Data C8, Midterm Exam

Summer 2023

Name: _____

Email: _____@berkeley.edu

Student ID: _____

Name of the student to your left: _____

Name of the student to your right: _____

Instructions:

Do not open the examination until instructed to do so.

This exam consists of **100 points** spread out over **6 questions** on **18 pages** and must be completed in the **110 minute** time period on July 14, 2023, from 10:10 AM to 12:00 PM unless you have pre-approved accommodations otherwise.

Note that some questions have circular bubbles to select a choice. This means that you should only **select one choice**. Other questions have boxes. This means you should **select all that apply**. Please shade in the box/circle to mark your answer.

There is space to write your student ID number (SID) in the upper right-hand corner of each page of the exam. **Make sure to write your SID on each page** to ensure that your exam is graded.

Honor Code [1 pt]:

As a member of the UC Berkeley community, I act with honesty, integrity, and respect for others. I am the person whose name is on the exam, and I completed this exam in accordance with the Honor Code.

Signature: _____

This page has been intentionally left blank.

1 What Would Python Do? [12 Points]

For each of the Python expressions below, write the output when the expression is evaluated. If the expression evaluates to an array, you should format your answer like so: `array([..., ..., ...])`. If an error occurs, write "Error".

```
from datascience import *
import numpy as np
```

(a) [2 Pts]

```
arr1 = make_array(1, 2, 3)
arr2 = make_array(4, 5, 6)

len(np.append(arr1, arr2)) != len(arr1 + arr2)
```

Solution: `True`

(b) [2 Pts] `sum((np.arange(131, 138) - 135) >= 0)`

Solution: `3`

(c) [2 Pts] `"o" + "skit".replace("t", "") + str(int(1941.561))`

Hint: `str.replace(old_string, new_string)` replaces each occurrence of `old_string` in `str` with the value of `new_string`.

Solution: `oski1941`

(d) [2 Pts] `make_array(8, 3, 4, 5) + np.arange(2, 11, 3)`

Solution: `Error`

(e) [4 Pts] Which of the following functions correctly returns the sum of the even numbers from 0 to n, inclusive of n? **Select all that apply.**

Hint: The expression a `%` b evaluates to the remainder when dividing a by b.

☐
```
def sum_cumulative_even(n):
    total = 0
    for i in np.arange(0, n+1, 1):
        if i % 2 == 0:
            total = total + i
    return total
```

☐
```
def sum_cumulative_even(n):
    return sum(np.arange(0, n+1, 2))
```

☐
```
def sum_cumulative_even_backwards(n):
    return sum(np.arange(n, 0, -2))
```

☐
```
def sum_cumulative_even_backwards(n):
    total = 0
    for i in np.arange(n, 0, -1):
        if i % 2 == 0:
            total = total + i
    return total
```

2 Dance, Dance, Data 8! [16 Points]

It's the annual Dance Dance Revolution (DDR) competition! In Dance Dance Revolution, participants take turns dancing to a song of their choice, with the goal of obtaining the highest score. There are two divisions in the competition: intermediate and advanced.

Data 8 Staff has nominated Kristen and Ethan to represent their team in this year's competition. Kristen, a novice competitor will enroll in the intermediate division, while her teammate Ethan will compete in the advanced division.

All competitors are provided with a table named `songs`. The first few rows are given below.

- `song` is the **string** title of the song
- `difficulty` is the **integer** difficulty rating of the song's dance, on a scale of $1 - 10$
- `bpm` is the **integer** number of beats per minute of the song
- `length` is the **integer** length of the song, in seconds

song	difficulty	bpm	length
Idola	8	190	150
Kiss Kiss Kiss	5	201	104
Clarity	2	128	110

In the following blanks, you can write as much code as is necessary that fits the skeleton code. Do not add any additional lines of code.

