Final Report Data Visualization

Vishal Sehgal* 2109374
Hieke van Heesch† 2135945

he report explores the dynamics of player performance during the FIFA World Cup 2022, focusing on how players from the same club perform when representing their national team. The analysis includes match data and player attributes, through data visualization techniques insights are provided into player and team dynamics. The created tool, Club-Dash, enables users to interactively analyze player performance through various visualizations such as radar plots, scatter plots, box plots, and cumulative bar charts. The tool helps to make intricate analysis accessible and engaging for soccer fans.

1 Introduction

One of the biggest sports events of 2022 was the FIFA World Cup, with 32 national teams playing in the heat of the Qatar desserts. From November 20th to December 18th, many soccer fans watched the different matches and saw their favourite players shine in their national teams. After the World Cup, a lot of data became available regarding player performance, including positions, scored goals, assists, passes, and cards received. Next to the national teams, the players have a club, teammates, and coaches they train with during the rest of the year. Fans all over the world might be interested to see how the players of their favourite clubs play against each other in the national teams. Being able to understand the dynamics between players and teams allows fans to appreciate the complexity of the sport even more.
Our visualization allows users to start exploring how players performed in the World Cup relative to their home club colleagues. The visualization aims to analyze player dynamics and strategies since players who are normally colleagues know each other's tactics. Through a visualization, it becomes clear how players who are accustomed to playing alongside each other adapt to facing one another during one of the biggest sports events of the year. The dynamics between positions on the team and other player performance is best suited to see in visualization since it allows for a better exploration of both the club and the national team simultaneously and analysis of the big data sets. Through interactive analysis the fans can analyze their favourite players just like they do in the FIFA game, making the tool familiar and enjoyable.

2 Data Analysis

In this project, we will focus on analyzing and interpreting the dataset provided by the professors, specifically the FIFA Qatar World Cup 2022 data. This dataset includes match data, player data, player images, predictions, team data, a Twitter dataset, historical World Cup data, and penalty shootouts. Later in the project, some additional datasets were added for better insights into the player's ways of football. The essence of our analysis is to leverage various types of data to create an insightful and meaningful predictive model about player and team dynamics.

*e-mail: V.sehgal@tilburguniversity.edu
†e-mail: h.j.m.vanheesch@tilburguniversity.edu

2.1 Data components

1. **Player Data**: This includes detailed information about Players Basic Information such as position, home club, age, and the number of games played, is mostly categorical data. Performance Stats consist of goals, assists, yellow cards, red cards, and various per 90-minute stats. Advanced Metrics include expected goals (xG), non-penalty expected goals (npxG), and expected goals assists. Both Performance stats and advanced metrics are quantitative sequential data that will be used to compare the players. Next up for goalkeepers, there are more advanced metrics specific to their position, which are all quantitative sequential data. Next up, is the data about how the players take defensive action and the possession metrics, which is again quantitative sequential data and can be used to create the dashboard to see what the normal style of the players is and if it is the same across different teams or clubs. Defensive and offensive metrics include tackles, interceptions, shot-creating actions, assists, and penalties. Through new dataset we found attributes like Acceleration, Ball Control, Aggression and Vision, all of these attributes are quantitative sequential data that will enhance the metrics on what players can be compared.

2. **Match Data**: This data tells us more information about the individual games that have been played. The date and time of the match are cyclic and the venue is included in the geometry data since this relates to the position of the stadium. The game statistics include expected goals, score, and the attendance of the fans, the last one being sequential quantitative data. The team performance and defensive actions are important aspects for analyzing how the teams react to each other, by linking the information together a dashboard can be created to compare players from the clubs. The linking of clubs, national teams, possession and attacks is necessary to investigate the difference in players' overall performance and per opponent.

3. **Team Data**: Team data includes details about which players are used in which game and how the rotation of players went. We will link the player data with the national team data and create a dataset per home club, to make the dashboard creation possible. Data on how the team as a whole had ball possession and some defensive and offensive metrics is included in the dataset. The metrics are the same as the metrics that were described under player data, but now it's cumulative.

