Project Milestone 1: Group 11

Jinghao Liu, Marwan Lloyd, Skylar Shafer, Max Zou

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 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. 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. We will be using player information across several major sports and the Olympics with datasets obtained from Kaggle. These datasets will include an NBA dataset of players from 1950-2019 scraped from basketball-reference, an all time Premier League Player Statistics dataset of 2020 Premier League players, NHL Player Data from 2004-2018 (links to all three can be found in the citations), and 120 years of Olympics history: athletes and results.

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 a table sub-section as well for easier diving into a specific professional player, if so desired. A general foundational mockup of this dashboard can be found below.

Chart, scatter chart

Description automatically generatedGraphical user interface, application

Description automatically generated

Table

Description automatically generated

Foundational mockup of the dashboard for viewing age data across different sports

Additionally, not shown in the present mockup, there would be a secondary dropdown 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 selections, we aim to provide a user with a broad spectrum of insights in order to make the different sports as comparable as possible.

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 making the data visualization as relatable as possible for all groups. As such, we plan to prioritize the filtering and customization, so that the visualization will present relatable data across a range of sports, and satisfy the interest of a particular user. 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. This normalization will allow for our display to change between sports without constant jarring visual changes impacting the user’s experience, and will hopefully allow users to still gain insight from sports outside their particular focus group.

When reviewing previous statistical research into this subject, as well as tangential ones, a few themes for 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, as shown in the screenshot below.

**Chart

Description automatically generated**

An example of a­ table elaborating on its relative visualization for weightlightfiting data (Huebner and Perperoglou)

Additionally, a lot of the literature has been interested in attempting to determine the “peak” age for each sport. It is intuitive to assume that the relative “peak” age is different when it comes to different sports, and literature supports this assumption. 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). Meanwhile in “... research shows that players tend to peak around the age of 27 or 28” (Steadman, 2021), while in soccer, average peak ages also vary from 26 for forwards to 30 for goalkeeper (Oterhals Geir, 2021). Thus, from these three 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. However, we are also prepared for the case that such a metric may not possible to develop, or in turn does not prove as insightful as we currently anticipate. In such a case, there are additional visualizations that we have seen in our research that we can draw inspiration from, so that insights can still be gathered across different sports without requiring standardization.

Chart

Description automatically generated with medium confidence

An example of a visualization comparing peak performance age in weightlifting across different geographic regions and genders (Huebner and Perperoglou)

Chart

Description automatically generated

A visualization comparing non-standardized performances b/w track & chess over age (Berthelot et al.)

To summarize, we have decided to choose a design studio on age in professional sports as our modality. We think this information would be useful for the decisions made by a number of organizations, and individuals within those organizations, associated with sports, to help them make accurate assessments of performance in the years to come. Through our research, we’ve encountered a number of pieces of literature on this kind of topic, and so we feel confident in our ability to form our design in a way that communicates the most valuable information possible. We also have a number of available data that we will be able to use to extract the key aspects of our data and visualize it in a way that is easy for our audience to understand. Overall, we think at this point that we are aptly prepared to best make our design studio efficient, insightful and useful for our target audience, based on the research we have completed.

**Citations:**

Berthelot, Geoffroy et al. “The age-performance relationship in the general population and strategies to delay age related decline in performance”, *Archive of Public Health,* 09 Dec. 2019, <https://archpublichealth.biomedcentral.com/articles/10.1186/s13690-019-0375-8>

Huebner, Marianne and Perperoglou, Aris. “Performance Development From Youth to Senior and Age of Peak Performance in Olympic Weightlifting.” *Frontiers*, Children’s Exercise Physiology, 27 Aug. 2019, <https://www.frontiersin.org/articles/10.3389/fphys.2019.01121/full>

Kaggle NBA Data: <https://www.kaggle.com/datasets/drgilermo/nba-players-stats?resource=download>

Kaggle NHL Data: <https://www.kaggle.com/datasets/xavya77/nhl04to18>

Kaggle Olympic Data: <https://www.kaggle.com/datasets/heesoo37/120-years-of-olympic-history-athletes-and-results>

Kaggle Premier League Data: <https://www.kaggle.com/datasets/rishikeshkanabar/premier-league-player-statistics-updated-daily>

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/.

Oterhals, Geir & Lorås, Håvard & Pedersen, Arve. (2021). Age at Nomination Among Soccer Players Nominated for Major International Individual Awards: A Better Proxy for the Age of Peak Individual Soccer Performance?. Frontiers in Psychology. 12. 661523. 10.3389/fpsyg.2021.661523.