DSCI401 - Homework 2

Due: September 23, 2023

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

Answer all questions below with R AND Python.

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 "CHN") hit at least 200 HRs in a season and what was the median number of wins in those seasons.
   (c) (10 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 have the most home runs been hit?
   (d) (10 points) Make a line plot of total home runs per season and stratify by league. Remove observations where league is missing.
   (e) (10 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. Use the nycflights13 package and the flights data frame to answer the following questions:
   (a) (10 points) What month had the highest proportion of cancelled flights? What month had the lowest? Interpret any seasonal patterns.
   (b) (10 points) Given that a delay is longer than an hour, what is the average time of the total delay by airport (i.e. origin)
   (c) (10 points) What is the average air time for all flights by carrier? Which carrier has the longest average air time on their flights?
   (d) (10 points) Keeping only flights that had a delay greater than 0, create a histogram for each month of the delay data.
   (e) (10 points) Create side-by-side boxplots of delay times for flights with delays 60 minutes or greater for the top