4. **Additional Data**: Lastly there is additional data that we would use to make the visualization more visually appealing like Player images. A new CSV file was added to the provided datasets in which correct pictures of the players were stored. We will possibly use the Twitter dataset, which contains network data and sentiments about the games per play. The sentiment could be interesting to use if there are national teams with a high amount of players from the same home club. Data in the Penalty shootout file and historical data could be used for team-to-team comparison this could be used for future work

2.2 Analytics Focus

Predictive modeling: Forecasting the performance of players from the same club when they compete against each other while representing their respective national teams. This approach also seeks

to understand how a player's expertise can influence the strategy of their national team during the World Cup. Comparing players who play for the same club that are now representing their national theme creates new insights into the complexity of player tactics.

2.3 Data Handling strategy

We haven't found any missing values, as we only used two files from the team's folder. One of these files required some adjustments, so we removed the serial number column, which contained some incorrect values and was not useful for our analysis. However, if issues arise in the future, we have decided to use data extraction parameters to ensure the data is clean and legitimate. During the creation of the dashboard, we found out that we couldn't compare the players to the intended idea. We decided to find additional datasets on Kaggle: called 2022 FIFA Worldcup Qatar FULL LIVE DATASET, we used the `Fifa_world_cup_2022_Groups` and `Player_data_2022Fifa22_official_data` that would help the user discover more about their favourite players' tactics than just show how many goals they scored in the tournament or the more obscure calculated Xg.

3 Task Analysis

3.1 Domain Specific Tasks

As a soccer fan, the FIFA World Cup is the moment to see players all around the world play against each other. Since players are not in the same league, players who would normally not play against each other would during the World Cup. These facts present intriguing opportunities to explore the interactions between national teams and home clubs, particularly in strategy and tactics. This is why we have devised several visualization tasks that will likely interest users:

1. **Player's Performance Comparison:**
    - **Question:** How does a player perform when playing against a club colleague in the World Cup?
    - **Complexity:** This involves a side-by-side comparison of the player statistics across national team matches and comparing them to players from different clubs.vision, aggression, long and short shots, and strength are important measures to explore the dynamics of colleagues turned opponents during the biggest tournament in football.
2. **Player's per team Comparison:**
    - **Question:** How does the player perform compared to other players of the national team?
    - **Complexity:** Requires analyzing of all players that are playing in the national team on a specific metric to analyze whether they are better in defense or attack.
3. **Similar Players per Position**
    - **Question:** Are their similarities to be found in football style amongst players in the same position?
    - **Complexity:** Requires looking at plenty of data to find similarities. Important metrics are aggression, speed, and ball control to explore how position impacts the player's strategy and see trends amongst players at the same position.
4. **Effect per Club**
    - **Question:** How do the clubs perform in the World Cup?
    - **Complexity:**Involves an analysis of how the different clubs from all over the world perform in the World Cup, are there certain clubs where players have more goals or get more yellow or red cards? This analysis will look at the higher level than the players and combine more data.

3.2 Task Abstraction

All tasks focus on comparing different datasets in their respective scenarios. In our visualization design, user interaction allows for easy comparison and exploration of the data. Our approach aligns with Munzner's principles [1], ensuring that the visualization is not only aesthetically pleasing but also functionally relevant and user-centric. By satisfying the effectiveness and expressiveness principles users directly understand what they are looking at and what features are most important for the analysis. Task 1 requires multidimensional correlation, which should support direct comparison and highlight deviations in a player's performance. By focusing on the exploration of trends between players, positions on a scale will be used to see e.g. how many goals a player scored in comparison to his colleague. On the other hand, in Task 2, visualization will simplify complex relationships and enable users to learn more about trends, patterns, and outliers in players' performances and how this influences the whole team. By comparing the teams to each other and the links between players throughout different clubs. This is a higher-level analysis of network relations and is expressed in terms of offense, defense, and penalties. The third task requires analysis further than the countries and clubs, but focuses on position, by looking at similarities in playing style, fans can get to know more about how their favourite players are training and what the qualities are of good people in a certain position. It is expressed in more than goals but it includes long and short passes, aggression, and strength of the players. The fourth and final task keeps the focus on the clubs and how their players perform in the world cup, fans will be able to see information about their favourite player's combined results. Through intuitive interaction with the visualization, users can decide what feature they want to see from their favourite clubs. The visualization requires a distillation from countries to clubs to subtract the information per feature per club

