Project 2 Final Report — INFO 3300

Jason Steiner (jns226), Pinxian Lu (pl444), Richard Berlinghof (rb788), Nixon Zuniga (nrz4)

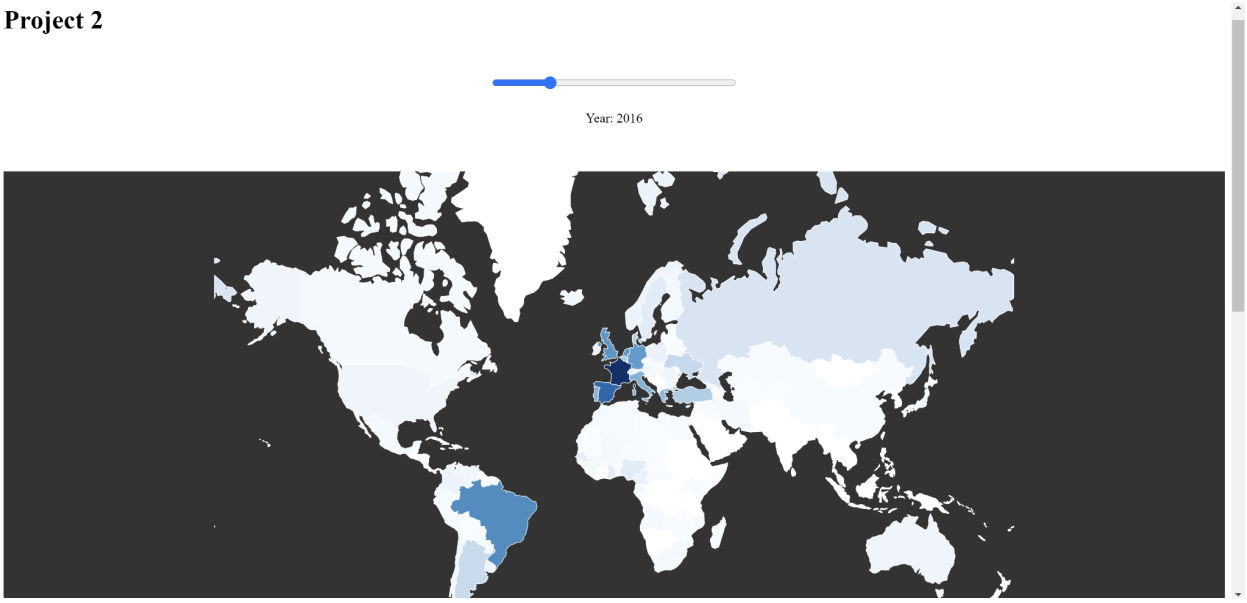
Data:

https://www.kaggle.com/datasets/davidcariboo/player-scores?select=appearances.csv

