INFOOTMATION

All the information you need about European football players in a single place!

Contact Information:

Yann Dubois: yanndubois96@gmail.com

Domain and task:

With an estimated 3.3-3.5 billion fans over the world [9], Football (soccer) is by far the most popular sport of our time. “The Beautiful Game”, as they call it, is being watch and played on every continent: it represents dream and joy for kids in countries as north-east as Japan and as south-west as Argentina but nowhere on earth is this game more popular than in Europe [4]. You would think that for a sport which attracts so much attention: data would be abundant and every possible visualisation already implemented. Unfortunately not, this is probably due to the fact that football also has the biggest number of professional leagues: with 170 in 2014 [6]. Each of these leagues has its own data that it is rarely released. As a result, getting the data is hard. Professional coaches and the media get the data from expensive sources and amateur are left with basic information. This is a big disappointment to football amateur, indeed who doesn’t like to brag about new (useless) stats he has recently seen? Who doesn’t like to get a sneak peak of the information coaches have at their disposition to see how they would act under the same circumstances? That is exactly what I would like to bring at the table. A visualisation method of players’ stats from the 11 biggest European Football Leagues. The visualisation is primarily intended to satisfy the curiosity of football fans that are eager to have an in-depth view of their favourite / most hated players. I will first try to support the following tasks:

- Compare the number of taken and made shots by type (right foot / left foot/ head / other).
- Compare a player skill to the average skill of other players.
- Discover the number of goals by competitions.
- Summarize basic players’ statistics.
- Discover the position of the player when he scored his goals.
- Compare the players’ performance depending on the opposing team.

Other tasks I will try to answer if I have the time, include (in order of priority):

- Appreciate the same visualisations stated above but with daily updated stats.
- Discover more advanced and useful stats of every player depending on their position (e.g. number of won duels for forwards, number of interceptions for midfielders, number of caused penalties for defenders, percentage of stopped shots for goalkeepers).
- Summarize the awards and team history of the player.
- See the salary of a player and his current “value”.
- Discover the number of tweets (and the main tweeting subject) on a player over time.

- Discover the average distance ran by a player and the positions on the field he goes the most to.
- Investigate the player to whom player X passes the most.
- Analyse the main direction of shots of a player and the balls final location.

The data:

The visualisation will be made possible thanks to the "European Soccer Database" which was obtained by web scrapping and released by Hugo Mathien on Kaggle in July 2016 [3]. The dataset contains information of 25'979 matches and 11'060 players from 2008 to 2016 seasons. Some of the interesting data that I will be using from the dataset includes the players: name, team, birthday, weight, height, potential, overall rating, preferred foot, rating of multiple skills (passing, shooting, attacking rate, shot power …), number of scored goals in each game and position when scored goal.

I will use this dataset for most of my project, but time-permitting I would like to extend the given dataset. Indeed, the data set is unfortunately not regularly updated. I will thus try in a second step to use the web scraping scripts written by Hugo Mathien to obtain a daily updated version of the data set. In a third step, that I will probably have to do after the end of the project, I would like to write new web scraping scripts to get additional information. The websites that could be of interest are: http://www.footytube.com/openfooty/ , https://openfootball.github.io/ , http://www.foxsports.com/soccer/ and http://www.squawka.com/players/ . I also wrote to sports newspaper and companies collecting sports data hoping that they would give me access to part of their more advanced and updated data.

Personal expertise:

My expertise in the domain is restricted to the fact that I have played 10 years of soccer. As most of my friends are big football fans, I can obtain a detailed feedback from my targeted users.

Concerning the technical background, I have neither web nor D3 experience, but I already started to learn them for the project. I do, at least, have a decent programming and data analysis background as well as the necessary SQL skills.

Proposed infovis solution and scenario of use:

The visualisation proposed, targets football fans that want to enjoy visualizing statistics of their favourite, most hated or other interesting players. It mainly tries to show all available statistics at the same place and in an enjoyable way. Below, I take every targeted task and I try to explain the proposed infovis solution as well as a scenario of use.

