IGR204: Visualization

Football Scouting

Group U:

BEN BOUBKER Mahmoud
MOUISSA Ismail
BAKEBECK Samuel
JAIT Fatima-Ezzahra

Table des matières

I- Dataset

The dataset represents the statistics of players playing in the 5 biggest European leagues (Spain, England, Italy, France and Russia) during the last six years.
The data used to the creation of this dataset can be found on the following website: https://understat.com/
The data has been scrapped thanks to the code from the following Github repository: douglasbc/scraping-understat-dataset

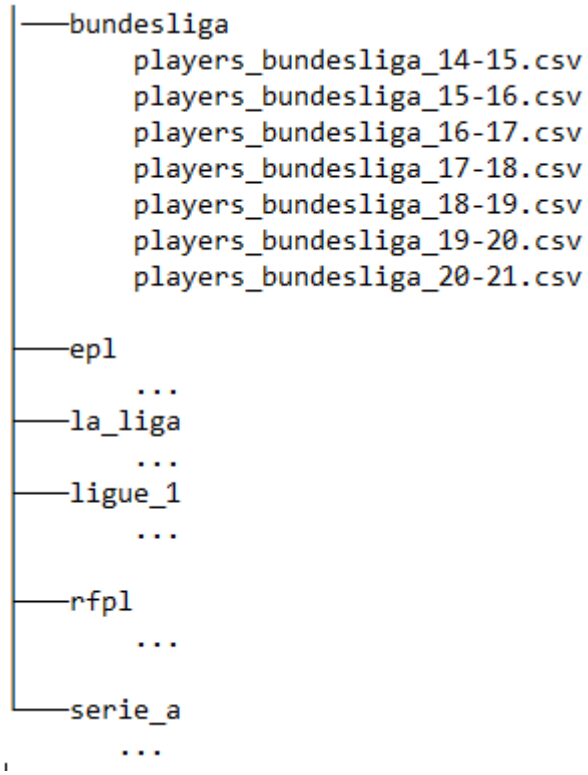
The data recovered by scraping has two different schemas:

- First schema: accumulated statistics by a player during a specific season in a specific league (number of games played, goals, assists ...)
- Second schema: Summary of actions carried out during a specific match during a specific season (goals, shoots, passes ...)

Thereafter, we will present the datasets as well as the transformations carried out for the different charts.

I-1- Players Dataset

I-1-1 Data Presentation



The recovered data is divided into 5 folders. Each folder represents a championship. Each folder is made up of 6 to 7 files. Each file contains statistics for a given season.
So, we have 5 x 6 = 30 data files.

Since the number of files is important, data preprocessing is necessary.

I-1-2 Data Preprocessing

The goal of this preprocessing is to have a single file (and over 30).

By observing a random file, we observe that a line is characterized by the player's id. This information cannot determine a row if we have the statistics of a player over several seasons or if the player left the championship for another during the same season.

Therefore, the new identifier of a row in the dataset is as follows: "idPlayer_season_league".

Other columns had to be transformed. The first is "position". It represents the position played by the player on the field. The values are as follows:

- S: Forward
- M: Midfielder
- D: Defender
- GK: Goalkeeper

However, some players can play multiple positions. Thus the "position" column is no longer atomic (example: F, S ...)

This is why new columns will be created. One for each position and the value will be Boolean (1 if the player plays in this position, 0 otherwise).

The same logic was applied for players who played in the same season at two different clubs. We create two columns: one for the first club and the second if the player plays in another club.

I-1-3 Schema of Data

	id	player_name	games	time	goals	xG	assists	xA	shots	key_passes	...	npg	npxG	xGChain
0	356_14-15_bundesliga	Alexander Meier	26	2209	19	15.395941	2	1.268694	80	22	...	16	12.364834	12.408789
1	227_14-15_bundesliga	Robert Lewandowski	31	2493	17	17.722183	5	4.644013	104	32	...	16	16.964407	23.315430
2	392_14-15_bundesliga	Arjen Robben	21	1681	17	10.364112	7	7.505970	88	50	...	15	8.848720	20.820983
3	158_14-15_bundesliga	Bas Dost	21	1532	16	11.341977	4	2.659186	43	15	...	16	11.341977	15.911891
4	318_14-15_bundesliga	Pierre-Emerick Aubameyang	33	2724	16	15.396760	6	6.378020	102	46	...	14	13.881207	20.852728

an overview of the data

The data table contains 20 columns and approximately 22,000 rows.