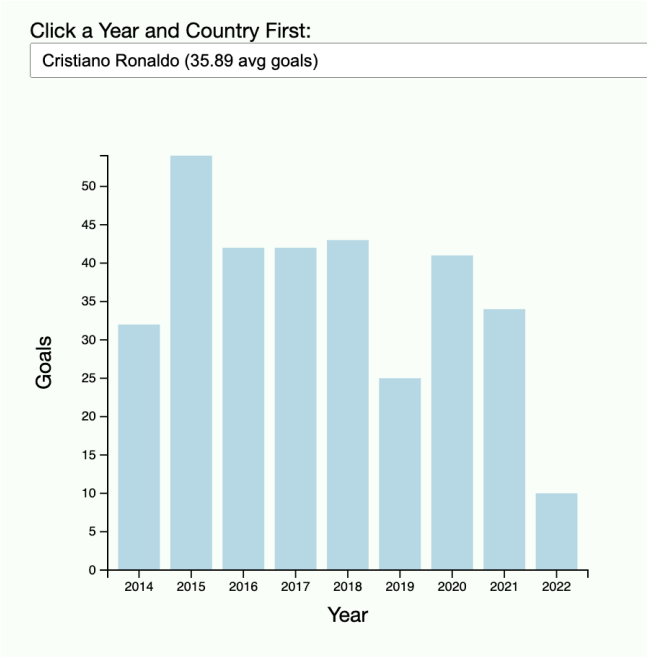
Our data is sourced from Kaggle datasets. The link to the kaggle page is above; it contains multiple .csv files that we compiled together. The main subset of variables that we used were player name, country of origin, goals, assists, highest_market_value_in_eur, and image_url. We felt that those metrics were most useful in providing a holistic and informative picture of the broader soccer landscape. We used multiple .csv files and combined them into a couple "master" datasets, namely a join on the datasets *appearances.csv* and *players.csv*. We grouped the player appearances by player id in order to aggregate the stats for each player and thereafter select the columns we wanted to capture. We were able to merge all the datasets together using the Python code below, using Google Colab to assist us in this effort. Even though we did not use every column captured by the Python program, the join on the two aforementioned datasets allowed us to have some extra information to consider while coding. We also used a geojson file to help us make the world map. An issue in making the geojson file work seamlessly was in the name discrepancies between countries—we had to change *United States of America* to *United States* and *United Kingdom* to *England* in order to match fully with the soccer datasets. This data helped us comprise the world chart, bar chart, and scatter plots. It's important to note that all of the datasets encapsulate the years 2014–2023. This means that all of our data is based on this 9-year window.

```python
import pandas as pd

appearances = pd.read_csv('appearances.csv', parse_dates=['date'])
players = pd.read_csv('players.csv', parse_dates=['date_of_birth'])

appearances['year'] = appearances['date'].dt.year

grouped_appearances = appearances.groupby(['player_id', 'year']).agg({
    'yellow_cards': 'sum',
    'red_cards': 'sum',
    'goals': 'sum',
    'assists': 'sum',
    'minutes_played': 'sum'
}).reset_index()

joined_data = pd.merge(grouped_appearances, players, on='player_id', how='inner')

selected_columns = [
    'player_id', 'year', 'yellow_cards', 'red_cards', 'goals', 'assists',
    'minutes_played', 'first_name', 'last_name', 'name', 'current_club_id',
    'country_of_birth', 'country_of_citizenship', 'date_of_birth',
    'sub_position', 'position', 'foot', 'height_in_cm',
    'market_value_in_eur', 'highest_market_value_in_eur',
    'contract_expiration_date', 'agent_name', 'image_url', 'url'
]
final_data = joined_data[selected_columns]

final_data = final_data.drop_duplicates(subset=['player_id', 'year'])

final_data.to_csv('yearly_aggregated_players.csv', index=False)

print(final_data)
```

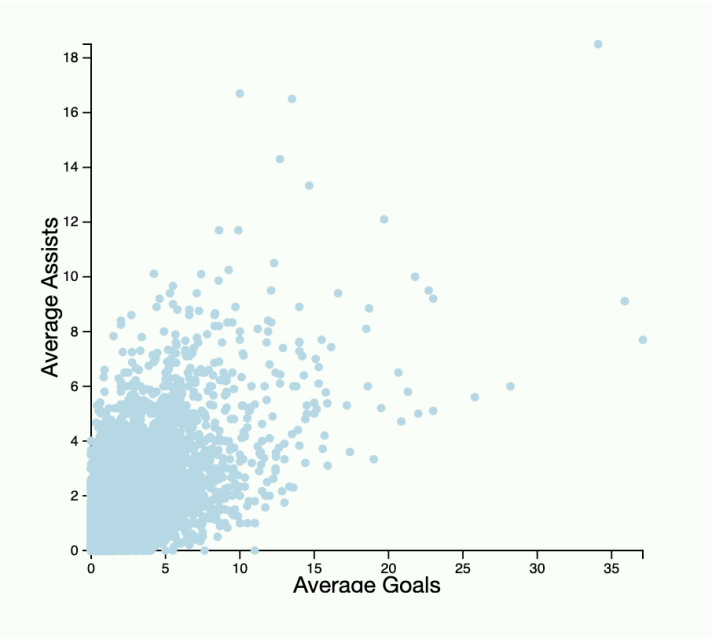
Visualization Overview:



