XB 0014: Data Wrangling

Project Report

Håvard Skjærstein and Valentin Guidon

VU id: hsk202

VU id: vgu208

February 6, 2023

Figure 1: This illustrates a fotball player.

1 Research questions:

What are the most optimal parameters for scoring a goal in a football match?

Subquestions:

- Where to place the ball
- Which bodypart to shoot with
- What is the best timing?
- What's the impact of penelties?

2 Data sources:

The research question was answered using a single dataset containing over 9000 matches across Europe and capturing over 900,000 unique events from 2011 to 2017.

The dataset can be found through this link: shorturl.at/kqEL6

3 Data wrangling methods:

In order to answer the research question we extracted the data from a CSV file into a dataset using the panda module, and then further filtering out the required information from that dataset, based on what question we were answering.

The events.csv file already had a well-organized structure and we did not have to organize it more. This made the processing task manageable for our two-person team. We applied filters to extract the relevant information, specifically seeking events with goals and analyzing time, position, and other parameters. To represent the data, we utilized seaborn and matplotlib libraries to create visualizations.

4 Conclusion

The limitations of our approach stem from the fact that we are not conducting statistical research. By solely relying on event occurrence and likelihood, there is a risk of reaching inaccurate conclusions regarding the optimal parameter. This is because these metrics can be misleading and fail to provide a complete picture of the underlying relationships and patterns in the data.

The conclusion of our analysis is that the optimal conditions for scoring a goal are during open play near the 90th minute, while taking into account the presence of cards and executing the shot with the right foot aimed at the center of the goal.