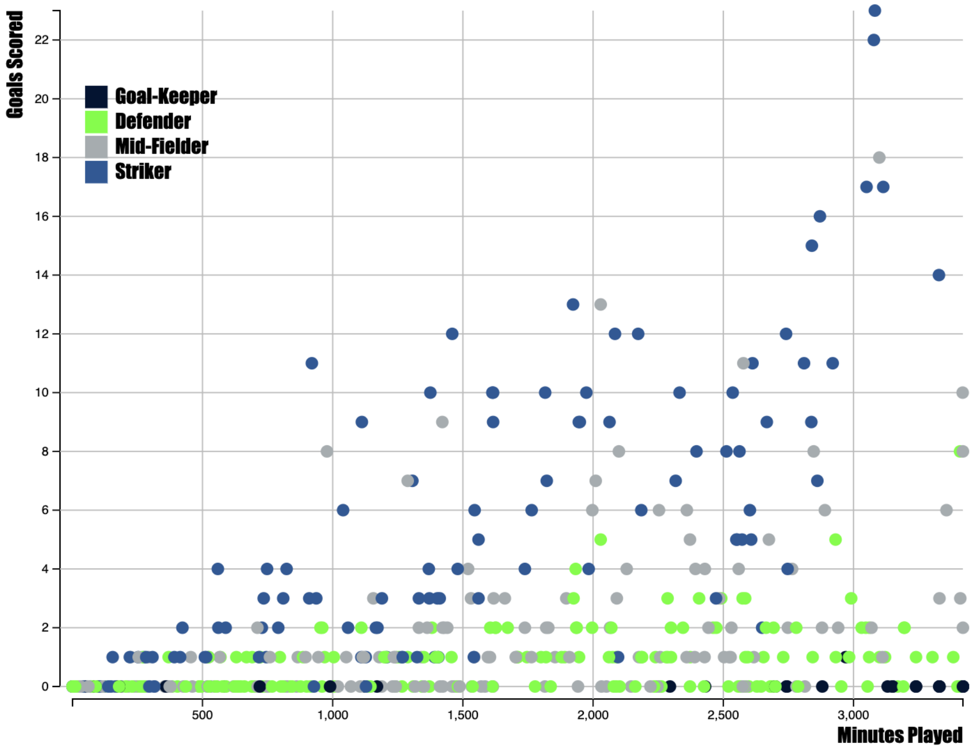
Project 1 Final Report

Team members: Mia Moon, Esther Han, Joseph Kim, Matthew Yang

a) **Chart 1: Soccer play performance observing minutes played and goals scored**
Chart 2: Average Goal + Assist contribution of players grouped by age

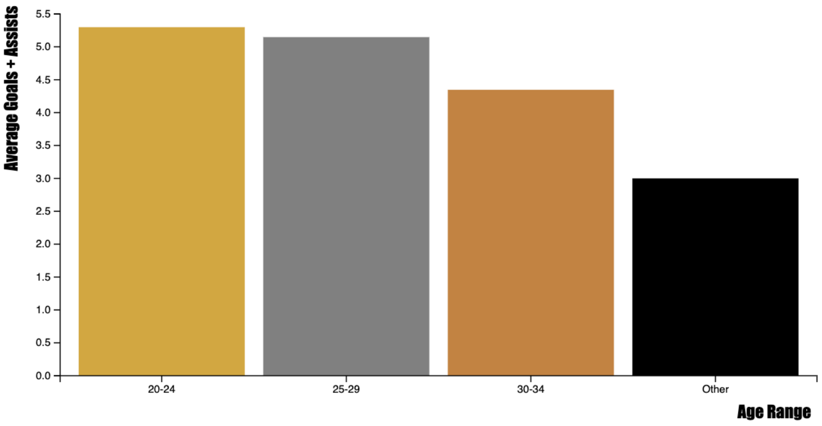
Soccer Player Performance Observing Minutes Played and Goals Scored

The first visualization shows us that there is a positive correlation between the minutes played and number of goals scored.