4 Club-Dash

1. **Choosing your players**

Figure 1: Initial interface to choose players

As described in all the sections above the main focus point is the interaction between the different players and how this influences team performances. In Figure 1 the first part of the dashboard is shown, where users can click on the player. Users have the option to choose from all other players in the tournament. Users can also pick one or multiple of the options provided. Through these steps, the user can see which players played in which national team and which other players of their favourite club they played against. Through linking all the datasets, users can decide where they want to start looking for certain interactions. To start the comparison, users can choose to compare the players based on the country, position, club or completely random. By seeing the options the freedom is given to compare any way users want.

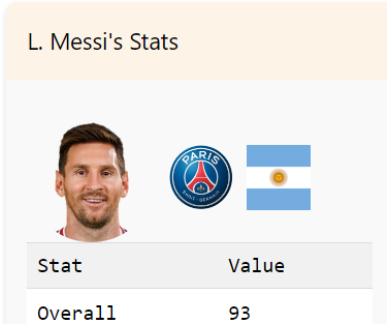


Figure 2: Messi shown as Selected Player

In 2, the effectiveness principle is satisfied, by seeing immediately a picture of the player, together with their club and home country. Recognition can help the user realize if they have the right person.

2. **List of attributes**
All the attributes to compare the players are shown under their picture with the respective score next to the attribute. The colour coding that is used to express which player is better confirms the expressiveness principle because the user sees the attribute as green if their selected player is better. The principle is also satisfied, due to having the overall points of the player as the highest-ranked attribute. Figure 3 shows the nine attributes that are shown first to the user, the colour coding depends on the other selected player, in this case, Messi vs Suárez.

Stat	Value
Overall	93
Club	Paris Saint-G
Crossing	85
Finishing	95
HeadingAccuracy	70
ShortPassing	91
Volleys	88
Dribbling	96
Curve	93

(a) Attributes Messi

Stat	Value
Overall	88
Club	Atlético de M
Crossing	80
Finishing	93
HeadingAccuracy	84
ShortPassing	83
Volleys	90
Dribbling	83
Curve	86

(b) Attributes Suárez

Figure 3: Attributes of the selected players

3. **Radar plot**
The radar plot shows the selected players in a multivariate visualization. The reason why we chose a radar plot is so the user can see the stronger and weaker points of the players in the plot, it draws the profile. The radar plot shows the weak spots of a player but also where they are better. By adding the radar plot users can intuitively see if the selected players have the same profile. By choosing blue and orange and having an amount of transparency, the user can distinguish between the two different profiles of the players. The radar plot helps see similarities and differences and helps discover trends amongst players in the same country, club or position.
The radar plot is better than the list since the multivariate data is shown in one graphical display. This allows users to identify relations between the attributes, as well as relations between the players. Most of the tasks are to compare players based on their tactics, the radar plot shows the relative weaknesses and strengths of the players. The user has selected the players based on the country, position or club they want to see and thus helps them identify which player is better in what aspect of the soccer game.
We chose the radar plot, Figure 4, to be shown as the first visualization because it shows the distribution of the values across all the attributes. The scale is 0-100 across all the attributes and since all the data is quantitative sequential data which is not ordered, the scale is consistent. The attributes are shown at the end of every axis and to reduce overlapping data not every point is shown in the list is also shown in the radar plot. Due to the still big amount of shown attributes, we decided not to add markers since the overall shape of the player is more important than the individual attributes since those are shown in the list under the player.

