1. Tables

You are given the following table called `pokemon`. For the following questions, fill in the blanks.

Name	Type	Total	HP	Attack	Defense	Sp. Atk	Sp. Def	Speed	Generation	Legendary
Bulbasaur	Grass	318	45	49	49	65	65	45	1	False
Ivysaur	Grass	405	60	62	63	80	80	60	1	False
Venusaur	Grass	525	80	82	83	100	100	80	1	False
VenusaurMega Venusaur	Grass	625	80	100	123	122	120	80	1	False
Charmander	Fire	309	39	52	43	60	50	65	1	False
Charmeleon	Fire	405	58	64	58	80	65	80	1	False
Charizard	Fire	534	78	84	78	109	85	100	1	False
CharizardMega Charizard X	Fire	634	78	130	111	130	85	100	1	False
CharizardMega Charizard Y	Fire	634	78	104	78	159	115	100	1	False
Squirtle	Water	314	44	48	65	50	64	43	1	False

... (790 rows omitted)

1. Find the name of the pokemon of type `Water` that has the highest HP.

```
water_pokemon = pokemon._____(_____,_____)
water_pokemon._____(_____,_____).column("Name").item(0)
```

2. Find the proportion of pokemon of type `Fire` in the dataset whose Speed is strictly less than 100.

```
fire_pokemon = pokemon._____(_____,_____)
fire_pokemon._____(_____,_____)._____ / _____
```

3. Create a table containing `Type` and `Generation` that is sorted in decreasing order by the average HP for each pair of `Type` and `Generation`.

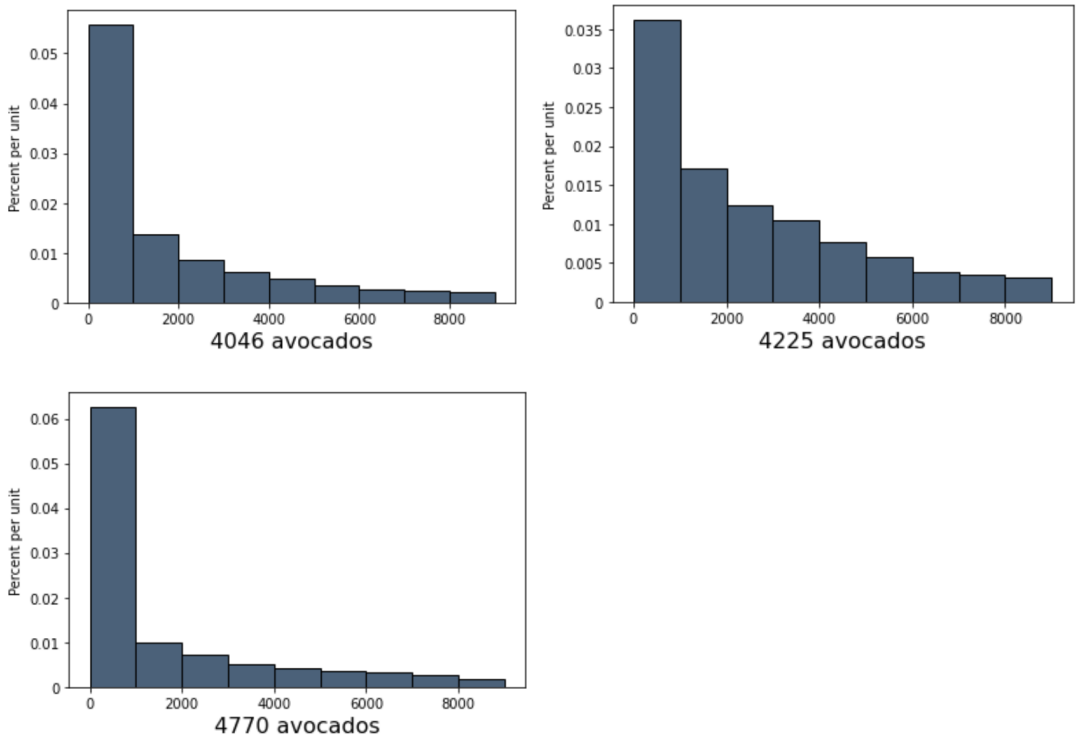
```
d = pokemon._____(_____,_____)
d.sort("HP mean",_____)._____(_____,_____)
```