Column	Type	Description
Id	Categorical	Identifies a season, player and league
Competition	Categorical	League (France, England, Italy, Russia, Spain)
Season	Ordinal	Season (or year) when the statistics were recorded
Games	Continuous	Number of games played
Time	Continuous	Number of minutes played
Goals	Continuous	Number of goals scored
xG	Continuous	Number of goals expected to be scored
Assists	Continuous	Number of assists made
xA	Continuous	Number of assists expected to be made
Shots	Continuous	Number of shots
Key passes	Continuous	Number of key passes
Yellow cards	Continuous	Number of yellow cards received
Red Cards	Continuous	Number of red cards received
Team Title 1	Categorical	Team in which the player is playing in
Team Title 2	Categorical	Second team in which the player is playing in (Optional, if the player changed team)
Npg	Continuous	Non-penalty goals
npxG	Continuous	Non-penalty goals expected
xGChain	Continuous	Expected goals after a possession in which the player is involved
xGBuildup	Continuous	
GK	Binary	Player is a goalkeeper
D	Binary	Player is a defender
M	Binary	Player is a midfielder
F	Binary	Player is a forward
S	Binary	Player is a striker

The next table presents the schema of the second dataset, shot dataset

Colonne	Type	Description
Minute	Continuous	Number of minutes played
result	str	Determines the result of the shot, this variable can take several values in our project we only focused on these 3 values: MissedShots, BlockedShots, SavedShots, Goal.
X	Discrete	Determine the abscissa of the player at the time of the shot
Y	Discrete	Determine the ordinate of the player at the time of the shot The X and Y values are set relative to the opposing team's half court.
xG	Continuous	Expected goal
player	Categorical	The name of the player who shot the ball
h_a	Binary	Indicates whether the shooter is playing at home ‘h’ or away ‘a’
player_id	Categorical	the shooter's id
situation	Categorical	Situation of the match at the time of the shot for example 'FromCorner'
shotType	Categorical	This variable takes 4 values: RightFoot, LeftFoot, Head and OtherBodyPart
player_assisted	Categorical	The player who touched the ball before the shooter

II- Users

Users:

Scouts, Head of scouting unit, Director of Football

Background and tasks:

A scout will supervise players performance and will carry out a report for his head of scouting. The head of scouting will analyze the needs of his team, assign players to be scouted and will forward the best reports to the Director of Football. The Director of Football will have to convince the board to recruit the players. He will also have to negotiate with the player's current team for the transfer fee and the contract with the player's agent.

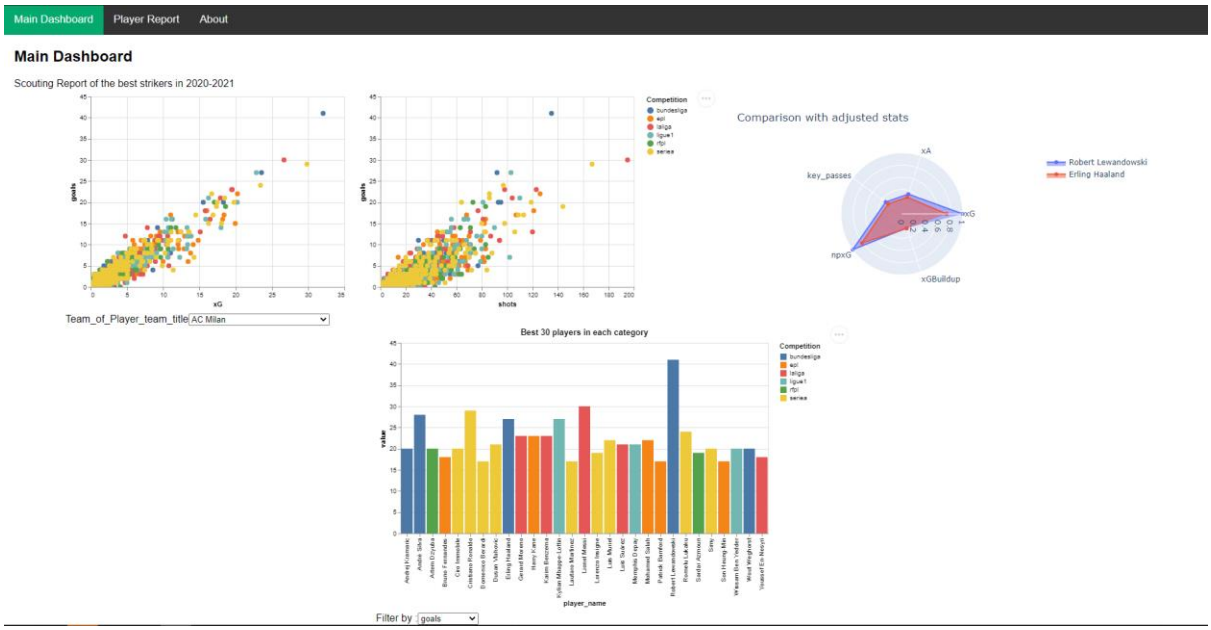
Purpose of the data:

