Data C8
Spring 2025

Foundations of Data Science Final

7:10-10:00PM, TUESDAY, MAY 13

Berkeley Honor Code [1 point]

"As a member of the UC Berkeley community, I act with honesty, integrity, and respect for others."

Initials: _____

Full Name: _____

Student ID Number: _____

Exam Location: _____

Name of person to your left: _____

Name of person to your right: _____

GSI/TA's Name (Write N/A if in self-service lab): _____

INSTRUCTIONS

- You may only have with you: a pencil(s), an eraser(s), your student ID, a water bottle, and your final reference sheet, unless you have received pre-approved accommodations otherwise.
- If you need to use the restroom, bring your phone, exam, reference sheet, and student ID to the front of the room.
- Do not open the exam until you are instructed to do so.
- Write your initials at the top of each page.
- There are **6** questions and **18** pages on this exam, including cover page. **Read the instructions and point values carefully** for each question, part and subpart.
- Multiple choice questions with bubbles ◯ have one correct answer. Multiple choice questions with squares ☐ have one or more correct answers.
- Where relevant, you may assume that all necessary Python modules have been imported. Use of any code which has not been taught in this iteration of the course is prohibited and it will not be graded.
- Where a written (English) answer is expected, you must use complete sentences. Your work will not be graded otherwise.
- **Each coding blank may include multiple arguments/methods/functions.** However, your solution must use every blank available.

1 General [21 Points]

Read each question carefully and answer as instructed.

a. (5 points) Public distrust of vaccines is often traced back to a now infamous paper published in 1998 by disgraced doctor Andrew Wakefield. The study claimed to have found a causal link between the measles, mumps and rubella (MMR) vaccine and a new type of disease indicative of autism. However, it was retracted in 2010 due to many concerns about its scientific integrity. For instance, Wakefield only studied children who had received the MMR vaccine; he did not observe the outcomes of children who were un-vaccinated. Furthermore, Wakefield carefully chose the children that participated. What obstacles to establishing causality between vaccines and autism did Wakefield introduce into his study? *Select all that apply.*

- ☐ The presence of confounding factors
- ☐ The lack of treatment and control groups
- ☐ The lack of random assignment

b. (4 points) Following two fatal aviation accidents in late January, incidents involving aircraft in the United States have received heightened media coverage. However, according to the National Transportation Safety Board, the U.S. government agency responsible for tracking aviation accidents in the country, there have been 252 accidents from the beginning of 2025 through April 25th; during the same period last year, 291 accidents had been recorded. *Fill in the blank:* This is an example of using the _____ facet of data science.

◯ Exploration ◯ Prediction ◯ Inference

c. (3 points) What will be output to the screen once the following Python code runs?

```
3 + make_array(1,4,9) / np.arange(1,4)
```

◯ `array([4, 5, 6])`
◯ `array([4.0, 5.0, 6.0])`
◯ `array([4.0, 3.5, 4.0])`
◯ `array([4, 4, 4])`
◯ The code will produce an error.
◯ The correct output is not listed here.

d. (3 points) The U.S. Department of Agriculture estimated the average retail prices of 93 fresh, frozen, canned and dried vegetables and found the mean and standard deviation of these prices to be 2 dollars and 1 dollars, respectively. *Fill in the blank with the largest number between 0 and 100 that satisfies the statement*: We can expect at least _____ percent of the prices to be within 0 and 4 dollars.

e. (2 points) In which of the following case studies across the course materials have we performed an A/B test? *Select all that apply.*

☐ Homework - Jade's face card game
☐ Lecture/text - smoking and birth weights
☐ Lab - vaccines with DeNero and Sahai
☐ Lab - The Great British Bakeoff
☐ Lecture/text - Mendel's pea plants
☐ Homework - Gender identity and age

f. (2 points) Cyrus performs a hypothesis test and obtains a p-value of 0.01. Which of the following statements about his p-value are true? *Select all that apply.*