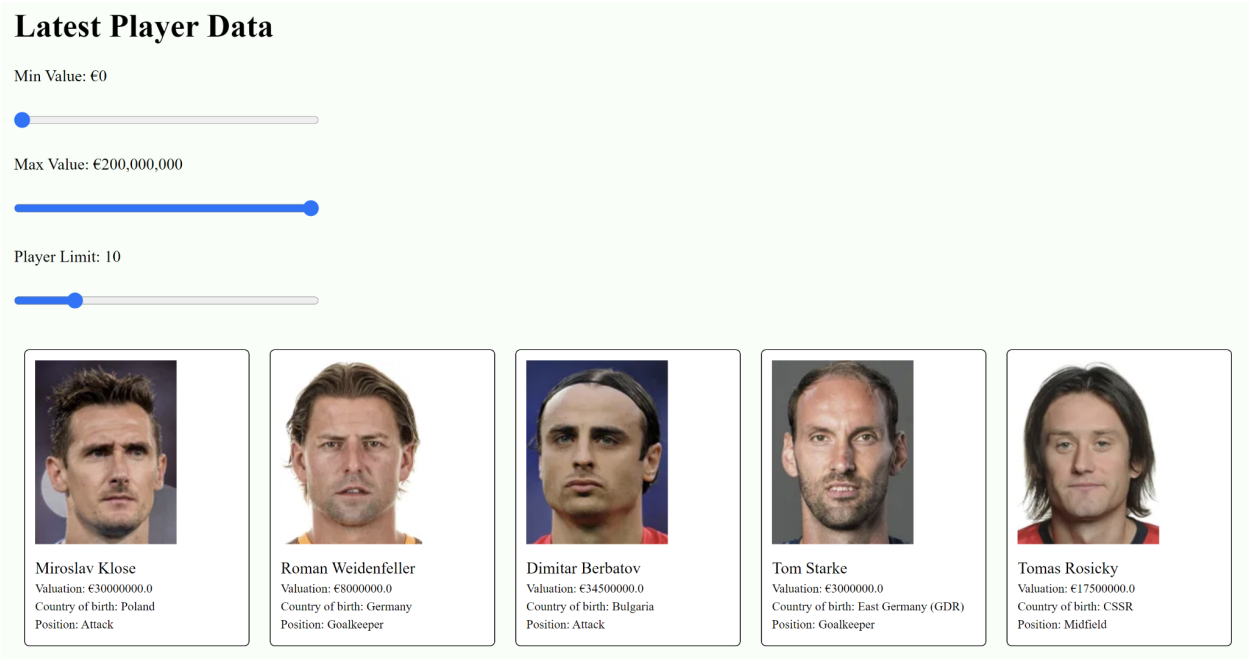
Our main visualization is the world map as shown above. Here, the number of players from a certain country is mapped to the saturation of that country. Countries that produce a lot of soccer players are more saturated, while countries with less players are less saturated, with white indicating no players being from that country. Hue remains constant, while we used a sequential color scale to create the variations in saturation. We employed sliders for data filtering, which encapsulates the years 2014–2023. We chose to give the map a dark background to create more contrast between the countries and oceans. We noticed this change also helped make the variations in saturation more pronounced. Position is another visual channel that is used in the map. It is used to get a sense of the distribution of geography between players.



In addition to the world map, we also implemented this supporting bar chart. The chart shows the number of goals a player had in the years they played between 2014–2023. It's important to note that when the user selects a certain country and year, the players from that country **who played in that selected year** will appear in the dropdown selector. For example, since Cristiano Ronaldo played from 2014–2022, he would not appear if the selected year were 2023. However, if the user selected him for any other year, they would find the corresponding time-series bar graph that shows his statistics from the years in which he played. In the dropdown, one can view his average number of goals between all the years he played (in this example 35.89). The main visual channel here is the length (or area) of the bars. Differences in this attribute show how a player has been performing over time. The main rationale behind this chart is that we wanted to be able to show player stats in addition to the geographic distribution that the map provides.



We also implemented this scatter plot that plots average goals against average assists for a player. This plot uses position as its main visual channel. Being further away from the origin on the x and y axes indicates a greater number of average goals and assists respectively. This contributes to our goal of holistically evaluating soccer statistics by visually representing opposing aspects of the same game and viewing the outliers among this representation of the data. The user can hover over any of the points to find the corresponding name, career average goals, and career average assists.

