Daniel Jackson

DSE5001

Module 01

September 3, 2023

1. Chapter 1 HW:  
   Consider Figure 1.1 (page 2) with respect to some particular scientific research study, such as a research project that the reader might be undertaking. What is involved in the data analysis of the study at all its stages, from processing raw data to eventual communication of results? How much time is likely to be needed at each stage? What specifically needs to be done at each stage? Will R be sufficient for all the stages of analysis of the study? What other tools are required?

Research data link:

<https://journals.ametsoc.org/view/journals/bams/104/5/BAMS-D-22-0235.1.xml#bib14>

This article highlights how rising temperatures in North America have led to more home runs hit in Major League Baseball (MLB). The raw data of this scientific research study used observations from 100,000 MLB games and 240,000 individual batted balls spanning from 1962 to 2019. The researchers used the raw data from these games and at-bats to visualize and identify a trend within the MLB, that when the game day temperature increases, home runs per game also increases. The researchers used the raw data to clean it up and create tidy data by game-time temperatures and home runs per game, so they could explore and visualize the trend between those factors. They even used ballpark gameday density to confirm that the higher temperatures have led to more homeruns, because higher temperatures mean lower air density.

After tidying the data using home runs per game, temperature and air density, the researchers then used a Poisson regression model which estimates the number of home runs in a game as a function of the temperature during the game. Using this model and coupling it with the results from the empirical estimate between day game temperatures and home runs, allowed the researchers to even project how many home runs may occur in the future with the constant rising of temperatures.

Using their modeling and analysis of the tidy data that they pieced together, the researchers found a 0.16 percentage point increase in home run probability per 1 degree Celsius, which leads to a anticipated 0.044 additional home runs per game. They used this data and other varying factors to communicate to the MLB that if they want to limit homeruns, they could convert all day games into night games, which reduces exposure to daily high temperatures. The researchers even acknowledged that the changes in technology, player skill, and player development will always shape their projections. However, the rising temperatures over time will affect player’s overall statistics, team’s acquisition strategies, and public perception of the game.

In this article, you can see the data science workflow in full effect. The researchers took the raw data from over 100,000 MLB games spanning almost 60 years, cleaned it up and focused on home runs per game based on rising temperatures, used visualization graphs to draw conclusions between those variables to explore what they were seeing, used a Poisson regression model to estimate future homeruns to then communicate what they found, and where home run totals are heading based on their analysis.