☐ Given that Cyrus observed the test statistic that he did, the probability of the null hypothesis being true is equal to 0.01.
☐ If Cyrus had chosen a p-value cutoff of 0.05 for his hypothesis test, he would reject the null hypothesis based on this p-value.
☐ If the null hypothesis is true, the probability of Cyrus observing the test statistic that he did, or a test statistic more extreme, is equal to 0.01.

g. (1 point) Mariel plots a histogram of TV show ratings (on a scale of 0 to 10 stars) and notices that the distribution resembles a normal curve centered at $x = 8.9$ stars, with points of inflection at approximately $x = 8.7$ stars and $x = 9.1$ stars. Not wanting to waste time on a poor quality show, Mariel calls a show "good" if it has a rating of 8.7 stars or higher. What proportion of Mariel's TV shows are "good"? *Write a number in the box between 0 and 1.*

h. (1 point) According to the Data 8 mid-semester feedback form, $\frac{2}{3}$ of students enrolled in regular lab this semester. Out of these students, $\frac{1}{3}$ ended up visiting office hours. Out of self-service students, $\frac{1}{9}$ ended up visiting office hours. Sophie wants to take Data 8 in Fall 2025 and is like a student drawn at random from the Spring 2025 semester. Given that Sophie ends up visiting office hours during the semester, which lab format is it more likely that she enrolled in?

◯ Regular lab
◯ Self-service lab
◯ Both regular and self-service are equally likely.
◯ Another lab format

2 Battle Bus College [22 Points]

The Battle Bus College at UC Berkeley has a graduating class of 10,000 seniors in the year 2025. The `graduates` table contains 1,000 graduating seniors across the UC Berkeley campus, along with their graduating GPAs.

Name	GPA
Riyya	3.94
Ella	3.74
Cai	3.83
Sam	3.97

...(996 rows omitted)

a. (8 points) Thomas and Ramisha are two of the graduating seniors in the Battle Bus College and have secretly obtained access to the `graduates` table. After viewing the data, Thomas believes that the average GPA of all graduating Battle Bus College seniors is 3.8, but Ramisha disagrees. They decide to perform a hypothesis test to test their competing beliefs.

(i) (3 points) Select an appropriate **null hypothesis** for this scenario.

- ◯ The average GPA of the seniors in the `graduates` table is 3.8.
- ◯ The average GPA of all graduating Battle Bus College seniors is 3.8.
- ◯ The average GPA of all graduating Battle Bus College seniors is different than 3.8.
- ◯ The average GPA of the seniors in the `graduates` table is different than 3.8.

(ii) (2 points) Select an appropriate **alternative hypothesis** for this scenario.

- ◯ The average GPA of the seniors in the `graduates` table is 3.8.
- ◯ The average GPA of all graduating Battle Bus College seniors is 3.8.
- ◯ The average GPA of all graduating Battle Bus College seniors is different than 3.8.
- ◯ The average GPA of the seniors in the `graduates` table is different than 3.8.

(iii) (3 points) In order to use the `graduates` table to conduct this hypothesis test, what must be true about the 1,000 graduating UC Berkeley seniors in the table? What also must be true about how the seniors in the table were chosen?

b. (5 points) Once Ramisha and Thomas confirm the data in the `graduates` table is appropriate for the hypothesis test, they set a p-value cutoff of 0.10 and compute a 90 percent confidence interval for the average GPA of all graduating Battle Bus College seniors using the bootstrap percentile interval method. Help them compute the confidence interval by specifying the process they use below.

(i) (1 point) Ramisha and Thomas obtain their bootstrap samples from:

- ◯ The 1,000 graduating seniors in the `graduates` table
- ◯ All 10,000 graduating Battle Bus College seniors
- ◯ All graduating UC Berkeley seniors

(ii) (2 points) Ramisha and Thomas take their bootstrap samples:

◯ With replacement ◯ Without replacement

(iii) (2 points) The size of each of their bootstrap samples is:

◯ 10,000 ◯ 2,000 ◯ 1,000 ◯ Any of these sizes are appropriate.

c. (4 points) Ramisha and Thomas store the simulated averages in an array called `grad_stats`. They then use the `percentile` function to find the lower bound and upper bound of the 90 percent confidence interval. *Select the two percentiles of `grad_stats` that Thomas and Ramisha will need to access, one of them corresponding to the lower bound and the other corresponding to the upper bound.*

- ☐ 0
- ☐ 2.5
- ☐ 5
- ☐ 10
- ☐ 90
- ☐ 95
- ☐ 97.5
- ☐ 100

d. (2 points) Ramisha and Thomas obtain a confidence interval of [3.85, 3.89]. Interpret this interval in the context of the problem.

