Name:

Email:

SID:

Data 8 Summer 2020 Midterm

Instructions

- The exam is worth 110 points. You have 110 minutes to complete it.
- If you lose internet connection during the exam, the page will give you a warning. If you cannot easily reconnect to the page please email data8su20@gmail.com as soon as you can for instructions.
- For all Python code, you may assume that the statements `from datascience import *` and `import numpy as np` have been executed. Do not use features of the Python language that have not been described in this course.
- In any part, you are free to use any tables, arrays, or functions that have been defined in previous parts of the same question, and you may assume they have been defined correctly.
- The blanks on fill-in-the-blank coding questions do not reflect the length of the correct expression.
- We strongly recommend that you have a piece of paper or writing device to help you work through the problems.
- The exam autosaves and you do not need to submit at the end

Honor Code

“As a member of the UC Berkeley community, I act with honesty, integrity, and respect for others. If there is any evidence that I worked with someone else, I used materials outside of the allowed resources, or I distributed exam materials online, my exam will not be graded and I will receive a failing grade in the course.”

Please sign your full name to acknowledge your agreement to the statement above.

_____

1. Arrays and Tables (27 points)

The `artists` table contains data about numerous artists who have work displayed in the Museum of Modern Art (MoMA). Each artist has a unique ID number that can be used to identify artwork, represented by the column `Artist ID`. A few rows are shown below:

Artist ID	Name	Nationality	Gender	Birth Year	Death Year
6174	Giuseppe Viviani	Italian	Male	1898	1965
4262	Virginia Nepodal	American	Female	1914	2009
1444	Niki de Saint Phalle	French	Female	1930	2002
612	Oscar Bluemner	American	Male	1867	1938
7468	Charles B. Kaufmann	American	Male	1890	1957

The `artworks` table contains some information about the art pieces displayed in MoMA, and a few rows are shown below:

Artwork ID	Artist ID	Date	Department	Classification	Height (cm)	Width (cm)
3	7470	1987	Architecture & Design	Architecture	40.6401	29.8451
2	6210	1896	Architecture & Design	Architecture	48.6	168.9
43161	229	1925	Photography	Photograph	16.5	21.2
65154	4823	1997	Prints & Illustrated Books	Print	133.35	85.7252
32	2964	1968	Architecture & Design	Architecture	113	167.6

Fill in the blanks with the corresponding Python expressions.

a. The nationality of the artist who lived the fourth longest. You can assume that each artist died at a different age. (4 pts)

```
age = artists.with_column("Age at Death", _______(i)_______)
age._______(ii)_______._______(iii)_______._______(iv)_______
```

Blank i:

Blank ii:

Blank iii:

Blank iv:

b. *The name of the artist that was born the earliest in the 20th century (1900s). In answering this question you can assume that no two artists are born in the same year and no artist is born after 1999. (4 pts)*

```
filtered_by_birth_year = __________(i)__________
filtered_by_birth_year.__________(ii)__________
```

Blank i:

Blank ii:

c. *The number of art pieces in MoMA that are classified as "Print" and were created after 1950. (4 pts)*

```
____(i)____.where("Date", ____(ii)____).____(iii)____.___(iv)____
```

Blank i:

Blank ii:

Blank iii:

Blank iv:

d. *The birth year of the artist that created the tallest (by height) piece of artwork. (5 pts)*

```
artists_and_artwork = artists._____(i)_____
artists_and_artwork._____(ii)_____._____(iii)_____._____(iv)_____
```

Blank i:

Blank ii:

Blank iii:

Blank iv:

e. *The highest mean height among all combinations of department and classification.(5 pts)*

```
_____(i)_____(_____(ii)_____._____(iii)_____._____(iv)_____)
```

Blank i:

Blank ii:

Blank iii:

Blank iv:

f. *Many artists have many pieces of work in MoMA. Create a table with two columns named “Name” and “Width (cm) average” that reflects the mean width of all pieces created by an artist. (5 pts)*

```
artists._____(i)_____._____(ii)_____._____(iii)_____
```

Blank i:

Blank ii:

Blank iii:

2. Data Type Pot Pourri (6 points)

a) Which pairs of table methods return the same type? Select all that apply. (3 pts)

i) .column() and .select() [ ]
ii) .relabled() and .group() [ ]
iii) .pivot() and .join() [ ]
iv) .take() and .where() [ ]
v) .drop() and .num_rows [ ]
vi) .apply() and .where() [ ]

b) Cynthia writes the following code:

```
early_losses = games.where('Opponent', 'UC Berkeley').column(2).num_rows
```