(a) [2 Pts] Dance Dance Revolution rules require that competitors in the intermediate division choose a song with a dance difficulty **between, and inclusive, of both** 5 **and** 7. Fill in the following line of code to help Kristen build a table of songs allowed in the intermediate division. This table should contain the same columns as the `songs` table.

```
intermediate_songs = songs._____(A)_____
```

Fill in the blank `(A)`

Solution: `where("difficulty", are.between_or_equal_to(5, 7))`

(b) [4 Pts] For her introductory dance, Kristen wants to choose a song with a low difficulty rating among the songs allowed in her division. Fill in the blanks to assign `kristen_songs` to an array of *any* of the 5 least difficult songs in `intermediate_songs`.

```
kristen_songs = intermediate_songs.___(A)___.___(B)___.__(C)__
```

Fill in the blank `(A)`

Solution: `sort("difficulty")`

Fill in the blank `(B)`

Solution: `take(np.arange(5))`

Which of the following choices is most appropriate for blank `(C)`?

◯ `select("song")`

◯ `take("song")`

◯ `column("song")`

◯ `item("song")`

(c) [2 Pts] Unfortunately, Kristen is having trouble picking which of these 5 songs to dance to. Her teammate, Ethan suggests that Kristen should choose a song at random. Write **one** line of code that evaluates to a song picked at random from `kristen_songs`.

Solution: `np.random.choice(kristen_songs)`

Meanwhile, Ethan is competing in the advanced division, and is given a table `advanced_songs`. This table resembles the `songs` table, but only contains songs with a dance difficulty of at least 8. Among these, Ethan will only consider dancing to the songs of his favorite artists, found below.

```
ethan_artists = make_array("NAOKI", "Junko", "Ryutaro")
```

Ethan is also given a second table `artists`. The first few rows are provided below.

artist	song
NAOKI	Kiss Kiss Kiss
Riyu	Honey Punch
Sota	Blew my Mind
NAOKI	Red Zone

(d) [4 Pts] Fill in the blanks to assign the variable `ethan_songs` to a table of songs that Ethan will consider dancing to. This table should contain the same columns as the `songs` table, along with an additional `artist` column.

Note: You may assume that there are no two songs with the same song title.

```
songs_and_artists = advanced_songs._____(A)_____
ethan_songs = songs_and_artists._____(B)_____
```

Fill in the blank `(A)`

Solution: `join("song", artists, "song")`

Fill in the blank `(B)`

Solution: `where("artist", are.contained_in(ethan_artists))`

For the final round, Ethan and Kristen will have to perform 3 complex dances together from the `songs_and_artists` table. One of these songs must have a dance difficulty of 8, the other a dance difficulty of 9, and the last a difficulty of 10. Kristen and Ethan decide to conserve their energy by choosing the song from each of these difficulty levels with the shortest length.

(e) [4 Pts] Fill in the blanks to create a table with **columns** corresponding to each difficulty level, **rows** corresponding to each artist, and **cells** containing the song titles with the shortest length.

```
def get_first(arr):
    _____(A)_____

songs_and_artists._____(B)_____._____(C)_____
```

Fill in the blank `(A)`

Solution: `arr.item(0)`

Which of the following choices is most appropriate for blank `(B)`?

◯ `select("length")` ◯ `where("length", min)`

◯ `sort("length")` ◯ `group("length")`

Fill in the blank `(C)`

Solution: `pivot("difficulty", "artist", "song", get_first)`

3 Heightened Histograms [15 Points]

Dylan, a Data 8 student, wonders if there is a difference in heights between his high school and college classmates. Among Data 8 students, he collects a random sample of 100 high school students and a random sample of 200 college students, and organizes his data in a table called `heights`. The first few rows are provided below.