- ◯ We are 90 percent confident that the average GPA of graduating seniors in the `graduates` table is between 3.85 and 3.89.
- ◯ There is a 90 percent chance that the average GPA of all graduating Battle Bus College seniors is between 3.85 and 3.89.
- ◯ We are 90 percent confident that the average GPA of all graduating Battle Bus College seniors is between 3.85 and 3.89.
- ◯ There is a 90 percent chance that the average GPA of graduating seniors in the `graduates` table is between 3.85 and 3.89.

e. (3 points) Based on the interval, which hypothesis will Ramisha and Thomas support?

◯ Null hypothesis ◯ Alternative hypothesis

3 Don't rain on my Glade! [18 Points]

Marissa loves visiting the campus Glade at 4pm to relax after a long day of class, but wants to make sure the weather is nice ahead of time. She has access to a table called `weather` which contains historical information on the weather in Berkeley, California at 4pm for the past 10,000 days. A three-row sample from the table lies below.

- **Precip** (integer): precipitation; the percent chance of receiving at least 0.01 inches of rain.
- **Humid** (integer): relative humidity; the percentage of possible water vapor that can exist in the air at the air's current temperature.
- **Cond** (string): weather conditions; the state of the weather (one of Rainy, Cloudy or Sunny).

Precip	Humid	Cond
82	73	Rainy
71	96	Cloudy
16	94	Sunny

...(9997 rows omitted)

a. (5 points) Marissa would like to use the `weather` table to build a predictive model for the weather conditions of an upcoming Berkeley afternoon (Rainy, Cloudy or Sunny). What type of prediction problem is Marissa working on?

◯ Regression ◯ Classification

b. (2.5 points) Complete the following code so that the `weather` table is split into two tables: one called `training_set`, and one called `testing_set`. The first eighty percent of the data in `shuffled_table` should be allocated to `training_set`.

```
shuffled_table = weather.sample(_______A_______)

training_set = shuffled_table._______B_______(_______C_______)

testing_set = shuffled_table._______D_______(_______E_______)
```

(i) (1 point) Blank `A`:

◯ `k = 10000, with_replacement = False` ◯ `k = 8000, with_replacement = True`
◯ `k = 10000, with_replacement = True` ◯ `k = 2000, with_replacement = False`
◯ `k = 8000, with_replacement = False` ◯ `k = 2000, with_replacement = True`

(ii) (0.5 points) Blank `B`:

(iii) (0.5 points) Blank `C`:

(iv) (0.25 points) Blank `D`:

(v) (0.25 points) Blank `E`:

c. (3 points) Marissa decides to use the k-nearest neighbors prediction algorithm. Which of the following formulae correctly computes a distance between a new observation (new) and a training point (train)?

○ $\sqrt{(\text{Precip}_{\text{new}} - \text{Precip}_{\text{train}})^2 + (\text{Humid}_{\text{new}} - \text{Humid}_{\text{train}})^2 + (\text{Cond}_{\text{new}} - \text{Cond}_{\text{train}})^2}$

○ $\sqrt{(\text{Precip}_{\text{new}} - \text{Precip}_{\text{train}}) + (\text{Humid}_{\text{new}} - \text{Humid}_{\text{train}}) + (\text{Cond}_{\text{new}} - \text{Cond}_{\text{train}})}$

○ $\sqrt{(\text{Precip}_{\text{new}} - \text{Precip}_{\text{train}})^2 + (\text{Humid}_{\text{new}} - \text{Humid}_{\text{train}})^2}$