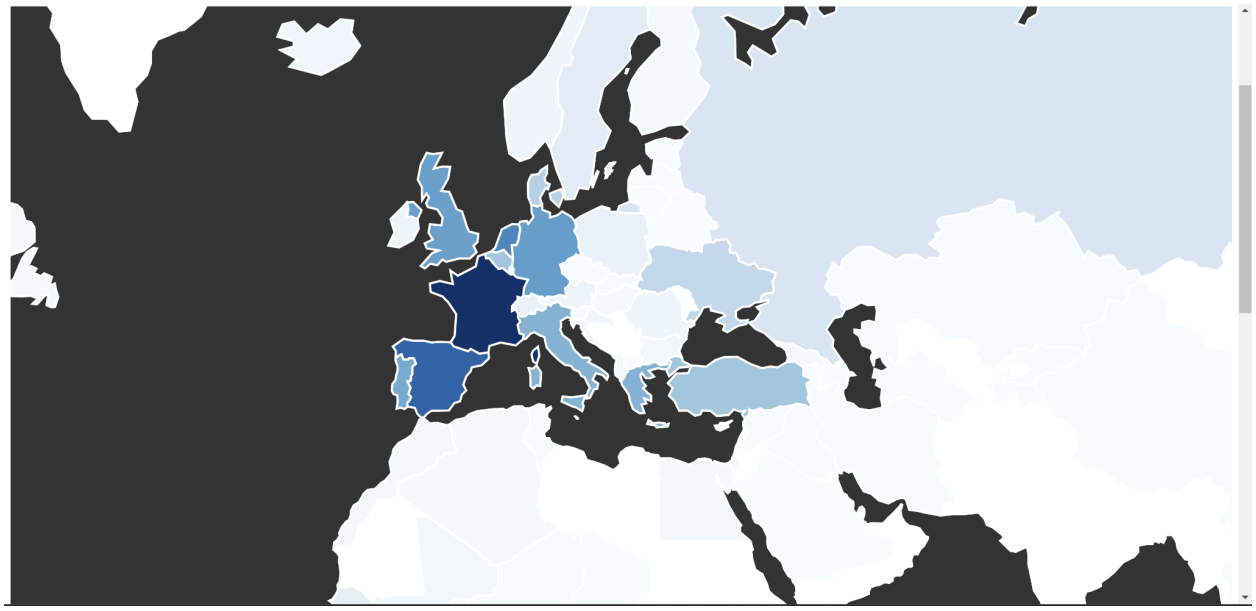


The final visualization we added is a player value section. The user can specify the maximum and minimum values of the returned players and how many players they want in view. In each player card, we chose to display the *Valuation*, *Country of birth*, and *Position*. This allows users to view where the most (or least) valuable players are from, valuable statistics associated with them, their pictures, etc. The main visual channels here are the positions of the sliders and the positions of the players. We decided to add this section because it supplements the world map visualization in an effective way, allowing users to get more details about players they may have initially analyzed through the other visualizations.

