Lab5

Vivian Yeh

Part I: Understanding the Context of the Data

1. **Descriptive question:** Which player participated in the most games in this year? / Which team won the most game in this year?

   Predictive question: Which player will have the most hits during the next year? / Which team will have the greatest probability to win the most games during the next year?

2. The unit of observation for the **Teams data set** is [one team's total performance in a specific year]

   The unit of observation for the **Batting data set** is [a player's total performance in a specific year]

3. **What is a question you could answer using the Teams data set but not the Batting data set.** What is the average win rate of each team?

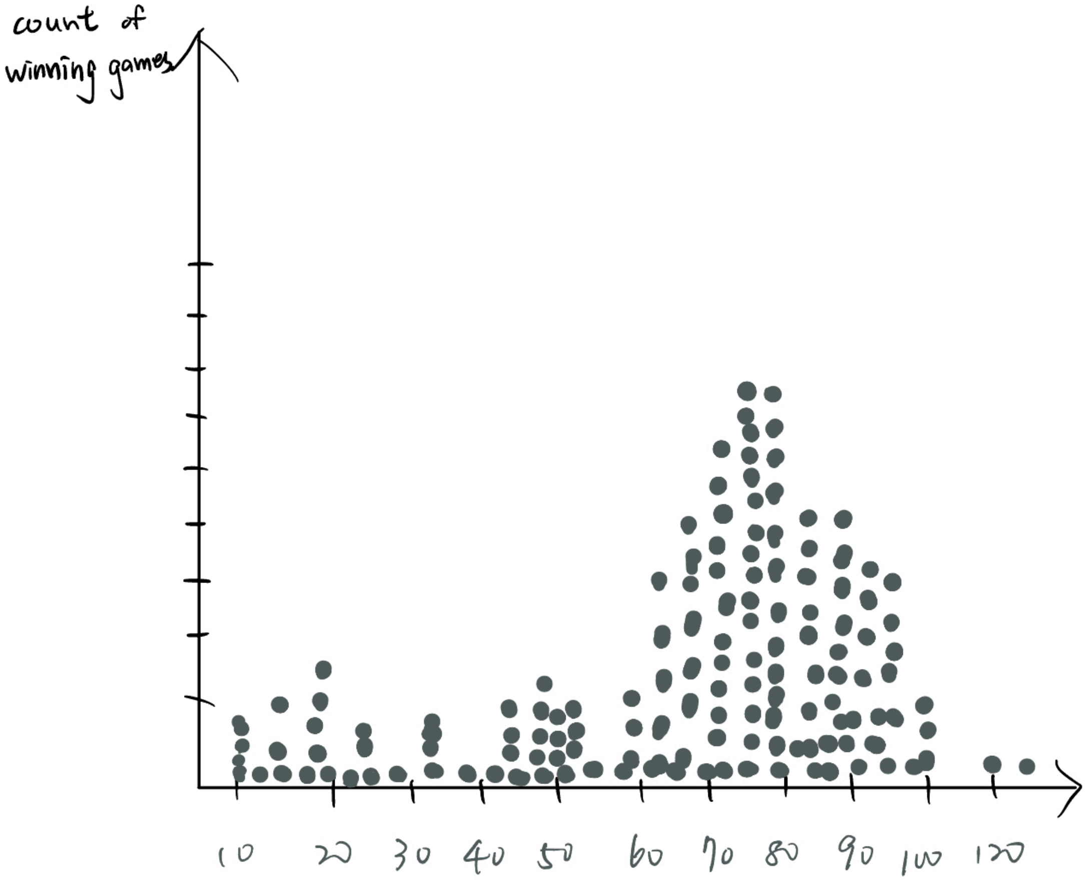
   What is a question you could answer using the Batting data set but not the Teams data set. How many games does player ID abadfe01 have participated in?

4. **What is a question that we would need more granular (measured on a finer/more specific part of the game) data than the Teams and Batting data set provide to answer?**

   How long does the specific player take to run a full round of the arena?

