Data Science Homework Problem

This assignment is intended to quickly assess general data science skills such as data munging (manipulation, reshaping, joining) and exploratory analysis.

Completion of the tasks should only require about an hour of your time. The goal is to showcase general skills, so it is not necessary to provide lengthy answers or polished results. An answer that simply addresses each question will suffice. If it takes you longer than an hour, that is okay, but we don’t want to take up too much of your time with this exercise.

The dataset that will be used for this exercise is related to American Major League Baseball statistics. This is intended to simulate scenarios where a data scientist must work with data that may be in an unfamiliar domain.

There are two data files. The main data file is **hits.csv**. Each record of this dataset pertains to an “at-bat”, where a ball is thrown to a batter and the batter has a number of tries to hit the ball. The dataset has records for 5,000 at-bats in the 2020 baseball season that resulted in a hit. The target in this dataset is whether or not the batter hits a “home run”, which is hitting the ball over the outfield wall. Variables recorded for each at-bat include the batter’s name and team, the pitcher’s name and team, the speed of the pitch, the angle the ball comes off the bat, etc. While no data dictionary is provided, the variables of interest in this dataset will be described in the exercises. Please provide the source code for your solutions. You can submit your solution however you would like – a Python script with comments, a Jupyter notebook, a Word document, etc.

## 1. Read in hits.csv

You should get something that looks like this:

**bip\_id game\_date home\_team away\_team batter\_team batter\_name pitcher\_name**   
 1 67045 2020-07-28 OAK COL OAK laureano, ra~ bard, daniel   
 2 116070 2020-07-26 NYM ATL ATL albies, ozzie oswalt, corey  
 3 8677 2020-09-08 HOU OAK OAK canha, mark greinke, zack  
 4 28003 2020-09-21 ATL MIA ATL d'arnaud, tr~ rogers, trev~  
 5 69185 2020-09-01 COL SF SF longoria, ev~ santos, anto~  
 6 88408 2020-07-24 BOS BAL BAL nunez, renato valdez, phil~  
 7 36269 2020-07-31 SEA OAK OAK laureano, ra~ walker, taij~  
 8 17990 2020-08-22 SF ARI ARI marte, starl~ anderson, ty~  
 9 69380 2020-09-15 LAA ARI ARI escobar, edu~ teheran, jul~  
10 68768 2020-08-25 CWS PIT CWS abreu, josé tropeano, ni~  
# ... with 4,990 more rows, and 18 more variables: batter\_id, pitcher\_id, is\_batter\_lefty , is\_pitcher\_lefty, bb\_type, bearing, pitch\_name, park, inning, outs\_when\_up, balls, strikes, plate\_x, plate\_z, pitch\_mph, launch\_speed, launch\_angle, is\_home\_run

# 2. Explore the number of hits over time

Using the game\_date variable, create a visualization that shows the number of hits over time. Do you notice anything strange about the distribution? If so, what might you do to further investigate? (Actual investigation not necessary - just a quick answer will suffice).

## 3. Create a univariate exploratory plot for launch speed

The continuous variable launch\_speed is the speed at which the ball leaves the bat after it is hit. Provide a simple univariate exploratory visualization of this variable.

## 4. Create a univariate exploratory plot for the bearing of the hit ball

The bearing of the hit ball is the direction that the ball travels with respect to the baseball field, one of either “right”, “center” or “left” field. Create a summary visualization of this categorical variable.

## 5. Visually compare the distribution of launch speed based on whether the hit was a “home run” or not

The variable is\_home\_run is if the hit ball goes over the outfield wall (called a “home run”), or if the hit ball does not go over the wall. Create a visualization that compares the distribution of launch speed for hits that are home runs and hits that are not.

## 6. Visually compare the distribution of bearing based on whether the hit was a “home run” or not

Similar to (5), create a visualization that compares the distribution of “bearing” and whether or not the hit is a home run.

## 7. Calculate a summary of the percentage of home runs for each batter

For each batter, denoted by the variable batter\_name, calculate the percentage of home runs that they hit. Which batter has the highest home run percentage? Explain for this batter whether you would trust the computed statistic or not.

**8. Read in park\_dimensions.csv**

The file **park\_dimensions.csv** contains information about every baseball park in American Major League Baseball, including how far away the left, center, and right field walls of the baseball park

are from the batter, and the height of the walls. It is thought that augmenting our hits dataset with some of this information could give us more insight into whether a hitter will hit a home run. The data should look something like this:

**park NAME Cover LF\_Dim CF\_Dim RF\_Dim LF\_W CF\_W RF\_W** 1 0 Chase Field Roof 328 407 335 8 25 8  
 2 1 SunTrust Park Outdo~ 335 400 325 6 8 16  
 3 2 Oriole Park at Camden Ya~ Outdo~ 333 400 318 7 7 21  
 4 3 Fenway Park Outdo~ 310 390 302 37 9 3  
 5 4 Wrigley Field Outdo~ 355 400 353 16 11 16  
 6 5 Great American Ballpark Outdo~ 328 404 325 12 8 8  
 7 6 Progressive Field Outdo~ 325 405 325 19 9 14  
 8 7 Coors Field Outdo~ 347 415 350 13 8 16  
 9 8 Guarantee Rate Park Outdo~ 330 400 335 8 8 8  
10 9 Comerica Park Outdo~ 345 420 330 6 8 8  
# ... with 20 more rows

## 9. Pivot the park dimensions from wide format to long format

The variables LF\_Dim, CF\_Dim, RF\_Dim provide the distance to the left field, center field, and right field walls, in feet. Create a new dataset from these variables that is pivoted from wide to long format, which should result in something that looks like this:

park bearing wall\_distance  
 1 0 left 328  
 2 0 center 407  
 3 0 right 335  
 4 1 left 335  
 5 1 center 400  
 6 1 right 325  
 7 2 left 333  
 8 2 center 400  
 9 2 right 318  
10 3 left 310  
# ... with 80 more rows

## 10. Join the wall distance with the hits data

For each record in the hits dataset, we have the baseball park ID stored in park and the bearing of the hit ball stored in bearing. Join your results from (9) to the hits dataset using these variables so the resulting hits dataset has a new column indicating the distance to the wall for the given baseball park and bearing of each hit.