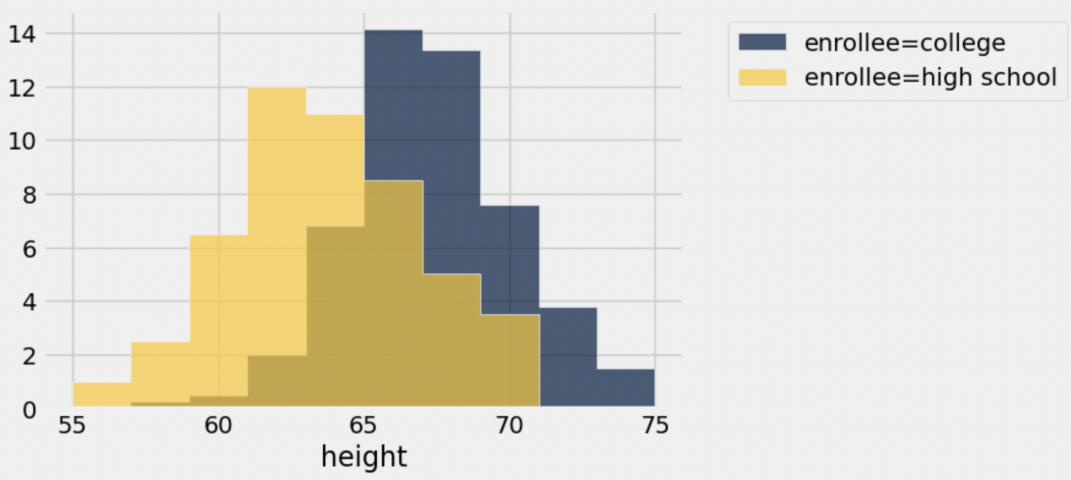
- `height` is the **float** height of an individual, measured in inches
- `enrollee` is a **string** containing either "high school" or "college"

height	enrollee
68.5	college
62.1	high school
66.5	college

(a) [4 Pts] Fill in the following blanks to create the histogram below.

Note: The x-axis of the histogram has a range of **[55, 75]**, and each bin is **2 inches** in width.

```
heights.__(A)__(_____(B)_____, bins=_____(C)_____, group=__(D)__)
```



(i) Fill in blank `(A)`:

Solution: `hist`

(i) Fill in blank `(C)`:

Solution: `np.arange(55, 76, 2)`

(ii) Fill in blank `(B)`:

Solution: `"height"`

(ii) Fill in blank `(D)`:

Solution: `"enrollee"`

(b) [2 Pts] Unfortunately, Dylan forgot to include the y-axis label in his histogram screenshot. Which of the following is an appropriate y-axis label? **Select one.**

- ◯ Percentage of students
- ◯ Percentage of students per height
- ◯ **Percentage of students per inch**
- ◯ Rate of change in height

(c) [2 Pts] Which of the following is closest to the total number of individuals in our table that are greater than or equal to 61 inches, but less than 63 inches in height? **Select one.**

- ◯ 14
- ◯ 24
- ◯ 28
- ◯ **32**

(d) [2 Pts] Using only the conclusions he can draw from his histogram, Dylan claims that he is shorter than **at least** 75% of high school students, but taller than **at least** one college student. Which of the following bins **may** Dylan belong in? **Select all that apply.**

- ☐ $[57 - 59)$
- ☐ $[59 - 61)$
- ☐ $[61 - 63)$
- ☐ $[63 - 65)$

(e) [2 Pts] Which of the following conclusions can you draw from the histograms in **part (a)**? **Select all that apply.**

- ☐ Assuming we know the height of each bin, we can calculate the percentage of individuals in any range of heights.
- ☐ **College Data 8 students are generally taller than high school Data 8 students.**
- ☐ The area of both histograms together sum to 100%, so the area of the blue and yellow histograms are $\frac{100}{300} \times 100$ and $\frac{200}{300} \times 100$, respectively.
- ☐ There is an association between a Berkeley student's height and age.
- ☐ College students enrolled in Data 8 are generally taller than high school students in Data 8 because they are older.

(f) [3 Pts] Let **x** and **y** be the heights of the first and second bin in the yellow histogram, respectively. Suppose we combine the first and second bin. Write a mathematical expression using **x** and **y** that evaluates to the height of the combined bin. If there is not enough information to answer the question, write "N/A".

Solution: `(x + y)/2`

4 Kayaks or Kanus [18 Points]

Kanu has been hired as a data scientist for Cal Rec Sports! His primary task is to analyze aquatic equipment rental data for the Berkeley Marina. Kanu's team presents him with a `rentals` table that contains the daily number of kayak and windsurfing board rentals, along with information pertaining to the day that the data were collected. Shown below are the first few rows of `rentals`.

- `date` contains the **string** date on which the rental data were collected
- `weekend` is a **boolean** corresponding to whether the day fell on a weekend
- `num_kayaks` is the **integer** number of kayak rentals on that particular day
- `num_boards` is the **integer** number of windsurfing board rentals on that particular day
- `wind` is the **integer** wind speed on that particular day, in miles per hour

date	weekend	num_kayaks	num_boards	wind
04/29	True	35	19	4
05/01	True	26	24	22
05/02	False	17	14	10

Choose which single visualization is most useful for answering each question below. **Select one**.