For the scenario of use, let's suppose that *Ben* is a huge Bayern Munich fan. Let's also suppose that *Sacha* who hates Bayern's best Player, Robert Lewandowski, goes to have a lunch with *Ben*. They start what seems to be a peaceful lunch but suddenly start to talk about football. *Sacha* states that Robert Lewandowski is far from being the current best forward player. In order to prove him wrong *Ben* decides to show him Lewandowski's

infootmation page[1]. They see a single page subdivided (using juxtaposed facets) into multiple visualisations that targets specific tasks.

First, I will concentrate on what they would see using only the available "European Soccer Dataset". Note: in the "What: Data" part, I always describe the final data that will be visualised, but most of the time extensive data processing will be needed to obtain required form (especially true for position).

Compare the number of taken and made shots by type

What: Data	A table containing 4 rows: left foot / right foot / head / other and their corresponding number of shots and goals from 2008 to 2016. The data will also have a column indicating the competition in which the shot was taken (for filtering, see below). The weight and height of the player. It therefore consists of quantitative and categorical attributes.
What: Derived	Computes the proportion of goals (simple division).
Why: Tasks	Presents the most effective type of shot of player X and compare it to previous seasons.
How: Encode	Size: number of shots. Hue: Percentage of goals. Position: Type of shot.
How: Reduce	Embed: superimpose. Filter: this chart will be linked to Fig.3. and its data will be filtered depending on the competition that is currently selected by hovering over.
How: manipulate	Change view over time.

[1] Note : every stated number is completely made up, but it gives an idea of possible numbers.

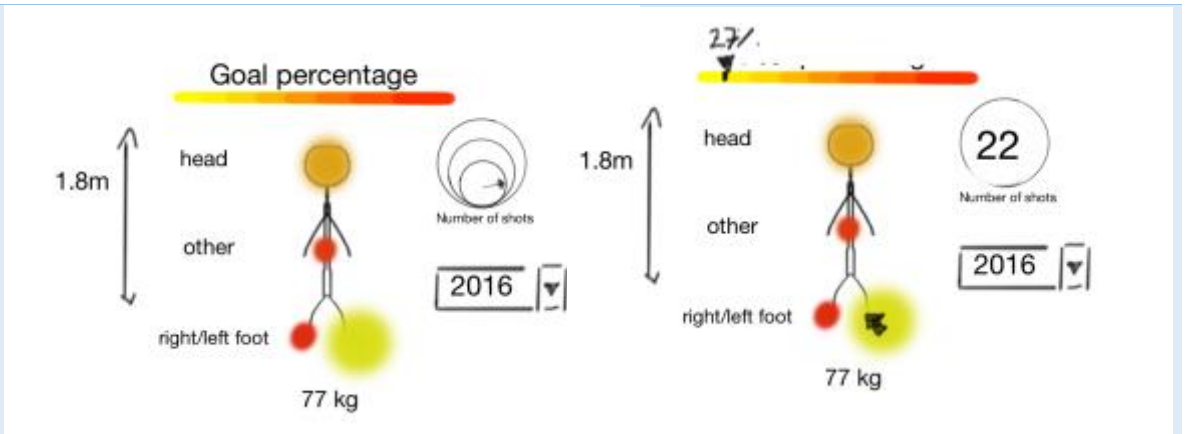

Figure 1: Compare the number of taken and made shots by type

Sacha argues that Lewandowski uses to much his left foot. He thus shows *Ben* the above visualisation (left image) that clearly shows that Lewandowski uses a lot his left foot although it has a bad goal percentage. To mark his point, he hovers over the left foot (right image): which gives the exact number of shots and goal percentage. *Ben* has to agree on the 2016 season, but he argues that this is only true in the selected season. He thus click (not shown) on Lewandowski's left foot. This initiates a changing visualisation that shows how the statistics changed over the 8 seasons.

Compare a player skill with the average skills of other players

What: data	A table containing a row for each 11'060 players and 35 columns containing a score for each EA GAMES player skills. Quantitative attributes.
What: Derived	Mean score for each skill of every player of a certain position.
Why: Tasks	Presents the skill of a player and compare with the average player at that position.
How: Encode	Position: coding the score value; radial layout. Hue: coding categorical data: average or player.
How: Reduce	Embed: superimpose. Filter: show only 5 important skills (to maximize readability). One can choose between general / best / worst and a custom one.