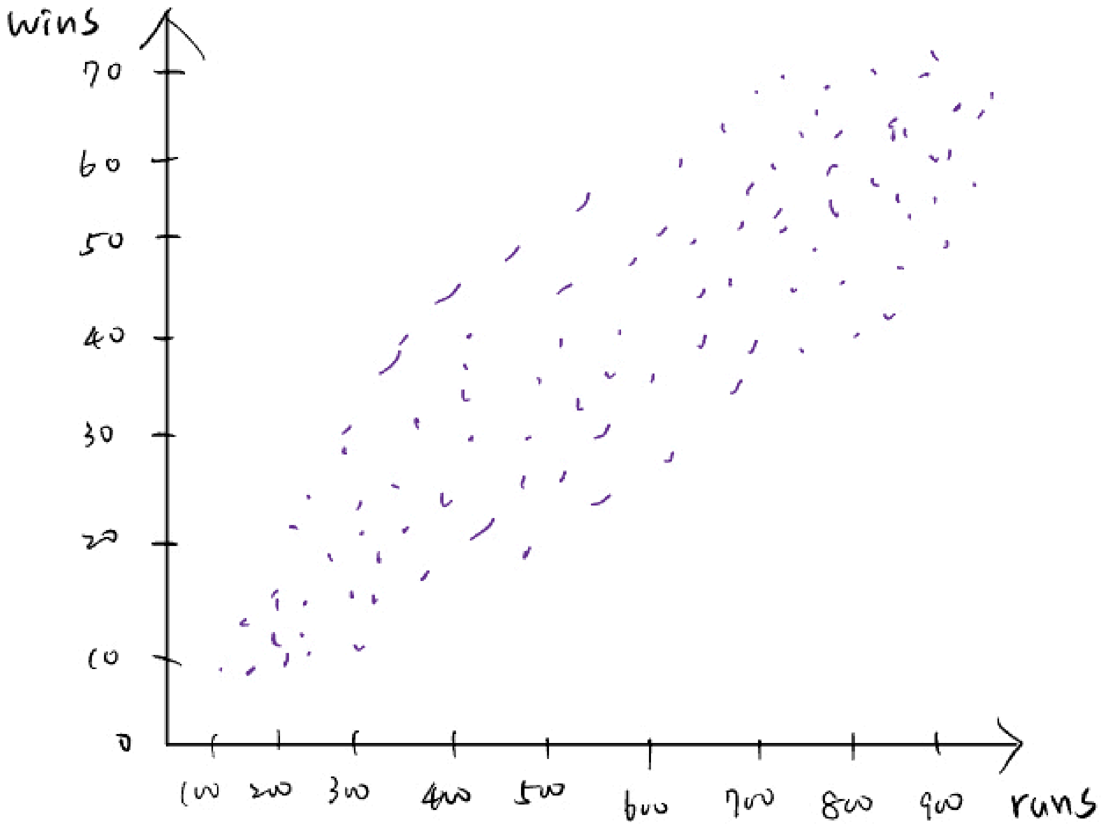
5. **Roughly since 1962 MLB teams have played 162 games in a season. What do you think the distribution of wins looks like? Sketch a plot and describe**

   I think the distribution of wins will looks like the below graph. I think most of the teams will win between 60 to 100 because there are 162 games in a season, and only a few of them win very less or very many times.



6. **What do you think the relationship between wins and runs looks like? Sketch a plot and describe.**

   I think the relationship between wins and runs will be a strong positive association because the more the players run, the more points they can get to win.



7. **Some people believe analytics is ruining baseball because teams are more cautious which makes the games less entertaining. Do you agree or disagree? Why?**

   Yes, I do agree that analytics is ruining baseball because teams are more cautious which makes the games less entertaining because the players are human not robot, we should not expect that every player can perform perfectly. Moreover, I think that we should focus on the competition itself not the data and the scores.

Part II: Computing on the Data

```
library(Lahman)
library(tidyverse)
```

```
-- Attaching packages --------------------------------------- tidyverse 1.3.2 --
v ggplot2 3.3.6      v purrr   0.3.4
```

```
v tibble  3.1.8      v dplyr   1.0.10
v tidyr   1.2.1      v stringr 1.4.1
v readr   2.1.2      v forcats 0.5.2
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
```

```
data(Teams)
```

8. **Subset the Teams dataset to only include years from 2000 to present day. What are the dimensions of this filtered dataset?**

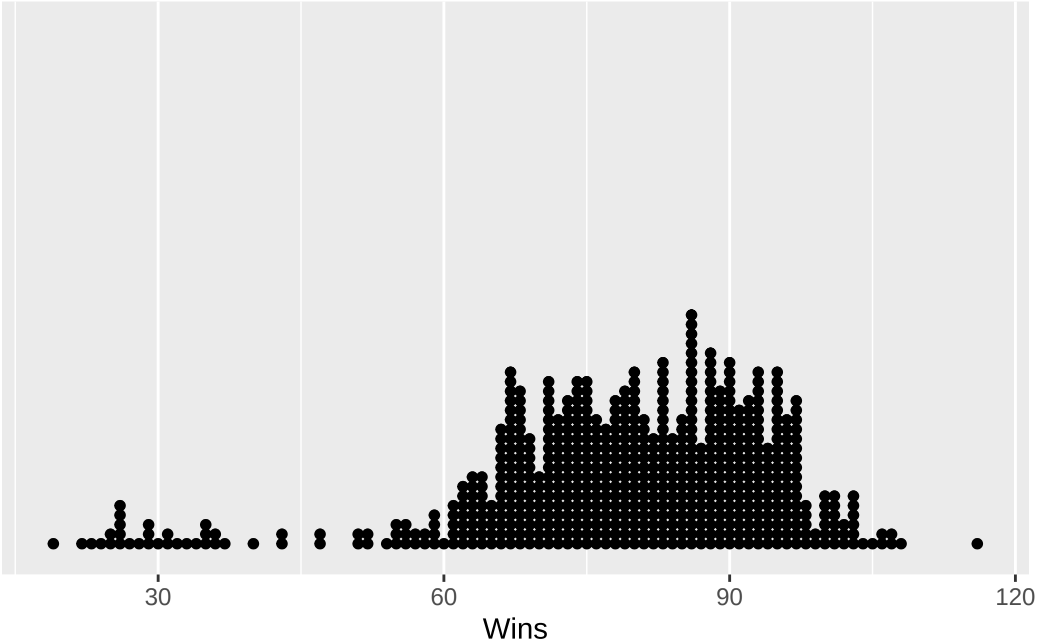
```
Teams <- filter(Teams, yearID > 2000)
dimension_new <- dim(Teams)
dimension_new
```

```
[1] 630  48
```

