Expected Performance in the Premier League

Floris Dobber*

Northeastern University

ABSTRACT

By comparing expected goals to actual performance, users will be able to see which soccer teams are over or under performing. This project aims to compile data related to the English Premier League and create visualizations to represent performance at the team and individual player level. These visualizations will be useful to bring attention to undervalued players, as well as to teams that are beating expectations. The information gained from the visualizations can be used by itself or as a precursor for further analysis.

Link to GitHub Repository

1 INTRODUCTION

Our project aims to visualize the performance and expected performance of teams in the English Premier League. Through this visualization we hope to gain a better understanding on how measures such as shots and possession correlate with on pitch performance. Currently, measures of expected goals are used at the surface level. We plan to further analyze and apply expected goals to thoroughly take advantage of the data and display to the user the data. We aim to exemplify what tactics and players currently are beating expectations.

2 RELATED WORK

In the EuroVis article, "MatchPad: Interactive Glyph-Based Visualization for Real-Time Sports Performance Analysis" [2] it discusses a visualization that shows on field performance in live action. The work talks about how currently, team analysts for a sports team will annotate match tapes and use that statistical analysis of the game to help the coaches with their general strategy. They go more into depth about their glyph-based visual design that allows coaching staff to analyze their team's performance in real time rather than after the fact. This is similar to our use case which is to better understand and visualize a team's on pitch performance. The main difference between the use cases is that their visualization is more directed towards the coaching staff to help make more informed decisions about their team's tactics whereas our visualizations focuses more on helping fans make better game predictions. An important part of their visualization that could be used to inform our work as a group is their ability to convey information to the end user in an engaging way. While we are not focusing on live-action predictions, both our and their visualizations focus on a visually-engaging way of showing the data to a user and I think that it could be helpful to look at this group's visualization to get a better idea at how to do that.

Similarly to the EuroVis article about sports data analytics, the article from frontiersin.org, "Data-Driven Visual Performance Analysis in Soccer: An Exploratory Prototype" [3] discusses the application of data visualization and analytics in modern sports, specifically soccer. This is relevant to our use case as it specifically discusses on-pitch performance of players and teams as a whole. This specific article can be used to better inform our work as it talks about how there are certain parts of a visualization or statistics in soccer that haven't yet been implemented because of how difficult the data is to get from games. A good example that they provide is the "use of movement patterns and complex inter-player coordination metrics to explain collective tactical behavior in soccer games" and how there are no interactive visualizations that convey this just based on the difficult on implementation and data collection. This is helpful to us because while we continue on in making a visualization to help convey the on pitch performance of a team, it's important to focus on statistics and data that is available to us and easy to draw real conclusions from.

From the IEEE article, "What are the topics in football? Extracting time-series topics from game episodes" [4], the visualization analyzes movements of players in game to identify patterns in the team behavior. The visualization that we will create does not analyze movement on field, however we can draw inspiration in the use of identifying patterns and apply this to the data sets we have to make predictions based off of historical performance. The visualization also makes good use of visuals representing the football pitch, and we can use this as a base for making more interesting visualizations that are catered specifically towards soccer.

In the IEEE article, "A Network Visualization of Sustainable Consumption Corridors" [1] a visualization is shown to take in user information in order to generate suggestions. If we develop our visualization to be particularly useful for individuals in cases such as betting, having user-inputted data can aid in creating more personalized analysis and aid.

In the book, Fundamentals of Data Visualization [5], there are many visualizations that work well to help users visualize uncertainty. These visualizations are effective in allowing a user to understand expected value, and can be directly adapted to our visualization.

3 USE CASE

As a fan, one of the most popular activities to partake in is sports betting. Through the use of our visualization tool, users will be able to apply existing expected goal data to make better game predictions. It will also augment the ability to digest the new datapoints now common in modern soccer.

4 DATA

Below we've listed several data sources that we can use individually or in combination with each other to create our visualization. All data sets record facts, and so therefore cannot be biased. Where our data sets might differ slightly is in the approach to use to calculate expected goals (xG) and how they define certain metrics. This caveats will be made clear when we decide what source to use in our final visualization.

- FB Ref

  FB Ref has aggregated xG, goals, shots, and other metrics for all players and clubs across the big 5 leagues across several seasons. They are generally regarded as the most reliable data source, and make all of their data public. They are also very transparent in their data collection techniques.

*e-mail: dobber.f@northeastern.edu

- Understat

  Understat provides xG, goals, assists, shots, games played, minutes played and other metrics for the big 5 leagues since 2014-2015. The data linked has been scraped directly from there. If we use this data, we will do spot checks to ensure the integrity of the data, but overall the data appears to be very reliable.

- Football Data

  Football Data is a website that is mostly used for training betting models. Their data includes basic metrics for all matches played in the big 5 leagues, as well as betting odds for several bookies. Overall, a great possible addition in case we want to add some additional metrics.

- Football xG

  Football xG provides tables for all big 5 leagues across the past 5 seasons with xG, goals and xPoints. They could be a great addition for their xPoints values, as well to compare different xG values across vendors. They are a little less transparent about their collection metrics, so we would verify the integrity of their data before integrating them into our project.

5 Task Analysis

5.1 Task Table

Task ID	Domain Task	Analyze Task	Search Task	Analytic Task
Soccer Scout	Use a visualization to see how players are actually performing in relation to how they are expected to play.	Consume $\longrightarrow$ Discover	Browse	Compare
Sports Bettor	Use a visualization to determine which teams I should bet on given current years form	Consume $\longrightarrow$ Discover	Explore	Identify
Sports Fan	See if my favorite teams and players are over or under performing this year along with other teams in the same and different leagues	Consume $\longrightarrow$ Enjoy	Explore	Identify
Sports Analyst	See and compare season and league based averages to use in my own data project	Consume $\longrightarrow$ Discover	Explore	Summarize

The primary use of our visualization will be for people looking to generally compare teams and players to other teams and players. For this reason, the primary consumer of our visualization will be scouts, fans, and analysts who are looking at broad trends to assert if a certain team or player is over or under performing.

This visualization will primarily be developed for the discover type of consumption. This is because it's main purpose is to help the user better understand, or discover, what a team's overall performance is and how they are doing relative to what is expected.The visualization is centered around supporting the user's desire to discover and explore data and its trends in soccer.

6 Implementation Plan & Preliminary Work

For our final project, we will be creating two linked visualizations. We will have a scatter plot on the left accompanied by a bar chart on the right. The scatter plot will have points and lines for marks, and position (horizontal & vertical), color and shape for the channels. The bar chart will have lines for marks and position (horizontal & vertical) and color for the channels. The graphs will also have associated filters using drop-down menus, which we will create using HTML.

Teams can be filtered by clicking or brushing on the scatter plot. This filtering will lead to players being filtered out in the bar chart. Currently not-included teams will have increased opacity. Clicking on players in the bar chart will highlight their respective team in the scatter plot. Many of the drop down filters will affect both visualizations, for example the season selector. Lastly, hovering over either a team or a player will show a pop-up displaying additional information.

The basic requirements for our visualizations will be the marks and the position. Color on both graphs is nice to have, and much of the info in the pop-up will also be nice to have. The drop-down filters are fairly essential, as many of the use cases will rely on that. Filtering will add additional use However, we do aim to implement all of the features mentioned above, as that will provide the most complete user-experience.