4. Return an array that contains ratios of legendary to non-legendary pokemons for each generation.

```
t = pokemon._____(_____,_____)
ratio = t._____(_____) / t._____(_____)
```

2. Histograms

Everyone knows Zoomers love their avocado toast, so it comes as no surprise that this Hass Avocado dataset was a popular selection among your peers! Each type of avocado (PLU 4046, PLU 4225, PLU 4770) has a corresponding histogram below; each data point in the histogram represents the number of avocados sold in one order. All bars are **1000 units wide**, but take note: the **density scale is different in each histogram.** There are **16812 orders shown in the PLU 4046 histogram** and **19572 orders shown in the PLU 4225 histogram.**



1. Calculate each quantity described below or write *Unknown* if there is not enough information above to express the quantity as a single number (not a range). Show your work!

a. The percentage of PLU 4225 orders sold of less than 1000 avocados is equal to the percentage of PLU 4046 orders sold of less than 1000 avocados.

b. The number of PLU 4046 orders containing at least 2500 avocados but less than 3000 avocados.

c. The number of PLU 4770 orders containing at least 1000 avocados but less than 2000 avocados.

d. The number of PLU 4225 orders that contained less than 8000 avocados.

2. If the PLU 4225 histogram were redrawn, replacing the three bins from 0-1000, 1000-2000, 2000-3000 with one bin from 0-to-3000, what would be the height of its bar?

3. Probability

1. A fair coin is tossed five times. Two possible sequences of results are HTHTH and HTHHH. Which sequence of results is more likely? Explain your answer and calculate the probability that each sequence appears.

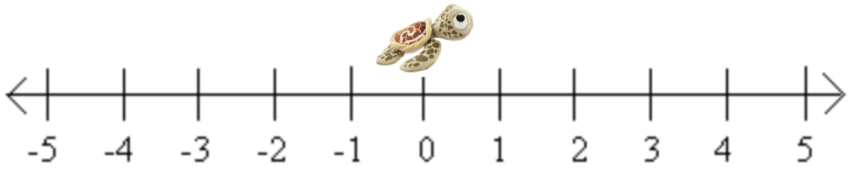
2. For questions 2 - 4, assume we have a biased coin such that the probability of getting heads is 1/5 and the probability of getting tails is 4/5. The coin is tossed 3 times. What is the probability that you get exactly 2 heads?

3. Once again, we toss the same biased coin 3 times. What is the probability you get no heads?

4. Again, we toss the same biased coin 3 times. What is the probability you get at least 1 heads?
Hint: There are two ways of calculating this probability. One is significantly easier to calculate than the other.

4. Simulation and Hypothesis Testing

Achilles the turtle sits on the number line. Achilles loves long random walks that last a total of 100 times steps. At each time step, Achilles moves based on the following scheme: He flips a coin and moves one step to the right if the coin comes up heads or one step to the left if the coin comes up tails.