The dimensions of their filtered data set is 630 * 48.

9. **Plot the distribution of wins. Describe the relationship.**

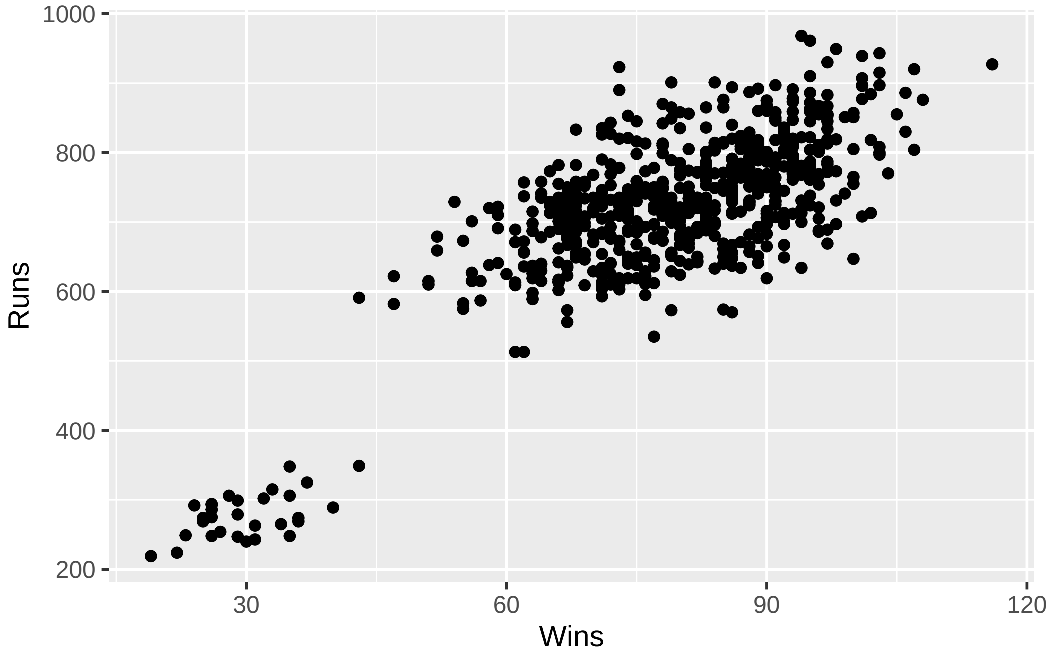
```
Teams %>%
  ggplot(aes(x = W)) +
  geom_dotplot(binwidth = 1) +
  scale_y_continuous(NULL, breaks = NULL) +
  xlab("Wins")
```



This data is a left-skewed data. Most of the teams win between 70 to 100 times in a year.

10. **Plot the relationship between runs and wins. Describe the relationship (form, direction, strength of association, presence of outliers).**

