Lab 4

Due on 4/12/25 at 11:59 pm

For this lab assignment you will have to create a data set called `sc_bip_small`. This data set should contain statcast data for all balls in play from 2017-2021. Statcast data can be obtained from this link:

https://uofi.app.box.com/file/1449126291821?s=we34tcz4wqdu063zpzuwjvb4r6u9j21s

There is a script called `create_sc_bip_small.R` in the **stat430sp25** repo to aid you in this task of creating the `sc_bip_small` data set.

Question 1 Do the following for a year of your choice with the exception of 2020:

(a) List the batters with the ten highest average exit velocities on batted balls.

(b) Plot the distribution of exit velocities across batters. Does exit velocity vary significantly across batters? Explain your reasoning.

(c) List the pitchers with the ten highest average exit velocities allowed on batted balls.

(d) Plot the distribution of exit velocities allowed across pitchers. Does exit velocity allowed vary significantly across batters? Explain your reasoning.

(e) Compute the correlation for exit velocity for batters between the first and second half of the season you chose.

(f) On the basis of your calculations, do you believe exit velocity us a batter skill? Explain.

Question 2 In this question we will try to predict next years slugging percentage using several variables including statcast variables. Load in the data set `sc_bip_small` and run the code below to calculate some possibly important variables that encode launch angle and exit velocity distributional information for each player.

```
foo = sc_bip_small %>%
  mutate(yearID = year(game_date)) %>%
  group_by(batter_name, yearID) %>%
  summarise(N = n(), launch_angle = launch_angle, launch_speed = launch_speed) %>%
  filter(N >= 10) %>%
  summarise(avg_la = mean(launch_angle, na.rm = TRUE),
            sd_la = sd(launch_angle, na.rm = TRUE),
            la10 = quantile(launch_angle, prob = c(0.10), na.rm = TRUE),
            la25 = quantile(launch_angle, prob = c(0.25), na.rm = TRUE),
            la50 = quantile(launch_angle, prob = c(0.50), na.rm = TRUE),
            la75 = quantile(launch_angle, prob = c(0.75), na.rm = TRUE),
            la90 = quantile(launch_angle, prob = c(0.90), na.rm = TRUE),
            avg_ev = mean(launch_speed, na.rm = TRUE),
            sd_ev = sd(launch_speed, na.rm = TRUE),
            ev10 = quantile(launch_speed, prob = c(0.10), na.rm = TRUE),
            ev25 = quantile(launch_speed, prob = c(0.25), na.rm = TRUE),
            ev50 = quantile(launch_speed, prob = c(0.50), na.rm = TRUE),
            ev75 = quantile(launch_speed, prob = c(0.75), na.rm = TRUE),
```

```
            ev90 = quantile(launch_speed, prob = c(0.90), na.rm = TRUE)) %>%
  rename(name = batter_name)
```

- Create a data frame for batters that contains slugging percentage (SLG) for each player. Call this data frame `bat_stat`. This data frame should contain the following variables: `name`, `yearID`, `teamID`, `AB`, and `SLG`. You can restrict attention to batters who had at least 200 ABs and who only played on a single team (`stint = 1` in the `Batting` data frame in the `Lahman` package).
- Merge `foo` into `bat_stat` using `inner_join` or a similar function. Run the following code which creates new variables `SLG_next` and `team_next` which are a player's slugging percentage and team for the next season. The code also creates a categorical variable `COL` which indicates whether the player's next season is with the Rockies. Note that you may have to change the `by` argument in the `inner_join` call below to get it to work.

```
bar = bat_stat %>% mutate(yearID = ifelse(yearID == 2021, 2020, yearID)) %>%
  group_by(name, yearID) %>%
  summarise(SLG, teamID) %>%
  mutate(SLG_next = SLG[match(yearID, yearID-1)]) %>%
  mutate(team_next = teamID[match(yearID, yearID-1)]) %>%
  mutate(yearID = ifelse(yearID == 2020, 2021, yearID)) %>%
  select(-SLG,-teamID)
bat_stat = inner_join(bat_stat, bar, by = c("name", "yearID")) %>%
  mutate(COL = ifelse(team_next == "COL",1,0)) %>%
  filter(complete.cases(.))
```