We plan to implement our visualizations using D3.js. We learned about it during class and thought it was more than sufficient to create the visualization that we have in mind. We can use basic HTML elements like ¡select¿ for our drop boxes, and the graphs can fully rely on D3. All the data we'll be graphing will be pre-computed and stored in CSV files, so no server components will be required.

7 Visualization Design

The final visualization tool design will be composed of two main visualizations. The two visualizations will be side-by-side and work together to represent data at a team and player level. The team data will be displayed in a scatter plot on the left while the player data will be displayed in a bar plot on the right.

The left side is composed of a scatter plot that represents performance data for teams. By using drop down menus, users will be able to select the team data that they want included in the visualization. These selections include *Season*, which determines the what specific seasons the data will be taken from and *Team*, which determines which teams are included in the visualization. We will also have a selector to flip between two over-arching views. The first option will be *xG Output*, while the second will be *Performance*. Changing this option will change both the x and y axes. When *xG Output* is selected, the *x-axis* will be xGc per 90 and the *y-axis* will be xG per 90. When *Performance* is selected, the *x-axis* will be (xGc per 90 - Gc per 90) and the *y-axis* will be (G per 90 - xG per 90). Once these options are specified, the visualization will populate with the data intersections given by the user and the axes will reflect the mode the user has selected. Points on the scatter plot will represent different teams through their respective team logos. When *xG Output* mode is selected, there is also a dotted line, *(y=x)*, that runs diagonally across the scatter plot. The top side of the line is shaded in a green gradient, and the bottom side of the line is shaded in a red gradient. This is so that users can determine if that team is creating more xG than conceding. When *Performance* mode is selected, all four quadrants of the graph will be shown. There will be a dotted line, *(y=-x)*, that runs diagonally across the scatter plot. The top side of the line is shaded in a green gradient, and the bottom side of the line is shaded in a red gradient. This is so that users can determine if that team is over or under performing. When hovering the mouse on a team logo (in either mode), a tool-tip will appear that displays the season and team name, along with the following data, xG per 90, xGc per

90, G per 90, Gc per 90, along with the record of the team over that season.

The visualization on the right side of the final visualization tool will be a bar graph that is linked to the scatter plot. The bar chart will reflect the filter combinations chosen by the user on the scatter plot. If the user clicks on a team logo in the scatter plot, the bar graph will be populated with player data from that respective team and season combo. If the user is in *xG Output* mode, the bar plot will display the top 10 xG per 90 players of the selected data on the scatter plot. If the user is in *Performance* mode, the bar plot will display the top 10 xG per 90 - G per 90 players of the selected data on the scatter plot. There is also the visual encoding element of color used here as bars will be darker the higher performing the player is. Lastly, when a user hovers over a bar, a tool-tip will appear with player data such as name, position, nationality, and performance statistics such as G per 90, xG per 90, A per 90, xA per 90, and 90s played.

REFERENCES

[1] M. Hoefer and S. Voida. ieevis.org: A Network Visualization of Sustainable Consumption Corridors. *IEEE VIS*, 2021.

[2] P. Legg, D. H. S. Chung, M. L. Parry, M. W. Jones, R. Long, I. W. Griffiths, and M. Chen. MatchPad: Interactive Glyph-Based Visualization for Real-Time Sports Performance Analysis. *Computer Graphics Forum*, 2012. doi: 10.1111/j.1467-8659.2012.03118.x

[3] A. B. Santos, R. Theron, A. Losada, J. E. Sampaio, and C. Lago-Peñas. frontiersin.org: Data-Driven Visual Performance Analysis in Soccer: An Exploratory Prototype. *frontiersin*, 2018.

[4] G. Shirato, N. Andrienko, and G. Andrienko. ieevis.org: What are the topics in football? Extracting time-series topics from game episodes. *IEEE VIS*, 2021.

[5] C. O. Wilke. *Fundamentals of Data Visualization*. O'Reilly Media, Inc., 2019.

8 Appendix A: Group Charter

8.1 Group Purpose

Our group came together due to our shared interest in soccer. We hope to create interesting visualizations that will help people gain additional insights into the game.

8.2 Group Goals

Our group aims to produce high quality work that is finished within the set due dates. Together we hope to achieve a grade of an A for this project.

8.3 Group Member Roles & Responsibilities

For our group to be successful we will need everyone to be invested in our success. We will each carry equal responsibility and we will aim to carry out a variety of roles throughout the project. These roles include project leader, who is responsible for dividing up work and creating a schedule to make sure we hand our work in on time, as well as formatter, who is responsible for making sure the document we turn in meets all the requirements. Other rules will include a sort of tech lead role when we get to the coding part, as well as someone who just focuses on getting work done. The switching of roles will facilitate personal growth, as people may have to perform duties that they are not typically responsible for.

8.4 Ground Rules

Our group will meet on as-needed basis. In between we will stay in contact through the group chat. We will handle decision making through popular vote. In case of strong disagreements, we will discuss until the vote is unanimous. Everyone will be held accountable by the rest of the group, and in the case that someone is not pulling their weight, it can be reflected in the peer assessment.

8.5 Potential Barriers & Coping Strategies

Communication is the most important action when it comes to effective group work. If a team member has another obligation and cannot contribute their work on a certain day, they should let the rest of the team know. In this case, the rest of the team will undoubtedly pick up the additional share of work. Throughout many projects in our educational career, many instances of stress is caused by this exact reason. If a team member does not speak up, the rest of the team cannot effectively plan around this additional work and may be rushed, leading to poor results.

8.6 Progress Update

In general we feel that we have been abiding the agreed-upon guidelines. We haven't missed any deadlines and have turned in work of decent quality. While we have not switched up the group roles as much as we initially said we would, everyone does feel comfortable with the roles they have had up until now. Our biggest struggle has been to effectively organize meetings, as we all have other classes and commitments, but we have all gotten much better at remote work. Going forward, we will try to step-up our communication even further so that we can avoid any time crunches as the project nears the end.

See below for a list of positive things that each of us have contributed to the project.

Floris: Very proactive in reaching out to the team, planning out deadlines, and initiating efforts. Great driving force for the team to stay on track.

Forrest: Did so well on the interview. Also did a fantastic job on the sketches, really bringing our ideas to life.

Henry: Did a great job formatting the task table so it fits nicely. Suggesting and implementing data scraping in Python so the team can go beyond project requirements. Comes up with new ideas for visualizations and ways of combining them

Sarah: Was so integral to our successful data collection. Working relentlessly to make sure we have clean data to base our project on, as well as helping with data scraping and making sure the project is completed to the best of the team's ability.

9 Appendix B: Data Exploration

9.1 Data Review

Our datasets consist of two data-frames (exported as csvs), one of team data from 2017-2022, and one of player data from 2017-2022. The team data-frame consists of 100 entries (row), which logically is sound as the Premier League has 20 teams and we have data from 5 seasons. There are 39 fields in this data-frame, which consists of team name, and 18 fields of team data, and 18 fields of opponent data (while playing the team in question). These fields include statistics such as goals scored, shots taken, penalties awarded, xG created, and xG per shot. For each of these fields, there is an "against" field which shows the opposition stats while player the team in question.

