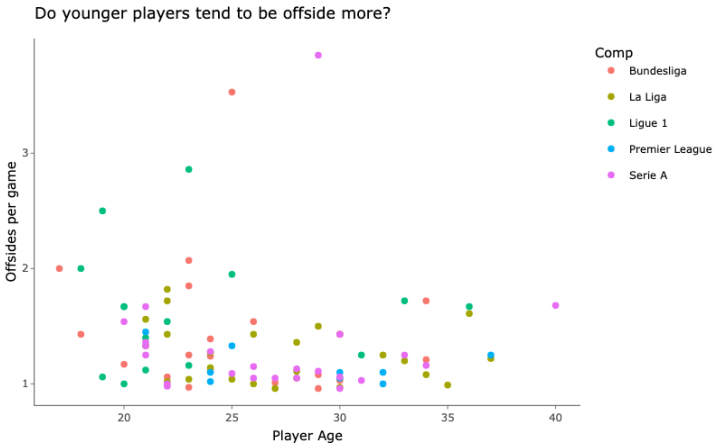
For Q1 the bar plot, which can be seen in figure 4, was used because I was comparing quantitative ratio data against categorical data hence a bar plot was more suitable than a histogram. By the bars being the same colour , with reasoning from Gestalts Grouping Principles, "similar elements are visually grouped", "They can be grouped by color, shape or size".[4] In addition I felt the bar plot is clear and cognitive processing time is instant at showing younger ages tend to dribble more.



(Figure 4)

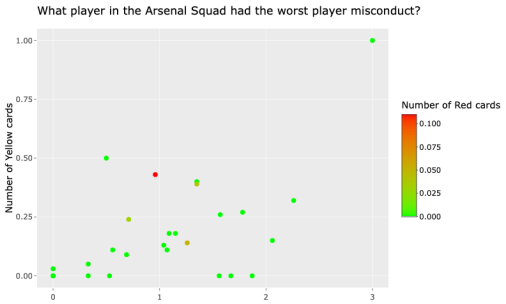
For Q2 I decided to use a scatter plot, as there were many points of data being shown yet a scatter plot is able to display them all. On the opposing side to the similar colours for the bar plot, different colours were used to show the League each point belonged to, I feel the legend gives extra clarity on this. The graph is also interactive, the ability to filter between

competition is a bonus as relations can be seen for each league separately or as a group, interactivity also gives the option to seek more information and potentially provide insight into another exploratory visualisation.The ability to zoom on interaction also helps in making the visualisation more understandable.



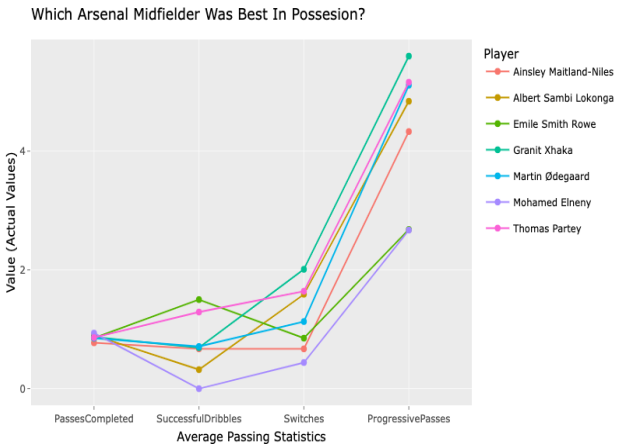
(Figure 5)

For Q3 although similar to the previous chart, Q3 is a multivariate scatter plot which uses the axis to compare committing fouls against the number of yellow cards acquired, shown below in Figure 5.I modified attribute names to give better information to the user for the interactivity.I used the Red Card attribute for colour visual encoding for a few reasons. Some rows have a column value of 0.0, therefore 0 can still be mapped to a colour gradient scale however if i were to use the red card attribute to be responsible for the dot size, 0.0*dot size would not work out as the dot would not be shown.Furthermore the gradient scale goes from green to Red, this was done on purpose in order to speed up cognitive processing as colour "has been found to play a significant role in enhancing memory performances"[5].Humans tend to associate Green with Good and Red as bad, it just so happens that the card in football for misconduct is red as well.

