HW 1: Baseball metrics using univariate and bivariate tools

Stats and sports class

Fall 2020

Preliminary notes for doing HW

1. All files should be knit and compiled using R Markdown. Knit early and often! I do not recommend waiting until the end of the HW to knit.
2. All questions should be answered completely, and, wherever applicable, code should be included.
3. If you work with a partner or group, please write the names of your teammates.
4. Copying and pasting of code is a violation of the Skidmore honor code

Part I

HW Grade

Return to Homework 0 and assign yourself a grade:

- 1-3 out of 5 points: Most questions attempted, minimal effort
- 4 of 5 points: All questions attempted, complete effort, graded questions incorrect
- 4.5 of 5 points: All questions attempted, complete effort, graded questions partially correct
- 5 of 5 points: All questions attempted, graded questions perfect

Solution to Q3: The three ways of evaluating metrics: Does it measure a contribution important to the team? Does it reflect players individual talent? Is there a better way to measure the same thing?

Part II

Return to the `Lahman` package in R, and we'll use the `teams_2016_batting` data frame that we organized in last week's lab.

```
library(tidyverse)
library(Lahman)
teams_2016 <- Teams %>% filter(yearID == 2016)
teams_2016_batting <- teams_2016 %>% select (yearID:teamID, R:SF)
```

1. Write code to select only the `teamId` and `H` variables
2. Which team led the league in home runs? Use the `arrange()` command to answer
3. Make an appropriate graph of team wins during this season. Is the distribution of wins skewed left, right, or symmetric?
4. Batting average is defined as hits divided by at bats. Make a new variable, batting average (you can call it whatever you want) in this data set.
5. Can you find the team with the lowest batting average?
6. Look at the following code:

```
ggplot(data = teams_2016_batting, aes(x = H, y = R)) +
         geom_point()
```

Is there an association between runs scored and hits? How would you describe it?

7. Describe the center, shape, and spread of the `X3B` variable – split by each league (lgID) – using an appropriate plot.
8. How can you change the x and y labels on your plots? How can you add a title? Use google to guide you, and update your plot in Question 3 with a new x-axis label, a new y-axis label, and a title. One trick: include `ggplot` in your google search.
9. Moneyball was based on which team-statistics most strongly correlated to runs. Though there are some variables that already exist in the data, the code below creates batting average, on base percentage, and slugging percentage.

```
teams_2016_batting <- teams_2016_batting %>%
  mutate(BA = (H/AB),
         OBP = (H + BB)/(AB + BB),
         SLG = ((H - X2B - X3B - HR)*1 + X2B*2 + X3B*3 + HR*4)/AB)
```

Using visual evidence (See Q6), find the variable that you think seems to boast the strongest association to runs (`R`).

10. *Estimate* the correlation between (i) slugging percentage and runs, (ii) on base percentage and runs and (iii) batting average and runs. Which would you prioritize as a coach using these results? Why?
11. Make both a histogram and a boxplot of `hits`. What features are apparent in the histogram that aren't apparent in the boxplot? What features are apparent in the boxplot that aren't apparent in the histogram?

Part III

Read Voros McCracken's "Pitching and Defense: How Much Control Do Hurlers Have?", provided here.

1. What is McCracken's primary finding?
2. Why would traditional baseball followers feel surprised with this result?
3. What might one consider to supplement McCracken's analysis?