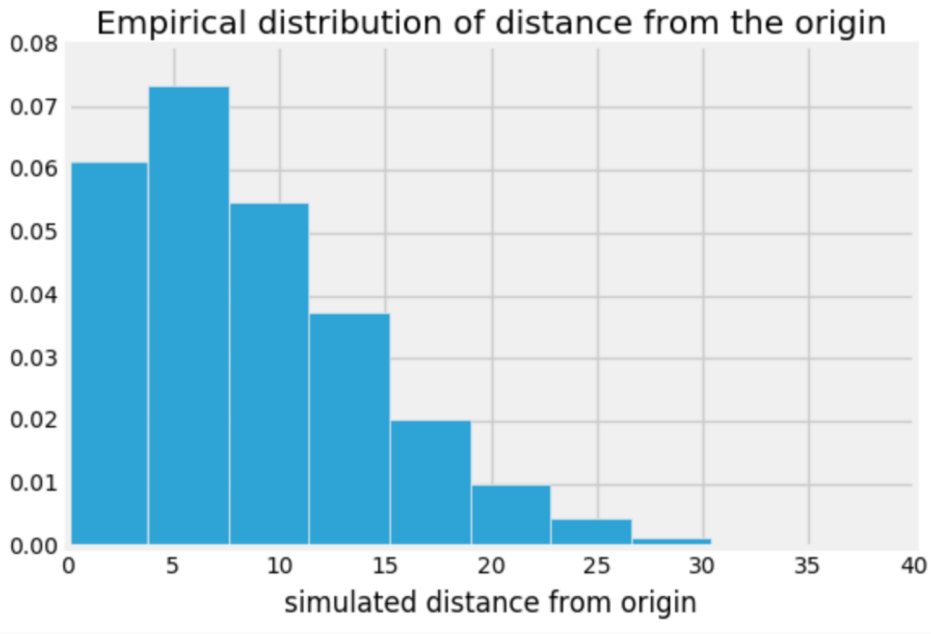


1. Assuming that Achilles' coin is fair, write a function called `one_walk` that simulates one random walk of 100 time steps and returns how far from the origin Achilles ends up at the end of his walk. You may assume that Achilles always starts from the origin.

```
def one_walk():
```

2. Assuming that Achilles' coin is fair, we would like to simulate what would happen if Achilles took 10000 different random walks. Complete the simulation below and keep track of how far Achilles ends up from the origin in each of his walks in an array called `distances`. The histogram shown below is an example of a histogram plotted from `distances`.

```
distances = make_array()

for i in np.arange(10000):

      new_distance = _____________________________________________________

      distances = _____________________________________________________
```



3. Achilles goes for a walk and claims that at the end of his walk, he ended up 30 steps away from the origin. You notice this is strange, so you want to run a hypothesis test to test whether or not Achilles used a fair coin. Fill in the blanks below for the null and alternative hypotheses and test statistics.

Hint: When considering your alternative hypothesis, note that we do not really care about whether the coin is biased towards heads or towards tails.

Null Hypothesis:

Alternative Hypothesis:

Test Statistic:

4. Write the code to calculate the p-value given the test statistic listed above and using a 5% p-value cut-off. Then, describe the different conclusions that you would arrive at depending on the p-value.

Hint: We simulated an array in part(b) of test statistics under the null hypothesis. Try to use the `distances` *array.*

```
p_value = _____________________________________________________
```

5. True/False

Respond with true or false to the following questions. If your answer is false, explain why.

1. In the U.S. in 2000, there were 2.4 million deaths from all causes, compared to 1.9 million in 1970, which represents a 25% increase. The data shows that the public's health got worse over the period 1970-2000.

2. A company is interested in knowing whether women are paid less than men in their organization. They share all their salary data with you. An A/B test is the best way to examine the hypothesis that all employees in the company are paid equally.

3. Consider a randomized control trial where participants are randomly split into treatment and control groups. We are 100% certain there will be no systematic differences between the treatment and control groups if the process is followed correctly.

4. A researcher considers the following scheme for splitting a people into control and treatment groups. People are arranged in a line and for each person, a fair, six-sided die is rolled. If the die comes up to be a 1 or a 2, the person is allocated to the treatment group. If the die comes up to be a 3, 4, 5 or 6 then the person is allocated to the control group. This is a randomized control experiment.

5. You are conducting a hypothesis test to check whether a coin is fair. After you calculate your observed test statistic, you see that its p-value is below the 5% cutoff. At this point, you can claim with certainty that the null hypothesis can not be true.

6. You roll a fair die a large number of times. While you are doing that, you observe the frequencies with which each face appears and you make the following statement: As I increase the number of times I roll the die, the probability histogram of the observed frequencies converges to the empirical histogram.

6. Multiple Choice

1. Gary is playing with a coin and he wants to test whether his coin is fair. His experiment is to toss the coin 100 times. He chooses the following null hypothesis.

Null Hypothesis: The coin is fair and any deviation observed is due to chance.