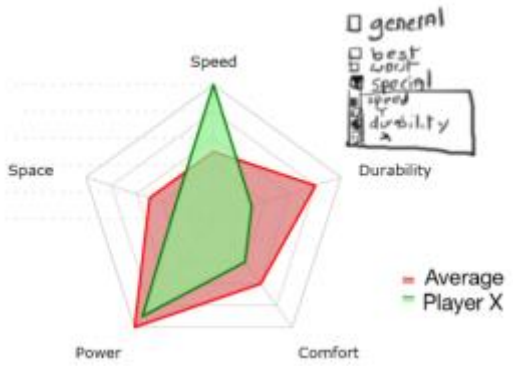

Figure 2: Compare a player skill with the average skills of other players. Image modified from: http://blog.scottlogic.com/archive/2011/09/radar_chart_visiblox.png

Initially the radar plot shows the 5 most general football skills depending on the players' position. As Lewandowski plays as a forward it shows his: speed, accuracy, dribbling skills, shot power and finishing score. *Sacha* quickly checks the "worst checkbox" which changes the shown skills to indicate Lewandowski's worst skills compared to other players of his position. *Ben* finds that unfair and thus clicks on the "best checkbox" that does the opposite. After going through some custom skills, they choose to go back to the general one, as it is summarizing the important skills. The data doesn't prove that Lewandowski is the best forward player, but it clearly shows that he's well above average for every important skill.

Discover the number of goals by competitions

What: data	A table containing a row for each competition (depends on the player and team) and the respective number of goals the player scored in it. It therefore consists of quantitative and categorical attributes.
Why: Tasks	Presents the number of goals scored by competition.
How: Encode	Length: proportional to the number of goals; radial layout. Hue: indicates the categorical variable of the competition.

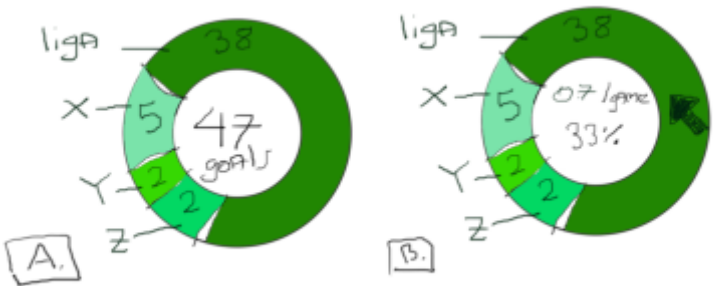

Figure 3: Discover the number of goals by competitions

Ben decides to show *Sacha* the number of goals scored by Lewandowski: scoring goals is indeed essential for being a good forward player. The diagram above clearly shows that

Lewandowski scores a lot of goals in every competition he plays in. *Sacha* doesn't agree with *Ben*: by hovering over the "liga" competitions, he sees that Lewandowski only scored 33% of his shot in this competition. He argues that the number of goals is a biased stat: you cannot only look at the goals without taking into account the number of shots!

Note: hovering over the competitions will also dnamically modify Fig.1. and Fig.5.

Summarize every basic players statistics

What: data	A table containing a row for each 11'060 players and columns showing multiple stats: number of goals, number of assists, number of interception …. Quantitative attributes.
Why: Tasks	Presents the most important stats depending on the position of the player. Investigate the correlation between these variables.
How: Encode	Simple table showing the real numbers. With colour coding to indicate how he compares to the average. Possibility to use a scatter plot to investigate the correlation between these variables. (Position)