(a) [2 Pts] Does a higher wind speed *lead* to a larger number of windsurfing board rentals?

◯ Histogram ◯ Scatter Plot

◯ Bar Chart ◯ Line Plot

◯ **None of the above**

(b) [2 Pts] Do days with more kayak rentals have fewer windsurfing board rentals?

◯ Histogram ◯ **Scatter Plot**

◯ Bar Chart ◯ Line Plot

◯ None of the above

(c) [2 Pts] How does the distribution of kayak rentals on weekdays compare to the distribution of kayak rentals on weekends?

◯ **Overlaid Histograms** ◯ Overlaid Scatter Plots

◯ Bar Chart ◯ Line Plot

In the following blanks, you can write as much code as is necessary that fits the skeleton code. Do not add any additional lines of code.

Suppose that Kanu is tasked with calculating the revenue generated from rentals for each day. One kayak rental generates $\$15$ in revenue; one windsurfing board rental generates $\$25$ in revenue.

(d) [2 Pts] Fill in the blank to complete the function `revenue_calc` that calculates the revenue from one day. `revenue_calc` takes in two arguments: `num_kayaks` and `num_boards` that represent the number of kayaks and windsurfing boards rented that day, respectively.

```
def revenue_calc(num_kayaks, num_boards):

    ____________________(A)____________________
```

Fill in blank `(A)`:

Solution: `return 15 * num_kayaks + 25 * num_boards`

(e) [4 Pts] Using the function in **part (d)**, fill in the blanks to add a column "`revenue`" to the `rentals` table. This column should contain the revenue collected for each day in `rentals`. You may assume that the function `revenue_calc` has been implemented correctly.

```
revenue_array = rentals.___(A)___(___________(B)___________)

rentals = rentals._____(C)_____(___(D)___, revenue_array)
```

Fill in blank `(A)`:

Solution: `apply`

Fill in blank `(B)`:

Solution: `revenue_calc, "num_kayaks", "num_boards"`

Fill in blank `(C)`:

Solution: `with_column`

Fill in blank `(D)`:

Solution: `"revenue"`

(f) [2 Pts] Kanu's team wants to know how Cal Rec Sports' revenue fluctuates from day-to-day. Write one line of code to create a visualization that best displays the daily trend in revenue.

Solution: `rentals.plot("date", "revenue")`

(g) [4 Pts] Kanu is curious to know if Cal Rec Sports collects more revenue, on average, over some day on the weekend as compared to some weekday. Fill in the following blanks to help him create a visualization that answers his question.

```
rev_tbl = rentals.select("weekend", "revenue")

avg_revenue_per_day_type_tbl = rev_tbl.___(A)___(_____(B)_____)

avg_revenue_per_day_type_tbl._____(C)_____(_______(D)_______)
```

Fill in blank `(A)`:

Solution: `group`

Fill in blank `(B)`:

Solution: `"weekend", np.average`

Fill in blank `(C)`:

Solution: `barh`

Fill in blank `(D)`:

Solution: `"weekend", "revenue average"`

5 Poke Probabilities [14 Points]

Emma gets poke for lunch every week and enjoys trying different combinations of proteins. She always gets 3 servings of protein in her bowl, and chooses each protein at *random with replacement*. For example, she could randomly pick salmon three times, and thus end up with three servings of salmon in her bowl.

The following table lists the possible protein options and the probability that Emma chooses each.

Protein	Probability
Salmon	0.35
Spicy Tuna	0.05
Shrimp	0.15
Chicken	0.1
Tofu	0.15
Edamame	0.2

(a) [2 Pts] What is the probability that Emma ends up picking items in this exact order: salmon, chicken, and edamame?

- ◯ $(1 - 0.35) \times (1 - 0.1) \times (1 - 0.2)$
- ◯ $0.35 + 0.1 + 0.2$
- ◯ $(1 - 0.35) + (1 - 0.1) + (1 - 0.2)$
- ◯ $0.35 \times 0.1 \times 0.2$

(b) [2 Pts] What is the probability that Emma does not get tofu in any of her 3 servings?

