DATA 8 Fall 2020

Sample Exam.

FINAL EXAM

INSTRUCTIONS

This is your exam. Complete it either at exam.cs61a.org or, if that doesn't work, by emailing course staff with your solutions before the exam deadline.

This exam is intended for the student with email address `<EMAILADDRESS>`. If this is not your email address, notify course staff immediately, as each exam is different. Do not distribute this exam PDF even after the exam ends, as some students may be taking the exam in a different time zone.

For questions with **circular bubbles**, you should select exactly *one* choice.

- ◯ You must choose either this option
- ◯ Or this one, but not both!

For questions with **square checkboxes**, you may select *multiple* choices.

- ☐ You could select this choice.
- ☐ You could select this one too!

You may start your exam now. Your exam is due at <DEADLINE> Pacific Time. Go to the next page to begin.

For fill-in-the-blank coding questions, you can put anything inside the blanks, including commas, parentheses, and periods.

The exam is worth 130 points.

In alphabetical order, the sections are as follows (it might not be this order on your exam):

Assistant - 15 points

Blackboard Erasers - 11 points

Boba - 15 points

Imposter - 19 points

Mask On - 30 points

Olympics - 12 points

Roommates - 7 points

Soup Dumplings - 21 points

There is also a Just For Fun section, worth 0 points, and a Last Words section, where you can state any assumptions you made on the exam, also worth 0 points.

If you encounter any logistical problems during the exam, please contact us at data8berkeley@gmail.com. We will not be answering any questions related to the contents of the exam.

(a) Your name:

(b) Your @berkeley.edu email address:

(c) The Berkeley Honor Code states: "As a member of the UC Berkeley community, I act with honesty, integrity, and respect for others." Do you agree to follow the honor code on this exam?

◯ Yes

◯ No

1. (12 points) Olympics

Tam and Ananya are project partners in a history class about the Summer Olympics. For their project, they've collected data about every athlete who competed between 1896 to 2016 and put it into a table called `OLYMPIC_TBL`.

The table has the following columns:

- *Name*: (string) the athlete's name
- *Sex*: (string) the athlete's sex ('M' for male, 'F' for female)
- *Age*: (int) the athlete's age in years
- *Team*: (string) the athlete's country
- *Sport*: (string) the sport in which the athlete competed
- *City*: (string) the location of the Olympics
- *Year*: (int) the year of the Olympics

Here are the first few rows:

Name	Sex	Age	Team	Sport	City	Year
Patimat Abakarova	F	21	Azerbaijan	Taekwondo	Rio de Janeiro	2016
Jerzy Adamski	M	23	Poland	Boxing	Roma	1960
Nathan Ghar-Jun Adrian	M	27	United States	Swimming	Rio de Janeiro	2016
Michael B. "Mike" Adam	M	24	Canada	Curling	Torino	2006

... (132,201 rows omitted)

For each question below, write Python code to answer the question using what we have taught you in this class. If we ran your Python code, it should evaluate to the answer to the question.

(a) (2 pt) How many countries participated in the 2016 Summer Olympics?

Recall: The `OLYMPIC_TBL` table has the following columns `['Name', 'Sex', 'Age', 'Team', 'Sport', 'City', 'Year']`.

(b) (3 pt) What was the name of the oldest SPORT_SUMMER08 in the 2008 Summer Olympics?

Recall: The `OLYMPIC_TBL` table has the following columns `['Name', 'Sex', 'Age', 'Team', 'Sport', 'City', 'Year']`.

(c) (4 pt) Which team had, on average, the youngest athletes at the Olympics?

Recall: The `OLYMPIC_TBL` table has the following columns `['Name', 'Sex', 'Age', 'Team', 'Sport', 'City', 'Year']`.

(d) (3 pt) Suppose you've created the following function, which returns the description of a single athlete in the form of a sentence:

```
def description(name, sport, year):
    return name + " competed in " + sport + " for the " + year + " Olympics."
```

Write a Python expression that will return an array containing the descriptions of each athlete in the table.

Recall: The `OLYMPIC_TBL` table has the following columns `['Name', 'Sex', 'Age', 'Team', 'Sport', 'City', 'Year']`.

2. (7 points) Roommates

Troy and Abed, residents of the new dorm called Greendale at UC Berkeley, are applying to become roommates. Assume that the chance they get selected as roommates each year is 1/5, regardless of whom they lived with previously.

(a) (3 pt) If Troy and Abed plan to stay in Greendale for three years, what is the probability that they are roommates at *at least one* of the three years they apply?

- ◯ 1/125
- ◯ 1 - 24/25
- ◯ 3/125
- ◯ 1 - 64/125
- ◯ 124/125
- ◯ Cannot be calculated with the information given

(b) (2 pt) If Troy and Abed plan to stay in Greendale for four years, what is the probability that they are roommates in their fourth year, given that they were roommates the first three years?