The player database includes 2613 (entries) and consists of all players who played in the Premier League from 2017/18 season to this current season (2021/22). These player stats are separated by season, meaning that if Player A played in all 5 seasons, there will be 5 different entries for the same player. This database includes player position, date of birth, country, games played, 90s played (how many sets of 90 minutes did they play), goals scored, assists scored, xG created, and yellow/red cards given, to name some of the 33 fields.

Within the team data, there are 40 attributes with two of them representing categorical data while the rest of them are quantitative data. These instances of categorical data exist under the team name as well as season. Season can also be represented as ordinal data.

Within the player data, there are 33 attributes with four attributes representing categorical data. These categorical attributes include the nation the player is from, the player's name, the player's position, and the player's team. Player rank is the only instance of ordinal data besides season year, which can also be represented as categorical data.

The following bullets outline some of the issues that we encountered while scraping the data:

- The original players dataset had some issues with the formatting of the country name. For example, instead of appearing as "ENG", it originally appeared as "eng ENG".
  - In order to fix this issue, we had to write two functions to help clean the code. In the first we had to iterate through each value in the 'Nation' column and return a substring of the original nation but only from the character after the space onwards. Therefore, instead of returning "eng ENG" it would just return "ENG". The second function then appended the cleaned version to a list and we then replaced the original column with the cleaned data from the list.
- The original players dataset also had some issues with the formatting of the player names. For example, instead of appearing as "Harry Kane", it originally appeared as "Harry Kane\1836939".
  - In order to fix this issue, we had to write two functions to help clean the code. In the first function we had to iterate through each player name in the 'Player' column. Similarly to the nation cleaning function, it returned a substring of the original player name but only from the character after the blackslash onwards. Therefore, instead of returning "Harry Kane\1836939", it now just

returned ”Harry Kane”. The second function also appended the cleaned version to a list and we then replaced the original column with the cleaned data from the list.

- After the initial data review, the team noticed that player names held the same values as country for just the 2018-2019 season. Therefore, instead of having one column with the player names and another with the country names, there were two columns with the player names.

  – When creating the list of nation names for the 2018-2019 season with the functions that I described above, we accidentally set the nation list for the season to the result list from the player cleaning function rather than the nation cleaning function. After just changing that one line of code from player to nation the data went back to how it was supposed to be.

9.2 Insights

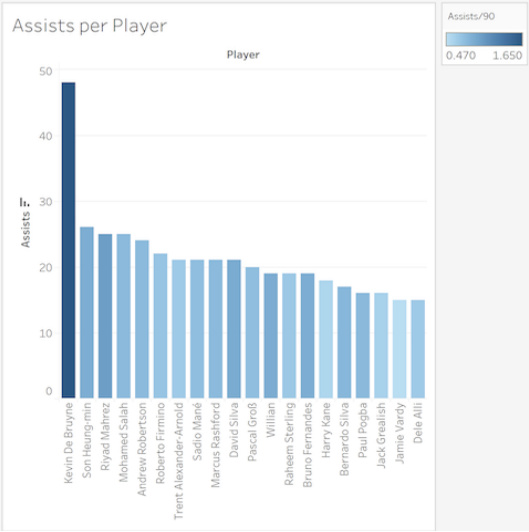
During our data exploration, we were most surprised with the quality of the data. Besides mistakes we made ourselves when scraping the data, the data itself appears to be near flawless. This is unusual for many of the data sources that we have worked with in the past, but a pleasant surprise. The most confusing aspect of the data is the column labels, which can be hard to understand initially. We will therefore ensure to either label them clearly or have clear explanations when we make them part of our visualization. This can be in the form of a *, a pop-up when someone hovers over the label or simply replacing it by a clearer label.

Besides that, we noticed a few things while exploring the data in detail. First of all, we were surprised to see the number of players fluctuate so much across seasons. We had assumed that each season there would be 500 players in the data set, 25 per squad in the league, but instead we found it to always be over 500. On second thought, this makes sense, as squad have a limit of 25 players per squad but they are free to use many more under 21 players as well as any players they bring in over the winter transfer season. Seasons did have a consistent 20 teams per season, which is what we expected. Any fluctuations there would have been harder to explain.

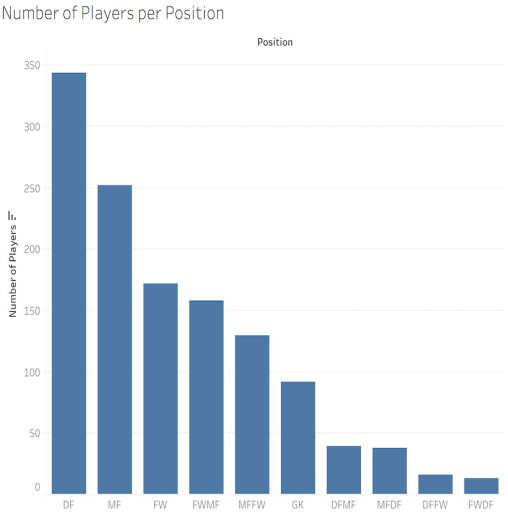
Something else we found interesting was the abundance of players from the UK, but also how many countries are represented in the Premier League. Over 200 players were from the UK, but if you look at the attached map below, you can see almost every country in South America is represented, many countries in Africa, and countries in Europe even have 30+ players in the Premier League. This really illustrated to us how global of a sport soccer is, and that the Premier League imports talent from all over the world. It was also interesting to see that while many European countries host their own domestic leagues, they each had 30+ players playing in the Premier League as well.

Lastly, we also took a quick look at the data most relevant to us, xG and goals. We plotted goals vs xG avg over the past 5 seasons per team. What was interesting to see was that better teams seem to outperform their xG where as worse teams seem to underperform on xG. Perhaps this points to a small flaw in the way FB-Ref calculates their xG values? We will have to do a deeper dive before we create our visualization, but it was an interesting pattern find so quickly.

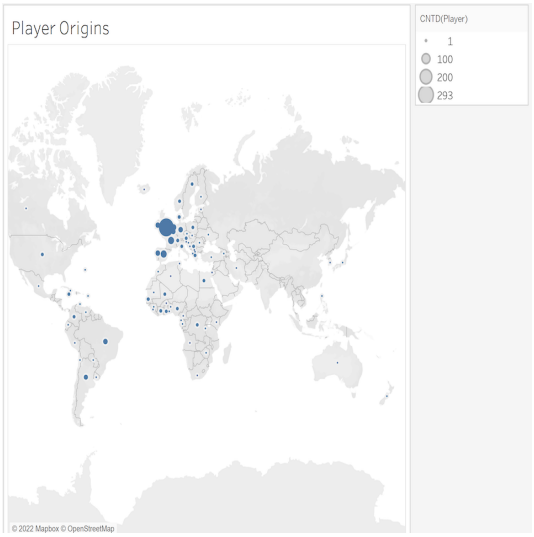
9.3 Screenshots



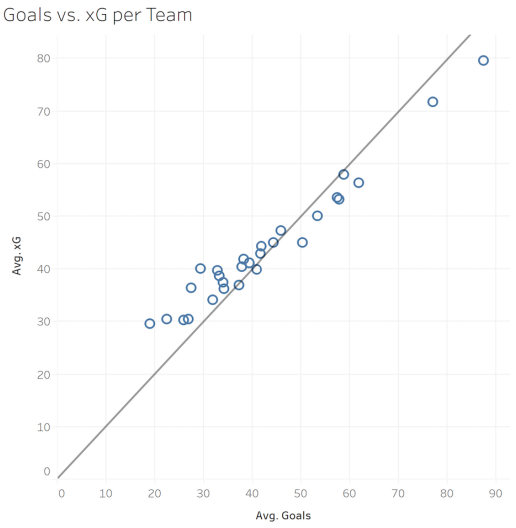
This visualization pulls data from the player-specific data set to show the assists per player in the form of a bar chart. The x-axis shows player names while the y-axis shows the total number of assists the player has. There is also the encoding element of color - the darker the bar is, the more assists a player has per 90 minutes of playing. It can be seen that Kevin De Bruyne has a noticeably higher number of total assists, as well as a higher number of assists per 90 minutes, when compared to other players included in the visualization.