- ◯ 0.15^3
- ◯ $1 - 0.15^3$
- ◯ $(1 - 0.15)^3$
- ◯ $3 \times (1 - 0.15)$

(c) [2 Pts] What is the probability that Emma gets either 3 servings of salmon or 3 servings of edamame?

- ◯ $0.35^3 + 0.2^3$
- ◯ $(1 - 0.35)^3 + (1 - 0.2)^3$
- ◯ $0.35^3 \times 0.2^3$
- ◯ $(1 - 0.35)^3 \times (1 - 0.2)^3$

(d) [2 Pts] What is the probability that Emma ends up with two servings of spicy tuna and one serving of chicken?

- ◯ $0.05^2 \times 0.1$
- ◯ $3 \times 0.05^2 \times 0.1$
- ◯ $0.05^2 + 0.1$
- ◯ $3 \times 0.05^2 + 2 \times 0.1$

(e) [2 Pts] What is the probability that Emma gets one or more servings of shrimp?

- ◯ $3 \times 0.15 \times 0.85^2$
- ◯ $(1 - 0.15)^3$
- ◯ $(0.15 \times 0.85^2) + (0.15^2 \times 0.85) + 0.15^3$
- ◯ $1 - (1 - 0.15)^3$

(f) [4 Pts] For this sub-part, assume that Emma picks **2 proteins** at *random* ***without*** *replacement*. In other words, Emma cannot pick the same protein option twice. What is the probability that Emma gets at least one serving of tofu?

Please show all your work in the provided blank as partial credit may be awarded. Write your final answer as an expression - you do not need to simplify. Please be sure to circle your final answer.

Note: This is a challenging question. Please ensure you have allocated enough time for the remainder of the exam.

Solution:

There are two equivalent ways of approaching this problem.

Approach 1:

```
P(Emma gets at least 1 tofu)
 = 1 - P(Emma does not get tofu on either serving)
 = 1 - (P(first serving salmon, second serving not tofu) +
   P(first serving spicy tuna, second serving not tofu) +
   P(first serving shrimp, second serving not tofu) +
   P(first serving chicken, second serving not tofu) +
   P(first serving edamame, second serving not tofu))

 = 1 - ((0.35)(0.5/0.65) + (0.05)(0.8/0.95) +
       (0.15)(0.7/0.85) + (0.1)(0.75/0.9) +
       (0.2)(0.65/0.8))

 = 0.3193
```

Approach 2:

```
P(Emma gets at least 1 tofu)
 = P(Emma gets tofu on her first serving or second serving)
 = P(tofu on first serving) + P(tofu on second serving)
 = 0.15 + (0.35)(0.15/0.65) + (0.05)(0.15/0.95) +
          (0.15)(0.15/0.85) + (0.1)(0.15/0.9) +
          (0.2)(0.15/0.8)

 = 0.3193
```

6 A (Corn)y Pizza Test [24 Points]

Every day after Data 8 lecture, Prasann the pizza enthusiast walks to Sliver to buy a slice of pizza. Sliver only serves one flavor of pizza on each day; they claim that the flavor is picked uniformly at random from all four possible flavors, and is independent of the flavor served on any other day. The four flavors are: *Mushroom*, *Tomato*, *Potato*, and *Corn*.

Curious of the validity of Sliver's claim, Prasann documents the pizza flavors of his next 100 orders.

Flavor	Counts
Mushroom	30
Tomato	20
Potato	18
Corn	32

Prasann notices that the number of corn pizzas in his sample is large. He wonders if this is due to chance, or if Sliver is serving corn pizza more frequently than advertised. He decides to put his Data 8 knowledge to use and begins to formulate a hypothesis test.

(a) [2 Pts] Please state a clear and complete **null** hypothesis for Prasann's hypothesis test.

Solution: The chance of Sliver serving a corn pizza is 25% each day and is independent of any other day. Any deviation is due to chance.

(b) [2 Pts] Please state a clear and complete **alternative** hypothesis for Prasann's hypothesis test.

Solution: The chance of Sliver serving a corn pizza on some arbitrary day is greater than 25%.

(c) [3 Pts] Which of the following test statistics are best for choosing between these null and alternative hypotheses? **Select all that apply**.