- ◯ 1/625
- ◯ 624/625
- ◯ 0
- ◯ 4/5
- ◯ 1/5
- ◯ Cannot be calculated with the information given

(c) (2 pt) What is the purpose of Bayes Rule? Select all that are correct:

- ☐ To quantify the impact of subjective probabilities on our predictions.
- ☐ To test whether there is a causal relationship between two variables.
- ☐ To evaluate the accuracy of a machine learning model on the population.
- ☐ To determine what percentage of our data lies within a certain number of standard deviations from the mean.
- ☐ To update our predictions with new information.

3. (11 points) Blackboard Erasers

Mr. White is a teaching Chemistry class in Latimer Hall. His class has 100 undergraduate staff (uGSIs, tutors, etc.).

Mr. White learns that one of his staff is stealing blackboard erasers from his lecture hall when the building is closed overnight. The only way the thief could access the building overnight is with a key card that belongs to them.

Suppose Mr. White discovers that 10 of his staff have a key card.

(a) (2 pt) If Mr. White randomly selects one of his staff, what is the probability that they are the eraser thief?

◯ (0.01 * 1) / (0.01 * 1 + 0.99 * 9/99)

◯ (0.1 * 1) / (0.1 * 1 + 0.9 * 1/99)

◯ 0.01

◯ 0.1

◯ 0.99 * 9/99

(b) (2 pt) If Mr. White randomly selects one of his staff, what is the probability that they are **not** the eraser thief and **have** a key card?

◯ (0.01 * 1) / (0.01 * 1 + 0.99 * 9/99)

◯ (0.1 * 1) / (0.1 * 1 + 0.9 * 1/99)

◯ 0.01

◯ 0.99

◯ 0.99 * 9/99

(c) (3 pt) At the next course staff meeting, Mr. White notices that one of his GSIs, Gus, has a key card sticking out of his wallet.

Given this information, what is the probability that Gus is the eraser thief?

◯ (0.01 * 1) / (0.01 * 1 + 0.99 * 9/99)

◯ (0.1 * 1) / (0.1 * 1 + 0.9 * 1/99)

◯ 0.01

◯ 1

◯ 1 - 0.99 * 9/99

(d) (4 pt) Mr. White is skeptical of his head GSI, Jesse. Prior to learning any information about Jesse's keycard access, Mr. White believes there is a 25% probability that Jesse is the eraser thief.

Suppose Mr. White later discovers that Jesse has a key card.

Given this new information, what is the probability that Jesse is the eraser thief?

◯ (0.25 * 1) / (0.25 * 1 + 0.75 * 9/99)

◯ (0.25 * 9/99) / (0.25 * 9/99 + 0.75 * 1)

◯ 0.25

◯ 0.1

◯ 1 - 0.75 * 9/99

4. (19 points) Imposter

Rick and Morty have been aboard a spaceship for 3 weeks among a crew of 100 people. Rick learns that there is an imposter among them, trying to sabotage their mission.

Rick remembers that other space missions recently have had imposters sabotaging their crew. He thinks he might be able to make some predictions about his mission based on how those other missions concluded.

Rick samples flight logs of 26 crew members from across various missions in the last year and notes down a couple attributes about each sampled person, including the following:

- *speed*: (float) how fast they can run in zero-gravity
- *num tasks*: (int) how many tasks they completed by Week 3 of their mission
- *outcome*: (string) whether they ended up being an Imposter or an innocent Crewmate

Rick is suspicious of Morty's sister, Beth, so he visualizes this sample data along with Beth's attributes (the red X) in the following chart:

Suppose Rick now wants to use a k-nearest neighbor classifier to predict whether Beth is an imposter based on just her *speed* and *num tasks*.

(a) (3 pt) Which of the following values of k would be appropriate for Rick to use for his classifier? Select all that are correct:

- ☐ 1
- ☐ 2
- ☐ 4
- ☐ 7
- ☐ 10
- ☐ 29

(b) (2 pt) If Rick uses k=3 for his classifier, what will his prediction for Beth be?

- ◯ Imposter
- ◯ Crewmate

(c) (2 pt) If Rick uses k=5 for his classifier, what will his prediction for Beth be?

- ◯ Imposter
- ◯ Crewmate

(d) (2 pt) Rick now decides to try his classifier on Morty. Suppose Morty has a *speed* of 5.2 and *num tasks* of 4.

The closest neighbor to Morty in the training set has a *speed* of 5.6 and *num tasks* of 3.

Which of the following Python expressions returns the distance between Morty and his closest neighbor in the training set?