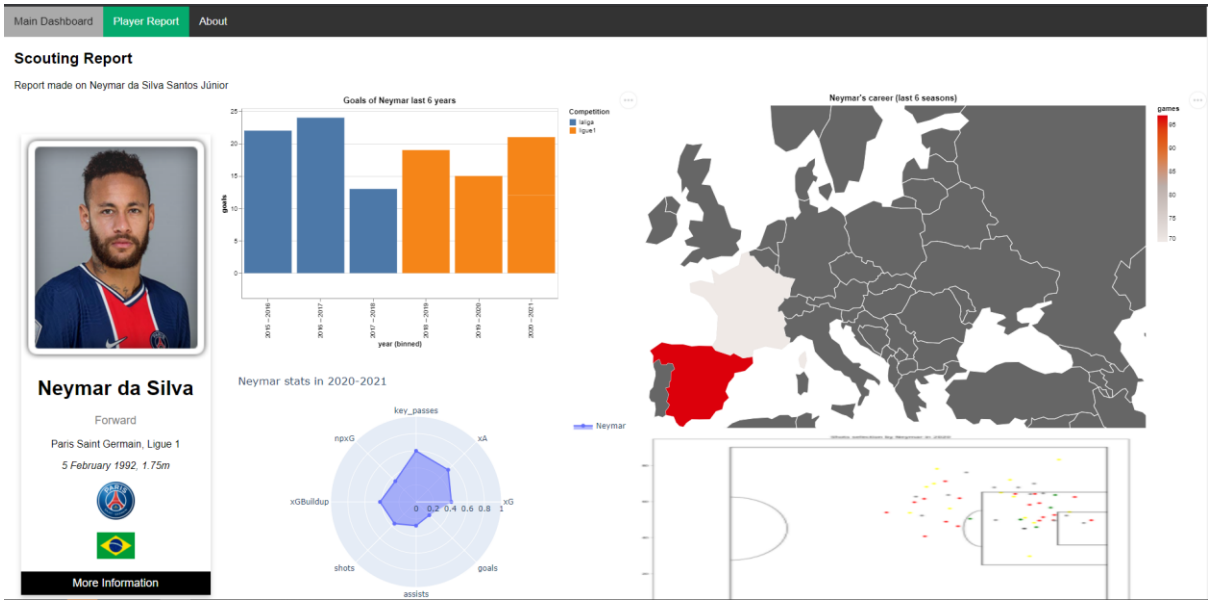
The visualization aimed primarily at exploring the data. The idea is to analyze and decipher the different methods (data exploration, data processing, data mining and data visualization) to help identify: - players with similar skills - potential market value of these players - position within the teams where these players would be the most effective.

The visualization can also be used to justify the purchase of players to a board that lacks football background.

III- Chosen Design

Our project is available on the web in this link





The two illustrations below represent the two interfaces of our dashboard.