○ $\sqrt{(\text{Precip}_{\text{new}} - \text{Humid}_{\text{new}})^2 + (\text{Precip}_{\text{train}} - \text{Humid}_{\text{train}})^2}$

○ $\sqrt{(\text{Precip}_{\text{new}} - \text{Precip}_{\text{train}}) + (\text{Humid}_{\text{new}} - \text{Humid}_{\text{train}})}$

○ $\sqrt{(\text{Precip}_{\text{new}} - \text{Humid}_{\text{new}}) + (\text{Precip}_{\text{train}} - \text{Humid}_{\text{train}})}$

d. (2 points) Before finding the k-nearest neighbors to a new observation, Marissa will first need to compute the distance between the new observation and every point in the training set. Which of the following table methods or coding techniques can be used to compute these distances? *Select all that apply.*

☐ `group`
☐ `join`
☐ Conditional statements
☐ Iteration
☐ `apply`
☐ `pivot`

e. (4 points) Marissa computes the distance between her new observation and each training point, and then creates a new version of the `training_set` table that includes these distances as an additional column called "Distance". Below is a three row excerpt from the updated table.

Precip	Humid	Cond	Distance
82	73	Rainy	5
71	96	Cloudy	2
16	94	Sunny	12

...(7997 rows omitted)

Fill in the skeleton code below to classify the new observation using the k-nearest neighbors prediction algorithm where $k = 7$.

```
nearest_nbrs = training_set._______A_______(_______B_______)._______C_______(_______D_______)

neighbor_votes = nearest_nbrs._______E_______(_______F_______).sort(_______G_______)

majority_class = neighbor_votes.column("Cond").item(0)
```

(i) (1 point) Blank `A`:

(ii) (1 point) Blank `B`:

(iii) (0.5 points) Blank `C`:

(iv) (0.5 points) Blank `D`:

(v) (0.5 points) Blank `E`:

(vi) (0.25 points) Blank `F`:

(vii) (0.25 points) Blank `G`:

f. (1.5 points) Marissa's friend, Ishani, wonders if $k = 7$ offers the best predictive performance, and wants to try out a few different values of k before finalizing the k-nearest neighbors prediction model. Ishani's plan is to try out different values of k and choose the one which performs best on the testing set. Does this method make appropriate use of the testing set?

◯ Yes ◯ No

4 Have a ball while taking this exam! [19 points]

The `nba_players` table consists of statistics on each basketball player who participated in the 2024-25 NBA regular season. Below lies a three-row excerpt of the table.

- **Player** (string): The player's name.
- **Steals** (integer): The (average) number of times during the game that the player takes the ball away from a player on the opposing team, rounded to the nearest integer.
- **Assists** (integer): The (average) number of times during a game that the player passed the ball to a second player who scored directly after, rounded to the nearest integer.
- **Minutes** (integer): The (average) number of minutes the player spent on the court per game, rounded to the nearest integer.
- **Points** (integer): The (average) number of points the player scored per game, rounded to the nearest integer.

Player	**Steals**	**Assists**	**Minutes**	**Points**
Jimmy Butler	1	5	32	17
Zeke Nnaji	0	1	11	3
AJ Johnson	0	2	22	8

...(566 rows omitted)

Below are some statistics Reynaldi calculated using the Steals and Points variables.

Statistic	Value
Correlation between Steals and Points	0.64
Mean of Steals	1
SD of Steals	0.5
Mean of Points	9
SD of Points	7

a. (5 points) Select the slope of the least squares regression line which predicts **Points**, using **Steals** as a predictor.

◯ 0.64

◯ $0.64 * \frac{7}{0.5}$

◯ $0.64 * \frac{0.5}{7}$

◯ The correct answer is not listed here.

b. (3 points) Next, Reynaldi tries out k-nearest-neighbors regression with $k = 5$ to predict **Points**, using **Steals** as a predictor. The table below contains the points scored per game of the five players in the training set considered the nearest neighbors to Zach Edey, a player in the testing set. How many points per game will Reynaldi predict Edey to score?