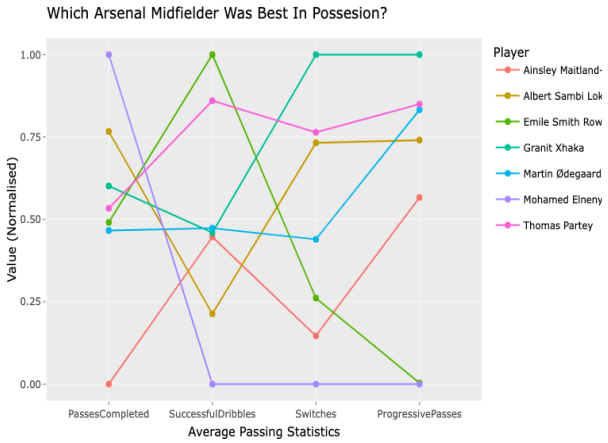


(Figure 6)

Q4 is a very effective multivariate visual as it allows 4 different attributes to all be compared.I decided to make 2 different charts, one showing precise values (fig 6) and another showing all the values normalised so they can be compared in relation to each other rather than their actual results, happens when the scale is at "globalminmax".The normalised chart (fig 7) is easier to interpret as the "best" midfielder is easier to identify through the straightness of a line and how high up the exis it is, this is due to setting the scale to "uniminmax". The colours were specifically meant to be different to each other as there is overlapping yet you can still clearly identify which player is which line.

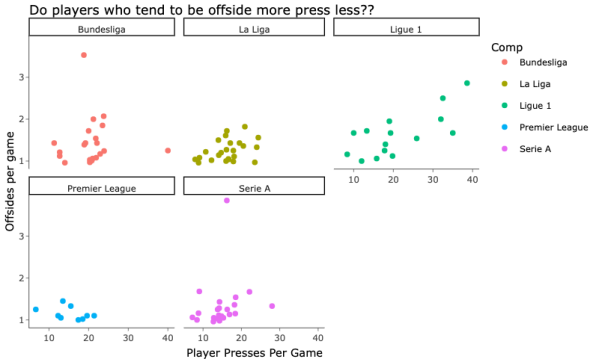


(Figure 7)



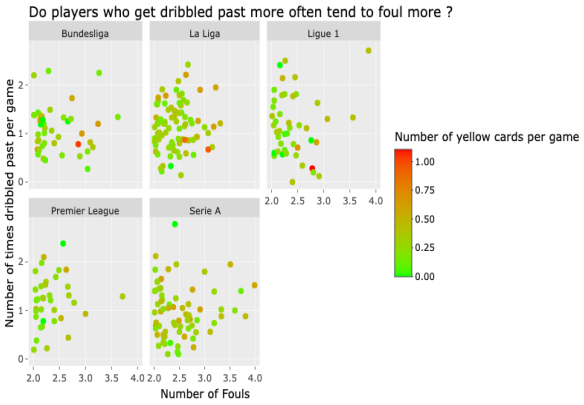
(Figure 8)

Q5 is the first exploratory design which follows from Q2.A scatter plot is used with the facet_wrap() function with the Comp inside creating a trellis Display chart so that there are graphs for each league, by creating small multiples this chart is “visually enforcing comparison of changes”[6].This also allows data to be less clustered and the ability to further explore the visual, potentially noticing trends within each League, for example in Figure 8 you can see a positive correlation which does not appear as clear in other leagues. I also feel the grouped cluster of colours makes processing easier about the separation of leagues.



(Figure 9)

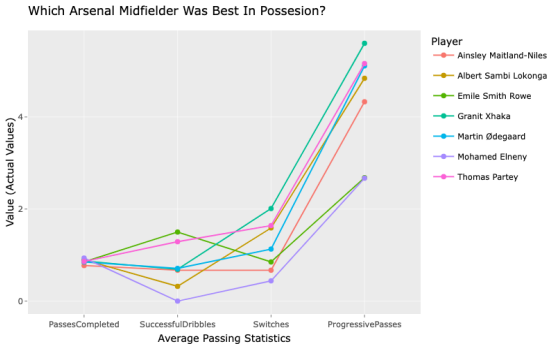
Q6 has a colour scale gradient to measure the number of yellow cards received per game, I feel the colour gradient is clear at showing what number it represents.Trellis display is used, in figure 9 you can see how it separates the chart from clustering and overloading with data.Interactivity provides all the clarity as the hover presents all the information that may not be so clear on first sight.