◯ `np.sqrt((5.2-5.6)**2 + (4-3)**2)`

◯ `np.sqrt((5.2-4)**2 + (5.6-3)**2)`

◯ `np.sqrt(abs(5.2-5.6) + abs(4-3))`

◯ `(abs(5.2-5.6) + abs(4-3))/2`

◯ `np.sqrt(((5.2-5.6)**2 + (4-3)**2)/2)`

(e) (4 pt) Morty is suspicious of Rick's classifier and decides to create his own classifier. He conducts a random sample of 100 crew members from various space missions in the past year, observing the same attributes like *speed* and *num tasks*. Before using his classifier to make predictions, Morty would like to estimate its accuracy.

Which of the following are appropriate ways to do this? Select all that are correct:

☐ Train the classifier on all 100 crew members in the sample, use the classifier to predict the outcome of each of the 100 members, and measure how often the classifier is correct.

☐ Train the classifier on 55 randomly selected crew members from the sample, use the classifier to predict the outcome of each of the remaining 45 members, and measure how often the classifier is correct.

☐ Train the classifier on 70 randomly selected crew members from the sample, use the classifier to predict the outcome of each of the remaining 30 members, and measure how often the classifier is correct.

☐ Train the classifier on all 100 crew members in the sample, use the classifier to predict the outcome of 50 randomly selected members in the sample, and measure how often the classifier is correct.

☐ Train the classifier on all 100 crew members in the sample, use the classifier to predict the outcome of just the imposters in the sample, and measure how often the classifier is correct.

(f) (2 pt) Morty is worried that his classifier is going to perform very well on his training set but won't be very accurate when making predictions on new crew members in the population. What is the name of this phenomenon?

◯ Overfitting

◯ Underfitting

◯ Overprediction

◯ Bayes Condition

◯ Chebyshev's Paradox

(g) **(2 pt)** Suppose Morty would instead like to predict a crew member's *num tasks* based on their *speed*. Which of the following charts would be most appropriate to visualize the relationship between these variables?

- ◯ Scatterplot
- ◯ Pivot Table
- ◯ Histogram
- ◯ Line Graph
- ◯ Bar Chart

(h) **(2 pt)** Suppose Morty wants to predict a crew member's *speed* based on their *num tasks*. Which of the following could be used to answer this question? Select all that are correct:

- ☐ Linear Regression
- ☐ Total Variation Distance
- ☐ Central Limit Theorem
- ☐ Classification
- ☐ A/B Testing

5. (30 points) Mask On

Peter, owner of the Sweetwater clothing store in Berkeley, notices that whenever the number of daily reported COVID-19 cases in Alameda county increases, his shop seems to sell more cloth face masks the following day.

Peter decides to sample a few days over the last year and puts this data into a table called `masks`. Here are the first few rows:

Date	Cases	Temperature	Sales
May 1, 2020	1848	68	45
June 15, 2020	2709	72	30
July 4, 2020	8878	75	65
August 26, 2020	5292	74	103

... (158 rows omitted)

The table contains the following columns:

- *Date*: (string) the month, day and year of a particular day
- *Cases*: (int) the number of new cases reported that day
- *Temperature*: (int) the average temperature that day, in Fahrenheit
- *Sales*: (int) the number of cloth face masks sold the following day

Peter also writes the following function in Python:

```
def mystery(x):
    return (x - np.mean(x))/np.std(x)
```

Note: You may use the function above for any of the following questions.

(a) (1 pt) Describe in one sentence what the function `mystery()` does.

(b) (3 pt) Peter wants to visualize the relationship between the number of reported cases and the number of masks sold, so he creates the chart below.

Write a set of python expressions that would generate this chart.

(c) (2 pt) What can Peter conclude about the sample from the above chart? Select all that are correct:

- ☐ Because the reported cases are from the day before the sales, we can conclude that an increase in reported cases causes an increase in face mask sales.
- ☐ There is a positive association between reported cases and face mask sales.
- ☐ There is a positive correlation between reported cases and face mask sales.
- ☐ There is a negative association between reported cases and face mask sales.
- ☐ There is no correlation between reported cases and face mask sales.

(d) (6 points)

For the next three questions, assume you know the following:

- the *Cases* column has a mean of 5000 and a standard deviation of 1000
- the *Sales* column has a mean of 50 and a standard deviation of 10
- the correlation between the *Cases* and *Sales* columns is 0.6

i. (2 pt) Suppose Peter wants to predict *Sales* from *Cases* and decides to fit a regression line.

If there were 15,000 new cases reported on November 30, what would the regression predict as the number of sales on December 1?

- ◯ 90
- ◯ 60
- ◯ 110
- ◯ 120
- ◯ 80

ii. (2 pt) Suppose Peter wants to instead fit a regression line to predict sales in standard units from cases in standard units.

What is the **slope** of this regression line?