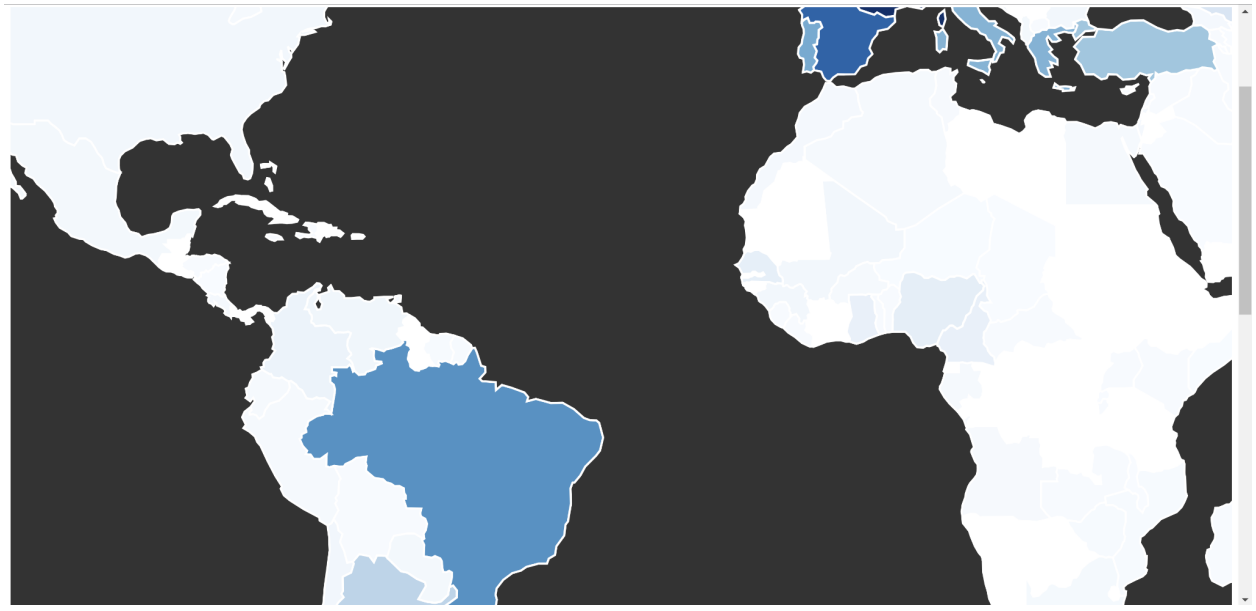
Interactive Elements:

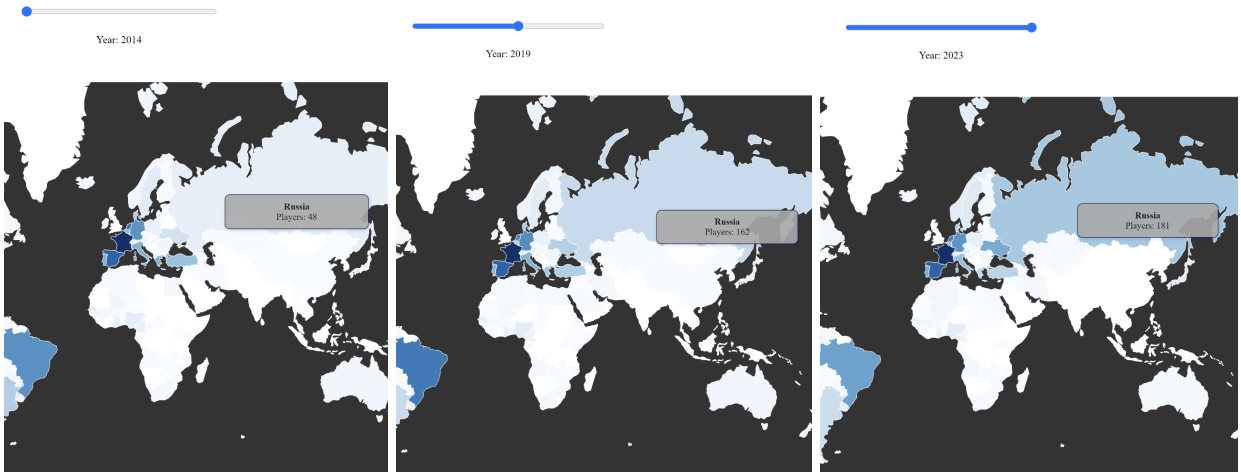
The interactive elements of our project are meant to enable the user to further explore the visualizations. Since the visualizations are meant to be interactive, we prompt the users to click on different countries and explore, increasing the discoverability of the interactions. Some visualizations, notably the bar graph, will not appear until the user clicks a country. While it may seem like this hides data from the user, we think that it actually encourages them to further explore. They are rewarded for their interaction by gaining access to more data and interactions. This keeps our interactions interesting. In terms of usability, most of our interactions involve mouse hovers and clicks. The mouse hovers are easily discovered and used by the user, and the user is explicitly prompted to do click events.