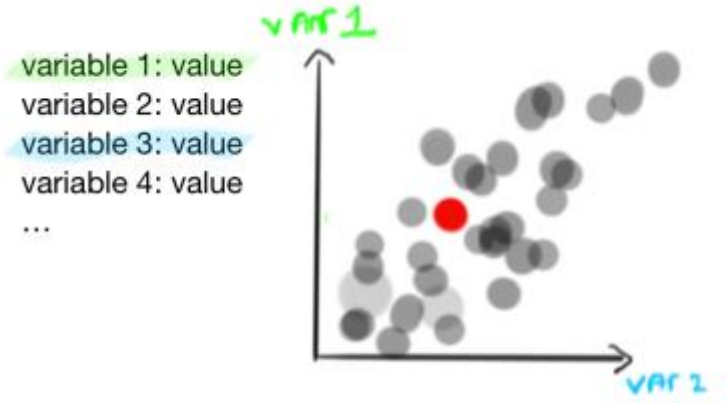

Figure 4: Summarize every basic players statistics

These basic numbers are a summary of the most important statistics you can find on internet. Both *Ben* and *Sacha* are happy to look at them, as it can make them seem more knowledgeable about Lewandowski. Neither of them remember these essential statistics and this allow them to concentrate on more advanced tasks without continuously looking up the basic statistics on internet.

Discover the position of the player when he scored his goals

What: data	A table containing a row for each goal scored by the player. The columns would give the X and Y coordinates, as well as the

	competition in which the goal was scored. Other columns would be used for filtering and giving additional information when hovered over: the year of goal, opponent team, competition and time of the game when the shot was taken. It therefore consists of quantitative and categorical attributes.
Why: Tasks	Presents the position at which the player scores his goals.
How: Reduce	Filter: this chart will be linked to Fig.3. and Fig.6., its data will thus be filtered out depending on the competition or team that is currently selected by hovering over it.
How: Encode	Position: the (X;Y) coordinates of the player when he shot the goal.

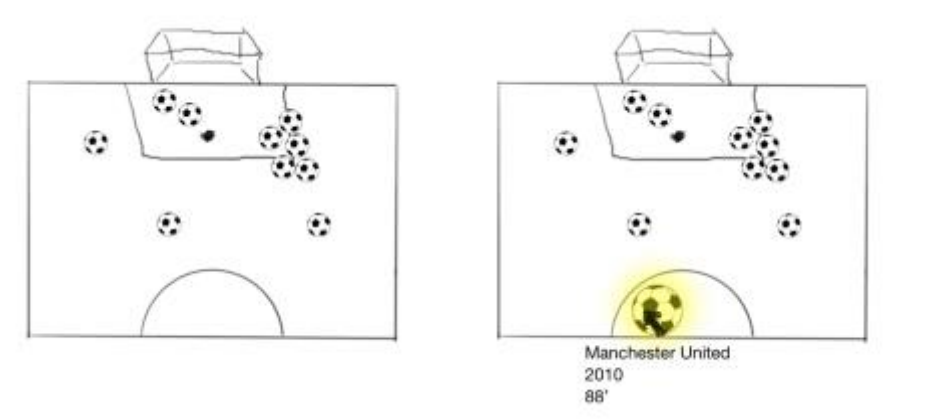

Figure 5: Discover the position of the player when he scored his goals

After arguing for a long time, they still can't agree on whether Lewandowski is the best forward player of their time. At that point they see an interesting visualisation showing the position of Lewandowski when he scored his goals. *Ben* asks *Sacha:* "Do you remember when he scored against Manchester United in 2010 from half court ": he hovers over the ball near half court and sees that he's right: the shot was made in 2010 against Man-U at the UEFA Champions League (88th minute of the game). They start playing quizzes trying to remember against who / when was that shot made.

Note: the data will be filtered by hovering over Fig.3. and Fig.6.but hovering over a ball will also highlight the competition in Fig.3. and the opponent team (if present) in Fig.6.

Compare the players' performance depending on the opposing team

What: data	A table containing a row for each team against which the selected player scored. The cells will simply contain the number of goals scored against the team. It therefore

	consists of quantitative and categorical attributes.
Why: Tasks	Compares the players' performance depending on the opposing team.
How: Encode	Vertical position: Line marks express number of goals with aligned vertical position, separate team are categorical attribute with a horizontal position.

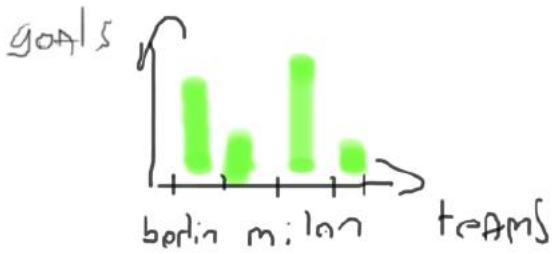

Figure 6: Compare the players' performance depending on the opposing team

As *Ben* and *Sacha* see that Lewandowski scored most of his goals against Milan, they are very happy. Indeed, both do agree on the fact that they hate Milan! They hover over the Milan bar to show only the goals in Fig.5. that were scored against Milan.