She receives the following error message: `'numpy.ndarray' object has no attribute 'num_rows'`*, and needs help debugging.*

What is the best explanation for this error? (3 pts)

i) `games` is a table with no rows [ ]
ii) `games` only has 2 columns [ ]
iii) the num_rows attribute only exists for rows and not for columns [ ]
iv) the num_rows attribute must follow a table type [ ]

3. GDP Growth Rates (14 points)

Congrats! You've just got your dream internship working as a research assistant for the Economics Department's illustrious Professor Olney! She wants you to analyze the GDP per capita growth rates of African countries. Let's get started.

Professor Olney provides you with the table below, called `gdp_africa`, which has two columns:

- `Country`: The name of the country
- `GDP per capita Growth(%)`: the GDP per capita growth rate (2018)

The first five rows are displayed below:

Country	GDP per capita Growth (%)
Angola	-5.28778
Burundi	-1.55741
Benin	3.97518
Burkina Faso	3.81291
Botswana	2.18307

To visualize the distribution of GDP per capita growth rates in Africa, you plan to generate a histogram using `gdp_africa`.

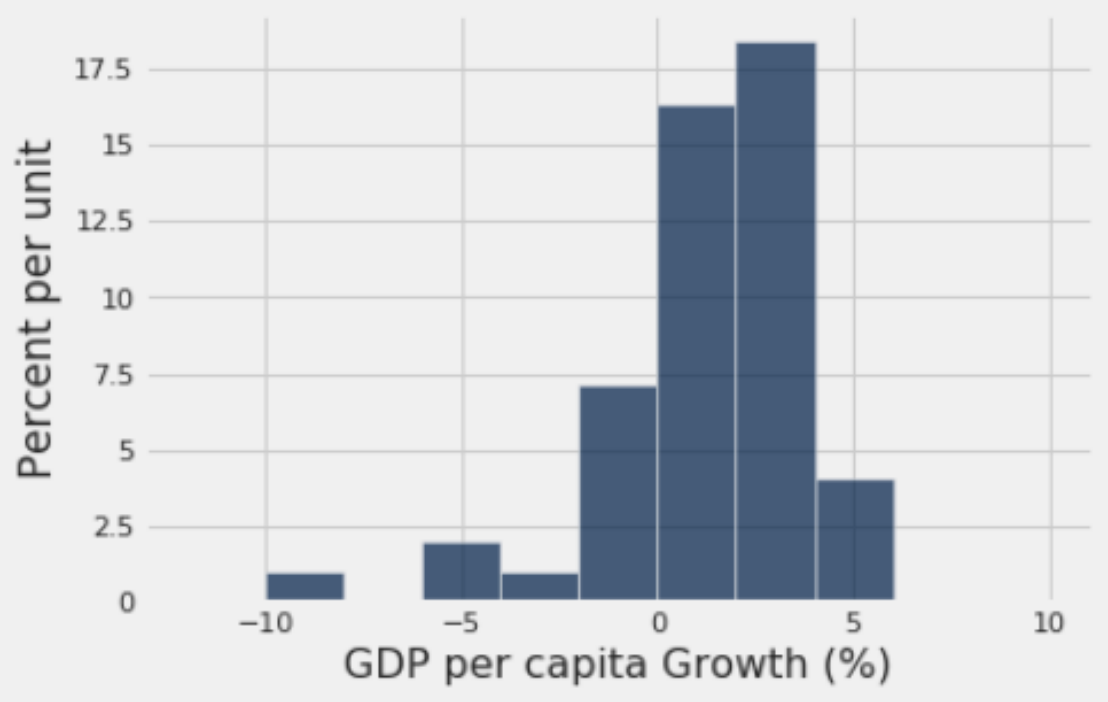
a) *Given that no country in Africa had a GDP per capita growth rate higher than 8% or lower than -10% in 2018, construct bins that reflect this range with bin widths of 2%. (2 pts)*

```
gdp_bins = _____________________________________________
```

b) *Write a line of code that would produce a histogram using the GDP per capita bins you just created. (2 pts)*

```
gdp_africa._____________________________________________
```

