Project Milestone 1: Group 11

Jinghao Liu, Marwan Lloyd, Skylar Shafer, Max Zou

**Modality Decision:** For our project, our group has decided to work with age data in sports and athletics at the professional level. More specifically, we are looking to analyze the impact of aging on athletic performance, seeing why age impacts athletes differently, both within the same sport and across different sports, and see what factors might contribute to these differences. **(NOTE: CHANGE/UPDATE THIS IF WE DECIDE ON SOMETHING ELSE)** We believe that the design studio modality will be the better option for this analysis, as it can allow users to tailor the data they are looking at to sports and athlete groups that they are familiar with and better able to relate to, thus providing them with greater understanding of the data they are viewing.

Chart, scatter chart

Description automatically generated When thinking about the dashboard that this data could produce, several key factors come to mind. These include, but are not limited to, options to switch between sport displayed, handling increased specificity if a user wants to narrow in on an age range, and may also include a table sub-section as well for easier diving into a specific professional player, if so desired.

Graphical user interface, application

Description automatically generated

Table

Description automatically generated

Additionally, there would be a secondary menu allowing a user to narrow in on what statistic within the sport they are looking at, this would include options like age distribution, performance rating per age across the sport, and the number of players that retire at each age in that sport. Thus, through the combination of these two statistics, we aim to provide a user with a broad spectrum of insights in order to make the different sports as comparable as possible.

**Motivation and Goals:** The possible uses for this data are widespread, with clients ranging from sports teams, athletic training groups, and individuals who are looking to stay athletic at an above average level into their 30’s, 40’s and beyond. Given the disparities between these groups, it will be a challenge to satisfy them all. However, doing this will require successful implementation of the customization tools described above. Most importantly will be having relatable data across a range of sports, as there is not large amount of insight that an aging golfer can gain from an aging soccer player. It may be a challenge to find the exact same data for every sport and so a normalization process will be required, most likely via the creation of a type of rubric, in order to provide meaningful comparisons on the factors that do stay similar across sports, like injuries. This normalization will allow for our display to change between sports without constant jarring visual changes impacting the user’s experience.

**Literature Review (NEED TO DEFINITELY FLESH OUT HERE MORE):** When reviewing previous statistical research into this subject, as well as tangential ones, a few themes of visualization became apparent. Firstly, the prevalence of numerical tables, it is very common to find a table with a subset of raw numbers accompanying the visualization. Typically, they serve to provide further elaborations of the data (e.g. by including information like sample size, n, or % differences between points) and support the images by ensuring that this information is easily available without reducing the image’s legibility. (**NOTE: PROVIDE SCREENSHOTS HERE OF A FEW TABLE/GRAPH COMBINATIONS IN THE LITERATURE)**

One of the goals of this project is to create a visualization of performance vs. age across different sports. It is intuitive to assume that the relative “peak” age is different when it comes to different sports. In fact, many literatures supports this assumption. Take basketball and football for example, according to the article “The NFL Is No League For Old Men” on FiveThirtyEight, “Data shows that players at all positions generally see performance declines by age 30, with players’ peak ages ranging from 24 for running backs to 28 for most offensive linemen and quarterbacks” (Salfino, 2018). And for basketball, “In general, research shows that players tend to peak around the age of 27 or 28” (Steadman, 2021). Thus, from these two literatures, one can expect the range and average of age distribution is different when looking at performance. A more confrontational and energy intensive sport is to be expected to have a younger age distribution. If a sport is more benefited from a developed knowledge of the game, for example, golf, an older distribution of age is expected.

One of the challenges is then to come up with a measurement of performance. According to Feng, “There is no formal definition of a player’s “peak,” but by using various metrics to measure a player’s performance, such as player efficiency rating (PER), box plus-minus (BPM), and win shares (WS)” (Steadman, 2021). Therefore, multiple metrics will be looked at and tested for this project to ensure the most accurate visualization result. It is also important to take into consideration of positions/roles of player. If looking at score itself, a player playing a defensive position is prone to score much less compares to a player at an offensive position. Another challenge is to standardize the performance measuring metrics so that a consistent visualization can be presented. It is obvious that using the average score each player scored at certain age is not ideal. Since for a sport like soccer the average player score is around single digits but for basketball is usually around 15 to 20. It is important to come up with a suitable rule of standardization of metrics across sports.

**Packages:** One major benefit of working with sports data is its widespread prevalence which significant helps make the overall data collection process easier. There are many publicly available datasets on Kaggle which provide pre-packaged information, however, as with any pre-made dataset, the risk with using these is that we do not entirely know what the data collection process was. As such, in addition to these pre-packaged datasets, our group will be able to utilize several packages in R that have been designed to easily scrape and aggregate sports data. These include nflscrapR and cfbfastR, which are packages for aggregating and scraping both pro & college level football information; nbastatR, hoopR, and wehoop, for NBA and NCAA levels of basketball; baseballR for the MLB; engsoccerdata for top level European soccer; and hockeyR for the NHL. With these available to our team, we will be able to verify and supplement pre-packaged datasets as necessary through data collection work that we can ensure matches our desired standards.

**Conclusion: (PUT STUFF HERE AT SOME POINT)**

**Citations:** (Add in Citations as we add in charts/images/screenshots, then once done put it on its own page at the end)

Salfino, Michael. “The NFL Is No League for Old Men.” *FiveThirtyEight*, FiveThirtyEight, 30 Aug. 2018, https://fivethirtyeight.com/features/the-nfl-is-no-league-for-old-men/#:~:text=Data%20shows%20that%20players%20at,is%20pronounced%20for%20all%20positions.

Steadman, Will, et al. “Peak Age in Sports.” *Dartmouth Sports Analytics*, Dartmouth Sports Analytics, 10 Nov. 2021, https://sites.dartmouth.edu/sportsanalytics/2021/11/10/peak-age-in-sports/.