If I have the time to make a script that makes a daily update of the given dataset, the same infovis solution as above would be used but the data would be updated every day. Below are the tasks that I will try to answer if I can get additional data through web scraping (in order of priority). Note: I'm not sure about the data I will be able to obtain: the description are thus more vague than for the first part.

Discover more advanced and useful stats of every player depending on their position

What: data	A table containing a row for each 11'060 players and columns showing multiple stats depending on the players position: number of won duels for forwards, number of interceptions for midfielders, number of caused penalties for defenders, percentage of stopped shots for goalkeepers …. Quantitative attributes.
Why: Tasks	Presents advanced stats depending on the position of the player.
How: Encode	Simple table showing the real numbers. With colour coding to indicate how he compares to the average. Possibility to use a scatter plot to investigate the correlation between these variables. (Position)

See Fig.4.

Summarize the awards and history of the player

What: data	A table containing a row for each 11'060 players and columns showing the most important awards won by the player as well as his previous teams. It therefore consists of quantitative and categorical attributes.
Why: Tasks	Summarizes the awards and history of the player.
How: Encode	Simple table showing the awards and previous teams with images. Ex: not writing the team but showing the logo.

See the salary of a player and his current "value"

What: data	A table containing a row for each 11'060 players and columns showing the salary of the player as well as the value his last transfer. Quantitative attributes.
Why: Tasks	See the salary of a player and his current "value".
How: Encode	Position: Line Chart which is a dot chart with connection marks between dots. Simple number for the value of his last transfer.

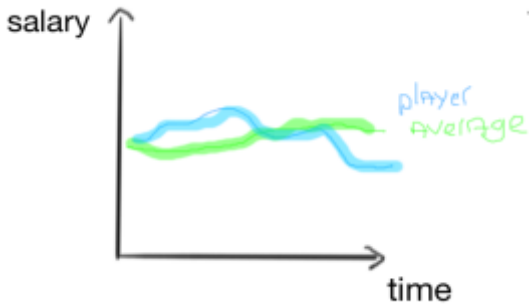

Figure 7: See the salary of a player and his current "value"

Discover the number of tweets on a player over time

What: data	A table containing a row for each 11'060 players and columns showing the number of tweets on the player as well as the most common words in the tweet (besides the players' name). It therefore consists of quantitative and categorical attributes.
Why: Tasks	Discovers the number of tweets on a player over time.
How: Encode	Position: Line Chart, Dot chart with connection marks between dots.

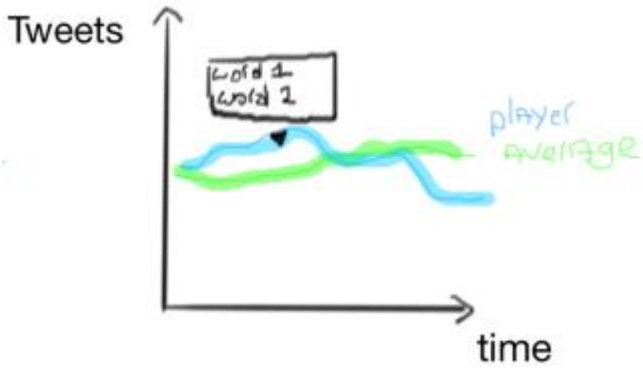

Figure 8: Discover the number of tweets on a player over time"

Discover the average distance ran by a player and the positions on the field he goes the most to

What: data	A table containing the time spent by XY coordinates. Quantitative attribute.
Why: Tasks	Discovers the average distance ran by a player and the positions on the field he goes the most to.
How: Encode	2D matrix alignment of area marks, diverging colormap.

