R: Homework 1
Fall 2019
Due October 18 @ 1159pm

Instructions

- Submit your assignment through Blackboard as a *.pdf document.
- Use full sentences to answer the questions.
- Include the code you used to produce your answers. Please don't copy and paste all the code you wrote (only paste the parts that are relevant).
- Your document should be easy to browse: your answers should be easy to identify, and it should be easy to know where an exercise starts and where it ends.
- When in doubt, err on the side of explaining what you did using your own words.
- If you have any questions, please let me know (my email is victor.pena@baruch.cuny.edu).

The Ultimate Halloween Candy Power Ranking

In this exercise, you will work with data that was collected for a fivethirtyeight article. You can download the data from a Github repository (link below). It is also accessible in library(fivethirtyeight), where it is available as data(candy_rankings). You can get some information on the variables by typing ?candy_rankings.

Article: https://fivethirtyeight.com/features/the-ultimate-halloween-candy-power-ranking/
Github repo: https://github.com/fivethirtyeight/data/tree/master/candy-power-ranking

1. Find the top 5 best rated and top 5 worst rated candy.
2. Plot winpercent against sugarpercent. Do you see any association? Now, plot winpercent against pricepercent. Do you see any association?
3. Consider all the logical-type variables in the dataset. For each logical variable, find the average difference in winpercent between the treats that satisfy the condition and the treats that don't satisfy it. Which logical variable seems to have the strongest effect on winpercent?

College admissions dataset

Read in the following dataset:

http://vicpena.github.io/admin.csv

It contains information about college admissions by gender and department at some college.

1. Find the percentage of men who applied and got in and the percentage of women who applied and got in. What do you see?
2. Now, find the percentage of men who applied and got in by department. Do the same with women. Compare the results with what you found in part 1.
3. Explain what is going on in this dataset. Do you see any evidence of gender discrimination?

Fandango movie ratings

Read the article https://fivethirtyeight.com/features/fandango-movies-ratings/. The dataset is accessible in library(fivethirtyeight). You can read it in with the command data(fandango). The variable names are reasonably self-descriptive, but you can get a more detailed description by typing in ?fandango.

1. Identify the Top 5 best rated and Top 5 worst rated movies in the dataset. Average over different platforms.
2. Visualize the difference between Fandango stars and actual Fandango ratings. Comment on what you see.
3. Some movies are loved by the critics, but hated by the audience (and sometimes, it’s the other way around). Given the data you have, create a metric to measure discrepancies between user and critic ratings. Create a table that contains the Top 5 movies that seem to appeal to critics but not the audience, and another table with the Top 5 movies that users seem to like more than critics do.

Lahman Baseball Dataset

Download the data from:

http://vicpena.github.io/sta9750/Teams1719.csv

Answer the following questions.

Some questions about home advantage

1. Create a statistic that quantifies “home advantage”. You’ll use this statistic for the next few questions. There is more than one reasonable choice here. Propose 2 different statistics and justify why you picked the one you’ll use from now on.
2. Find home advantage statistics for the American League (AL) and National League (NL) in the 2017-2019 period. Comment on the results. Do you see any differences between leagues? Do you see any evidence of home advantage at all? What are the years where there seems to be more of a home advantage, and those where the effect might not be as strong (or doesn’t seem to be there)?
3. Find the teams that had the highest and lowest home advantage effect by league in 2017, 2018, and 2019 separately. Comment on the results.
4. Which franchise had the highest average home advantage in the 2017-2019 period? Which one had the lowest average home advantage effect?
5. After completing these exercises, what did you learn about home advantage effect in the MLB? You’re welcome to try out a few new queries to illustrate your points.

In this exercise, you’ll work with the Lahman Baseball datasets, which you can access after installing library(Lahman). After installing the package, you can type in ?Lahman to get some information on the structure of the datasets and see what’s available. If you want to do a class project with baseball data, you’re welcome to use this resource.

Aging in pitchers and batters

1. Let’s consider data from 2018 only and look at the subset of pitchers who pitched more than 250 outs. Plot the earned run average (ERA; small values are good and big ones are bad) of the pitchers against their age. Do you see any patterns? Now, find a table with the average ERAs by age. Do you see any patterns?
2. Again, let’s look at pitchers who pitched more than 250 outs in 2018. Identify the top 5 best and worst pitchers, in terms of ERA.

3. Consider the best pitcher (in terms of ERA) that you found in part 2. Find his ERA by season throughout his career. Based on this alone, do you think he’s already “peaked”? If you like baseball, you’re welcome to share your opinion here as well.
4. Let’s do a similar exercise, but now with batting average (BA; more is better). Use the battingStats function in Lahman to find BAs. Consider data from 2018 only and look at players that have more than 200 at bats (AB). Plot BA against age. Do you see any patterns? Find a table with average BAs by age. Explain what you see.
5. Again, let’s look at players with more than 200 ABs in 2018. Find the top 5 best and worst players in terms of BA.
6. Consider the best player (in terms of BA) that you found in part 5. Find his BA by season throughout his career. Based on this alone, do you think he’s already “peaked”? If you like baseball, you’re welcome to share your opinion here as well.