Homework 1

Data Analysis and Lab.

Deadline: 2023-10-25, 23:59

Problem 1.

In the **Lahman** package, we will use the Major League Baseball (MLB) batting statistics of players that are stored in the **Batting** data table. Also, we need to use the **People** data table to identify the players' names. Our questions are about the batting average (`BA`) that is defined by

$$BA = H/AB$$

where `H` and `AB` mean hits and at bats respectively. Though `H` and `AB` can be found in **Batting**, `BA` is not included. `BA` can be misleading especially when `AB` is too low, so we focus on only players with `AB` ≥ 400.

(a) Our interest is whether the batting average `BA` is decreasing over time since 1957. Find the top 20 players with the highest `BA` in each year and plot them over time. By adding a smooth trend line to the plot, explain what you think about the `BA` trend over time.

(b) Find the top five players with the highest `BA` since 1957. Provide (i) player's name, (ii) year, and (iii) BA (to three decimal places).

In MLB, there are two leagues, American League (AL) and National League (NL). In each season/year and each league, a player wins the "batting title" for having the highest `BA`. The NL winner is known as the "Tony Gwynn NL Batting Champion".

(c) Find the three best seasons for Tony Gwynn (in terms of `BA`) and report those years' batting averages for him.

(d) Tony Gwynn won 8 NL Batting Champion titles. Find what years he won the title.

(e) Find who had the highest `BA` in each league since 2001. You need to find two players, one for AL and the other for NL.

(f) Find how many players had `BA` ≥ 0.300 in 2021. Guess who won the AL/NL Batting Champion titles.

(g) In 2021, Trea Turner actually won the NL Batting Champion title. Since he changed the team from *Washington Nationals* to *Los Angeles Dodgers* during the season, he had two entries in the data. He was initially excluded since we used the restriction of `AB` ≥ 400. Identify his records in 2021, compute the total values of `AB` and `H`, and compute his true `BA` in 2021.

Problem 2.

We use the same **Batting** data table to see how the total number of home runs varies over time.

(a) Compute the total number of home runs (`HR`) by year, and plot the total home runs by year (year before 1919 i.e., $\leq$ 1918). What can you tell from the plot?

(b) We expect more home runs as the total number of games increases. So, compute the total number of games (`G`) that players played. Use the home run ratio (i.e., total home runs divided by total games) and plot the ratios over year. Any difference from (a)?

(c) Let's consider all years. Plot the total home runs by year and add a loess curve. Any interesting patterns?

(d) Plot the home run ratios by year and add a loess curve on it. Are the patterns the same as (c)? If not, what are the differences?