Executive Summary for Sports Statistics in 2019

3Lattes: Deontavius Harris, Adam Blumnoff, and Helen Mak

Insights into Topic

As avid or casual sports watchers, we thought it would be interesting to see the statistics of some sports during 2019, the year before Covid hit. We wanted to specifically explore the statistics of Basketball (NBA), Baseball (MBL) and Football (NFL) players during 2019. The question we hope to address is what the relationship between general in-game statistics looks like for each sport and the statistics individual players relative to their peers. Thus, our Shiny App was designed to make the relationships between the statistics open to explore.

Data

For basketball, we used data from the Basketball reference website: (https://www.basketball-reference.com/leagues/NBA_2020_per_game.html#per_game_stats), where we used the 2019-20 NBA player statistics. Besides the player's name, we chose variables with a "per game" denominator that standardizes the statistics. Additionally, we merged players with multiple observations to account for players who were in more than one team throughout the season.

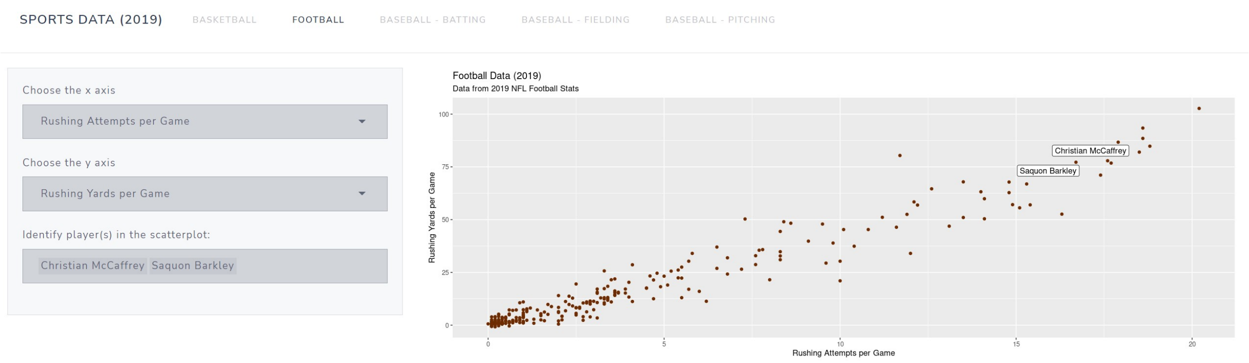
For football, we collected data from Pro Football Reference website :https://www.pro-football-reference.com/years/2019/scrimmage.htm#receiving_and_rushing where we used the 2019 NFL Scrimmage statistics. Similar to basketball we chose the player's name and variables that were standardized to a "per game" basis.

For the baseball data, we used three different baseball data sets from the Lahman Package: Batting, Pitching, and Fielding. To add the player's names we used inner_join() to match the names from the People dataset with the 3 datasets. Then, for each individual dataset, we merged multiple observations to account for some players who change teams during the season.

For more about our data wrangling process, you can view the "data-wrangling.rmd" file under the "data-wrangling" folder.

Shiny App

Our Shiny App has 5 tabs that output a scatterplot for a specific sport (Baseball has 3 separate tabs). On the left side of each panel is 3 widgets: one to select the x-axis variable, another to select the y-axis variable, and a search bar to select a player and highlight them on the scatterplot. These widgets allow the user to explore the relationships between each given statistic and figure out where the player of interest lies in relation to other players.



Choose the x axis: Rushing Attempts per Game

Choose the y axis: Rushing Yards per Game

Identify player(s) in the scatterplot: Christian McCaffrey Saquon Barkley

Figure 1: Football Data Panel. The default setting is Games on the x-axis and Receptions per Game in the y-axis. When switching the x-axis to Rushing Attempts per Game and Rushing Yards per Game in the y-axis we get the following scatterplot. We selected the players Christian McCaffrey and Saquon Barkley whose name tags show up next to their respective points in the scatterplot.

Results

Because our app is designed to explore data throughout the three sports, we explored some exciting trends and cases in each of the sports that aren't obvious. In basketball, we discovered a decent correlation between assists per game and turnovers per game, which makes sense because if you make more passes you are more likely to make mistakes/turnovers. A good player to explore in this plot is Russell Westbrook. He is one of the most electrifying players in the league and has a lot of assists, yet he also has a lot of turnovers because he takes a lot of risks with his passes. Therefore, the plot shows that Russell Westbrook has one of the highest assists per game as well as turnovers per game in 2019.

Also, in football, we looked at the relationship between receiving yards per game and rushing yards per game. This plot does not correlate because they are two completely different parts of the game. However, one outlier who is proficient in both is Christian McCaffrey. This relationship makes a lot of sense because he runs and catches the ball in the Panther's offense, and he is good at both.

Finally, in baseball, we explored how the number of triples affects home runs. Intuitively, one would think that there would be a strong correlation between the two variables because they both involve good batting. However, we found that there is actually no correlation between the two, yet when we looked at doubles instead of triples, there is a much stronger correlation. Because triples are so rare and random (most of them happen on errors) in baseball, it is understandable that there is no correlation. On the other hand, there is a stronger correlation with doubles because they happen more frequently and are less random.

We learned about the relationship between specific variables within a sport and particular players within those relevant comparisons.