The main interactive element of our visualization is the slider at the top. This slider allows the user to change what year the visualization is displaying and lets them see how the number of players from a country has changed over time. We implemented this in order to let the user explore the data over time, seeing how some countries gained soccer ability or the other way around. To support this idea, hovering over a country with the mouse will display the country's name as well as exactly how many players originated from there. This adds to how the user can explore the visualization, and it rewards users with more information as they explore more. We also implemented pan and zoom features that allow the user to zoom in on the map and look around. Below, you can view screenshots of when a user would use the interactive elements (they are cropped for the sake of this document):

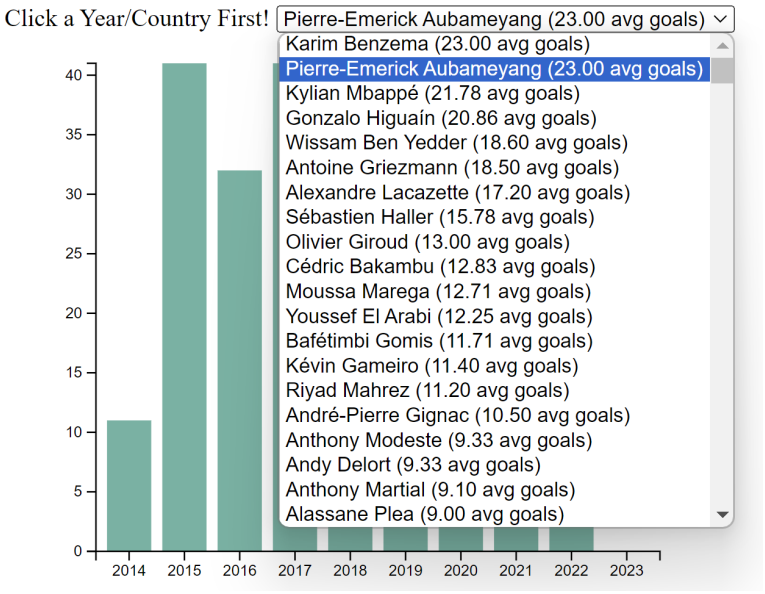


(User zoomed into France and then panned over to Brazil.)