Now that you've run your histogram code, we find that the histogram for 2018 GDP per capita growth rates for African countries looks like this:



c) *What percentage of countries in Africa had a GDP per capita growth rate between -2% and 0% in 2018? (2 pts)*

(i) Between 6% and 7.5% [ ]
(ii) Between 12% and 15% [ ]
(iii) Between 15% and 17.5% [ ]
(iv) None of the above [ ]

d) *How many countries in Africa had a GDP per capita growth rate of between 2% and 3% in 2018? (2 pts)*

(i) 18 [ ]
(ii) 36 [ ]
(iii) 40 [ ]
(iv) Cannot tell with the information given [ ]

e) Assume the heights of each bar are as follows:

GDP Growth Range	[-10, -8)	[-8, -6)	[-6, -4)	[-4, -2)	[-2, 0)	[0, 2)	[2, 4)	[4, 6)	[6, 8)
Height (%)	0.93	0	1.85	0.93	7.41	16.67	18.52	3.70	0

Using the information above, write a mathematical expression that evaluates to the approximate number of countries in Africa that had a GDP per capita growth rate of between 0% and 6% in 2018. There are 54 countries represented in the table `gdp_africa`*. There is no need to round. (4 pts)*

Type your answer here:

f) Select all statements that are true. (2 pts)

(a) This distribution is left-skewed [ ]
(b) Most countries had a positive GDP per capita growth percentage in 2018 [ ]
(c) The GDP per capita growth percentage's median is smaller than its mean [ ]

4. Probability (22 points)

On a hot summer day, Jo decides to drop by the local creamery for ice cream. The local creamery has four flavors: vanilla, chocolate, strawberry, and the flavor of the month, orange. Jo is worried that he had his craving too late in the day and the creamery might have ran out of his favorite flavor combination: vanilla and orange swirl.

On any given day, the chance that the creamery runs out of vanilla, chocolate, or strawberry is equal to 1/6. Since orange is a speciality flavor, the probability that the creamery runs out orange is ⅓. These events are independent of one another.

Please leave your answers as an unsimplified expression. For example, your answers may take the form of 7/10, (7/10) * (5/8), or (7/10) * (5/8) + (6/10) * (1/8).

Show all your work in the provided area.

a. *If Jo visits the creamery right before they close, what is the probability that Jo gets what he wants? (3 pts)*

   Type your answer here:

b. *Suppose that Jo would be happy with vanilla and orange swirl, or a strawberry cone. If there are no vanilla and orange swirls left, Jo would order a strawberry cone. Given that Jo is happy, what is the probability he got the vanilla and orange swirl? Recall that flavors run out independently of each other. (5 pts)*

   Type your answer here:

c. *Suppose that if the creamery runs out of chocolate ice cream, the chance that they also run out of vanilla ice cream by the end of the day increases to ¼. If you aren't sure if the shop has run out of chocolate yet, what is the probability that Jo gets what he wants? (5 pts)*

Type your answer here:

When Jo gets to the creamery, he finds out that they're having a special promotion! The store has a box full of coupons that customers can draw from. The discount listed on the coupon drawn is applied to the customer's purchase. After the transaction is complete, the coupon is placed back into the box.

There are 50 coupons in the box: 30 coupons are for a 25% discount, 10 coupons are for a 50% discount, 8 coupons are for a 75% discount and 2 coupons are for a 100% discount.

d. *What is the chance that Jo gets at least 50% off his ice cream? (3 pts)*

Type your answer here:

e. *Suppose that Jo brought his friend Jackie to the creamery with him. What is the chance that at least one of them got 100% off their ice cream? (3 pts)*

Type your answer here:

f. *Jo forgot to put his ticket back before Jackie drew from the box! Given this information, what is the chance that at least one of them got 100% off their ice cream? (3 pts)*

Type your answer here:

5. Sampling and Iteration (21 points)

Suppose you and your friends are playing a special game that involves 50 cards, numbered from 1 to 50.

On a player's turn, they draw 5 cards with replacement and are assigned a score that is equal to the sum of the cards drawn. If the score is divisible by 3, the player is awarded 33 additional points.

