DSCI401 - Homework 4

Due: October 8th, 2024

Homework should be submitted as an R Markdown file with links to Google colab notes where necessary. Homework should be turned in on Sakai.

1. Using the Teams data frame in the Lahman package:

   (a) (10 points) Create a data frame that is a subset of the Teams data frame that contains only the years from 2000 through 2009 and the variables yearID, W, and L.

   (b) 10 points) How many years did the Chicago Cubs (teamID is ") hit at least 200 HRs in a season and what was the median number of wins in those seasons.

   (c) (20 points) Create a factor called election that divides the yearID into 4-year blocks that correspond to U.S. presidential terms. The first presidential term started in 1788. They each last 4 years and are still on the schedule set in 1788. During which term were the most home runs been hit?

   (d) (10 points) Summarize the total home runs per season by league. Remove observations where league is missing. Plot the changes over time.

   (e) (20 points) Create an indicator variable called "winning record" which is defined as TRUE if the number of wins is greater than the number of losses and FALSE otherwise. Plot a scatter plot of Runs (R) vs Runs against (RA) with the color of each point showing whether that team had a winning record or not.

2. Using the penguins data fram from the PalmerPenguins package:

   (a) (10 points) Create a summary data frame that calculates the average body mass for each penguin species in the dataset. Which species has the highest average body mass?

   (b) 10 points) Find out which island has the highest number of recorded penguin observations. Create a bar plot to visualize the counts of observations for each island.

   (c) (10 points) Create a filtered dataset that includes only male penguins. Calculate and display the average bill length for this subset. How does this compare to the average bill length of female penguins in the dataset?