# Open-ended Practice Test (BaseballHits)

Here is the name of the dataset in DataCamp: BaseballHits

This is a dataset for 30 Major League Baseball teams from the 2010 season

1. What are the categorical variables in this data set? Which of these R functions could help you figure that out? Check all that apply: favstats(), tally(), select(), head(), str()
2. Arrange the data from teams with the least to most wins, and save this arrangement as a new data frame called **WinView**. Write the R code here.
3. Find out which team had the second-fewest wins in Major League Baseball in 2010.
4. Are the contents of this dataset sample data or population data?
5. Make histogram of the variable **Wins**. Write the R code here.
6. The default number of bins in a histogram is 30. What is the R code to see only 8 bins in your histogram?
7. How would you describe the frequency distribution of the variable **Wins**? What is the shape, center, and spread? Do you notice any weird things?
8. Could we model the distribution of **Wins** with a normal distribution? Why or why not?
9. Wins constitute an important outcome variable for analyzing baseball data. What would the empty model predict for SEA (Seattle Mariners)? How about for NYY (New York Yankees)?
10. There are data from teams from two leagues (AL and NL) in this data frame. Is **League** a good explanatory variable for **Wins**? Make a visualization to explore this idea. Also write this as a word equation.
11. Write R code to find the average Wins for AL (American League) and NL (National League). On average, which League wins more games?
12. Does this prove that being in one of the leagues causes a team to win more games?
13. Why is the grand mean a good model for the distribution of **Wins**?
14. Fit a complex model that predicts **Wins** based on **League**. Interpret the resulting numbers.
15. What would be the GLM notation used to represent the **League** model? Which numbers would go with which part of the GLM notation? What does the Y and X mean?
16. Which has a bigger SS: the empty model or the **League** model?
17. Which R function would you use to compare the **League** model to the empty model?
18. The PRE from the **League** model is .0024. What does the PRE mean?
19. Why is the degrees of freedom for the **League** model equal to 1?
20. Let’s try to explain variation in Wins using Walks. Make a visualization to explore this idea of the DGP. What would be the word equation for this idea?
21. Create a variable that splits the teams up into two groups, those with few walks and those with more walks. Call it **Walks2group** and make it a factor.
22. Create a visualization to explore whether **Walks2group** could explain some of the variation in **Wins**. Just from exploring the visualizations, which explanatory variable is better at predicting **Wins**: **Walks2group** or **League**?
23. Create a model of **Walks2group**: specify the model in GLM notation, fit the model using R, and then interpret the numbers.
24. What would the Walks2group model predict for NYY’s wins? What would it predict for SEA’s wins?
25. Let’s say we run supernova on both the League model and the **Walks2group** model. They will both have the same SS Total. Why?
26. Let’s say we run supernova on both the League model and the **Walks2group** model. Which one has a bigger PRE? Why?