Each player gets one turn. The player with the highest score at the end of the game wins. If there is a tie, no one wins.

a. *Write a function one_turn() that simulates one player's turn. The function should take no arguments and returns the player's score at the end of their turn. (5 pts)*

```
def one_turn():
    cards = np.arange(1, 51, 1)
    draw = np.random.choice(_____(i)_____,  _____(ii)_____)
    score = _____(iii)_____(_____(iv)_____)
    if _____(v)_____:
        score = score + 33
    return score
```

Blank i:

Blank ii:

Blank iii:

Blank iv:

Blank v:

b. *Write a function* `play` *that simulates one play of the game. The function takes in one argument* `num_players`*, the number of players in the game. It should return the number of the player that wins (If the first player to go wins, return 1. If the second player wins, return 2. etc.) If no one wins, return 0. (7 pts)*

```
def play(num_players):
    scores = make_array()
    winning_score = 0
    winning_player = 0

    for i in np.arange(_________(i)_________):
        score = _________(ii)_________
        if score > _________(iii)_________:
            winning_score = score
            winning_player = _________(iv)_________
        scores = _________(v)_________
    if _________(vi)_________ > 1:
        return 0
    else:
        return winning_player
```

Blank i:

Blank ii:

Blank iii:

Blank iv:

Blank v:

Blank vi:

c. *You soon get bored of playing the game. Your group of friends decides to eat lunch instead, and you conclude that in the time you spend eating, you could have played 100 games. Write a function that simulates 100 rounds of the game. The function takes in one argument, `num_players`, the number of players in the game, and draws a bar chart that visualizes the number of times each player won. (9 pts)*

```
def a_hundred_games(num_players):
    winners = _________(i)_________
    _________(ii)_________:
        winner = _________(iii)_________
        _________(iv)_________= _________(v)_________
    win_count_tbl = _________(vi)_________._________(vii)_________
    _________(viii)_________._________(ix)_________("Winners")
```

Blank i:

Blank ii:

Blank iii:

Blank iv:

Blank v:

Blank vi:

Blank vii:

Blank viii:

Blank ix:

6. Hypothesis Testing (20 points)

Chloe is a big fan of Trader Joes' frozen mac n cheese, and has noticed that the cheese used varies from box to box. A Trader Joe's employee provides her with some data describing the four different cheeses used:

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

*Only one type of cheese is used in any given box of mac n cheese.

Chloe is suspicious of this distribution. For example, Velveeta is much cheaper than Gruyère. Also, she has never bought a box that uses Monterey Jack.

Chloe decides to conduct a hypothesis test and buys a random sample of 100 mac n cheese boxes. She finds that 20% of her boxes have velveeta, 30% have Gruyère, 45% have Sharp Cheddar, and 15% have Monterey Jack.

a. *What null hypothesis, alternative hypothesis and test statistic should Chloe use? (6 pts)*

Null Hypothesis

Alternative Hypothesis

Test Statistic

b. *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. (4 pts)*

```
observed_proportions = make_array(0.2, 0.3, 0.45, 0.05)
employee_proportions = make_array(0.05, 0.55, 0.25, 0.15)

def one_simulated_test_stat():
      sample_prop = _________(i)_____________
      test_stat = _________(ii)_____________
      return test_stat
```

Blank i:

Blank ii:

c. *Generate 10,000 simulated statistics using* `one_simulated_test_stat`*, and store them in* `simulated_stats`*. Calculate the empirical p-value, and assign it to the* `p_value` *variable. Assume that the test statistic computed from the sample is provided by the* `observed_stat` *variable. (6 pts)*

```
simulated_stats = ____________(i)____________
repetitions = 10000

for ____________(ii)____________:
      one_stat = ____________(iii)____________
      simulated_stats = ____________(iv)____________

p_value = ____________(v)____________
```

Blank i:

Blank ii:

Blank iii:

Blank iv:

Blank v:

d. *Chloe defined a cutoff of 0.10 for her hypothesis test prior to conducting the analysis, and she finds that* `p_value` *is equal to 0.02. Is the data more consistent with the null hypothesis or alternative hypothesis? (2 pts)*

   i. Consistent with the null hypothesis [ ]
   ii. Consistent with the alternative hypothesis [ ]
   iii. Can't say [ ]

e. *Suppose the true distribution of mac n cheese flavours and the distribution provided by the employee are identical. If Chloe were to repeat this entire sampling and testing process 2000 times using a pre-specified cutoff of 10%, she should expect to reject the null hypothesis roughly _____ times. (2 pts)*

   i) 0 [ ]
   ii) 100 [ ]
   iii) 200 [ ]
   iv) 900 [ ]
   v) 800 [ ]
   vi) Can't say [ ]