- ◯ 0.6
- ◯ 100
- ◯ 0.01
- ◯ 0.006
- ◯ 600

iii. (2 pt) What is the **intercept** of the regression line in the above question?

- ◯ 0
- ◯ 20
- ◯ 2,000
- ◯ 0.02
- ◯ 0.2

(e) (3 pt) Suppose you've created a `predict()` function to predict *Sales* from *Cases*. Assume that this function takes an array of cases (in original units) as input and returns an array of predicted sales (in original units).

Write a Python expression that returns the `masks` table with a new column named *Resids* that contains the residuals of the linear regression.

(f) (3 pt) Suppose you make a chart plotting *Resids* against *Cases*. Which of the following **must** be true about this chart? Select all that are correct:

- ☐ The correlation between the *Resids* and *Cases* is exactly 0
- ☐ The correlation between the *Resids* and *Cases* is exactly 0.6
- ☐ There is a positive association between *Resids* and *Cases*
- ☐ There are an equal number positive residuals as there are negative residuals
- ☐ The sum of the positive residuals equals the absolute value of the sum of the negative residuals

(g) (2 points)

Peter, excited by the positive slope of the regression line between *Sales* and *Cases*, believes he's made a groundbreaking discovery.

However, his daughter, Dolores, wonders if the association they observed in the sample was just due to chance.

i. Provide a null hypothesis Dolores could use to assess her claims.

ii. Provide an alternative hypothesis Dolores could use to assess her claims.

(h) (8 pt) Dolores decides to construct a PERCENT_CONF% confidence interval for the true slope of the regression line between *Sales* and *Cases* by bootstrapping the regression line 10,000 times.

To assist with this, she first creates a function called `get_slope`, which takes in a table and two column names as input and returns the slope of the least squares regression line.

She then writes the following partially completed code to create the confidence interval for the slope (as an array):

```
slopes = make_array()
for i in _______:
    BOOT_TBLNAME = masks._______(_______)
    boot_slope = get_slope(BOOT_TBLNAME, 'Sales', 'Cases')
    slopes = _______(_______, _______)
left = percentile(_______, _______)
right = percentile(_______, _______)
_______(left, right)
```

Copy/paste the code above and fill in the blanks.

(i) (2 pt) Suppose that the value of `left` in part (8) is `0.03`. Which of the following are conclusions that Dolores can make?

☐ The data are consistent with the alternative hypothesis.

☐ The data are consistent with the null hypothesis.

☐ The data suggest that the association observed between *Sales* and *Cases* is due to chance alone.

☐ The data suggest that the association observed between *Sales* and *Cases* is *not* due to chance alone.

☐ The data suggest that *Sales* and *Cases* have no association.

6. (15 points) Boba

Every day, Angela gets a boba drink from Asha in Berkeley. She believes the machine that's used to add the boba pearls to the drink is calibrated to put an exact amount of boba in each drink, with some variability due to chance.

To get a sense of how the amount of boba in a drink varies, Angela plans to randomly sample customers throughout November and record the weight (in grams) of the boba pearls in each customer's drink.

(a) (2 pt) Suppose that Angela wants to use a sample of size 100 to create a confidence interval for the true **population mean** of boba weight per drink.

Which of the following could be used to help her create this confidence interval? Select all that are correct:

☐ Central Limit Theorem

☐ Bootstrapping

☐ Nearest Neighbors

☐ Linear Regression

☐ Classification

(b) (2 pt) Suppose that Angela wants to use a sample of size 100 to create a confidence interval for the true **population median** of boba weight per drink.

Which of the following techniques could be applied to help her create this confidence interval? Select all that are correct:

☐ Central Limit Theorem

☐ Bootstrapping

☐ Nearest Neighbors

☐ Linear Regression

☐ Classification

(c) (2 pt) Suppose that Angela wants to use her sample to create a **95%** confidence interval for the true **population mean** of boba weight per drink and she knows that the population SD is 2 grams.

What is the minimum sample size she needs to create a confidence interval that has a width of 0.4 grams?

◯ 1600

◯ 400

◯ 200

◯ 100

◯ 25

◯ Not enough information to tell

(d) **(2 pt)** Suppose that Angela wants to use her sample to create a **68%** confidence interval for the true **population median** of boba weight per drink and she knows that the population SD is 2 grams.

What is the minimum sample size she needs to create a confidence interval that has a width of 0.4 grams?

- ◯ 1600
- ◯ 400
- ◯ 200
- ◯ 100
- ◯ 25
- ◯ Not enough information to tell

(e) **(2 pt)** Suppose that Angela observes an average of 30 grams of boba per drink from a random sample of 100 customers and she knows that the population SD is 2 grams.

What is her **95%** confidence interval for the true **population mean** of boba weight?