The purpose of the dashboard is to aid the exploration of the user in his search to find potentially useful players who could not have been found directly by the raw data.

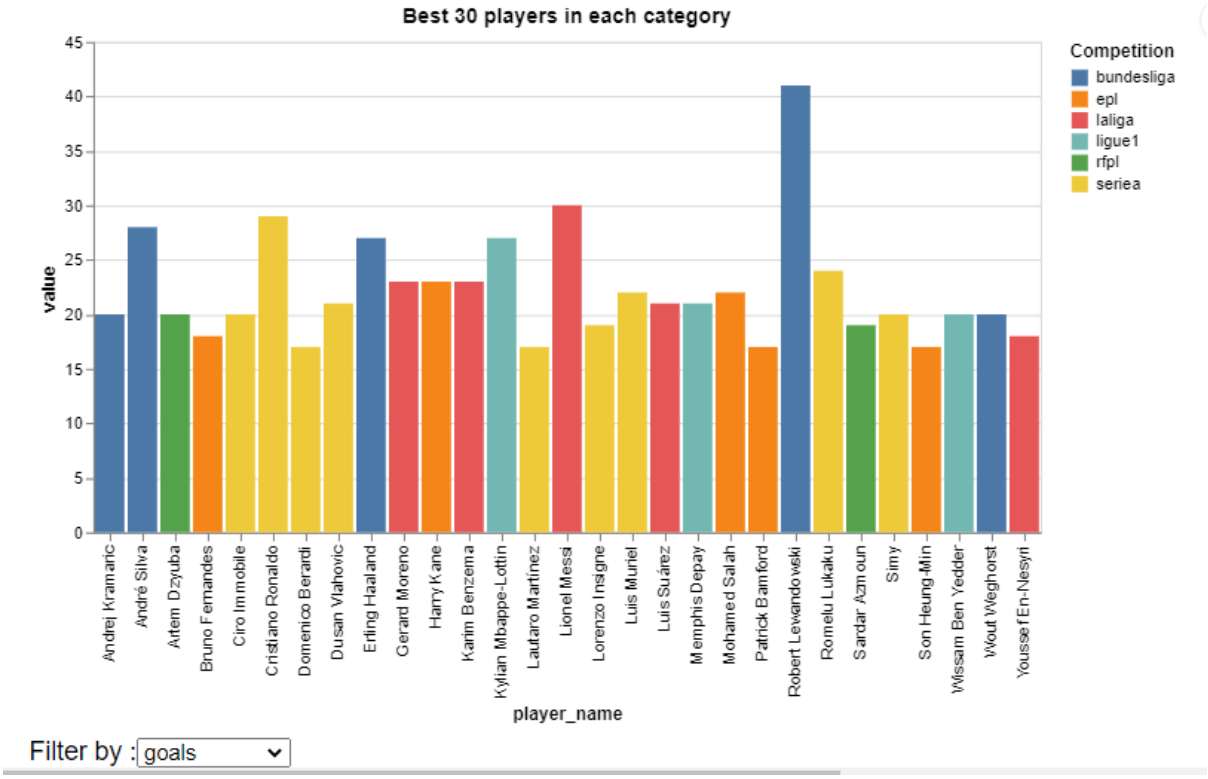
The first illustration represents the first dashboard which is also the main one. The user has access to data on multiple players. He can confront them, compare them.

The user can select a player from one of the graphics.

Subsequently, the second interface is displayed. This time around, the graphics are specific to the chosen player. This interface therefore allows for a more in-depth analysis of this player.

In the next section we will provide information about every chart from the two dashboards.

III-1 First Dashboard



What is your chosen design?

- What representations did you choose to use?

For this first visualization we have chosen to show histograms which give the ranking of the different players over a season according to a well-defined criterion.

- What interaction does the design handle? /Can you relate these design decisions to your chosen users, data, and tasks?