For each of the alternative hypotheses listed below, determine whether or not the test statistic is valid.

a. **Alternative Hypothesis:** The coin is biased towards heads.
Test Statistic: # of heads

b. **Alternative Hypothesis:** The coin is not fair.
Test Statistic: # of heads

c. **Alternative Hypothesis:** The coin is not fair.
Test Statistic: |# of heads - expected # of heads|

d. **Alternative Hypothesis:** The coin is biased towards heads.
Test Statistic: |# of heads - expected # of heads|

e. **Alternative Hypothesis:** The coin is not fair.
Test Statistic: 1/2 - proportion of heads

7. Fun with Functions

1. Write a function called `compute_pvalue` that, given an empirical distribution in the form of an array and the observed value of your test statistic, calculates the p-value for that test statistic. You may assume that large values of your test statistic provide evidence against the null hypothesis.

```
def compute_pvalue(empirical_dist, observed_ts):
```

2. Now write a function called `is_significant` that takes in an empirical distribution, the observed test statistic and a p-value cutoff, returns `True` if the p-value of the observed test statistic is statistically significant based on the cutoff provided and `False` otherwise.
Hint: Use the function you defined in Question 1!

```
def is_significant(empirical_dist, observed_ts, cutoff):
    _____
    return _____
```

8. More Hypothesis Testing

Chloe is a big fan of Trader Joes' frozen mac n cheese, but she noticed that the cheese used in it varies from box to box. A Trader Joe's employee provides her with some data about the 4 different cheeses used and the probability of them being used in each box:

Cheese	**Probability**
Velveeta	0.05
Gruyère	0.55
Sharp Cheddar	0.25
Monterey Jack	0.15

Chloe is suspicious about this distribution. After all, Velveeta is much cheaper to use than Gruyère, and she has also never bought a box that uses Gruyère. Chloe decides to buy many boxes throughout the next month and tracks the type of cheese used in each box. She uses this to conduct a hypothesis test.

1. Write the correct null hypothesis for this experiment

- Null Hypothesis:

- Alternative Hypothesis:

```
observed_proportions = make_array(0.2, 0.3, 0.45, 0.05)

employee_proportions = make_array(0.05, 0.55, 0.25, 0.15)
```

The array `observed_proportions` contains the proportions of cheese that Chloe observed in 20 boxes of Mac n Cheese.

2. Chloe wants to use the mean as a test statistic, but Katherine suggests that she uses the TVD (total variation distance) instead. Which test statistic should Chloe use in this case? Briefly justify your answer. Then write a line of code to assign the observed value of the test statistic to `observed_stat`.

```
observed_stat = _____________________________________________________________
```

3. Define the function `one_simulated_test_stat` to simulate a random sample according to the null hypothesis and return the test statistic for that sample.

```
def one_simulated_test_stat():

      sample_prop = _____________________________________________________

      return _____________________________________________________________
```

4. Chloe simulates the test statistic 10,000 times and stores the results in an array called `simulated_stats`. The observed value of the test statistic is stored in `observed_stat`. Complete the code below so that it evaluates to the p-value of the test:

```
_____________________(simulated_stats _____ observed_statistic) / _________________
```

5. Given that the computed p-value is 0.0825, which of the following are true? Select all that may apply.
   a. Using an 8% p-value cutoff, the null hypothesis should be rejected.
   b. Using a 10% p-value cutoff, the null hypothesis should be rejected.
   c. There is an 8.25% chance that the null hypothesis is true.
   d. There is an 8.25% chance that the alternative hypothesis is true.

9. A/B Testing

1. Choose True/False for each of the statements below, and explain your answer.

a. A/B testing is used to determine whether or not we believe two samples come from the same underlying distribution.

b. To conduct a permutation test, you should sample your data with replacement with a sample size equal to the number of rows in the original table.

c. A/B testing is the same as using total variation distance as a test statistic for a hypothesis test.

2. Kevin, a museum curator, has recently been given specimens of caddisflies collected from various parts of Northern California. The scientists who collected the caddisflies think that caddisflies collected at higher altitudes tend to be bigger. They tell her that the average length of the 560 caddisflies collected at high elevation is 14mm, while the average length of the 450 caddisflies collected from a slightly lower elevation is 12mm. She's not sure that this difference really matters, and thinks that this could just be the result of chance in sampling.

a. What's an appropriate null hypothesis that Kevin can simulate under?

b. How could you test the null hypothesis in the A/B test from above? What assumption would you make to test the hypothesis, and how would you simulate under that assumption?

c. What would be a useful test statistic for the A/B test? Remember that the *direction* of your test statistic should come from the initial setting.

