

The data that we use in our visualizations come from two different places. The first place is from an excel sheet with data about the football teams in the “Super liga” (now referred to as the liga), which we were sent by Prozone (the company that is giving us access to their football data). That data is already somewhat processed and it is very convenient to work with it in R, as one don’t have to do a lot of API calls. The second place that we get our data from is from an API that we have received by Prozone. The API works in this way: you go to their website and browse the different things you can request. When you have found what you need, you get a URL where you can put arguments in, for example specify a team id or the encoding format (XML or JSON). When that is done, you need to generate a signature, which is the current unix time plus a shared key and an API key which we have been given by Prozone. That string is then hashed using SHA-256, and appended onto the request URL along with the shared key unhashed. Then you simply send a HTTP request, and you receive the data. The request link is only valid for a short period of time, meaning that if anything interesting has to be done, the requests have to be generated automatically. For this purpose, a function has been implemented in R which automatically generates the request URL and returns a data frame, when given the first part of the URL which describes what data is wanted. The function is especially useful when one has to retrieve large amounts of data from the API, as some of the data is very tedious to get from the API.

The data that we have access to through the API is things such as a list of all football teams in the liga, their rankings and so on and data from specific matches. An example of what one has to do to get all the goal attempts throughout the season is: First get a list of all football teams in the liga. Then for each football team get a list of the matches that they have participated in. Then for each match retrieve the events in that match, and then put all of those events into one big data frame, and then filter the events such that only goal attempts are left. What is important to note about the API is that one cannot simply say for example “give me all matches where there is atleast 5 goal attempts within 10 minutes”. One has to programmatically make multiple calls to the API and based on that data, decide what other calls should be made.

A short list of some of the data we can get through the API:

- Football teams in the liga
- Statistics about a football team (goal attempts, points, number of players, city of origin etc.)
- Players on a team
- Information about a player (height, weight, age etc.)
- Matches a team has participated in
- Statistics about a match (outcome, injuries, event data etc.)