- ◯ (29.8, 30.2)
- ◯ (29.6, 30.4)
- ◯ (28, 32)
- ◯ (26, 34)
- ◯ (20, 40)
- ◯ Not enough information to tell

(f) **(2 pt)** Suppose that Angela observes an average of 30 grams of boba per drink from a random sample of 100 customers and she knows that the population SD is 2 grams.

Which of the following is **guaranteed** to be true? Select all that are correct:

- ☐ At least 68% percent of the customers in the population will have a boba weight that is between 2 grams below and 2 grams above the population mean.
- ☐ At least 75% percent of the customers in the population will have a boba weight that is between 4 grams below and 4 grams above the population mean.
- ☐ At least 75% percent of the customers in the population will have a boba weight that is between 26 grams and 34 grams.
- ☐ At least 68% percent of the customers in Angela's sample have a boba weight that is between 28 grams and 32 grams.

(g) (3 pt) Suppose that Angela samples 500 customers and her **95%** confidence interval for the true **population mean** is (29, 31).

Which of the following can be concluded from this confidence interval?

- ☐ If Angela repeats this process 1000 times, she can expect that roughly 95% of the intervals she creates will contain the true population mean.
- ☐ If Angela gets boba once a day throughout November, she can expect to get between 29 and 31 grams of boba on roughly 95% of the days.
- ☐ If Angela gets boba once a day throughout the year, she can expect to get between 29 and 31 grams of boba on roughly 95% of the days.
- ☐ If you sample 100 Asha boba customers in November, you can expect roughly 95% of them to get between 29 and 31 grams of boba.

7. (21 points) Soup Dumplings

In late 2013, food writer Christopher St. Cavish visited 52 soup dumpling restaurants in various neighborhoods of Shanghai, China to quantitatively determine the neighborhood with the best soup dumpling.

At each restaurant, Christopher randomly sampled 8 dumplings and measured the following about each dumpling:

- *Restaurant*: (string) the restaurant's name
- *Soup*: (float) the amount of soup, in grams
- *Meat*: (float) the amount of meat, in grams
- *Skin*: (float) the thickness of the dumpling skin, in millimeters
- *Type*: (string) the type of meat used in the dumpling
- *LOCATIONCOL*: (string) the neighborhood where the restaurant is located

He put this data into a `dumpling` table. Here are the first four rows:

Restaurant	Skin	Meat	Soup	Type	LOCATIONCOL
JiaJia TangBao	1.15	8.60	5.78	Pork	Puxi
NaxXiang ManTou Dian	1.40	11.68	5.02	Crab	Puxi
Paradise Dynasty	1.12	12.65	4.83	Pork	Pudong
Din Tai Fung	1.04	9.39	5.02	Shrimp	Pudong

... (412 rows omitted)

Chirstopher also created a "Soup Dumpling Index" to objectively measure the quality of a soup dumpling. Here is the formula:

$$\text{Soup Dumpling Index} = \frac{\text{Meat Weight} + \text{Soup Weight}}{\text{Skin Thickness}}$$

(a) (2 pt) Write a Python expression that returns the `dumpling` table with a new column named *Soup Dumpling Index* that contains the soup dumpling index for each dumpling in the table.

(b) (3 points)

Your friend Meghan, a Puxi native, believes that the soup dumplings in Puxi are superior to those of Pudong. She cites that the average Puxi dumpling has a **higher** soup dumpling index by 2.1 than the average Pudong dumpling.

However, you believe that this difference is just due to chance.

Provide a null hypothesis, alternative hypothesis and test statistic that Meghan could use to assess her claims.

i. (1 pt) Select the null hypothesis that Meghan should use to assess her claims.

- ◯ On average, dumplings made in Puxi have a higher *Soup Dumpling Index* by 2.1 than dumplings from Pudong.
- ◯ Dumplings made in Puxi have the same *Soup Dumpling Index* distribution as those made in Pudong.
- ◯ On average, dumplings made in Puxi have a higher *Soup Dumpling Index* than those made in Pudong, due to chance.
- ◯ On average, dumplings made in Puxi have a higher *Soup Dumpling Index* than those made in Pudong.
- ◯ Dumplings made in Puxi have a different *Soup Dumpling Index* distribution than those made in Pudong.

ii. (1 pt) Select the alternative hypothesis that Meghan should use to assess her claims.

- ◯ On average, dumplings made in Puxi have a higher *Soup Dumpling Index* by 2.1 than dumplings from Pudong.
- ◯ Dumplings made in Puxi have the same *Soup Dumpling Index* distribution as those made in Pudong.
- ◯ On average, dumplings made in Puxi have a higher *Soup Dumpling Index* than those made in Pudong, due to chance.
- ◯ On average, dumplings made in Puxi have a higher *Soup Dumpling Index* than those made in Pudong.
- ◯ Dumplings made in Puxi have a different *Soup Dumpling Index* distribution than those made in Pudong, due to chance.

iii. (1 pt) Select all of the valid test statistic(s) that Meghan could use to assess her claims.