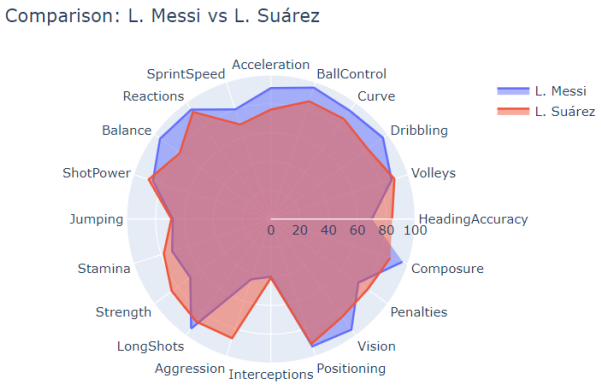


Figure 4: Radarplot Messi vs Suárez

4. **Scatter plot** The scatter plot shows the relative position of the player compared to all other players who participated in the World Cup. By using a location as the mean to express how good a player is in the selected attributes, users can see their favourite players compared to all other players. We included the scatter plot because we wanted the user to see more than just the two selected players.
In this view, users also get more autonomy over how they want to view the data and what specific trends they want to see. With a scatter plot users can see positive or negative trends for example on aggression and overall performance. By hovering over the scatter plot users see the score and the name of the player that is represented by each dot marker. When zooming in users can look in detail at which players are better than others and can start. Users can change what attributes will be shown by clicking on the drop-down menus.
A dot represents each player and when hovering over you can see which dot is who. Due to the data being quantitative sequential and there being a lot of data points, we decided to make the plot more transparent and colour scale from a dark to a lighter hue. This is not because there are no categories, but the hue makes it easier to differentiate between the dots. Choosing the colours this way, makes the scatter plot comply with the expressiveness principle.
The axis of the scatter plot do not start at 0, since no players that are playing in the biggest soccer tournament on earth have 0 skill at one of the attributes. Going from 20 to 100 points makes the data more evenly distributed and less white space is shown in the graph.
Figure 5 shows Messi vs Suarez based on their overall score across all participants of the World Cup and their position of

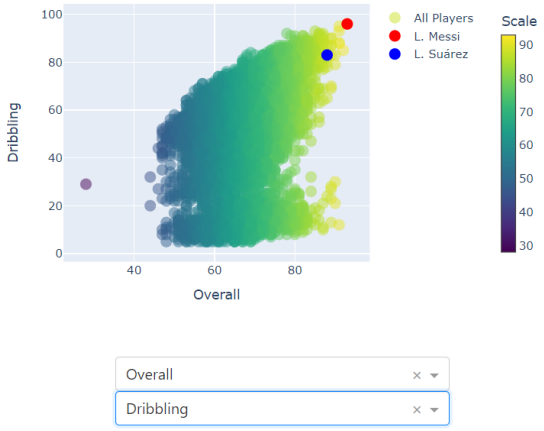


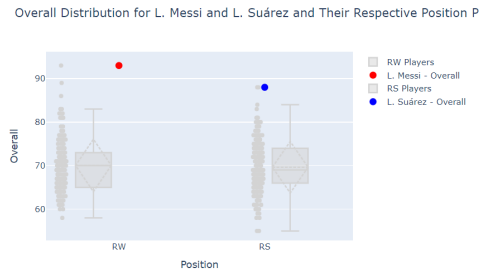
Figure 5: Scatterplot Overall score over Dribbling

how well they can dribble.

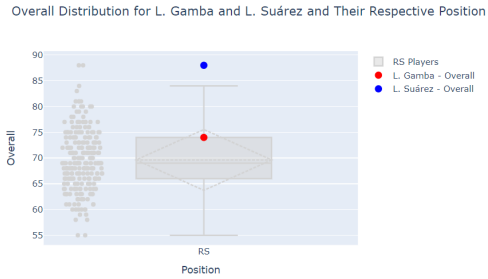
5. **Box plot per position**
The box plot, figure 6, shows the relative position of the selected player amongst all other players that play the same position. As we want to help users see more information about their favourite players, the box plots show if someone is really better than average. A box plot shows outliers very well, thus the best players are easily distilled. Again the users can decide on what basis they want to compare the players. A use-case scenario of the box plot is if fans want to combine all of their favourite players to make a fantasy team, the box plot can help them choose who is a better player at a certain position on the field. As well as, seeing what is missing in the team, since some players might perform better or worse on certain attributes.
If there are two players selected that do not play the same position, the user is presented with two box plots. If the selected players play the same position, the user can see the distribution amongst all players based on two different color dots. Since all the data we use to compare players is sequential quantitative data, the box plots will give another layer to the exploratory analysis.