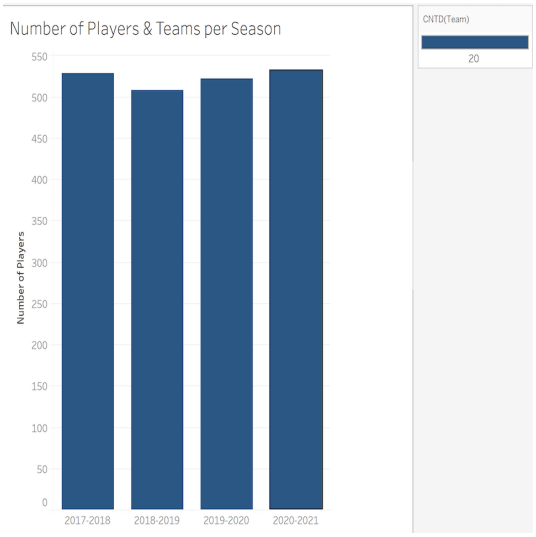
This visualization pulls data from the player-specific data set to show the number of players per position in the form of a bar chart. The x-axis shows the names of the positions, while the y-axis shows the total number of players who play this position. It can be seen that the most common position played is defense, which has nearly 350 players dedicated to this position. The visualization also shows that the most uncommon position played is a combination of forward and defense.



This visualization pulls data from the player-specific data set to show the origins of players across the globe. The visualization shows a map of the world, with circles atop the origin countries of players from the data set. The main encoding element is the size of the circle, with larger circles representing a larger number of players that hail from the specific country. Overall, the visualization gives a good idea of the spread of players from around the world, and also shows the extreme amount of players who originate from European countries such as the U.K., France, Spain, and Portugal.



This visualization pulls data from the team-specific data set to show the actual goals per team compared to the xG (expected goals) per team. The visualization a scatter chart with average goals on the x-axis and average xG on the y-axis. The main encoding element are the circles that represent each team, as well as a line of fit that represents the number of goals compared to the xG for the overall league. Teams that are on the left size of the line of fit have a higher average xG value compared to their average goals, while teams on the right side of the line have a higher number of average goals compared to their average xG. Overall, the visualization gives a good idea of what teams are over or under performing their expectations.



This visualization pulls data from the team-specific data set to show the number of players and teams per season in the form of a bar chart. The x-axis shows the season while the y-axis shows the total number of players in the specified season. There is also the encoding element of color - the darker the bar is, the more teams there are for the specific season. It can be seen that the number of players per season has remained steady as there have been between 500 and 550 players per season. The 2018-2019 season has noticeably less players than the other seasons, however the number still remains above 500.

9.4 Data Snippet

Year	Rank	Nation	Player	Position	Team	Age	Born	MP	Starts	Min	90s	Goals	Assists	Goals-PKs	PKs	PKatt	Yellows	Reds	Goals per 90
2017-2018	1	NED	Patrick van Aanholt	DF	Crystal Palace	26	1990	28	25	2184	24.3	5	1	5	0	0	7	0	0.21
2017-2018	2	ENG	Rolando Aarons	FWMF	Newcastle Utd	21	1995	4	1	139	1.5	0	0	0	0	0	0	0	0
2017-2018	3	ENG	Tammy Abraham	FW	Swansea City	19	1997	31	15	1726	19.2	5	1	5	0	0	0	0	0.26
2017-2018	4	SCO	Charlie Adam	MF	Stoke City	31	1985	11	5	411	4.6	0	0	0	0	0	1	0	0
2017-2018	5	ESP	Adrián	GK	West Ham	30	1987	19	19	1710	19	0	0	0	0	0	2	0	0
2017-2018	6	NED	Ibrahim Afellay	MF	Stoke City	31	1986	6	1	166	1.8	0	0	0	0	0	1	0	0
2017-2018	7	COD	Benik Afobe	FWMF	Bournemouth	24	1993	17	5	611	6.8	0	0	0	0	0	0	0	0
2017-2018	8	ARG	Sergio Agüero	FW	Manchester City	29	1988	25	22	1963	21.8	21	6	17	4	4	2	0	0.96
2017-2018	9	NED	Nathan Aké	DF	Bournemouth	22	1995	38	37	3353	37.3	2	3	2	0	0	5	0	0.05
2017-2018	10	ENG	Marc Albrighton	MFFW	Leicester City	27	1989	34	30	2538	28.2	2	7	2	0	0	5	1	0.07
2017-2018	11	BEL	Toby Alderweireld	DF	Tottenham	28	1989	14	13	1178	13.1	0	0	0	0	0	3	0	0
2017-2018	12	ENG	Trent Alexander-Arnold	DF	Liverpool	18	1998	19	18	1573	17.5	1	1	1	0	0	3	0	0.06

Cleaned Player Data Sample

Season	Squad	# Pl	90s	Gls	Sh	SoT	SoT%	Sh/90	SoT/90	G/Sh	G/SoT	Dist	FK	PK	PKatt	xG	npxG	npxG/Sh	G-xG	np:G-xG
2017	Arsenal	30	38	73	590	227	38.5	15.53	5.97	0.12	0.3	17.4	30	4	5	68.4	64.6	0.11	4.6	4.4
2017	Bournemouth	22	38	45	473	153	32.3	12.45	4.03	0.09	0.27	17.5	13	3	3	44.8	42.5	0.09	0.2	-0.5
2017	Brighton	24	38	33	381	113	29.7	10.03	2.97	0.07	0.25	17.9	3	5	7	37.2	31.8	0.09	-4.2	-3.8
2017	Burnley	24	38	35	388	126	32.5	10.21	3.32	0.09	0.28	15.9	16	0	0	34.5	34.5	0.09	0.5	0.5
2017	Chelsea	26	38	60	602	209	34.7	15.84	5.5	0.09	0.27	17.4	19	3	3	57.9	55.7	0.09	2.1	1.3
2017	Crystal Palace	28	38	44	466	133	28.5	12.26	3.5	0.08	0.27	17	20	8	10	56.1	48.9	0.11	-12.1	-12.9
2017	Everton	30	38	42	359	108	30.1	9.45	2.84	0.1	0.34	16.9	8	5	8	41.5	35.2	0.1	0.5	1.8
2017	Huddersfield	25	38	25	362	105	29	9.53	2.76	0.06	0.22	18.1	13	2	2	31.7	30.2	0.09	-6.7	-7.2
2017	Leicester City	27	38	53	418	137	32.8	11	3.61	0.11	0.35	17	12	5	6	48.1	43.5	0.11	4.9	4.5
2017	Liverpool	27	38	82	642	228	35.5	16.89	6	0.12	0.35	17.1	22	2	3	72.8	70.5	0.11	9.2	9.5
2017	Manchester City	25	38	103	661	255	38.6	17.39	6.71	0.15	0.38	17.4	31	6	8	86.4	80.4	0.12	16.6	16.6
2017	Manchester Utd	27	38	67	513	176	34.3	13.5	4.63	0.13	0.38	17.6	22	1	3	60.7	58.2	0.12	6.3	7.8

Cleaned Team Data Sample

10 Appendix C: Interview

10.1 End User Personas

Persona 1

Our first end user persona is a soccer scout. They are looking for undiscovered talent by using xG data. Through their line of work they are very familiar with many different types of visualizations, and they are familiar with all common metrics used to evaluate players. Their main use case for the visualization is to see how players are performing in relation to their expected goals. Without the tool, they have to use Excel sheets and sort them. The tool make their process easier as it provides a visual interface with quick filtering options that make it easy for them to filter on a variety of metrics.