- ☐ Mean Soup Dumpling Index for Puxi dumplings minus mean *Soup Dumpling Index* for Pudong dumplings.
- ☐ Mean *Soup Dumpling Index* for Puxi dumplings minus 2.1.
- ☐ Absolute difference of mean *Soup Dumpling Index* for Puxi dumplings minus 2.1.
- ☐ Absolute difference of mean *Soup Dumpling Index* for Puxi dumplings minus mean *Soup Dumpling Index* for Pudong dumplings.
- ☐ 2.1
- ☐ Mean *Soup Dumpling Index* for all dumplings

(c) (2 pt) Meghan decides to bootstrap her sample to determine whether the observed difference of 2.1 is just due to chance. Suppose she stores the differences from each bootstrap sample in an array called `boot_diffs` and runs the following code:

```
make_array(percentile(2.5, boot_diffs), percentile(97.5, boot_diffs))
```

which returns `(-2.5, 2.7)`.

If her p-value cutoff is 5%, which of the following can she conclude about the hypothesis test in part (2)? Select all that are correct:

☐ The data are consistent with the null hypothesis.

☐ The data are consistent with the alternative hypothesis.

☐ The data are consistent with neither hypothesis.

☐ The null hypothesis is true.

☐ The null hypothesis is false.

(d) (10 points)

Suppose you learn that online customer reviews seem to be higher for restaurants in Pudong than in Puxi.

Meghan suspects that this might actually just be due to the fact that the restaurants in Pudong are more likely to adopt the latest experimental trends of adding crab or shrimp to their soup dumplings (Meghan, a Pork purist, is not a fan).

To test this claim, Meghan wants to compare whether Puxi and Pudong have the same distribution of soup dumpling *Type*.

i. (1 pt) Select the null hypothesis that Meghan should use to assess her claims.

◯ Dumplings made in Puxi have the same *Type* distribution as those made in Pudong.

◯ On average, dumplings made in Puxi have a lower *Type* than those made in Pudong.

◯ On average, dumplings made in Puxi have a higher *Type* than those made in Pudong.

◯ Dumplings made in Puxi have a different *Type* distribution than those made in Pudong.

◯ Dumplings made in Puxi have a different *Soup Dumpling Index* distribution than those made in Pudong.

ii. (1 pt) Select the alternative hypothesis that Meghan should use to assess her claims.

◯ Dumplings made in Puxi have the same *Type* distribution as those made in Pudong.

◯ On average, dumplings made in Puxi have a lower *Type* than those made in Pudong.

◯ On average, dumplings made in Puxi have a higher *Type* than those made in Pudong.

◯ Dumplings made in Puxi have the a different *Type* distribution than those made in Pudong.

◯ Dumplings made in Puxi have a different *Soup Dumpling Index* distribution than those made in Pudong.

iii. (8 pt) Meghan decides to use the total variation distance (TVD) between the *Type* distributions of Puxi and Pudong dumplings as her test statistic.

