Data 8 **Name:** _____
Summer 2018
Final Review Worksheet
8/4/2018
Conceptual Office Hours

This worksheet can serve as a general overview of materials you have learned this semester and give you an opportunity to practice solving them in an exam-like format. NOTE: This worksheet is a collection of problems that come from resources all students have access to, resources only staff members of previous iterations of the course have access to, and some that I have made up myself. This is not necessarily representative of what will be tested on the actual final. For any issues or questions, contact Robert Sweeney Blanco at robertsweeneyblanco@berkeley.edu.

For all problems, you may assume you have imported datascience and numpy as np.

The table *soccer* has the score of every international soccer game since 1872. For every game, there is a home team and an away team. The home team's nation is not always where the game is being played, that would be the host nation.

date	home_team	away_team	home_goals	away_goals	tournament	city	host_nation
1872-11-30	Scotland	England	0	0	Friendly	Glasgow	Scotland
1873-03-08	England	Scotland	4	2	Friendly	London	England
1874-03-07	Scotland	England	2	1	Friendly	Glasgow	Scotland
1875-03-06	England	Scotland	2	2	Friendly	London	England
1876-03-04	Scotland	England	3	0	Friendly	Glasgow	Scotland
1876-03-25	Scotland	Wales	4	0	Friendly	Glasgow	Scotland
1877-03-03	England	Scotland	1	3	Friendly	London	England
1877-03-05	Wales	Scotland	0	2	Friendly	Wrexham	Wales
1878-03-02	Scotland	England	7	2	Friendly	Glasgow	Scotland
1878-03-23	Scotland	Wales	9	0	Friendly	Glasgow	Scotland

... (38351 rows omitted)

1. Write a line of code that calculates the number of times the home team won and set it equal to *home*.

   home = np.count_nonzero(soccer.column("home_goals") > soccer.column("away_goals"))

2. Write a line of code that finds the proportion of times the home team has won. You may use the *home* variable you defined in the previous problem.

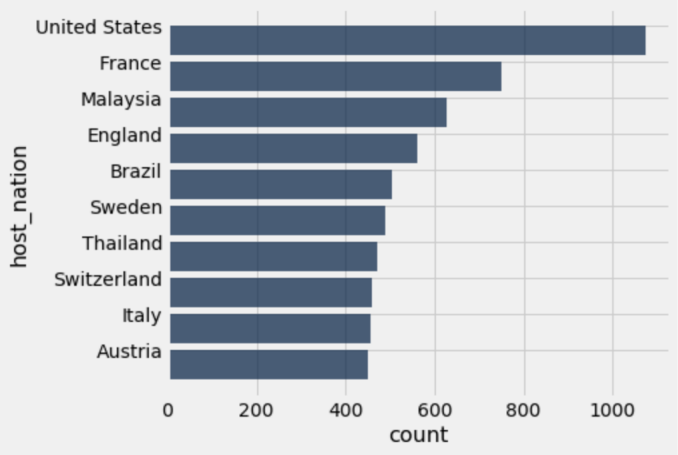
   home / soccer.num_rows

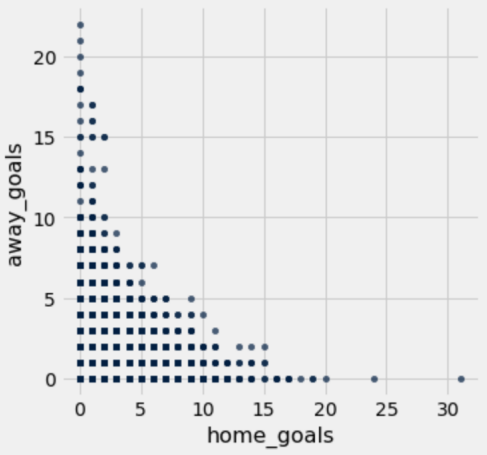
3. Create a bar graph of the top 10 nations that have hosted the most games.

   top_hosts = soccer.group("host_nation").sort("count", descending=True)

   top_hosts.take(np.arange(10)).barh("host_nation", "count")

The diagram on the left is the result if you ran the previous problem. The diagram on the right is a scatter plot comparing the number of goals the home team scores vs the away team.