Persona 2

Our second end user persona is a bettor. They are looking for teams and players who have been over- or under-performing on their xG values and so might regress to the mean in the near future. They are not very familiar with visualizations, and this is their first attempt at using advanced statistics to improve their betting returns.

Without this tool they tend to go by their gut and general knowledge of soccer, a generally unsophisticated approach. With this tool, the hope is that they can improve their returns by basing their bets on statistical knowledge. They can easily see which teams and players might be due for a reversal of their current form, which will also help speed up their decision making process.

10.2 Interview Script

Questions

1. What is your occupation?
2. What is your educational background?
3. Can you explain your workflow to me in 5 steps?
4. How often do you use visualizations in your daily work?
5. What takes up most of your time on a day-to-day basis?
6. What is 1 thing you really like about your current process?
7. Can you name 3 elements that would be essential in a helpful visualization?
8. Can you list 5 insights that you would want to get out of the visualization?

Potential Follow-ups

1. What is one specific task our visualization could simplify for you?
2. Do you have any examples of the visualizations you currently use?
3. How experienced are you with using visualizations?

10.3 Interview Notes

Notes for interview with persona 1

Questions

1. What is your occupation?
    a. Soccer scout
2. What is your educational background?
    a. High school
3. Can you explain your workflow to me in 5 steps?
    a. Talk to club about needs -> identify potential targets -> watch footage -> visit games -> make final recommendation
4. How often do you use visualizations in your daily work?
    a. Daily basis
    b. Using radars for player stats main use
5. What takes up most of your time on a day-to-day basis?
    a. Finding potential targets and watching footage
    b. *Follow up: How often do you watch footage of players you don't like?*
        i. A lot, would save a lot of time if more precise with selections
6. What is 1 thing you really like about your current process?
    a. The combination of stats and video
7. Can you name 3 elements that would be essential in a helpful visualization?
    a. A quick overview of which players are "the best"
    b. An easy way to filter on multiple metrics
    c. A high level of responsiveness
8. Can you list 5 insights that you would want to get out of the visualization?
    a. Which players that are overperforming their xG
    b. Which players that are underperforming their xG
    c. If over/under performance is related to club performance
    d. Those are the three most important ones

Potential Follow-ups

1. What is one specific task our visualization could simplify for you?
    a. Finding players to scout

Notes for interview with persona 2

Questions

1. What is your occupation?
    a. Sports better
2. What is your educational background?
    a. High school
3. Can you explain your workflow to me in 5 steps?
    a. Identify popular games to bet on -> Look through historical match data -> Determine probable outcome of game -> Determine amount of money to bet -> Place the bet
4. How often do you use visualizations in your daily work?
    a. Not often
    b. They show up sometimes when looking up data
5. What takes up most of your time on a day-to-day basis?
    a. Analyzing team and match data.
6. What is 1 thing you really like about your current process?
    a. It is very straightforward
7. Can you name 3 elements that would be essential in a helpful visualization?
    a. For a match: a display of the likely winner of the match
    b. For a player: something to track performance over time.
8. Can you list 5 insights that you would want to get out of the visualization?
    a. Accurate predictions of what team will win a game
    b. What teams are on a hot/cold streak
    c. What players are having bad games
    d. What is a likely return on investment
    e. Did not name a 5th

Potential Follow-ups

1. How consistent is your process on a day-to-day business?
    a. Pretty consistent
2. What do you not like about your current process?
    a. Very time consuming to analyze all of the data.
    b. Sometimes he cannot go through all of it, especially if there is more than one game going on.

10.4 Interview Results

Interview with Persona 1 (Soccer scout)

1. What is your occupation?

   I am a soccer scout for a variety of premier league teams.

2. What is your educational background?

   I completed my high school degree.

3. Can you explain your workflow to me in 5 steps?

   The first step is to find out what a club is looking for in terms of position, value, style, etc. Next, I go through the data and identify potential targets. Once I have selected a few, I find footage of them online and see if the stats match the footage. After that, I go and watch a few of their games. Once that is done, I can make my final recommendation.

4. How often do you use visualizations in your daily work?

   On a daily basis really. My most common use of visualizations right now is to use radar plots to get an idea of how good a player is, what their strengths and weaknesses are, and how they compare to other players.

5. What takes up most of your time on a day-to-day basis?

   My biggest time drain is finding potential targets and watching footage. It can be really hard to find targets that fit the exact need of a club. Then watching footage can take a long time as well, especially if it's not immediately clear if they're a good fit. Watching footage of players I don't end up scouting is the biggest waste of my time, so if I could make better choices in who to watch that would be really helpful

6. What is 1 thing you really like about your current process?

   I really like how I am able to combine statistics and the "eye-test" to find new players. I find that aspect very exciting and bringing an element of real world application to the statistics.

7. Can you name 3 elements that would be essential in a helpful visualization?

   First of all, I think a quick overview of which players are the best would be super helpful. Just a quick sorting with their names really, with some intuitive measure of how much better they are. Secondly, an easy way to then filter on this list to find exactly the right match for what the club is looking for. This can be really hard right now. Lastly, my current process can be really slow to load, so if it is very responsive that would be amazing!

8. Can you list 5 insights that you would want to get out of the visualization?

   I think most importantly, being able to see overperformance of xG would be a great help. And then secondly underperformance as well. Both of those are big indicators of future success. And then thirdly, how that performance relates to the club they are currently at. If the club as whole is trending one way, then perhaps it's not surprising their players are doing the same. I think those three are the biggest things, I'm not sure that anything else I add would be as valuable to me.

9. What is one specific task our visualization could simplify for you?

   Finding players to scout would be very helpful. I currently do it all in an Excel file, so being able to see all the data visually and have clickable filters would be so helpful.

Interview with Persona 2 (Sports Bettor)

1. What is your occupation?

   I am a sports better with a focus on soccer matches.

2. What is your educational background?

   I have a high-school diploma.

3. Can you explain your workflow to me in 5 steps?

   The first step in my workflow is to determine what soccer game I want to make a bet on. Usually the most popular game for the day will net the biggest possible returns. My next course of action is to do some research on historical match, team and player data. I look at previous matches to see compare the win/loss ratio of each team in the current season. I also look at the key players on each team to see if they are performing consistently. With this combination, I determine which team will probably win. Then based off how strongly I feel about my prediction, I will bet money accordingly.