In this Visualization three interactions are possible:

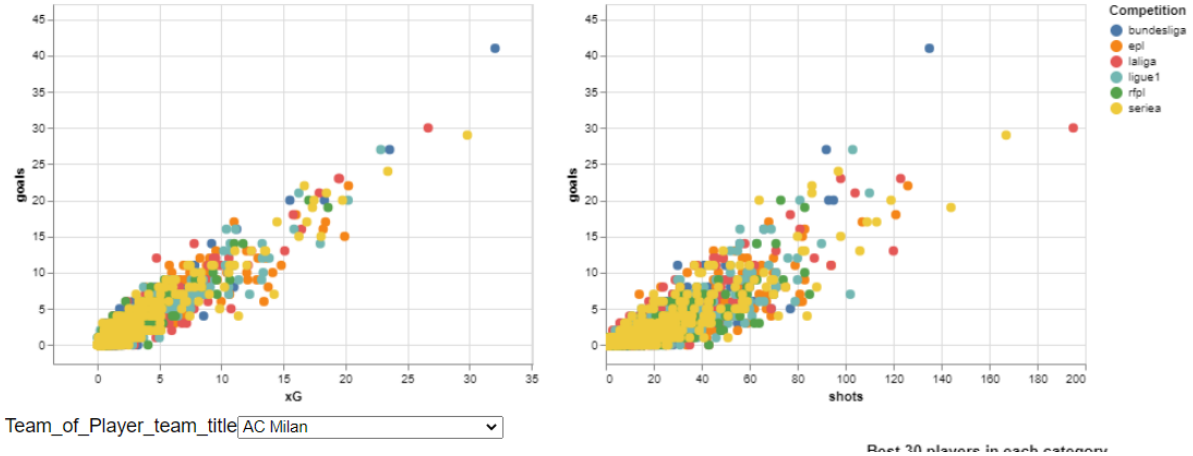
- The choice of the selection criterion: Indeed, we offer the user several choices according to his objective. If the user seeks to recruit an attacker, he can select the 'goal' or 'shots' criterion to find the 5 best players in this position, while if his goal is to recruit a midfielder, the 'Assists' criterion is the more adapted.

- The second interaction makes it possible to highlight the players of certain championships, by selecting the championship in question in the legend. Indeed, certain criteria are more relevant according to the championship, in particular the number of goals which vary for example between the German and French championship.

- The third and last interaction concerns the selected players. We display their stats individually and this in a box that displays when you hover the mouse over the histogram of a particular player.

- What does the design do well?

What this visualization does well is that it provides an overview of the different player rankings with the different criteria. But this visualization has two flaws the first is the impossibility of choosing the position which would have been a key criterion for the recruiter and the second is related to the proposed interaction which does not allow access to the individual pages of the players



- What representations did you choose to use?

This chosen visualization allows three statistics to be compared: goals, shots and expected goals. This visualization thus makes it possible to relativize the number of goals scored by using the number of shots and xG to observe effective players.

- What interaction does the design handle? /Can you relate these design decisions to your chosen users, data, and tasks?

In this Visualization three interactions are possible:

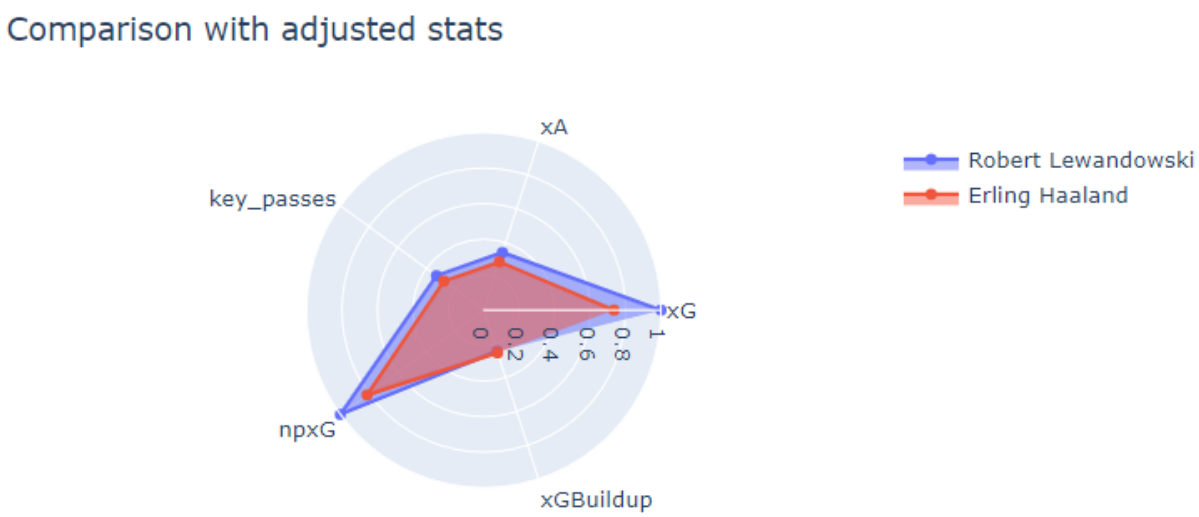
- The first selection criterion is to display the players of a particular choice. In addition to using a color scheme for the championships, the user can click on one of the colors in the legend to display only the players of the color chosen for the two connected graphs.
- The second interaction is also a filter. using a drop-down menu, you can choose to display the players playing in the selected team.
- The third and last interaction concerns the selected players. We display their stats individually and this in a box that displays when you hover the mouse over a specific player's point.