d. Assume `flies` refers to the following table:

Elevation	Specimen length
High elevation	12.3
Low elevation	13.1
High elevation	12.0

(1007 rows omitted)

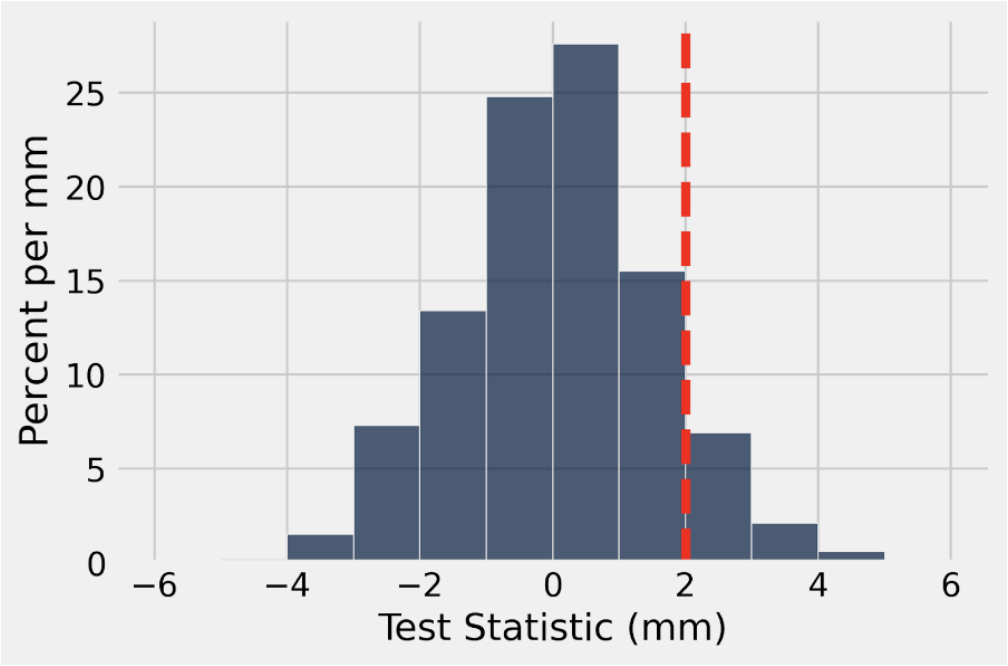
Fill in the blanks in this code to generate one value of the test statistic under the null hypothesis.

```
def one_simulation():
    shuffled_labels = flies._____.column('Elevation')
    shuffled_flies = flies.with_columns(_____, _____)
    grouped = shuffled_flies._____(_____,_____)
    means = grouped.column('Specimen length mean')
    statistic = _____
    return statistic
```

e. Fill in the code below to simulate 10000 trials of our permutation test.

```
test_stats = _____
repetitions = _____
for i in np.arange(_____)
    one_stat = _____
    test_stats = np.append(test_stats, one_stat)
test_stats
```

f. The histogram of `test_stats` is plotted below with a vertical red line indicating the observed value of our test statistic. If the p-value cutoff we use is 5%, what is the conclusion of our test?



g. Suppose that the null hypothesis is true. If we ran this same hypothesis test 1000 times, each time from our `flies` table and with a p-value cutoff of 5%, how many times would we expect to incorrectly reject the null hypothesis?

h. What effect does *decreasing* our p-value cutoff have on the number of times we expect to *incorrectly reject* the null hypothesis?

Past Exam Review

Grabbing Socks - Sp19 Midterm Q5

Professor Fithian stays up late laundering his socks and sleeps in one day when he is going to give his Data 8 lecture. While running out the door in a state of panic, he grabs two socks completely at random from the dryer, which contains 28 total socks: 16 black socks, 10 white socks, and 2 lime green socks.

The socks are not in pairs and haven't been touched since they were "shuffled" by the dryer the night before, so the two socks are like two draws at random without replacement.