4. How often do you use visualizations in your daily work?

   I don't use them often as I am looking for numbers and statistics to compare, but sometimes they come up when I am looking for match statistics. I have found some to be insightful, but I am not familiar with creating my own.

5. What takes up most of your time on a day-to-day basis?

   Finding and analyzing data takes up a huge portion of my time. Often times there are different games and different league. As I am trying to maximize my intake, I have to manually compare data for each set of teams. Given several leagues as well, this process is especially time-consuming.

6. What is 1 thing you really like about your current process?

   My process is something that I have developed for a while it. By now, it's a process that is very intuitive to me and I am very comfortable with going through the steps.

7. Can you name 3 elements that would be essential in a helpful visualization?

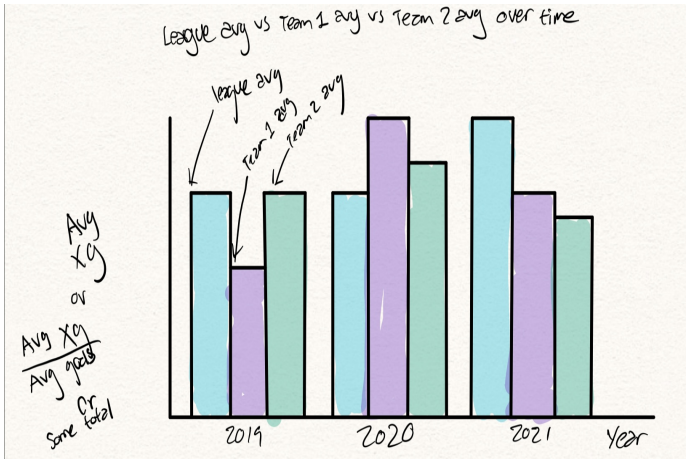
   The most important thing would be for it to save me time when it comes to analyzing. Keeping it simple will probably be best, as I want to easily see prediction results as I pretty much always bet on the outcome of the game.

   I also want to be able to see something that compares player performance so I can apply the data to player-centered bets.

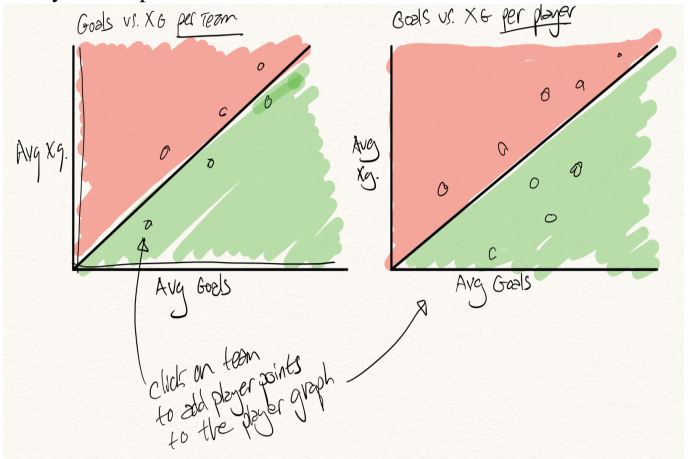
8. Can you list 5 insights that you would want to get out of the visualization?

   Most importantly, I want to be able to have a better prediction of what team will most likely win a given game. Otherwise, I want to be able to see what teams have been performing well in the season. Another thing I would also want is some insight regarding how much money I should most likely bet. Lastly, I would also request similar visualizations but to compare player versus player performance for better predictions.

11 Appendix D: Design Sketches

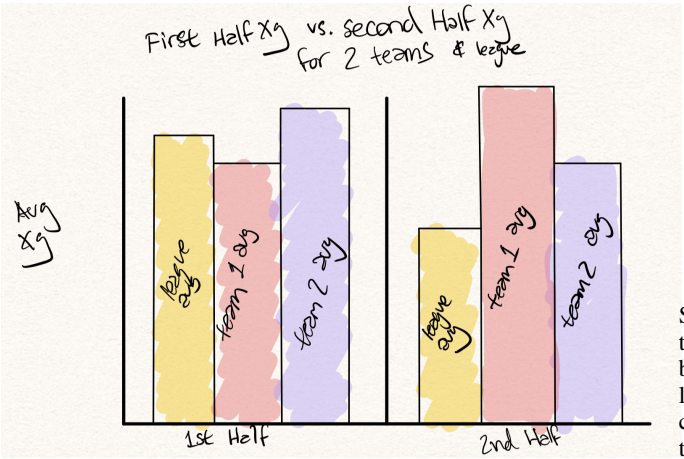


Sketch 1 by Forrest Meng.The marks in this visualization are the bars in the bar chart, while the channels are the different colors of the bars. This encoding can be used to compare the averages of several different xG related statistics across teams in the league. This addresses the an analyst's tasks of seeing and comparing season and league based averages, a soccer enthusiast's task of seeing if their favorite teams are over or under-performing over the years compared to other teams in the league, as well as a sport bettor's task of determining which team is more successful and therefore how they should place a bet.

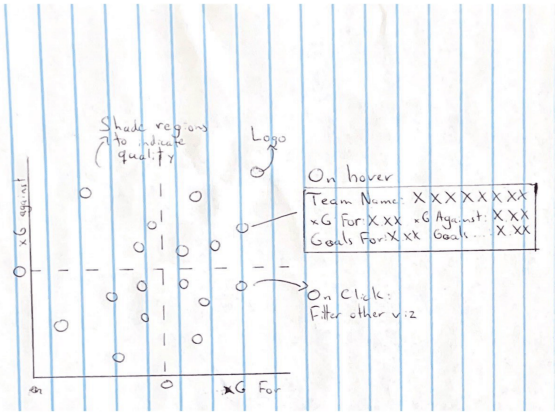


Sketch 2 by Forrest Meng.The marks in this visualization are the

points and diagonal line in the scatter plot. The channels are the different colors on each half of the diagonal line. Size of the points could also be a potential channel if the team decides to implement it. This encoding can be used to compare how well teams and players are playing to how well they are expected to play. This addresses the soccer scout's task of seeing how players are performing versus how they are expected to play, an analyst's tasks of seeing and comparing season and league based averages, a soccer enthusiast's task of seeing if their favorite teams are over or under-performing compared to other teams in the league, as well as a sport bettor's task of determining which team is more successful and therefore how they should place a bet.

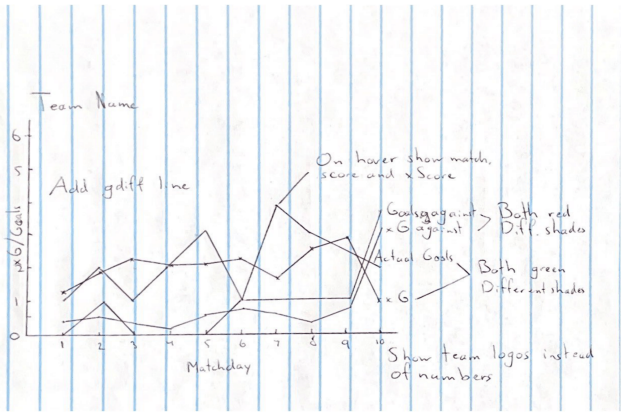


Sketch 3 by Forrest Meng.The marks in this visualization are the bars in the bar graph. The channels are the different colors of each bar. This encoding can be used to compare how well teams in a league are expected to play. This addresses an analyst's tasks of seeing and comparing season and league based averages, a soccer enthusiast's task of seeing if their favorite teams are over or under-performing compared to other teams in the league, as well as a sport bettor's task of determining which team is more successful and therefore how they should place a bet. The inclusion of two separate halves for each match are also designed for team performance to be analyzed. For example, a soccer analyst or bettor may use this visualization to help them determine the likelihood of a comeback.