- What does the design do well?

This graphic has a strong added value. The "goals" statistic is sometimes very overestimated if it is not used alone. We can think that a player who has scored a lot of goals is necessarily strong. However, this can be put into perspective by combining with the other two statistics.

In addition, this graph provides a global vision because it allows to see certain outlier players (whose point is far from the scatter plot) and to see that championships have more good players than others.

In addition, other interactions could have been added but for reasons of competence or technical limitation it was not possible. The Altair library can only display a maximum of 4000 points. We had to be content to display that the scorers (so not the attackers and midfielders)



- What representations did you choose to use?

This representation is a radio chart allowing to obtain several statistics of a player. The usefulness of this graph is also the comparison. A player with a larger area than another player may be considered better.

- What interaction does the design handle? /Can you relate these design decisions to your chosen users, data, and tasks?

It is possible to add or remove players in the radio chart to make a comparison.

By browsing the graph with the cursor, a box is displayed with more details about the player.

- What does the design do well?

Comparing players is quite complex because it depends on several parameters.

The radio chart can display many statistics for a single player. In addition, it is possible to monitor player statistics and therefore compare them by superimposing layers of statistics.

III-2 Second Dashboard

- What representations did you choose to use?

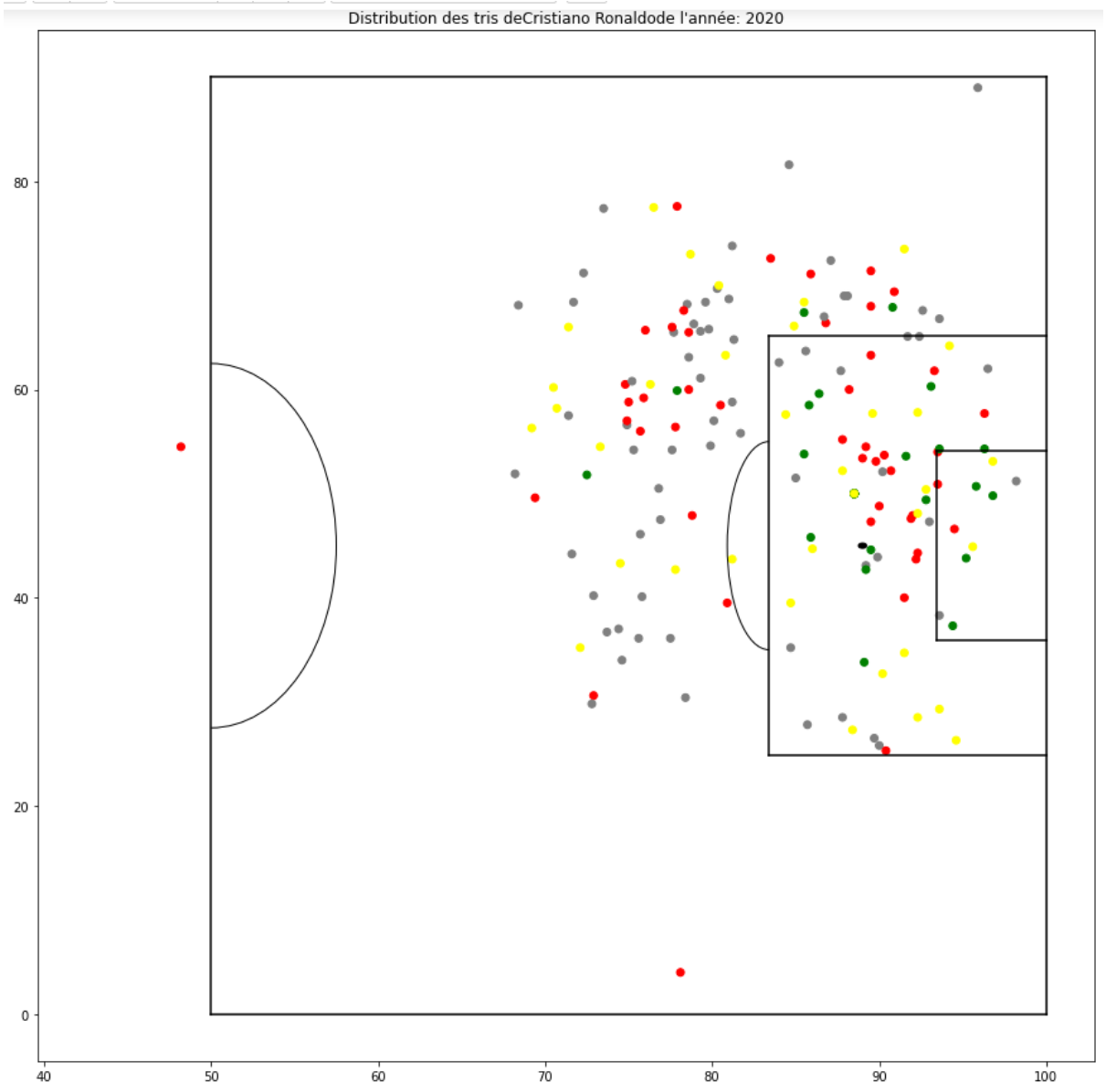
This graph is a map of Europe containing information about a player's career over the past six years. The scale used is the number of goals scored in the countries where he has played.

- What interaction does the design handle? /Can you relate these design decisions to your chosen users, data, and tasks?

By browsing the graph with the cursor on every none-gray country, a box is displayed with more details about the player and the statistics made in that country.

- What does the design do well?

In Europe, there are a lot of players but also many leagues in many countries. These countries do not have the same level of football. So, with this graph we can observe the evolution of the player over the years.



- What representations did you choose to use?

  For this second visualization we have chosen to show a shot card of a player for a given year, in this visualization the dots indicate the position of the player at the time of the shot while the color indicates the result of the shot.Le code couleur est le suivant :

- Red: The shot is blocked by a defender
- Yellow: The shot is blocked by the goalkeeper
- Green : Goal

- What interaction does the design handle?