Find the following probabilities. **Show your work!** Your answers should be math calculations, not Python code, but **you do not have to simplify any arithmetic**.

a. The probability that both of the socks are black:

b. The probability that at least one sock is lime green:

c. The probability that one sock is black and one is white:

d. The probability that the two socks are not the same color:

Cookie Factory - Su17 Midterm Q4

Cookie Monster owns a cookie factory. His factory produces cookie flavors in the following proportions:

Sugar	Chocolate Chip	Trash
0.2	0.6	0.2

Sam is unhappy that his box of 10 cookies had 4 Trash-flavored cookies in it. However, he doesn't remember

whether he got his cookies from Cookie Monster's factory or another factory. You decide to run a hypothesis test to figure out whether Sam's box came from Cookie Monster's factory or not.

a. Circle the correct option from each set of choices to state an appropriate null hypothesis for this problem.

Null hypothesis: Sam's box **[did] / [did not]** come from Cookie Monster's factory.

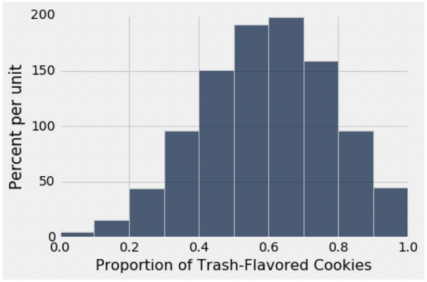
Thus, the chance that any one cookie in Sam's box was Trash-flavored is **[0.2] / [computable but not 0.2] / [unknown given this information]**.

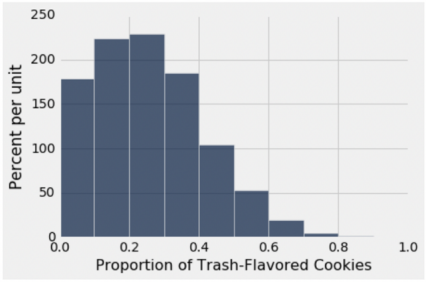
Because **[of random chance] / [the box didn't come from Cookie Monster's factory]**, Sam got a higher number of Trash-flavored cookies in his box.

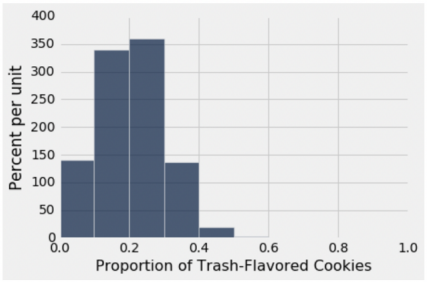
b. As a test statistic, you choose to compare **the proportion** of Trash-flavored cookies in boxes of 10 cookies. Complete the lines below to compute a single test statistic using a random sample from the population, storing the value into `test_stat`. You may write expressions as long as you need to on each line, but you may not use extra lines.

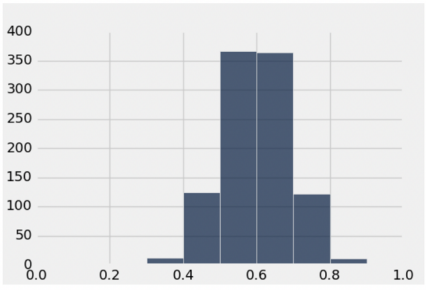
```
outcomes = make_array(_____________________________________________________)

sample = np.random.choice(outcomes, _____________________________________)

test_stat = _____________________________________
```

c. Given that your final computed p-value is less than 0.05, select all histograms corresponding to the possible probability distributions of the test statistic.

☐ 

☐ 

☐ 

☐ 

Briefly justify your answer below:

Python Practice - Fa20 Midterm Q1 (adapted)

a. Write a line of code to compute an array containing the first 100 squares, i.e., 1, 4, 9, 16, . . . , 10000. The square of a number is that number to the power of 2.

b. In one sentence, please describe what the following function returns.

```
def mystery(x):
    for i in np.arange(1, x+1):
        if i*i >= x:
            return i
    return 0
```

c. Assume you are given a function `simulate_once()` that, when called, returns the result of a single simulation. Write code that creates an array containing the result of 1000 simulations and assigns it to `results`.

d. Write a function `odd_sum` that, given an array of numbers `nums`, returns the sum of only the odd numbers in the array.

```
def odd_sum(nums):

   running_sum = 0
```

Blood Pressure - Sp19 Midterm Q7 (adapted)