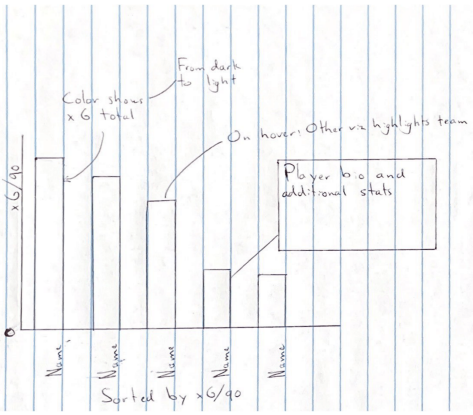


Favorite

Sketch 1 by Floris Dobber. I chose points and lines as my marks because I felt like it clearly conveyed the information. I chose shape, color, and position, both horizontal and vertical, as my channels because they conveyed the information in the right order. Position is the most important aspect, which is noticed. Color is interpreted second, and gives users reinforcement on a teams performance. Shape is interpreted last, which is a detail for when the user wants to identify which team the points represents. The visualization supports a sports fan who's seeing a team is over or under performing.

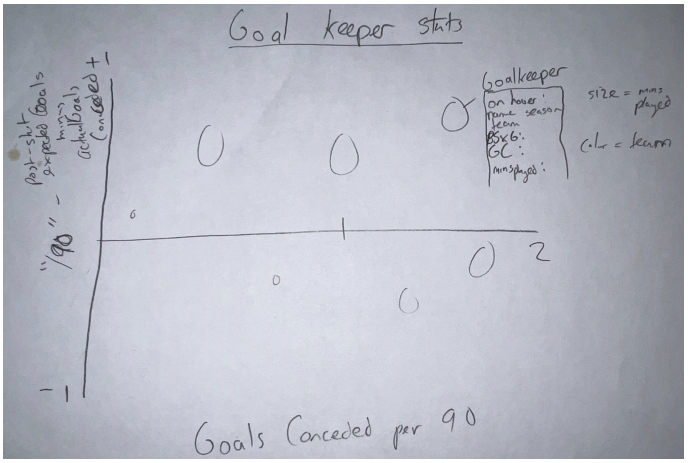


Sketch 2 by Floris Dobber. I chose the lines as my marks because they do a great job of portraying trends over time. I chose position, both horizontal and vertical, and color as my channels because I felt like both were intuitive to understand and would make the graph clean and easy to look at. It is intended to support a sports better trying to identify which teams are over or under performing, and their lack of experience with visualizations should be taken into account by making the graph easy to interpret.



Favorite

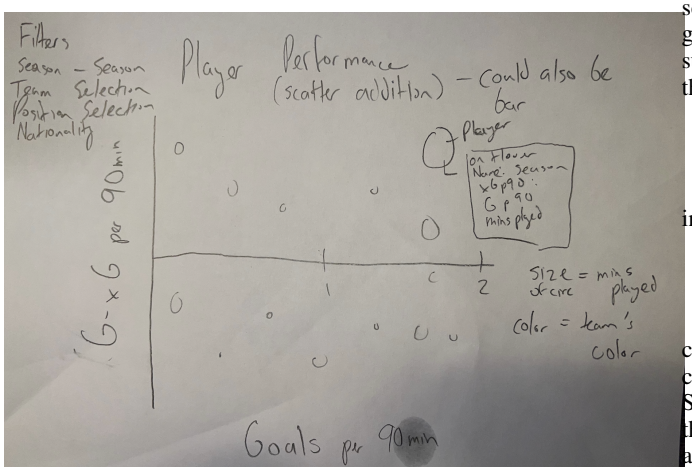
Sketch 3 by Floris Dobber. I chose lines as my marks because I felt like they very clearly linked to their name label. I chose position, both horizontal and vertical, and color as my channels because of their easy comparative value. Position immediately jumps out to the user, especially given the ordering, and then color provides some additional details when it is interpreted second. The visualization is intended for a soccer scout trying to determine which players are performing the best.



Favorite

Sketch 1 by Henry Dench. I chose points/circles and a dotted line as my marks to effectively show information. I used color as a mark to signify what team the points/circles correlate to. Position is the key mark used along with the dotted line to quickly display to the user whether the team is under or over performing and there is an idea to add a subtle gradient to the background to help with that as well. Size is also used to show if the team is under or overperforming their xG, using G/xG as the height of the circle and xGc/Gc for the width of the circle. When hovering over the circles, more information about the team, season, and key stats will be present.

This visualization focuses on the tasks specified by the sports better, as it shows them under and over performing teams compared to the rest of the league and possibly, their next opponents.



Favorite

Sketch 2 by Henry Dench. I chose points/circles as my marks to effectively show information. I used color as a mark to signify what team the player is on. Position is the key mark used along with the x axis to quickly display to the user whether the player is under or over performing with goal creation and finishing (xG and G). Size is also used to show how many minutes the player has played that season, as the goal is to highlight both over/under performance along with consistency. When hovering over the circles, more information about the player, season, team, and key stats will be present.

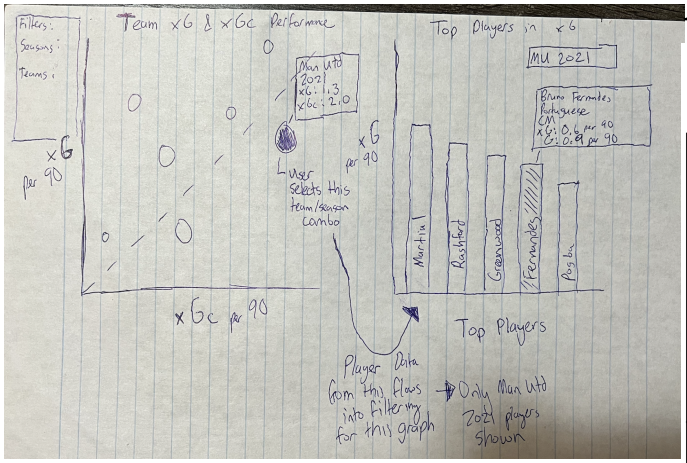
This visualization focuses on the tasks specified by both the sports scout and a sports fan, as it shows them under and over performing players compared to the rest of the league.

