HW 1: Baseball metrics using univariate and bivariate tools

Stats and sports class

Fall 2019

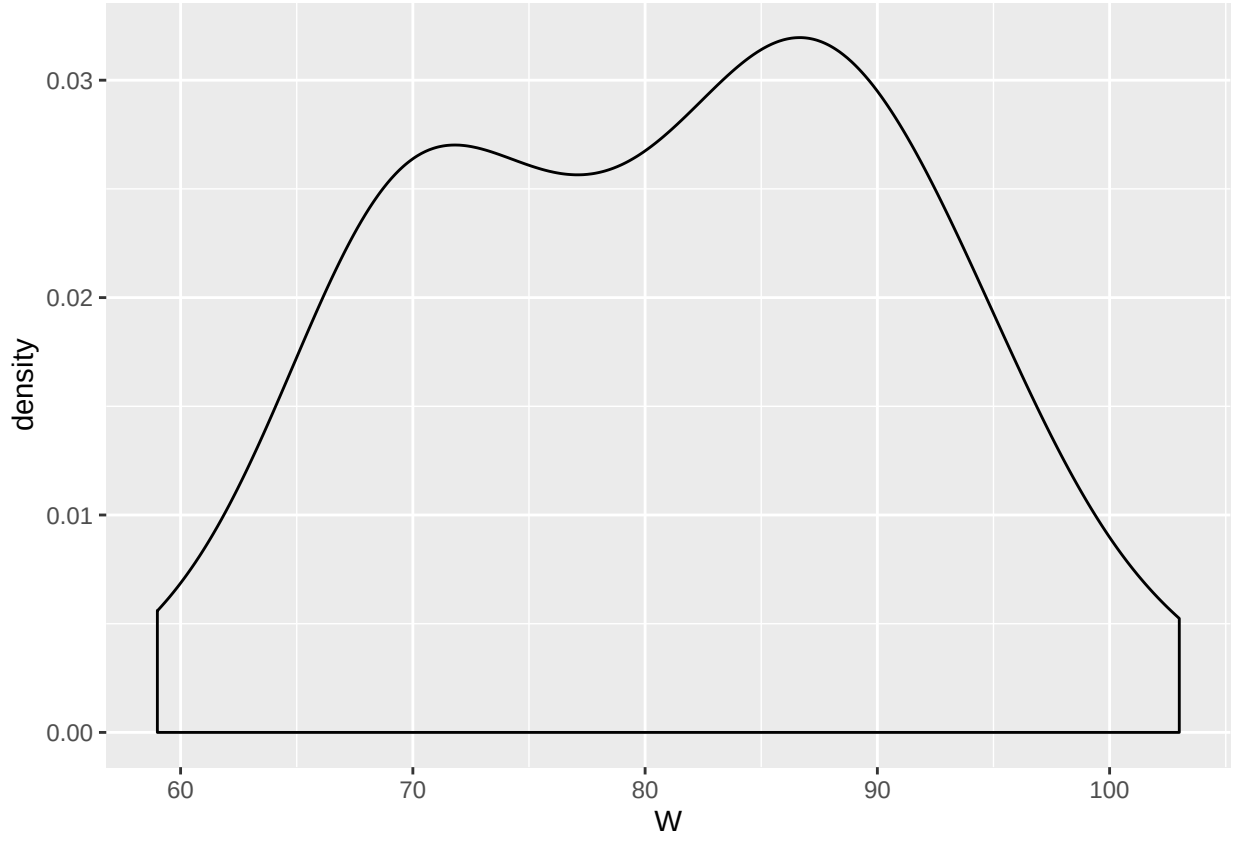
```
library(tidyverse)
library(Lahman)
teams_2016 <- Teams %>% filter(yearID == 2016)
teams_2016_batting <- teams_2016 %>% select (yearID:teamID, R:SF)
```

1. Make an appropriate graph of team wins during this season. Is the distribution of wins skewed left, right, or symmetric?