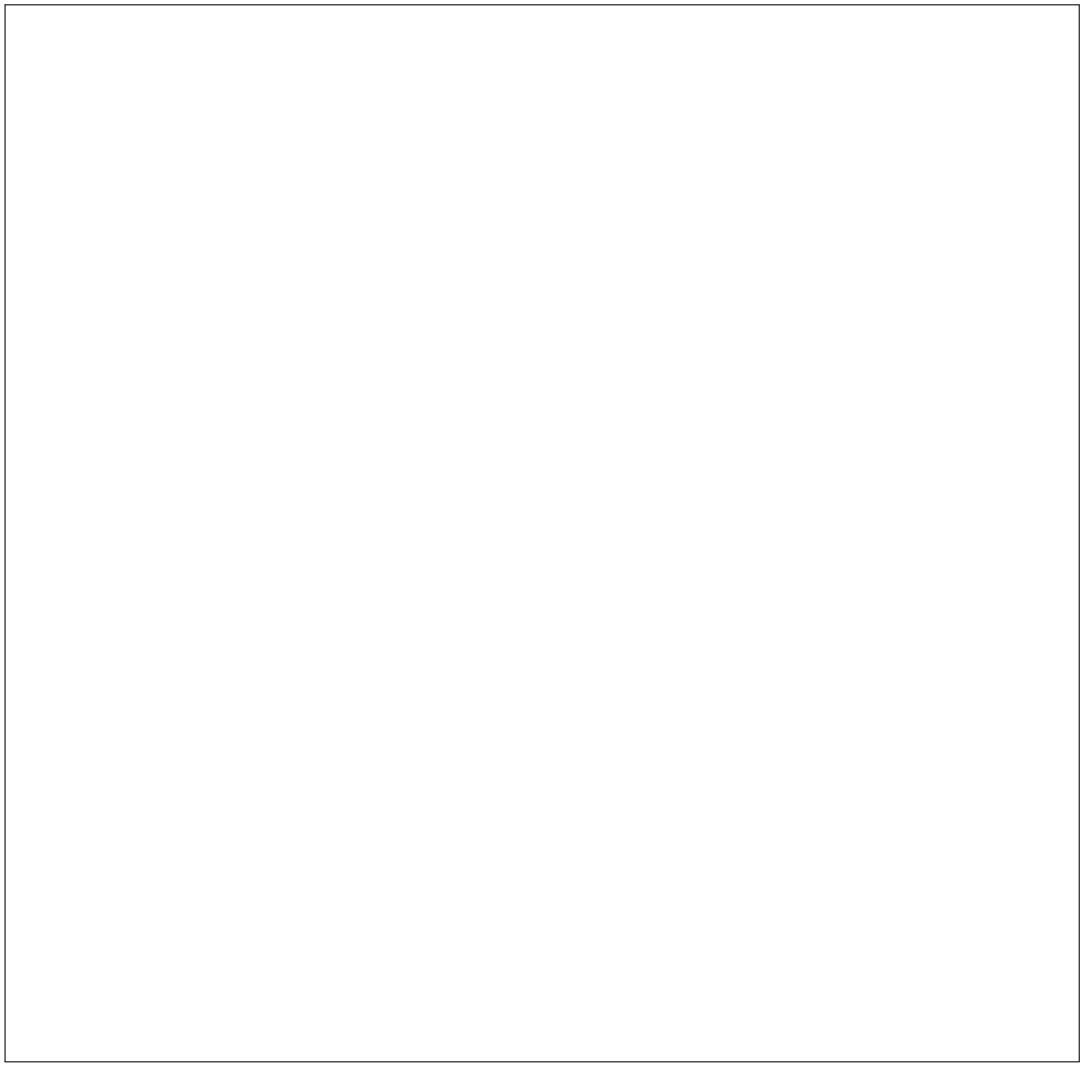
She writes the following partially completed code to help her simulate one value of the test statistic under the null hypothesis:

```
def label_proportions(table, neighborhood, label):
  label_dist = table.where('LOCATIONCOL Shuffled',_________)._________
  label_counts = label_dist.column(1)
  return label_counts / _________

def one_test_stat():
  shuffled_LOCATIONCOLs = dumpling._________._________
  shuffled_table = dumpling.with_column('LOCATIONCOL Shuffled', shuffled_LOCATIONCOLs)

  props_puxi = label_proportions(_________,_________,_________)
  props_pudong = label_proportions(_________,_________,_________)

  return _________ * np.sum(_________)
```

Copy/paste the code above and fill in the blanks.

Recall: The `dumpling` table has the following columns:

- *Restaurant*: (string) the restaurant's name
- *Soup*: (float) the amount of soup, in grams
- *Meat*: (float) the amount of meat, in grams
- *Skin*: (float) the thickness of the dumpling skin, in millimeters
- *Type*: (string) the type of meat used in the dumpling
- *LOCATIONCOL*: (string) the neighborhood where the restaurant is located



(e) (2 pt) Meghan uses the function above to simulate 1,000 values of the test statistic and stores these in an array called `test_stats`. Suppose the observed TVD is 0.34.

Select the Python expression(s) that return the p-value for this hypothesis test.

- ☐ np.count_nonzero(test_stats >= 0.34)/len(test_stats)
- ☐ np.count_nonzero(test_stats <= 0.34)/len(test_stats)
- ☐ np.count_nonzero(test_stats == 0.34)/len(test_stats)
- ☐ np.count_nonzero(test_stats >= 0.34)
- ☐ np.count_nonzero(test_stats <= 0.34)
- ☐ np.count_nonzero(test_stats == 0.34)

(f) (2 pt) Meghan then takes the `test_stats` array above and runs the following code:

```
percentile(75, test_stats)
```

which returns a value of `0.45`.

If her p-value cutoff is 5% and the observed TVD is 0.34, which of the following can she conclude? Select all that are correct:

- ☐ The data are consistent with the null hypothesis.
- ☐ The data are consistent with the alternative hypothesis.
- ☐ The data are consistent with neither hypothesis.
- ☐ The null hypothesis is true.
- ☐ The null hypothesis is false.

8. (15 points) Assistant

Dwight is trying to hire a new Assistant (to the) Regional Manager at Dunder Mifflin, a paper company with multiple branch locations in the Northeastern United States. He hopes to use machine learning to help him speed up the hiring process.