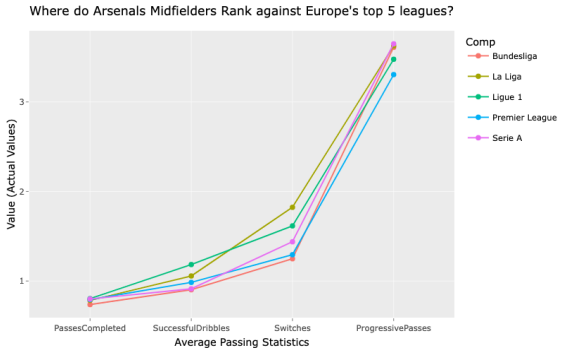


(Figure 10)

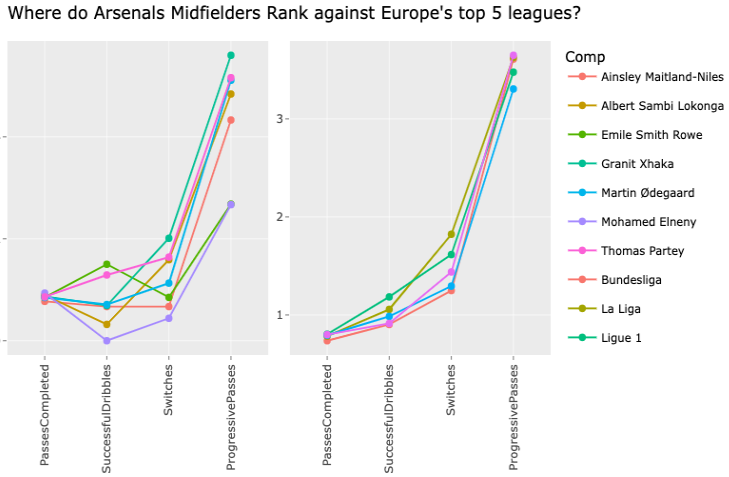
Q7 works in the same way as question 4, except both graphs are also shown in parallel using the **subplot**() method.This encourages comparison of values . Looking at the separate graphs (Fig 10 & 11) you're able to colour coordination of variables on the side, letting you know they are different.The symmetry of both charts plotted (fig12) eases cognitive processing when comparing graphs as both axis are the same, you simply just compare the line shape and position.The titles along the x-axis were plotted vertically so there is no

overlapping of them, this makes it clearer for people to see.





(Figure 11 & 12)



(Figure 13)

Evaluation

Overall I am satisfied with my visualisations, I believe I performed the correct preprocessing and processing methods before plotting my data, I found the plotly package useful due to its compatibility with ggplot when making interactive visualisations. I feel my visualisation choice was acceptable but am wary more variation could be used, I attempted to make a Chernoff Faces chart when answering q3 with regard to player misconduct.My vision was for the colour of a face to represent the number of red cards, the eyebrow slant to represent yellow cards and the frown degree to be represented by Fouls committed.It would've been an accurate representation of the player misconduct, the angrier the face the more misconduct, I would then plot these faces on a Map in each country where the competition is (La Liga = Spain, Bundesliga = Germany)..I'd say some visualisation lessons that were relearnt over the course of this project was the emphasis to keep my visualisations simple but informative, I now understand the importance of interactivity to make this possible.In terms of future work I could use another platform or software to make the Chernoff faces chart if there is the ability to alter features yourself.I also can take the idea from Figures 12, 6, 7 to purse my hopes of becoming a sports analyst by creating unofficial "scout reports" on current football club players or potential signings . I want to also have a go at working at a dataset with some time element, this

would allow me to make animations showing the effect of time on a given dataset.

References

- [1] Turkey J (1977) Exploratory Data Analysis
- [2] Hannah Williams () Data Visualization Basics - Infogram [https://infogram.com/blog/tutorial/finding-the-data/the-basics-of-data-visualization/]
- [3]Ben Fry (2007) Visualizing Data
- [4]By Cameron Chapman()Toptal - Exploring the Gestalt Principles of Design [https://www.toptal.com/designers/ui/gestalt-principles-of-design]
- [5]Mariam Adawiah Dzulkifli and Muhammad Faiz Mustafar(2013)- The Influence of Colour on Memory Performance: A Review[https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3743993/]
- [6]Edward Tufte() Visual Display of Quantitative Information

Appendix

- [A] https://www.kaggle.com/datasets/vivovinco/20212022-football-player-stats (Football player stats 21/22)
- [B] Comp3021 Notebook [20265875code.Rmd]
- How to View Code
    - Open up NoteBook
    - Run each chunk at a time (make sure to install packages if need be)
    - You can view Visualisations there
- [C] Web Page Showcasing Visualisations
    - How to view
        - Access FIV_HTML folder
        - Open index.html in any browser to view visualisations
        - (If Visualisations do not appear at first, open up the charts folder and view the charts, then go back onto index.html)