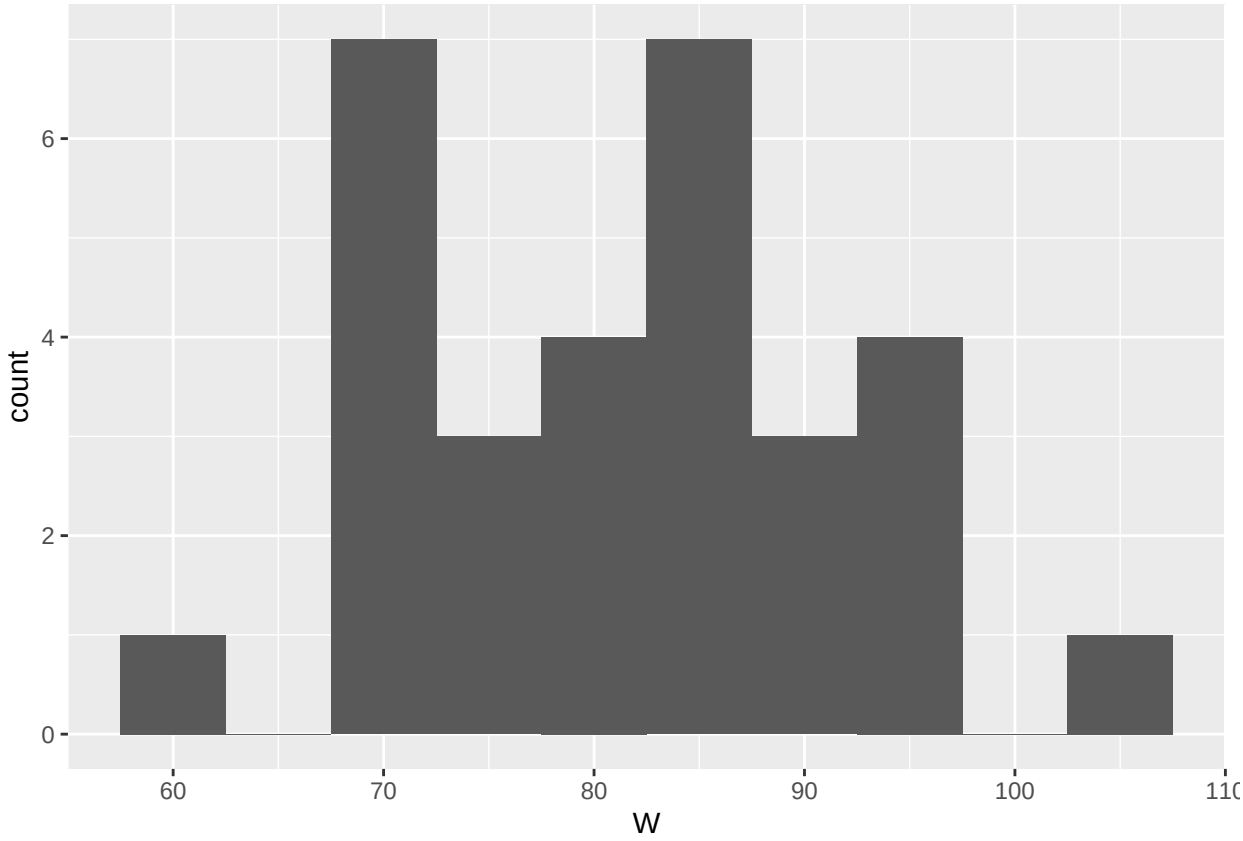
Answers

A histogram or density plot are likely the most appropriate plots.

```
ggplot(teams_2016, aes(W)) + geom_density()
```




```
ggplot(teams_2016, aes(W)) + geom_histogram(binwidth = 5)
```



Wins is roughly symmetric – answering bimodal is reasonable. There does not seem to be any skewness.

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`)

Answers

```
teams_2016 <- teams_2016 %>%
  mutate(win_p = W/162)
teams_2016 %>% filter(teamID == "CHN")
```