6. **Clubs**
In order to compare how the players of the clubs play, we decided to add another tab to the dashboard where clubs can be compared to each other, figure 7. By cumulatively adding up attributes like the goals, red cards and how many minutes were played from the respective players in the clubs, a bar chart is shown to compare the clubs. Users can decide which clubs to add in their analysis and what attribute they want to see.
The clubs are ordered in alphabetical order for easy searching for the club (figure 8). By hovering over the bars, the user can also see the club and the specific amount. We choose to have a uniform colour for every bar since we think the brushing and alphabetical order is enough to see the specifics. By adding colour we were afraid that the bar graph would become visually noisy.

7. **Extras** In this section, some extras will be shown that are not explicitly showcased in the screen recording or mentioned elsewhere. First of all, we included the score of a match (figure 9) if the selected players played against each other in the World Cup.This is to also see how players from the same



(a) Different positions



(b) Same position

Figure 6: Box plots

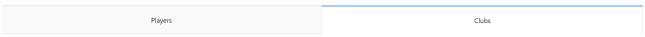


Figure 7: Player and Club View

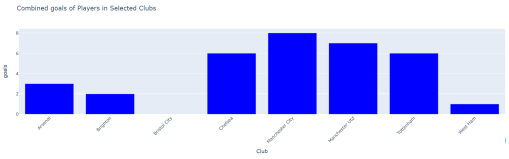


Figure 8: English clubs based on Goals

team played against each other, or from same position to see if the aggression or speed of one of the players helped the team more to get the points.

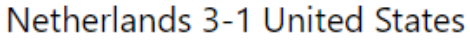


Figure 9: Result of a match

In the scatter plot, it is possible to turn the scatter plot towards the axis with numbers, in order to see the score of a certain attribute better. In the scatter plot, box plot and bar chart it is possible to hover over the marks to see more information about the players or clubs, as well as zoom in and select pieces of information interesting for the analysis of the user.

5 Implementation

We are currently utilizing the Dash framework within Python, which incorporates Flask on the server side and React on the client side. Our implementation involves libraries such as `dash_core_components`, which provide graphs, dropdowns, and other interactive elements, as well as `dash_html_components` that

enable HTML code integration within the Python environment. Additionally, we use Pandas for cleaning, reading, and modifying data. The Dash Dependencies, including Input and Output, are essential for defining callbacks. Moreover, we employ Plotly to enhance the dashboard's interactivity and aesthetics. `re index.html` was used for the structure of the webpage dashboard. `os` was imported in order to structure our file system. `style.css` was used for separate CSS code to make the dashboard.

6 Use Cases

How does Wijnaldum perform compared to one of his teammates, for example, Gueye? By filling in Wijnaldum and clicking on compare to players of the same club we get an extended list of possible players (Figure 10).

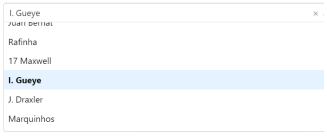


Figure 10: Club Colleagues of Wijnaldum

Comparing Wijnaldum to Gueye leads to the radar plot in figure 11. Their overall score is respectively 84 and 82 and we see that the Netherlands won 2-0 of Senegal. However, based on positioning, we see that Wijnaldum is a substitute for the dutch team, despite having a higher score than Gueye who is a Left Central Mid (LCM). This might be because the Dutch team is better than the Senegal team overall, it might be that good players are substitutes.

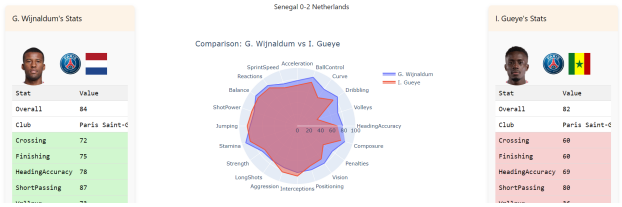


Figure 11: Wijnaldum vs Gueye

Were the Dutch players better at the position of LCM and therefore won the Match? By going back to the selection procedure we will find a Dutch player with the same position as Gueye and start the comparison. Gravenberch is a Dutch player playing LCM to who we will compare Gueye to see if that position had something to do with the Netherlands winning the match. In Figure 12 it can be seen that Gueye is overall a better player, however as a LCM it is important to have strength and Gravenberch has that more as seen in figure 13.

We will remember Gueye but start the analysis with Gravenberch, to look for more players in the Dutch team that are better than Gueye. Soon we find De Jong, who is a very well-known and good player from the Netherlands as can be seen in figure 14.