Player	Points
Darius Garland	5
Bismack Biyombo	6
Haywood Highsmith	5
Pete Nance	10
DeAndre Jordan	9

◯ 5 ◯ 6 ◯ 7 ◯ 8 ◯ 9 ◯ 10

c. (3 points) Reynaldi's friend, Simone, then decides to use a multiple linear regression model to predict **Points**, using both **Steals** and **Assists** as predictor variables. Once simplified, the regression equation that Renata finds is roughly:

$$\text{estimate of } \textbf{Points} = (4 * \textbf{Steals}) + (2 * \textbf{Assists}) + 2$$

Interpret the coefficient belonging to the **Steals** variable in the context of the problem.

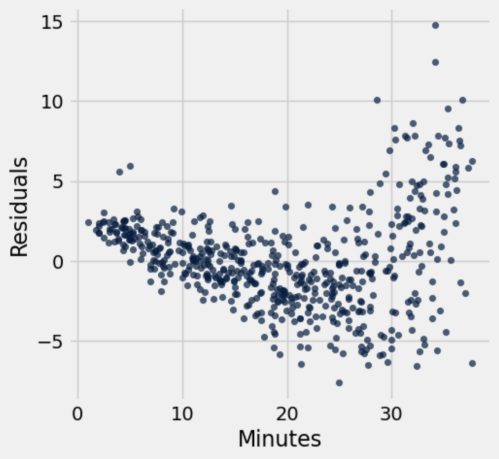
◯ For every steal per game, a player is expected to score an additional four points per game.

◯ Among players who contribute the same number of assists per game, an additional steal made means that a player is expected to score an additional four points per game.

◯ For every four steals per game, a player is expected to score one additional point per game.

◯ Among players who contribute the same number of assists per game, four additional steals mean that a player is expected to score one additional point per game.

d. (5 points) Reynaldi's other friend, Richard, attempts to fit a linear model to predict **Points** with **Minutes**, and produces the following scatter plot to evaluate the fit. Is the linear model suitable for this prediction problem?



◯ Yes ◯ No ◯ More information is needed to determine an answer.

e. (2 points) Richard decides to use a second model to predict **Points** with **Minutes** which has the formula given below.

$$\text{estimate of } \textbf{Points} = a * \textbf{Minutes}^2 + b * \textbf{Minutes} + c$$

To choose the values of a, b and c, Richard wants to find the values which minimize root mean squared error (RMSE) and starts work on the `rmse` function below. Complete the skeleton code below for Richard's function.

```
def rmse(a, b, c):
    x = nba_players.column('Minutes')
    y = nba_players.column('Points')

    estimate = ___________________A_____________________


    return _______________B___________________
```

(i) (1 point) Blank `A`:

(ii) (1 point) Blank `B`:

f. (1 point) Finally, Richard calls the `minimize` function on the `rmse` function he defined above. What is the length of the array that this call to the `minimize` function returns?

◯ 1 ◯ 3

◯ 2 ◯ 4

5 Read the questions on this exam very carefully! [18 points]

The `books` table contains physical measurements and other information on 42 best-selling books from the online platform Amazon as of May 2025. A three-row excerpt of the table lies below.

- **Material** (string): One of either "Paperback" or "Hardcover".
- **Style** (string): One of either "Fiction" or "Nonfiction".
- **Children** (integer): A numeric code; 1 if the book is part of the Children's Books section on the Amazon website, 0 otherwise.
- **Pages** (integer): How many pages long the book is.
- **Volume** (float): The volume of the book, measured in cubic inches.
- **Weight** (float): The weight of the book, measured in pounds.
- **Price** (float): The book manufacturer's suggested price of the book (before tax), measured in U.S. dollars.