In a randomized controlled experiment, 300 patients are randomized into the treatment group (Group A) and 200 to the control group (Group B). At the end of the experiment, the blood pressure of each patient is measured. The research team wants to test the null hypothesis that the treatment has no effect versus the alternative hypothesis that the treatment has an effect. As their test statistic they decide to use the absolute difference between the average blood pressures of Group A and Group B. The table data has 500 rows, one for each patient in the experiment. The table has just one column. The column is labeled 'bp' for blood pressure, and contains a numerical measurement of blood pressure for each patient. All measurements are in the same units. The table data has no other information. If this is possible, fill in the code below. If this is not possible, give a brief explanation as to why.

```
shuffled = data.sample(with_replacement = False)

# Two tables

group_A_shuffled_table = shuffled._____(_____)

group_B_shuffled_table = shuffled._____(_____)

# Two averages

shuffled_mean_A = np.average(group_A_shuffled_table.column('bp'))

shuffled_mean_B = np.average(group_B_shuffled_table.column('bp'))

# Test statistic

abs(shuffled_mean_A - shuffled_mean_B)
```

Visualizations - SP 21 Midterm Q7 (Modified)

During shelter-in-place, many Data 8 students explored new hobbies. Angela takes a poll about their hobbies and puts her findings in the hobbies table. Angela wonders how to visualize the collected data for presentation at her next discussion.
The first few rows of the hobbies table are shown below:

favorite color	favorite hobby	hours of sleep	wake up time	favorite food
blue	baking	7	early	crepe
green	running	10	late	boba
purple	tv shows	8	late	pizza
blue	basketball	7	early	mango
blue	hiking	6	both	pizza

. . . (282 rows omitted)
(a) (1.0 pt) Which of the following variables are categorical? (Select all that apply):

☐ Favorite Food

☐ Hours of Sleep

☐ Favorite Hobby

☐ Wake up time

☐ Favorite Color

(b) (1.0 pt) Select all that are correct:

☐ Whether a student attended discussion on a given day is a numerical variable.

☐ The size of attendance at a particular discussion section on a given day is a categorical variable.

☐ The size of attendance at a particular discussion section on a given day is a numerical variable.

☐ Whether a student attended discussion on a given day is a categorical variable.

(c) (1.0 pt) Angela wants to choose a finger food to send to the students in her discussion section. She uses the level of support (i.e., number of rows in the table) for each favorite food (Bagels, Chicken Nuggets, Samosas, Trinidad Cod Fritters, and Wontons) as the basis for her selection. Ultimately, she selects the food option with the most support (favorite).(Select all that apply):

☐ Favorite food option is a numerical variable.

☐ Level of support for a food option is a categorical variable.

☐ Level of support for a food option is a numerical variable.

☐ Favorite food option is a categorical variable.

(d) (1.0 pt) The best visualization to understand the distribution of the top 5 most popular foods in this discussion is

Ⓐ Scatter Plot

Ⓑ Histogram

Ⓒ Bar Chart

Ⓓ Line Plot

(e) (1.0 pt) Given the hobbies table, which methods would you use to help you plot the visualization you chose in part d? (Select all that apply):

- ☐ `.sort`
- ☐ `.join`
- ☐ `.group`
- ☐ `.apply`
- ☐ `.pivot`
- ☐ `None of the above`

(f) (1.0 pt) The best visualization to understand the association between favorite hobby and wake up time in this discussion is (Choose only one.):

- Ⓐ Scatter Plot
- Ⓑ Overlaid Bar Chart
- Ⓒ Bar Chart
- Ⓓ Line Plot
- Ⓔ Overlaid Histogram

(g) (1.0 pt) Which visualization would best display the association between hours of sleep and early or late wake up time? (Choose only one.)

- Ⓐ Scatter Plot
- Ⓑ Overlaid Bar Chart
- Ⓒ Bar Chart
- Ⓓ Line Plot
- Ⓔ Overlaid Histogram

(h) (1.0 pt) What visualization is impossible to make without modifying the hobbies table? (Select all that apply):

- ☐ Line Plot
- ☐ Bar Chart
- ☐ Overlaid Histogram
- ☐ Overlaid Bar Chart
- ☐ Scatter Plot