Since De Jong is a Right Central Mid (RCM) it means that Gueye was eye-to-eye with De Jong during the game, which might be one of the reasons why the Netherlands won the match against Senegal. For Midfielders it is important to have ball control to be able to pass the ball around to the whole field, so to complete our analysis of the intricacies between colleagues, countries, and positions, we will look at the players compared to all participants of the World Cup. In the scatter plot users can hover over the dots to find other players and as Gueye has a score 82 overall and 79 for ball control, users would have been be able to find him in the data, shown in figure 15.

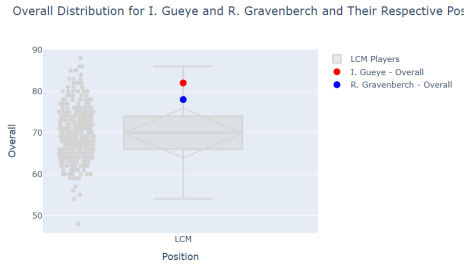


Figure 12: Box plots comparing Gueye vs Gravenberch based on Overall performance

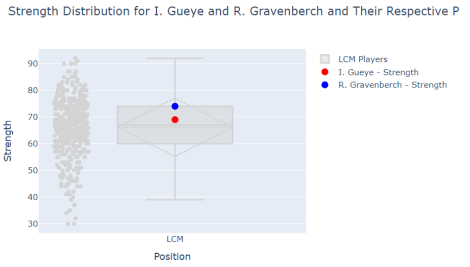


Figure 13: Box plots comparing Gueye vs Gravenberch based on Strength

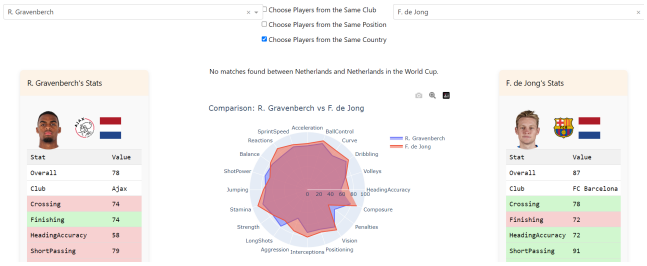


Figure 14: Gravenberch vs de Jong

However it looks like it he has been overwritten by another player, this is something to look for during future work.

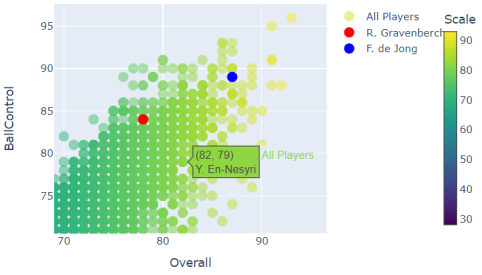


Figure 15: Gravenberch vs de Jong Ball control

The last task that the tool can do is the comparison of different clubs to see how well their players have performed in the World Cup. We will use the clubs of the players we have analyzed for this use case and see that the players from Paris Saint-Germain (PSG) have scored the most goals during the tournament, as seen in figure 16.

So PSG can be happy with their amazing players like Wijnaldum and Gueye and the user is happy to see the players of their favourite club perform well in the World Cup.
Comparing different players in the way that is described above reveals insights into the dynamics within countries, clubs, and positions. Users are able to customize how they want to analyze their favourite players and clubs and have full freedom of exploration of the dataset, making it an insightful yet fun experience.

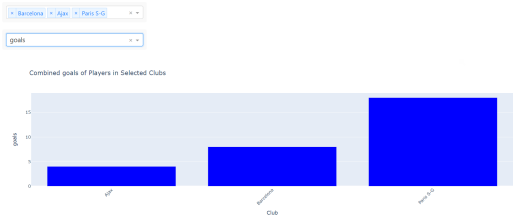


Figure 16: Goals scored by players of Ajax, FC Barcelona and PSG

7 Reflection

As shown in Figure 17 , during the midterm we had a different way of letting the users choose who they wanted to view, after making the dashboard with the drop-down menus, we found out that you too quickly do not get any matches in the data. By replacing drop-down menus with buttons for easier player selection and switching from country-based to club-based comparisons led to enhancing the overall user experience.