4. You might not have expected the U.S to be hosting so many soccer games. Write a line of code that finds the American city that has hosted the most games.

   us_cities = soccer.where("host_nation", "United States").group("city")

   us_cities.sort("count", descending=True).column("city").item(0)

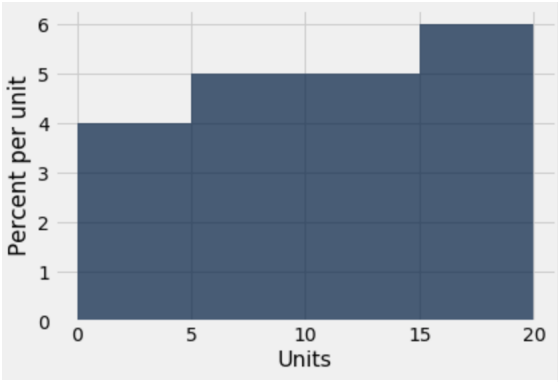
5. Write the line of code to generate the plot on the right.

   soccer.scatter("home_goals", "away_goals")

6. Wow, losing 31-0 must be tough! Write a line of code to find out which country lost that game.

   soccer.where("home_goals", 31).column("away_team").item(0)

Suppose you have the following histograms with bins made by bins=make_array(0,5,15,20).



7. If you were to split the middle bin in half, is it possible for the height (density) of the [5,10) bin to be 9?

   ◯ Yes: It is possible that 45% of the data in in that bin

   ◯ No:

8. If your answer to the previous question was yes, what would be the height of the [10,15) bin then?

   1, since that bin would have 5% of the data.

9. If you were to split the original histogram in half, is it possible for the height (density) of the [10,15) bin to be 11?

   ◯ Yes:

   ◯ No: Since the [5,15) bin has 50% of the data, it is impossible for the [10,15) bin to have 55% of the data.

10. If your answer to the previous question was yes, what would be the height of the [5,10) bin then?

Chef Ramsay has started selling his handmade cookies in grocery stores. He sends you, a government official, a dozen cookies to measure the sugar content. Your results are stored in the table *provided_cookies* on the left. The average amount of sugar in a cookie in that batch is 3.975 grams. You suspect that Chef Ramsay sent you a batch with a low sugar content to make them seem healthier. You go to the store, buy a dozen, and record their sugar content in the table *store_cookies* on the right. You record an average of 4.2 grams of sugar per cookie. You ask yourself if this could this have been by random chance and decide to run a hypothesis test.

Cookie	Provided Cookies Sugar
Cookie 1	3.8
Cookie 2	4.1
Cookie 3	3.9
Cookie 4	4.2
Cookie 5	4.1
Cookie 6	3.5
Cookie 7	4
Cookie 8	3.9
Cookie 9	4.2
Cookie 10	4
Cookie 11	3.9
Cookie 12	4.1

Cookie	Store Cookies Sugar
Cookie 1	4.2
Cookie 2	4.3
Cookie 3	3.9
Cookie 4	4.1
Cookie 5	4.3
Cookie 6	4.3
Cookie 7	4.1
Cookie 8	4.3
Cookie 9	3.8
Cookie 10	4.5
Cookie 11	4.2
Cookie 12	4.4

11. What kind of test do you need to do?

Permutation test

12. What is your Null Hypothesis?

The cookies come from the same distribution, and any difference is due to random chance.

13. What is your Alternative Hypothesis?

The cookies were sampled from different distributions. The population of the store cookies tends to have more sugar than the population of the given cookies

14. What is your Test Statistic?

average sugar of store cookies - average sugar of provided cookies

15. What is your Observed Test Statistic (You can leave your answer as a Python expression)?

4.2-3.975

Suppose you aggregate your data into a single table called *all_cookies*. The first column *Source* is 0 if the cookie was from the *provided_cookies* table and 1 if the cookie was from the *store_cookies* table.