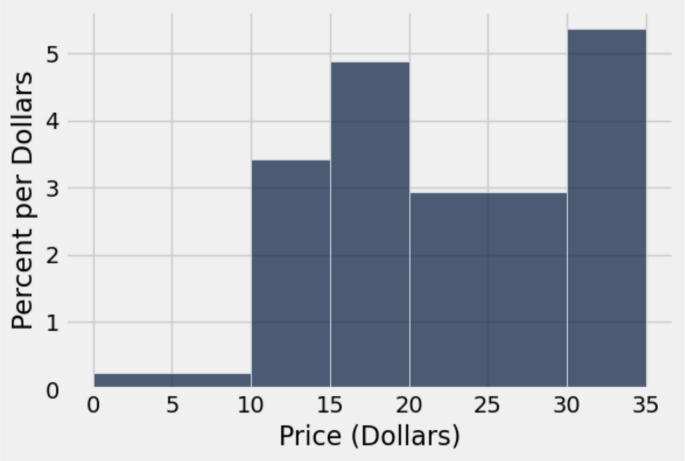
Material	Style	Children	Pages	Volume	Weight	Price
Paperback	Nonfiction	0	464	49.90	0.90	19
Hardcover	Nonfiction	0	336	60.13	1.25	29.99
Hardcover	Fiction	1	40	23.82	0.7	10.99

...(39 rows omitted)

a. (5 points) How many of the variables in the `books` table are categorical?

◯ 0 ◯ 1 ◯ 2 ◯ 3 ◯ 4 ◯ 5 ◯ 6 ◯ 7

b. (3 points) Andrew uses the `Table.hist()` method to create the following histogram, which displays the distribution of book prices. Select the statement below regarding the histogram that is true.



◯ Roughly 15% of books are priced between 20 and 25 dollars.

◯ There are more books priced between 15 and 20 dollars than there are books priced between 20 and 25 dollars.

◯ Over 50% of books are priced less than 20 dollars.

◯ None of the statements are true.

c. (5 points) Dagny is interested in visualizing the average price of a book on Amazon depending on its material construction. Complete the skeleton code below to create a visualization where the materials of the books are displayed on the vertical axis and the average prices of books made using each material are displayed on the horizontal axis using bars.

```
avg_price_plot = books.select(______A______).____B____(______C______).____D____(______E______)
```

(i) (1 point) Blank `A`:

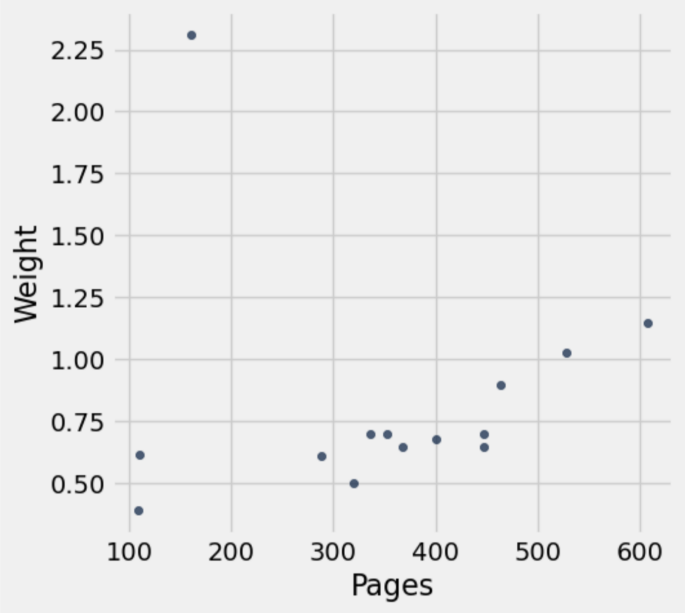
(ii) (1 point) Blank `B`:

(iii) (1 point) Blank `C`:

(iv) (1 point) Blank `D`:

(v) (1 point) Blank `E`:

d. (3 points) Dylan creates a scatter plot of the weight and pages of all paperback books in the `books` table. *The Four Agreements* by Don Miguel Ruiz is the book which is extremely heavy for the number of pages it has. Write **one line** of code in the box below which returns all of the information contained in the `books` table on *The Four Agreements* as a one-row table.




```
```

e. (2 points) Andy would like to find the difference in the average price of hardcover and paperback children's books, as well as the difference in the average price of hardcover and paperback books meant for adults. He decides to use a cross-classifiying table method to help him. Which of the cross-classifiying table methods that we have discussed is more efficient to complete this particular task?