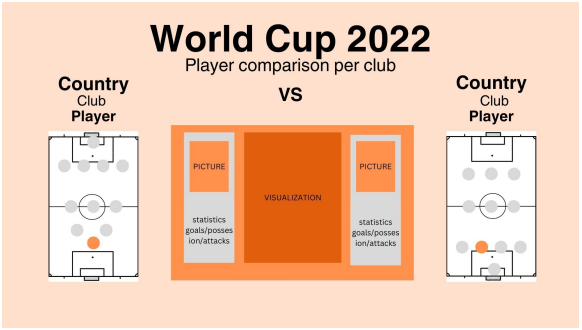


Figure 17: Midterm dashboard

Another part of our midterm solution was to compare the teams based on country, however, we decided to change to the clubs (figure 18). This is because we found out that there are not many players playing against each other in the World Cup actually from the same club in the different teams, this would make the solution less usable. Therefore we came up with the idea to add more information about the clubs and how good or bad their players are performing
There were multiple reasons why we added a couple of extra data sets. First of all, the pictures in the provided data set had multiple options per player making and often faulty pictures of players' family members.
Next to that, we wanted a more in-depth analysis of the player's tactics during games, however, the provided databases could not give us that. We had thought that still insightful facts would show but tactics and comparing the players was hard on mostly Xg data. By adding more datasets our tool got slower and the dashboard could react faster if we did more extensive data handling strategies.
After testing a prototype on novice users, we decided to change the dropdown menu to choose the players to buttons where users can decide on what similar level they want to compare players.

Figure 18: Team comparison

8 Future Work

Future enhancements could focus on increased interactivity, allowing users to link graphs and compare players more intuitively. This would enable use cases like creating fantasy teams and provide deeper insights into player tactics, further improving the tool's utility for soccer fans.
We believe that additional datasets and predictive models would make the dashboard even better for seeing the impact of trainers, colleagues and positions on the selected players.
During the project we did some user testing on novice users, like our friends. In the future, a robust testing framework would increase the usability of our dashboard to entertain more soccer fans all over the world.
Due to the addition of datasets and focus on the design of the dashboard, the data handling strategies and reduction of clutter within the dataset were not top priorities. Therefore our dashboard got slower over time, for a smooth interaction with the dashboard the speed could be improved. During the creation of the use cases we found out that in the scatterplot some players get overwritten when they have the same statistics, this is something that could be improved in the future by the creation of a list of players that have the exact same statistic.
Since we are talking about the creation of fantasy football player teams, a great addition to the tool would be the possibility of selecting certain players and explore their possible team statistics. In this way the dashboard is not only focused on analyzing but also becomes a form of entertainment.

9 Conclusion

The FIFA World Cup 2022 provided data for showcasing the talents and strategies of the world's best soccer players. Through Club-Dash fans all over the world can interactively analyze their favourite player performances. Club-Dash focuses on exploring the intricate relations between position, club and national teams.

Through the various visualizations: radar plot, scatter plot, box plot and the cumulative bar chart, users are free to decide how they want to view the tactics and attributes of players. The ability to compare players based on multiple attributes, player profiles, and all the players in the World Cup contributes to having more insights into the complex game. During the creation of the dashboard, the focus was on making a visualization that comply to effectiveness and expressiveness principles as well as [1] visualization designs. By keeping the end user and tasks in mind we made intrinsically understandable visualizations. Users can see in one eye view how two players' football profiles compare to each other and by scoring down they can personalize the ways to view the data more.

Our reflection on the development process mentions the incorporation of additional datasets and refining the user interface after testing with novice users. Through weeks of iterative processes, the tool became more robust and accessible. For future work, a couple

of steps can be taken including enhancing interactivity and enabling features like fantasy team creation will help future users analyze more thoroughly and enjoy the tool better.

All in all, Club-Dash tries to create a useful platform in data visualization for sports analytics. Through the provided dashboard, users can explore the game and players' dynamics based on country, club, or position. Hopefully, Club-Dash will make football fans appreciate the depth of football even more.

REFERENCES

[1] T. Munzner. *Visualization analysis and design*. CRC press, 2014.