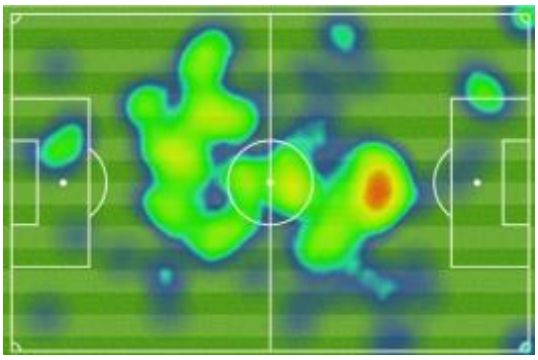

Figure 9: Discover the number of tweets on a player over time". Image from: http://www.adriaandefraeije.com/blog/2015/04/05/heat-maps-for-a-shot-analysis-in-badminton/

Investigate the player to whom player X passes or gets the ball from the most

What: data	A table containing a row for each 11'060 players and columns showing the number of passes between the player and each of his teammates. It therefore consists of quantitative and categorical attributes. It is a weighted graph.
What: Derived	The 3 maximum values of the data per player.
Why: Tasks	Investigate the player to whom player X passes or gets the ball from the most
How: Encode	Size: Nodes linked with connection marks.

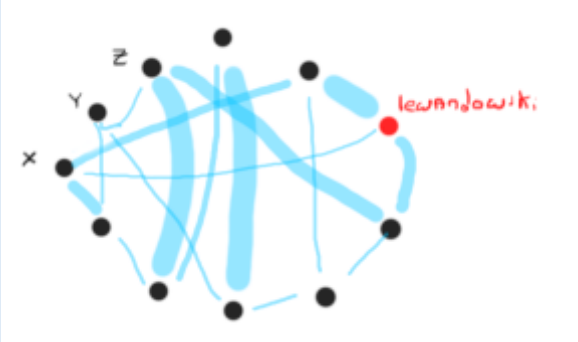

Figure 10: Investigate the player to whom player X passes or gets the ball from the most

Analyse the direction of most shots of a player and the balls final location

What: data	A table containing a row for each shot taken by the player. The columns would give the X and Y coordinates of the players position and of the ball landing, as well as the competition in which the shot was taken and if it resulted in a goal. Other columns would be used for filtering and giving additional information when hovered over: the year of goal, opponent team, competition and time of the game when the shot was taken. It therefore consists of quantitative and categorical attributes. It is a weighted graph.
What: Derived	Computes the proportion of goal (simple division).
Why: Tasks	Analyse the direction of most shots of a player and the balls final location.
How: Reduce	Filter: this chart will be linked to Fig.3. and Fig.6., its data will thus be filtered depending on the competition or team that is currently selected by hovering over.
How: Encode	Position: the (X;Y) coordinates of the player when he shot the ball. Position: the (X;Y) coordinates of the balls landing position. Hue: Goal percentage. Saturation: number of shots taken from that position.

Figure 11: Analyse the direction of most shots of a player and the balls final location

Note: the most probable direction of a ball shot by position Z can be visualised by hovering over position Z.

Proposed implementation approach:

I will be building the entire visualisation from scratch with D3.js. I intend to make it public through github pages for which I will need to use rudimentary CSS, HTML and Javascript. I will use MySQL to store the data that is currently in a .sqlite extension. The basic necessary analysis will be made with R. Time permitting, the web scrapping scripts will be written in Python.

Milestone and schedule:

Note: the unequal workload is due to the fact that I have taken into account my other projects, exams, assignments and activities.

- *March 6th: Proposal.*
- March 9th: Get a good knowledge of the data and make a rudimentary website.
- March 10-13th: Basic implementation of part 1.
- March 20th : Make part 1 nice and clean
- March 25-27th : Finish part 1 if there were some issues. Start part 2 (daily updates) if there were no issues.
- *March 31st: Interim write-up.*
- *April 8-10th: Write the initial paper.*
- *April 20th: Prepare presentation.*
- *April 25th: Final Presentation.*
- *April 27th: Finalise the paper.*
- *April 28th: Final Paper due.*

Previous work:

Football visualisation is a very new field, both for amateurs and coaches. Indeed, on one hand the public doesn't have access to the amount of free data you can find for other sports. On the other hand, the soccer coaches do have the data but have always been more reticent to use data than in other sports [11]. This is changing, coaches are starting to trust

data analyst and some pioneers already started the soccer analysis. We can thus find football visualisation techniques for professionals but they mostly focus on analysing a single game in details [2] [5] [7].