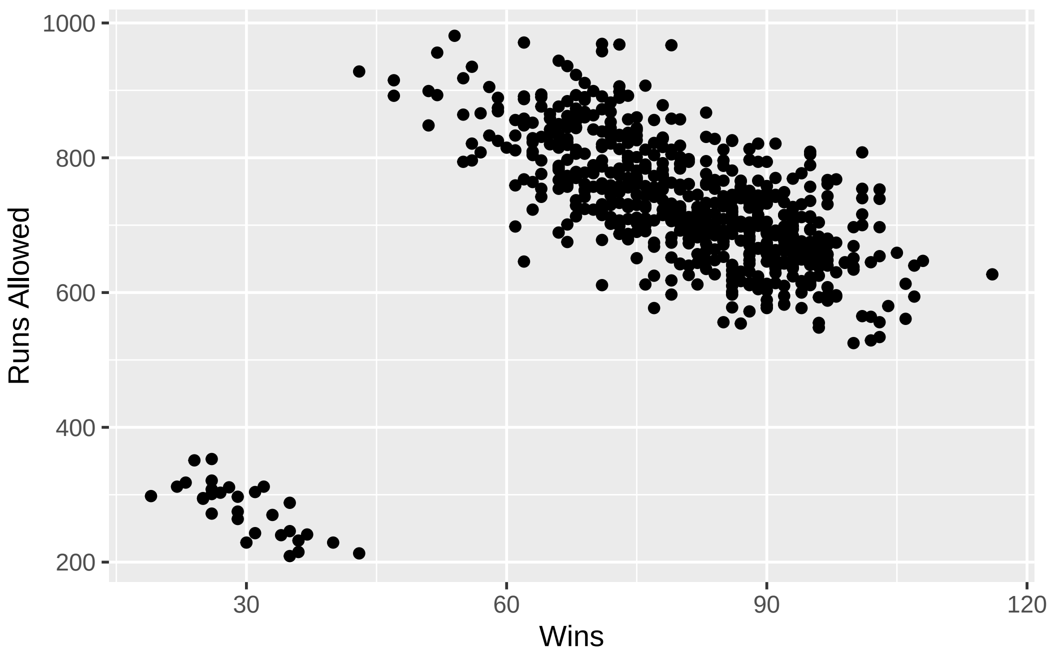
```
Teams %>%
  ggplot(aes(x = W, y = R)) +
  geom_point() +
  xlab("Wins") +
  ylab("Runs")
```



The relationship between Wins and Runs has a positive strong linear association. The outliers are located at the left bottom corner, which has less wins with less runs; however, these points are still on the trend of the data.

11. **Plot the relationship between runs allowed and wins. Describe the relationship. How does it compare to the relationship between runs and wins?**

```
Teams %>%
  ggplot(aes(x = W, y = RA)) +
  geom_point() +
  xlab("Wins") +
  ylab("Runs Allowed")
```



The relationship between Wins and Runs Allowed has a negative strong linear association. The outliers are still located at the left bottom corner, which has less wins with less runs allowed, and these points are not in the trend of the data anymore.

12. **Fit a simple linear model to predict wins by runs. Write out the equation for the linear model report the** R2**.**

```
lm12 <- lm(W ~ R, data = Teams)
library(broom)
r12 <- glance(lm12)%>%
  select(r.squared)
r12
```

```
# A tibble: 1 x 1
  r.squared
      <dbl>
1     0.632
```

The R^2 is 0.6118625.

13. **What is the average number of season runs and wins? Based on the previous model, how many games would you predict a team that scored the average number of runs would win? What about a team that scored 600 runs? What about 850 runs?**

```
Teams %>%
  summarize(avg_run = mean(R), avg_win = mean(W))
```

```
   avg_run  avg_win
1 713.9429 78.54127
```

```
avg_run <- mean(Teams$R)

new_data1 <- data.frame(R = avg_run)
new_data2 <- data.frame(R = 600)
new_data3 <- data.frame(R = 850)

y_hat_1 <- predict(lm12, new_data1)
y_hat_2 <- predict(lm12, new_data2)
y_hat_3 <- predict(lm12, new_data3)
c(y_hat_1, y_hat_2, y_hat_3)
```

```
       1        1        1 
78.54127 66.93755 92.39706 
```

The average number of season runs is 719.3258 and wins is 78.65.

Based on the previous model, 78.65 games I would predict a team that scored the average number of runs would win.

And, I would predict a team that scored 600 runs would win 66.9237 games.

I would predict a team that scored 850 runs would win 91.49153 games.

14. **Fit a multiple linear regression model to predict wins by runs and runs allowed. Write out the equation for the linear model and report the** R2. **How does this model compare to the simple linear regression from the previous question?**

```
lm14 <- lm(W ~ R + RA, data = Teams)

r14 <- glance(lm14)%>%
  select(r.squared)
r14
```

```
# A tibble: 1 x 1
  r.squared
      <dbl>
```

```
1     0.796
```

The R-squared of this multiple linear regression is 0.7850586.

Compared to the simple linear regression from the previous question, this multiple linear regression can explain the prediction of Wins and Runs better than the previous simple linear regression.

15. **Fit a multiple regression model to predict wins using at least two other variables in this data set. How does the** R2 **value change? Do you think the new model you created predicts wins better?**

```r
lm15 <- lm(W ~ R + RA + SF + CS + BBA + HA, data = Teams)

r15 <- glance(lm15)%>%
  select(r.squared)
r15
```

```
# A tibble: 1 x 1
  r.squared
      <dbl>
1     0.892
```

The R-squared increased from 0.7850586 to 0.8882266, and this means that the current formula can better predict the Wins. Based on the increasing R-squared value, I believe that the new model I created predicts wins better.