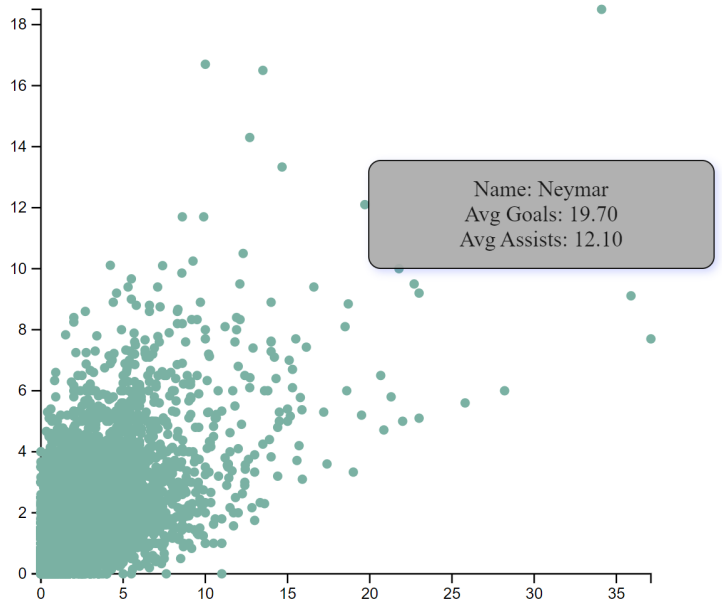




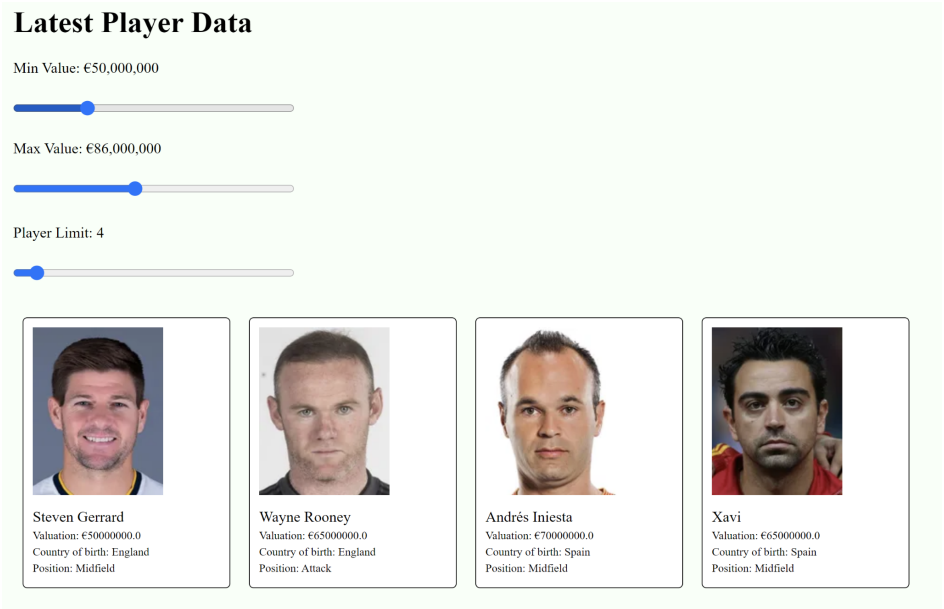
Our supporting graph also allows for interactivity. The dropdown menu allows users to select their favorite players from any country they click on. For example, clicking on England propagates the dropdown selector with all the players from England. The user can then compare and contrast their favorite players from a certain country, and accordingly, the players in the dropdown are sorted by the highest career average for goals scored. Hovering over a bar shows the year and number of goals the player has scored, making the chart easier to read. We wanted to implement these interactive features to tie together the main and supporting visualizations. This allows the user to further explore questions they might have about the map.



Our scatter plot also allows for further interactive exploration into the soccer players. When hovering over a point on the plot, users are shown the corresponding player as well as their career average in goals and career average in assists. This interaction can be seen below.



Finally, users can change the sliders in the player valuation chart to see different players. An example of this is shown below, where the user adjusted the sliders to fit their search criteria.



Player valuation is a popular topic among fans and the audience. This interactive element puts users' curiosity at the forefront and allows them to find the players that have the highest valuation. They could simply do this by adjusting the *Min Value* to a higher value and viewing the remaining players. After such interaction, users can confirm their interaction by looking at the exact valuation value of the player. They can then learn more about the player by checking out their country of origin and position in soccer matches. On the other hand, if the user is interested in discovering less popular players, they may adjust the *Max Value* to a lower value. If they need more of the playercards, they may simply increase the *Player Limit* value and see more players that match their selection. In summary, we designed the interactions by putting ourselves in a user's perspective and adding individualization options for users of various purposes to customize their experiences.

The Story

Soccer players come from all over the world. While this fact is quite simple, seeing it laid out on a world map really changes our perspective. According to the map, a lot of the players are from Europe and from South America, but outside of those regions, the total player count is quite sparse. The data we used is sourced from leagues that all take place in Europe. Hundreds of players leave their homes hoping to play soccer. Being from America—a country that has produced 0 professional soccer players—we easily underestimate the size and reach that this sport has on the world. It is surprising to see that most countries play this sport.

It is also easy for users to see who the best players are. Combining this with the geographic part allows users to track where the best players come from. Users can get a sense of which countries produce good players.

Team Contributions Outline

Our team met via Zoom on a tri-weekly basis. In these meetings we discussed how the project was going, highlighting what went well, and any difficulties we were having. We also used the time to break down the work into smaller components and then assign members to different tasks. Large parts of our discussions involved decisions on who would tackle each portion of the project, and we were able to do so successfully and in a way that brought everyone to agreement. The breakdown of tasks and how long they took are listed below:

Jason Steiner : Python data joining and data processing (2 hours); final report (1 hour); map, bar, and scatter plot visualizations (19 hours)
Richard Berlinghof : Data processing (3 hrs) and writeup (3 hrs).
Pinxian Lu : Visualization (4 hrs) and formatting (1 hrs) and write-up (1 hrs)
Nixon Zuniga : Player card data visualization (6 hrs) and touch ups (1hr)