HW 1: Baseball metrics using univariate and bivariate tools

Stats and sports class

Fall 2019

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

Homework questions

Part I

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. Make an appropriate graph of team wins during this season. Is the distribution of wins skewed left, right, or symmetric?
2. Teams play 162 games. Create a new variable, `win_pct`, which identifies the percent of games won by each team. Then, use the `filter` command to identify the winning percentage for the Chicago Cubs (teamID, `CHN`)
3. Describe the center, shape, and spread of the `X3B` variable – split by each league (lgID) – using an appropriate plot.
4. 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.
5. 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, find the variable that you think seems to boast the strongest association to runs (`R`).

6. *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?
7. Create a new variable for whether or not a team won 85 games or more. You can call this variable whatever name you want. How many teams won 85 games or more?
8. A coach is curious if teams that steal more bases also hit more home runs. Make and describe a scatter plot of team home runs versus stolen bases. Then, add a title to your plot. Finally, add a smoothed trend line: you can do this by adding (`+geom_smooth()` to the end of your code.) You only need to show the final graph.
9. 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?
10. Using a combination of the `arrange()`, `filter()`, and `select()` commands, find all teams since 2000 that were hit by a pitch (`HBP`) more than 100 times.

Part II

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

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?