Avergae Goals + Assists Contribution of Players Grouped by Age

The second visualization illustrates that the age group of 20-24 generated the highest number of average goals and assists among all age groups.



b) This data is from a dataset called English Premier League (2020-21) from Kaggle. It contains data points about each Premier League soccer player, including minutes played in the season, number of goals scored, team, position, and other various statistics. In this particular chart, we used the minutes played, goals scored, and position. We verified that none of the data points had missing or negative values for minutes played and goals scored, and for players with more than one position listed, we took only the first position listed into account. Additionally, for our bar graph to see how age is correlated to the number of goals+assists without it being skewed by outliers, we wanted to utilize only players who scored at least one goal or assist; this was to eliminate data points where the player may have been inactive, didn't play enough games, etc.

c) For this chart, we placed the minutes played on the x-axis and goals scored on the y-axis to visualize the relationship between the two variables. Each data point is a circle on the chart, and we used varying colors to represent the different positions the players played. We considered adding mixed positions for players with more than one position, but we felt that this would complicate the visualization and require the viewer to decipher what the various mixed positions were and what that would imply. Thus, we decided to take only the primary position of the players to visualize the data. This variation was important because we also wanted to see the spread of the positions when looking at the minutes played vs. goals scared relationship. Moreover, we used aligned positions for both vertical and horizontal so that the points were placed on the same axes and the data points could be compared with one another.
For our second visualization, we chose to use a bar graph with age groupings so that we could group age ranges together (instead of having a line/bar for each age) to see general trends in the contributions a player makes (through goals+assists) on average by age. We created an "other" group for ages not in any of the ranges (younger than 20, older than 34). We placed the age range on the x-axis and average goals+assists to be able to see that correlation with each bar. Starting with filtering the data for only eligible players with >= 1 goal+assist, we then used those players' age, goals, and assists, to calculate with all other eligible players, the average goals+assists for each of the three age ranges. We then constructed the bar graph and the results were that ages 20-24 had the highest average goals+assists, with the "other" being the lowest average. One channel that we employed was the use of varying color hues for the three bars. We used gold, silver, and bronze (in the order of highest average to lowest) because it makes it visually easy to identify the meaning and the significance of the data. The black was used to signify the "other" group which contained outliers relative to the main data we were presenting.

d) The first visualization shows us that there is a positive correlation between the minutes played and number of goals scored. We can see that the number of goals increases as the number of minutes played by a player increases. Another observation we made was that strikers generally had the most number of goals regardless of minutes played, which was expected due to their position on the field. One surprising insight we see from this dataset is that there seems to be a peak in the number of goals scored at 3,100 minutes played, and after this value the number of goals start to decrease. This may imply that there is an optimal amount of time a player should play to maximize performance based on goals scored. To confirm or refute this idea, a future step could be to integrate datasets from past years to gather more data and look at the relationship again.

The second visualization illustrates that the age group of 20-24 generated the highest average number of goals and assists among all age groups. This specific group accumulated an average of 5.3 combined goals and assists, whereas the other age groups were at 5.1 and 4.3. Finally the "other" group had an average of 3.0 This observation aligns with the conventional wisdom that soccer players typically peak in performance in their early twenties before age/injuries start to negatively impact their efficacy. Consequently, this visualization holds significant relevance for teams when they are considering player signings, as it highlights the increased offensive productivity associated with players in their prime age range.

Team contributions:

We decided to split up into sub teams for the two visualizations we created:

Esther and **Mia** created the scatterplot, each contributing and working collaboratively on the chart and the write-up. The time spent together and separate working on this was about 5 hours each. Understanding how we wanted to show the data on the graph, such as the colors, axes, and design took the longest out of the phases of the project.

Joseph and **Matthew** worked on the bar graph - in the filtering of data, creation of the graph, and writing of the document. We worked together, sharing ideas and design decisions to code the graph together. The hardest part was coming up with how to effectively portray the data - our first graph didn't take into consideration that there would be a varied number of players in each age range, so we made a second iteration using averages and thresholds for filtering data. The time spent working/ideating individually and collaborating/combining work was about 5 hours each.