◯ `group` ◯ `pivot`

6 Probability and simulation [18 points]

The gastroenterology unit at the John Bragado Outpatient Center in Berkeley performs **three colonoscopies a day**. However, not all of them are successful due to inadequate patient preparation or technical issues that might arise during a procedure. For this question, assume that each colonoscopy takes place independently of other colonoscopies during the day, each patient is like the next, and that **each procedure has a 0.10 probability of being unsuccessful**.

a. (2 points) Below is a partially completed table which contains the probability distribution of the number of unsuccessful colonoscopies that may occur during the day. **Fill in the empty cells** in the table below. Show your work inside the table. **Only work in the table will be graded**. You do not need to simplify.

Failed procedures per day	Bragado probabilities (you do not need to simplify)
0	
1	
2	$\frac{27}{1000}$
3	

b. (2 points) Below is a partially completed function called one `one_day_bragado` which simulates one day of colonoscopies at the John Bragado clinic and **returns the number of unsuccessful procedures performed.** Based on the information provided in the problem statement, complete the function.

```
def one_day_bragado():


    return _____A_____ * sample_proportions(_______B_______).item(0)
```

(i) (1 point) Blank `A`:

(ii) (1 point) Blank `B`:

c. (4 points) A second clinic, Diya Garg Outpatient Center, also performs three colonoscopies a day, each independently of the next, and with each patient like the one that came before. Below is the probability distribution for the number of failed procedures the Garg Center can perform during the day.

Failed procedures per day	Garg probabilities
0	$\frac{64}{125}$
1	$\frac{48}{125}$
2	$\frac{12}{125}$
3	$\frac{1}{125}$

Consider a day where both the Garg and Bragado centers are performing three colonoscopies each. Given the function `one_day_bragado` and a similar function called `one_day_garg`, which returns the number of unsuccessful procedures in a day of three colonoscopies at the Garg clinic, complete the function below called `successful_day`. This function returns `True` if there were no failures across both clinics, and `False` otherwise. *Hint: What type of data do the `one_day_bragado` and `one_day_garg` functions return? Can you combine these outputs?*

```
def successful_day():


    ______A______:

        ________B________


    ______C______:

        ________D________
```

(i) (0.5 points) Blank `A`:

(ii) (1 point) Blank `B`:

(iii) (1.5 points) Blank `C`:

(iv) (1 point) Blank `D`:

d. (5 points) The function `one_year` simulates one year (365 days) of colonoscopies at both the Bragado and Garg clinics and returns the **number of days having no failed procedures** across the entire, simulated year. Complete the function below. You may use functions that have previously been defined. You may also assume that the clinics' combined performance on a given day is independent from their combined performance on other days.

```
def one_year():
```

e. (3 points) Bing runs the `one_year` function 1,000 times in order to obtain 1,000 simulated **annual totals of successful days**, and then displays these totals on a histogram. What shape should we expect the distribution of totals to take? State the shape and the mathematical result discussed in this course that allows us to state the shape.

(i) (2 points) State the shape:

(ii) (1 point) State the result:

f. (2 points) Noah would like to perform a hypothesis test to determine whether the Garg probability distribution is the same as the Bragado probability distribution. Which of the following test statistics is appropriate for this test?

- ◯ Absolute difference in proportions
- ◯ Difference in proportions
- ◯ Total variation distance
- ◯ None of these test statistics are appropriate.

7 Congratulations! [0 points]

You have now completed the Final Exam. If you have not been told otherwise, you may bring all of your testing materials (reference sheet and this test paper), as well as your student ID, to the front of the room. Once you have been checked off, you may leave quietly.

- Please make sure that you have written your initials on each page of the exam. **You may lose points on pages where you have not done so.**
- Please make sure you have filled in bubbles and squares completely rather than having used a check mark, cross or any other mark.
- Double check that you have not skipped over any questions!

Below, you may draw and caption your favorite Data 8 experience or staff member!