Dwight starts by collecting a sample of past applications, recording various information about each applicant and storing it in a table called `results`. Here are the first few rows:

Name	Experience	Performance	Branch	Hired
Yanay	1	8	Scranton	0
Ruhi	4	7	Utica	1
Emily	0	9	Scranton	0
Grace	5	10	Stamford	1

... (446 rows omitted)

The table contains the following columns:

- *Name*: (string) the applicant's name
- *Experience*: (int) the number of years the applicant has worked in the paper industry
- *Performance*: (int) the applicant's most recent performance score (out of 10), as rated by their manager
- *Branch*: (string) the branch where the applicant worked at the time of applying
- *Hire*: (int) whether the applicant was hired as an assistant when they applied (1 means 'Hired'; 0 means 'Not Hired')

Name	Recommender

(a) (2 pt) Dwight first tries to see if he can hire applicants based on just years of experience alone, so he makes the following scatterplot from his sample data.

Note: A small amount of random noise has been added so you can see all overlapping points.

Suppose Dwight wants to write a `predict` function that takes in an integer `years` as input and returns `True` if the applicant should be hired or `False` if they should not. He writes the following partially completed code:

```
def predict(years):
    return _____________
```

If you had to fill in the blank above, which of the following lines of code would result in a prediction function that has 100% accuracy on the sample data? Select all that are correct:

☐ `years >= 3.5`

☐ `years >= 3.1`

☐ `years <= 3.5`

☐ `years > 3.5`

☐ `years > 4`

(b) (12 points)

Suppose Dwight also discovers historical records of recommendation letters, stored in a table called `recs`. Here are the first 4 rows:

Name	Recommender
Yanay	Swupnil
Tam	David
Ananya	Swupnil
Parham	None

... (1024 rows omitted)

The table contains the following information, with exactly one recommender per applicant:

- *Name*: the name of the applicant who applied
- *Recommender*: the name of the applicant's recommender

Dwight wants to use the *Recommender* column as a feature for his hiring classifier, but he first needs to manipulate the data to make this feature numerical.

To accomplish this, Dwight writes a function called `encode` which takes in a table (with 2 columns, *Name* and *Recommender*) as input and replaces the *Recommender* column with multiple columns, one for every possible value that the *Recommender* can take. The value in each column for a given applicant is equal to 1 if the applicant was recommended by that recommender, and equals 0 otherwise.

For example, the Python expression `encode(recs.take(np.arange(4)))` returns the following table:

Name	Swupnil	David	None
Yanay	1	0	0
Tam	0	1	0
Ananya	1	0	0

Name	Swupnil	David	None
Parham	0	0	1

i. (3 pt) Suppose the partially completed `encode` function looks like this:

```
def encode(table):
    return table.___________
```

Write a Python expression to fill in the blank.

Hint: The function should be able to return the correct result even if the inputted table has 100 different recommenders!

ii. (2 pt) Dwight takes the numerical encoding table and makes the following scatterplot of the *Swupnil* and *None* features, coloring the points based on whether the applicant was hired.

Note: A small amount of random noise has been added so you can see overlapping points.

Does it seem like a recommendation from Swupnil helps applicants get hired?

◯ True

◯ False

iii. (7 pt) Dwight is now ready to prepare all of the features for his classifier. At his disposal, he has the following:

- `recs` table with columns `['Name', 'Recommender']`
- `results` table with columns `['Name', 'Experience', 'Performance','Branch','Hired']`
- `encode` function from above

Dwight wants to use these to create a single table called `all_data` that has one row per applicant, along with the following columns (in this order):

- a *Name* column containing the applicants' names
- columns for all of the features in `results` besides *Hired* and *Branch*
- columns for all of the numerically encoded *Recommender* features
- an *Outcome* column containing whether each applicant in the `results` table was hired (`True` if hired, `False` if not hired)

He writes the following code to create the `all_data` table:

```
all_data = results.__________.__________
all_data = all_data.__________
all_data = all_data.with_column(__________)
```

Copy/paste the code above and fill in the blanks.

(c) (1 pt) Do you think Dwight should use machine learning to hire applicants? What are some ethical or privacy concerns?

Please limit your answer to 1 sentence. We will not read anything beyond that.

9. (0 points) Just for Fun :)

(a) Prof. Sahai hasn't seen a single episode of one of the following shows. Which is it?

- ◯ Silicon Valley
- ◯ The Office
- ◯ Westworld
- ◯ Breaking Bad
- ◯ Community

10. (0 points) Last Words

(a) If there was any question on the exam that you thought was ambiguous and required clarification to be answerable, please identify the question (including the title of the section, e.g., Experiments) and state your assumptions. Be warned: We only plan to consider this information if we agree that the question was erroneous or ambiguous and we consider your assumption reasonable.

No more questions.