Source	Sugar
0	3.8
0	4.1
0	3.9
0	4.2
0	4.1
0	3.5
0	4
0	3.9
0	4.2
0	4

... (14 rows omitted)

16. Fill in the code below

```
reps = 5000
stats = <1>
for i in np.arange(<2>):
    shuffle = all_cookies_tbl.<3>.column(<4>)
    tbl_with_shuffle = all_cookies_tbl.with_column("Shuffle", <5>)
    avg_array = tbl_with_shuffle.group(<6>, <7>).column(<8>)
    stat = <9> - <10>
    stats = np.append(stat, stats)
```

13.1) make_array()

13.2) reps

13.3) sample(with_replacement=False)

13.4) "Sugar"

13.5) shuffle

13.6) "Source"

13.7) np.mean

13.8) "Shuffle mean"

13.9) avg_array.item(1)

13.10) avg_array.item(0)

17. Write a line of code that evaluates the p-value.

np.count_nonzero(stats >= observed_value)/len(stats)

18. Suppose the p-value was 0.0052. Using a p-value cutoff of .05, what is your conclusion?

Reject the Null Hypothesis.

Suppose you have an array of integers called x. The standard deviation of the numbers in x is 2. What is the standard deviation of the array if you multiply every element by 5. In other words what is the SD of 5*x? Data 8 has not covered how to do this explicitly so lets break down some steps. Let y be the array when multiplying each element of x by 5 (i.e y=5*x).

19. What is the slope of the best fit line for predicting y based on x. In other words, if you make a scatter plot with x on the x-axis and y on the y-axis, what is the slope of the regression line? (HINT: Do not use any formulas, think intuitively or draw it out)

5

20. What is the correlation coefficient between x and y?

- ◯ 5
- ◯ 1
- ◯ 0
- ◯ -1
- ◯ -5
- ◯ Other: _____

21. What is the SD of y? (HINT: Look at your previous two answers and use a familiar formula)

$5 = 1\frac{SD_Y}{2} \Rightarrow SD_Y = 10$

Suppose instead you multiply each element by -5 (so y=-5*x). Lets find the SD of y.

22. What is the slope of the best fit line for predicting y based on x. In other words, if you make a scatter plot with x on the x-axis and y on the y-axis, what is the slope of the regression line? (HINT: Do not use any formulas, think intuitively or draw it out)

-5

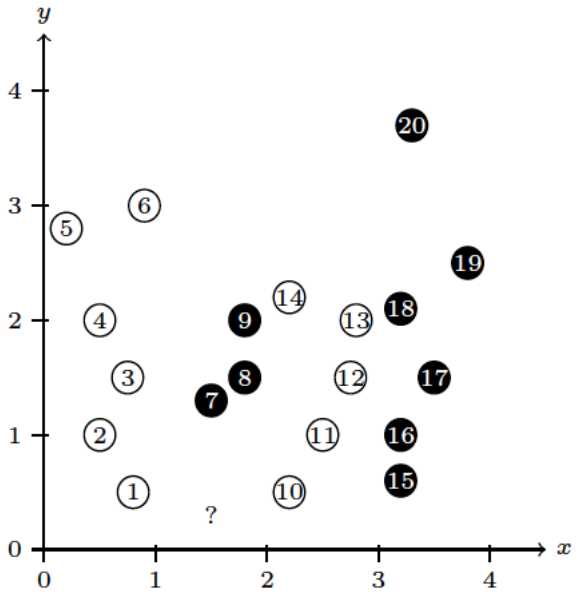
23. What is the correlation coefficient between x and y?

- ◯ 5
- ◯ 1
- ◯ 0
- ◯ -1
- ◯ -5
- ◯ Other: _____

24. What is the SD of y? (HINT: Look at your previous two answers and use a familiar formula)

$-5 = -1\frac{SD_Y}{2} \Rightarrow SD_Y = 10$

Suppose you want to classify the '?' either white or black in the picture below using KNN with k=3. The data that holds the coordinates to these points has x-values in the first column and y-values in the second column.



25. What would it be classified if you used the following as your distance function?

```
def distance(point1, point2):
    return np.sqrt(np.sum((point1 - point2)**2))
```

white

26. What would it be classified if you used the following as your distance function?

```
def distance(point1, point2):
    return np.sqrt(np.sum((point1.item(0) - point2.item(0))**2))
```

black

27. What would it be classified if you used the following as your distance function?

```
def distance(point1, point2):
    return np.sqrt(np.sum((point1.item(1) - point2.item(1))**2))
```

white