- ☐ Absolute difference between the sampled proportion of corn pizzas and expected proportion of corn pizzas under the null hypothesis.
- ☐ **Difference between the sampled proportion of corn pizzas and expected proportion of corn pizzas under the null hypothesis.**
- ☐ The expected proportion of corn pizzas under the null hypothesis.
- ☐ **The proportion of corn pizzas in the sample.**
- ☐ None of the above.

Looking at his data again, Prasann notices an even deeper systematic difference. The observed distribution of pizza flavors doesn't seem to follow the expected distribution under the null hypothesis. To investigate this larger difference in distributions, Prasann frames the following hypotheses:

Null hypothesis: Each of the 4 pizzas flavors are all equally likely to be served by Sliver.

Alternative hypothesis: Each of the 4 pizzas flavors are not equally likely to be served by Sliver.

(d) [3 Pts] Which of the following test statistics are best for choosing between these new null and alternative hypotheses? **Select all that apply**.

- ☐ **Total variation distance between the sampled flavor distribution and the expected flavor distribution under the null hypothesis.**
- ☐ **Sum of the absolute differences between the sampled flavor distribution and the expected flavor distribution under the null hypothesis.**
- ☐ Sum of the differences between the sampled flavor distributions and the expected flavor distribution under the null hypothesis.
- ☐ None of the above.

(e) [4 Pts] Suppose we use the total variation distance as the test statistic for our experiment. Complete the code below to simulate one value of the test statistic under the null hypothesis.

```
def simulate_test_statistic():
    expected_probabilities = _____(A)_____
    simulated_proportions = ___(B)___(_____(C)_____)
    difference = expected_probabilities - simulated_proportions
    simulated_tvd = _____(D)_____
    return simulated_tvd
```

Fill in the blank `(A)`

Solution: `make_array(1/4, 1/4, 1/4, 1/4)`

Fill in the blank `(B)`

Solution: `sample_proportions`

Fill in the blank `(C)`

Solution: `100, expected_probabilities`

Note: Part `(D)` can be found on the next page.

Fill in the blank `(D)`

Solution: `sum(abs(difference))/2`

(f) [2 Pts] Suppose that you construct an array called `simulated_tvds` that contains 1000 simulated test statistic values. Additionally, suppose you calculate your observed test statistic, `observed_tvd`. Write one line of code that computes the p-value of your experiment.

Solution: `np.mean(simulated_tvds >= observed_tvd)`

(g) [3 Pts] Suppose you choose a p-value cutoff of 5% and obtain a p-value of 0.062. Which of the following statements can you conclude? **Select all that apply.**

- ☐ The null hypothesis is true.
- ☐ The alternative hypothesis is true.
- ☐ **Your test results support the null hypothesis.**
- ☐ There is a 0.062 chance that the null hypothesis can be rejected.
- ☐ **There is a 6.2% chance that the total variation distance simulated under the null hypothesis is equal to or more extreme than the observed test statistic.**
- ☐ None of the above.

(h) [3 Pts] Which of the following can we conclude from our p-value and p-value cutoff above? **Select all that apply.**

- ☐ In Prasann's observed sample of data, Sliver made more corn pizzas than any other flavor of pizza.
- ☐ **Prasann can reasonably believe that Sliver chooses a pizza flavor uniformly at random every day.**
- ☐ Prasann is more likely to see other types of pizza the next time he goes to Sliver.
- ☐ None of the above.

(i) [2 Pts] In efforts to eat healthier, Prasann has decided that next semester, he will only go to Sliver every Friday rather than every day (thus yielding $\frac{1}{5}$ the amount of data). Suppose Prasann repeats this same experiment, and assuming the null hypothesis is true, he claims:

"Due to the Law of Averages, the probability of not getting a corn pizza on some arbitrary day is higher next semester than it is during this current semester."

Mark the statement above **True or False**.

- ◯ True
- ◯ False

7 Congratulations [0 Pts]

Congratulations! You have completed the Midterm.

- **Make sure that you have written your student ID number on *each page* of the exam.** You may lose points on pages where you have not done so.
- Also ensure that you have **signed the Honor Code** on the cover page of the exam for 1 point.

[Optional, 0 pts] Draw your favorite Data 8 Experience or TA!