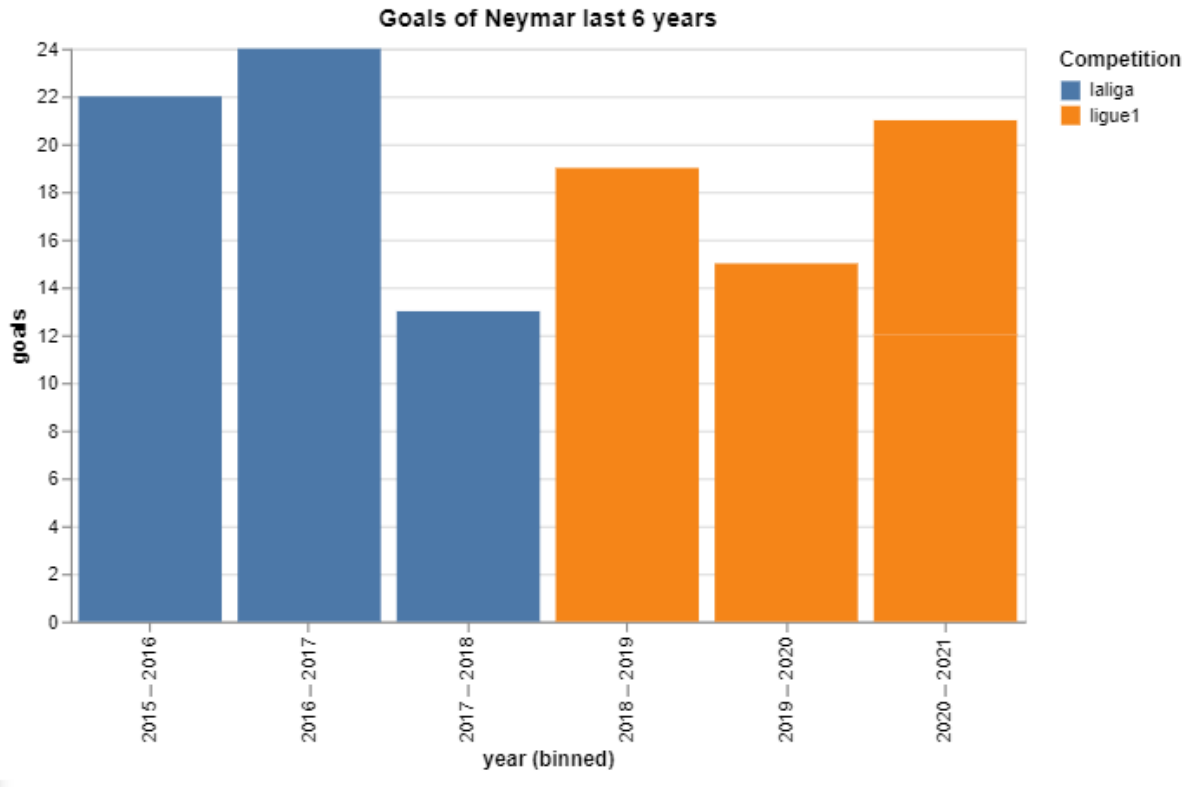
The possible interactions with this visualization are player choice and season choice.

- Can you relate these design decisions to your chosen users, data, and tasks?

This visualization is particularly important for a recruiter who is looking for a player at the post of a center forward, or the criterion of the number of but of is not enough but it is also necessary to know the effectiveness of the player. The dataset we have is completely suitable for this type of visualization because we have all the necessary information.

- What does the design do well? / What does it not do well?

While our design allows to display a player's shot card over a season with a color code that lets you know the outcome of the action, it lacks some interactions such as displaying more information on the game. shot (date, match, action before the shot) when you hover the mouse over this shot, we can also imagine a filter that only displays successful shots or those blocked



- What representations did you choose to use?

This graph is a histogram containing information about a player's career over the past six years. The scale used is the number of goals scored in the countries where he has played.

- What interaction does the design handle? /Can you relate these design decisions to your chosen users, data, and tasks?

By browsing the graph with the cursor on histogram, a box is displayed with more details about the player and the statistics made during that year.

- What does the design do well?

This graph allows you to see the evolution of the player: see if the player is improving or that his statistics are declining.

IV- Technologies

The main programming language used is Python to

- Scrap the data from a website.
- Preprocess the data.
- Create the charts.

The graphical library was using is **Altair**. However, it has certain limitations:

- it is not possible to create radio charts with this library. One of the open-source contributors admitted it. The alternative was to use Plotly to create this type of chart.
- Some charts cannot exceed more than 4000 rows of data. Therefore, for some illustrations, we had to reduce the size of our dataset.
- It is not possible to combine charts with different diagrams. This limits our visualizations. It is not possible to retrieve data from one selector and use it manually in another. Indeed, the selector is converted into Javascript and it is not possible to retrieve it in python to use it manually
- To create a web dashboard, you must use Vega. However, this library is much more complicated than Dash accompanying Plotly. So we exported in html the charts generated by altair.

For the web dashboard, we used plain HTML, CSS and JavaScript.

The code source of the project is available on Github using this link: https://github.com/MahmoudBenboubker/Data_Visualisation_Football

The dashboard is on this website: https://mahmoudbenboubker.github.io/