- We are going to use a simple procedure to assess predictive performance. Run the code below to split `bat_stat` into a model training data set `train` and a model testing data set `test`.

```
set.seed(13)
ind = sample(1:nrow(bat_stat), size = 400, replace = FALSE)
train = bat_stat[ind, ]
test = bat_stat[-ind, ]
```

- Fit and compare the following models. Which model would you select for predicting slugging percentage (root mean squared prediction error is a good metric for assessing predictive performance)? Are statcast variables important for predicting slugging percentage? Explain. Try to find a model which offers better predictive performance than the best model below. Comment on the success of your efforts.

```
m_big = lm(SLG_next ~ SLG + avg_la + avg_ev + team_next + sd_la + sd_ev +
           sd_la*avg_la + sd_ev*avg_ev +
           la10 + la25 + la50 + la75 + la90 +
           ev10 + ev25 + ev50 + ev75 + ev90,
         data = train)

m_small = lm(SLG_next ~ SLG + avg_la + avg_ev + COL,
         data = train)

m_smaller = lm(SLG_next ~ SLG  + COL, data = train)
```

Question 3 The 2021 San Francisco Giants certainly surprised a lot of people when they won 107 games with a rotation led by Kevin Gausman, Logan Webb, Anthony DeSclafani, and Alex Wood. Coming into the 2021 season, I think it is fair to say that this is a shaky rotation. One commentator said that the Giants have developed reputation as an organization that can make players better, but that reputation will be tested with a risky experiment in 2021. Let's investigate the success of the 2021 San Francisco Giants. In this question we will look at the 2021 San Francisco Giants pitching staff from a recent historical perspective. In the next question we will examine specific Giants pitchers. As an aside, anyone can go down the rabbit hole that your

professor went down as this problem was developed:

- Logan Webb, As Advertised
- What the Heck Is a Flat Sinker, Anyway?
- The Giants Took a New Angle With Sinkers
- The Seam-Shifted Revolution Is Headed for the Mainstream
- Pitch Movement, Spin Efficiency, and All That
- Prospectus Feature: All Spin Is Not Alike
- Determining the 3D Spin Axis from Statcast Data

First create a data frame that contains the following variables on team pitching statistics: `yearID`, `teamID`, `frac_junk`, `ERA`, `HAp9`, `HRAp9`, and `WAR`. This data frame only needs to be created for the 2017, 2018, 2019, and 2021 baseball seasons. The variables `HAp9` and `HRAp9` are, respectively, hits allowed per 9 innings and home runs allowed per 9 innings. The variable `frac_junk` is the fraction of team pitches that are sinkers (SI), splitters (FS), sliders (SL), or change ups (CH) for a given season. Calculation of `frac_junk` involves the use of all statcast data for the 2017, 2018, 2019, and 2021 baseball seasons. The statcast data set is massive and a Rmd document might not compile if this data set is directly loaded in and manipulated. I recommend that you first perform your data manipulations to statcast data in an active R session, then save a much smaller data set which contains your data manipulations onto your computer, then load this smaller data set into the Rmd document corresponding to your lab assignment. The code below obtains pitching WAR.

```
bwar_pit =
  readr::read_csv("https://www.baseball-reference.com/data/war_daily_pitch.txt",
                  na = "NULL") %>%
    filter(year_ID >= 2017) %>%
    select(team_ID, year_ID, WAR) %>%
    rename(teamID = team_ID, yearID = year_ID)
```

Use the data set that you created to study the 2021 pitching season. Were the 2021 Giants successful? Did they perform similarly to other teams that throw a lot of junk balls where junk is defined as sinkers, splitters, sliders, and change ups? Elaborate.

Question 4 For each pitch type thrown by Kevin Gausman, Logan Webb, Anthony DeSclafani, and Alex Wood, compute annual averages of `release_spin_rate`, `effective_speed`, `plate_x`, `plate_z`, `pfx_x`, `pfx_z`, `release_x`, and `release_z`, and compute the annual pitch type percentages for each of these pitchers. Now display a graphic showing how these annual averages change over time for each of these pitchers. It is best to display all nine plots for each pitcher in a single grid of plots rather than printing off nine separate plots. This can be achieved using the `grid.arrange` function in the `gridExtra` package. Comment on how the approach of these pitchers changed over time with an emphasis on any changes made in 2021. Comment on any commonalities or differences between these pitchers. What are some of the reasons for the pitching success of 2021 San Francisco Giants pitchers?