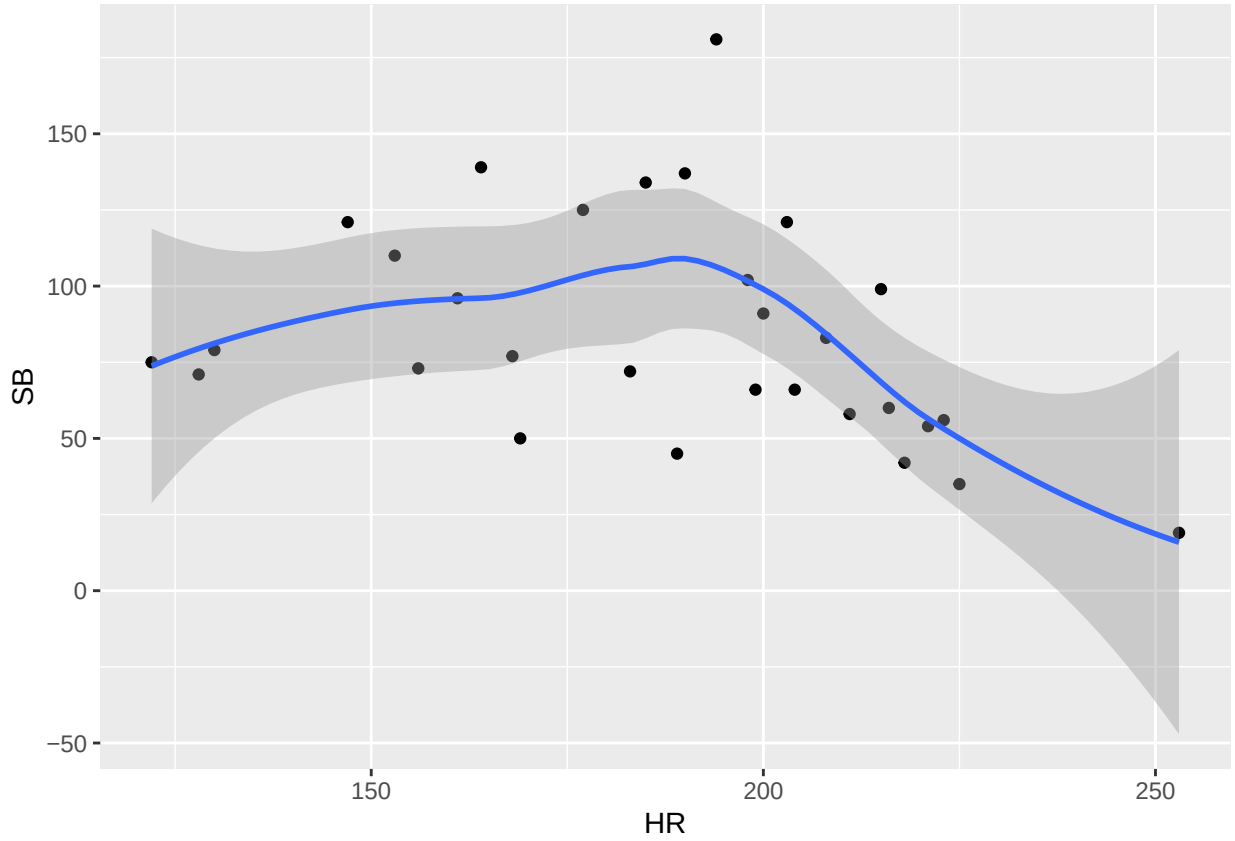
```
##   yearID lgID teamID franchID divID Rank   G Ghome   W  L DivWin WCWin
## 1   2016   NL    CHN      CHC     C    1 162    81 103 58      Y     N
##   LgWin WSWin   R   AB    H X2B X3B  HR  BB   SO SB CS HBP SF  RA  ER  ERA
## 1     Y     Y 808 5503 1409 293  30 199 656 1339 66 34  96 37 556 511 3.15
##   CG SHO SV IPouts   HA HRA BBA  SOA   E  DP    FP         name
## 1  5  15 38   4379 1125 163 495 1441 101 116 0.983 Chicago Cubs
##            park attendance BPF PPF teamIDBR teamIDlahman45 teamIDretro
## 1 Wrigley Field    3232420  95  93      CHC            CHN         CHN
##       win_p
## 1 0.6358025
```

The Cubs had a winning percentage of 63.6 percent.

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.

Answers

```
ggplot(teams_2016, aes(HR, SB)) +
  geom_point() +
  geom_smooth()
```



There seems to be a U-shaped curve! Teams that hit fewer or more home runs tend to have fewer stolen bases, while teams around 200 home runs tend to have more stolen bases.

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?

Answers Boxplots explicitly show the median, as well as any extreme outliers. Histograms can make it a bit easier to discover shape/skewness.

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?

Answer McCracken's main finding is this – "There is little if any difference among major-league pitchers in their ability to prevent hits on balls hit in the field of play." That is, pitchers control walks, home runs, strikeouts, but for balls in play, most of it is luck whether or not the ball ends up as a hit.

2. Why would traditional baseball followers feel surprised with this result?

Answer Primarily, leaving "luck" as the main driver of hits on balls in play reduces the amount of skill that is generally assumed to exist in baseball.

3. What might one consider to supplement McCracken's analysis?

Answers will vary