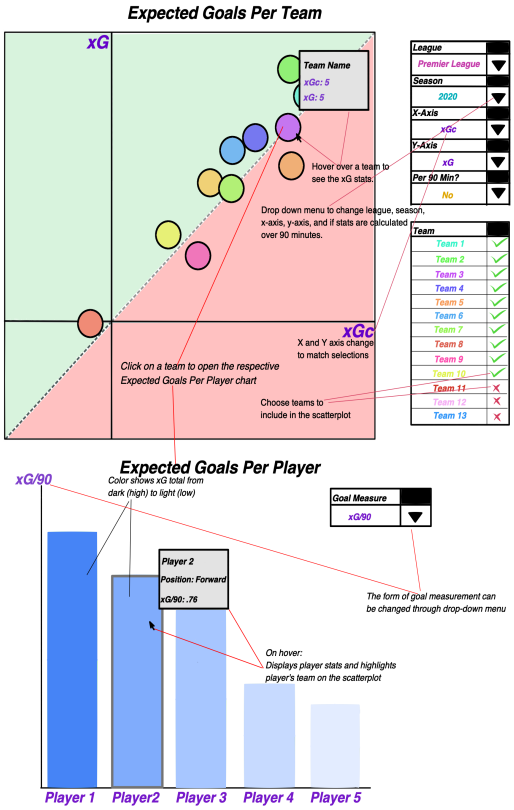
Sketch 3 by Henry Dench. I chose points/circles as my marks to effectively show information. I used color as a mark to signify what team the player is on. Position is the key mark used along with the x axis to quickly display to the user whether the goalkeeper is under or over performing using the comparison of xGSOT (shots on target) faced and goals conceded (per 90). Size is also used to show how many minutes the goalkeeper has played that season, as the goal is to highlight both over/under performance along with consistency. When hovering over the circles, more information about the goalkeeper, season, team, and key stats will be present.

This visualization focuses on the tasks specified by both the sports scout and a sports fan, as it shows them under and over performing goalkeepers compared to the rest of the league. This visualization is super important because xG usually focused on the attacking side of the game but it can also be used to show a goalkeeper's effectiveness.

Note: Contacted prof. Mosca and got approval to submit fewer individual sketches.

Favorite Selection: We selected the three favorites based off the criteria we wanted for an interactive view, and then modified and combined some aspects for each. Both Sketch 1 by Floris and Sketch 1 by Henry are scatterplots that easily relay information to the user using position and dotted lines as base points. Position is a key reason this graph works so well. Users can easily identify if a team is doing better than average and that is key when a scout is looking for players of overperforming teams. We wanted to create a modified version of the two, taking inspiration from both. This also helps with the domain task associated with sports fans who want to consume and enjoy data. This visualization also supports the browsing of soccer scouts, especially when combined with out next chosen visualization. Sketch 3 by Floris was selected as it is a bar chart with a very simple style which is easily digestible by the user. This bar chart ranks players based on their xG per 90 and is the perfect fit for an interactive visualization we can pair with our Team Scatterplot Visualization. Once again, the position (ordering) of this bar chart again focuses on the domain tasks associated with scouts, fans, and analysts. The visualization helps the users consume, explore, and identify the key statistic with little noise distracting from the data.





The two visualizations are linked by brushing and data filtering. If a user was to select a Team/Season combo from the Team Scatterplot, only players with those Team/Seasons combos will be shown on the bar chart. If the user was the select a group of points on the scatterplots, the bar chart would update to only contain the data of the brushed objects. The ease of selection using the Team Scatterplots to impact the Player Performance Bar Chart augments the tasks it focusing on completing. By allowing the user to use the Scatterplot as a general overview of team data, the user can then dive deeper into browsing, consuming, and exploring such data using brushing and linking. Our visualizations combine to create a fast and easy way to dive into the second level data that a scout, fan, or analyst would want to see. We have to have the brushing and linking flow from the Scatterplot to the Bar Chart since they are differing levels of data (Teams vs Players). It would not make sense to have brushing from the bar chart link to the scatterplot in anyway more than highlighting the point of the player's team in the scatterplot. Since this example of brushing and linking is very simple, it does not involve any data manipulation or filtering like the previously discussed brushing and linking.

12 Appendix E: Digital Sketch

A soccer scout may use the visualization by first determining what team they want to scout a player from. They can do this by adjusting the options to the right of the scatter plot so that they choose from the correct league, season, and measures of performance. The soccer scout will then have a scatter plot with points representing each team in the league. The scout can then easily determine what teams are performing better than expected, and use the filter to deselect teams that they aren't interested in. The scout can hover over any team's point to see more statistics, and can go further by clicking on the point to initiate the associated Expected Goals Per Player chart. This chart gives important player data that the scout can use to determine the best player for the position that they are trying to fill.

GitHub Classroom-Generated Repository

13 Appendix F: Usability Testing

13.1 Preparation

The visualization tools intended use is to discover insights into the performance of teams and players in the English Premier League. It shows the performance and expected performance across a variety of metrics such as goals and assists. By comparing actual performance versus expected performance it can be determined if any teams and players are wrongly valued in either the betting or transfer market.

1. Identify which team is the best performer this in the 2018 season.

   This is a core purpose of the tool, and can be done without touching any buttons. If users don't immediately recognize that this is possible, perhaps additional text should be added to explain how to use the tool. I will be looking at speed of completion and the number of clicks to get the result.

2. Who has the most xG for Manchester City in 2019?

   This task requires a few clicks and a better understanding of the tool, and so it is a good test of how easy to use to tool is. Following users completing the task, I will ask them how easy they thought this was to complete on a scale of 1 to 5. 1 will be hard, 5 will be easy. Speed of completion will also be used as a non-objective measure of how easy this task is, with lower speeds indicating greater ease of use.

3. Which team is over-performing in 2021?

   This task is good to test the completeness of the visualization. If a user is able to find the answer, it will mean that the visualization can be used for this core purpose. I will be mostly focused on whether a user can complete the task, and if the visualization performs as one would expect throughout the task. So the only outcome I will be measuring is completion, and I will record any bugs that pop up.

13.2 Results

Overall I think testing was a positive experience. I think no major issues were highlighted, which was a positive sign that the visualization is on the right track. A minor issue that quickly became obvious is that all elements that ”look” interactive, look a drop down menu, should be interactive or replaced by non-interactive looking elements. This can lead to confusion otherwise. The participants really liked the logos of the teams, as it quickly provided them with an idea of which team is which. The lack of a legend or color-scale for the bar graph was something they would like to see, to get a better idea of what the shading means.

The first task highlighted a lot of positive aspects of the visualization. Users were quick to complete the task as they could recognize the team by the logo and found the axes on the scatter plot to be intuitive. All users only used one click, and many did not feel a need to even hover over the team. I think this test shows that perhaps additional hints on the scatter plot to aid interpretation aren't necessary, as users were able to quickly interpret the graph.

The second task was less positive, immediately highlighting the need for linking between the two graphs. Without it, users had to hover over each bar and hope to find a Manchester City player. This process caused them to be quite slow, and find the task much harder than it has to be. I think once the planned linking is included in the visualization, users will perform much better on the task and find it a more pleasant experience.

My last task was not possible at the time of testing. Once completed, I will ask my test users to complete the task and see if they are able to do it. Since I was only looking for completion in this task, it is clear that the visualization is not feature complete and still needs some work. This was expected, and did not provide me with much additional information unfortunately. I did describe how users would be able to complete this task in the future, and my test users were generally impressed with the idea. They specifically liked the visual way in which users would be able to observe over-performance, and get additional details on hover.

I think the biggest modification that came out of the user testing is that I might have under-estimated my users. They were able to understand the graphs intuitively much quicker than I expected. Adding additional information might only increase the cognitive load and not do anything to enhance their understanding of the graphs. Besides that, I think once the graph is in a more final form, the other tasks will have similarly positive results as the first one.