On the web, we also can find some soccer visualisation but most of them aim to visualise World Cup games [10]. The only visualisation that I have found, which try to visualise multiple up to date player stats are Pointafter [12] and Squawka [8] . The former is relatively complete but only shows player of the premier league (British). It also focuses either on the career stats or on the game stats of a player and not on the season data. All the data is also shown with tables or bar charts which takes up a lot of space. Squawka on the other hand is much more comparable to what I would like to implement with infootmation. The data is available by season and for every big European league. Different interesting charts are being used to convey the data. One of the major problems I see, is that they only show a single chart at a time. As we know this is not optimal, comparing data charts is indeed easier if we see them side by side. There is some interactivity but due to the single facet view, the links between plots is not smooth which doesn't make it enjoyable to play around with. Finally, it has a lot of surrounding text, which I find is disturbing.

In other sports, one visualisation I really enjoy playing around with as it is simple but informative is Buckets from Peter Beshai [1].

Bibliography:

1. *Buckets: NBA Shot Visualization. (2017). Buckets.peterbeshai.com. Retrieved 6 March 2017, from http://buckets.peterbeshai.com/*
2. Charles Perin, Romain Vuillemot, Jean-Daniel Fekete. SoccerStories: A Kick-off for Soccer Visual Analysis. Proceedings of the IEEE Transactions on Visualization and Computer Graphics (InfoVis'13), Oct. 2013, Atlanta, GA, USA. IEEE
3. *European Soccer Database | Kaggle*. (2017). *Kaggle.com*. Retrieved 5 March 2017, from https://www.kaggle.com/hugomathien/soccer
4. Kits, F., Kits, 2., 1, F., Updates, F., Updates, M., & Open, A. et al. (2017). *25 World's Most Popular Sports (Ranked by 13 factors)*. *TOTAL SPORTEK*. Retrieved 5 March 2017, from http://www.totalsportek.com/most-popular-sports/
5. Laszlo Gyarmati, Mohamed Hefeeda. Analyzing In-Game Movements of Soccer Players at Scale (Submitted on 11 Mar 2016). arXiv:1603.05583.
6. Ley, R. (2015). *THE WORLD'S STRONGEST NATIONAL LEAGUE 2014 | IFFHS*. *IFFHS*. Retrieved 5 March 2017, from http://iffhs.de/the-worlds-strongest-national-league-2014/
7. Rein, R. & Memmert, D. (2016). Big data and tactical analysis in elite soccer: future challenges and opportunities for sports science. Springerplus, 5(1). doi:10.1186/s40064-016-3108-2
8. Robert Lewandowski | Football Statistics | Form Guide | Squawka.com . (2017). Squawka.com. Retrieved 6 March 2017, from http://www.squawka.com/players/robert-lewandowski/stats#avg.-pass-length#fc-bayern-münchen-(current)#german-bundesliga#22#season-2016/2017#682#all-matches#1-22#average

9. Russell, B. (2013). *top 10 Most Popular Sports | Most Followed Sports | Most Watched Sports*. *Sporteology*. Retrieved 5 March 2017, from http://sporteology.com/top-10-popular-sports-world/
10. Skau, D. (2014). A Look at the Most Interesting World Cup Visualizations and Champion Predictions - ScribbleLive. ScribbleLive. Retrieved 6 March 2017, from http://www.scribblelive.com/blog/2014/06/11/visualizing-predicting-world-cup/
11. *Three At The Back: Accelerating the Pace of Soccer Analytics - MIT Sloan Analytics Conference*. (2017). *MIT Sloan Analytics Conference*. Retrieved 6 March 2017, from http://www.sloansportsconference.com/content/three-at-the-back-accelerating-the-pace-of-soccer-analytics/
12. Zlatan Ibrahimovic. (2017). Premier-league-players.pointafter.com. Retrieved 6 March 2017, from http://premier-league-players.pointafter.com/l/753/Zlatan-Ibrahimovic