DATA 8 Foundations of Data Science

Spring 2022

Midterm Solution

INSTRUCTIONS

You have 1 hour and 50 minutes to complete the exam.

- The exam is closed book, closed notes, closed computer, closed calculator, except the provided midterm study guide.
- Mark your answers on the exam itself in the spaces provided. We will not grade answers written on scratch paper or outside the designated answer spaces.
- If you need to use the restroom, bring your phone and exam to the front of the room.

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!

Preliminaries

You can complete and submit these questions before the exam starts.

(a) What is your full name?

(b) What is your student ID number?

(c) Who is your lab GSI? You may write *self-service* if you have no lab GSI.

(d) Sign here to confirm that all work on this exam is your own (or type your name if online).

1. (41.0 points) Basketable

The `teams` table contains one row for each of the 30 teams in the National Basketball Association (NBA) league. Columns exist for the team's `name`, `division`, `conference`, and home `arena` capacity. Each team has its own arena. The first five rows are:

name	division	conference	arena
Celtics	Atlantic	Eastern	18642
Lakers	Pacific	Western	18997
Nets	Atlantic	Eastern	17732
Pistons	Central	Eastern	20491
Rockets	Southwest	Western	18055

The `players` table contains a row for each of the 528 players in the 2020 NBA season. Columns are the player's `name`, 2019 salary (`2019`), 2020 salary (`2020`), 2019 team name (`19team`), and 2020 team name (`20team`). For players who joined in 2020, their `2019` value is 0 and their `19team` value is `No Team`. The first three rows are:

name	2019	2020	19team	20team
Stephen Curry	37457154	40231758	Warriors	Warriors
Dwight Howard	5337000	1620564	Wizards	Lakers
Zion Williamson	0	9757440	No Team	Pelicans

(a) (4.0 points)

This partially completed expression evaluates to the **name of the team** (a string) with the smallest arena capacity. Assume no two arenas have the same capacity.

```
teams._________.________.________
        (a)       (b)       (c)
```

i. (1.0 pt) Fill in blank (a).

```
sort('arena')
```

ii. (2.0 pt) Fill in blank (b).

```
column('name')
```

iii. (1.0 pt) Which of these could fill in blank (c)?

- ○ `min()`
- ○ `max()`
- ● `item(0)`
- ○ `item(1)`

(b) (3.0 points)

This partially completed expression evaluates to a **table with one row per division in the Eastern conference** that has two columns: the `division` and the `count` of the number of teams in that division.

This expression should evaluate to the following table.

division	count
Atlantic	5
Central	5
Southwest	5

```
teams._________.group(_________)
         (a)             (b)
```

Reminders:

- The `teams` table has columns `name`, `division`, `conference`, and `arena`.
- The `players` table has columns `name`, `2019`, `2020`, `19team`, and `20team`.

i. (2.0 pt) Fill in blank (a).

```
where('conference', 'Eastern')
```

ii. (1.0 pt) Which of these could fill in blank (b)?

- ○ `'count'`
- ○ `'conference'`
- ● `'division'`
- ○ `'name'`

(c) (6.0 points)

This partially completed expression evaluates to a **table with one row per division** that has two columns: the `division` and the total 2020 salary for all players in that division. Any label for the second column is acceptable.

```
teams._________.select('division', '2020')._________
        (a)                                  (b)
```

Reminders:

- The `teams` table has columns `name`, `division`, `conference`, and `arena`.
- The `players` table has columns `name`, `2019`, `2020`, `19team`, and `20team`.

i. (4.0 pt) Fill in blank (a).

```
join('name', players, '20team')
```

ii. (2.0 pt) Fill in blank (b).

```
group('division', sum)
```

(d) (12.0 points)

Write an expression that correctly computes each of the following quantities.

You may use `t` for `teams`, `p` for `players`, `a` for the arena column of the `teams` table, and `np` for NumPy.

```
import numpy as np
t = teams
p = players
a = teams.column('arena')
```

Reminders:

- The `teams` table has columns `name`, `division`, `conference`, and `arena`.
- The `players` table has columns `name`, `2019`, `2020`, `19team`, and `20team`.

i. (3.0 pt) The largest **increase in salary from 2019 to 2020** (an integer) of any player.

```
max(players.column('2020') - players.column('2019'))
```

ii. (3.0 pt) The **number of players in 2020** (an integer) who played for the same team in 2019 and 2020.

```
sum(players.column('19team') == players.column('20team'))
```

iii. (2.0 pt) The **number of teams** (an integer) that have an arena size that is above average.

```
sum(a > np.average(a))
```

iv. (4.0 pt) Select **all** of the quantities below that can be computed from these two tables.

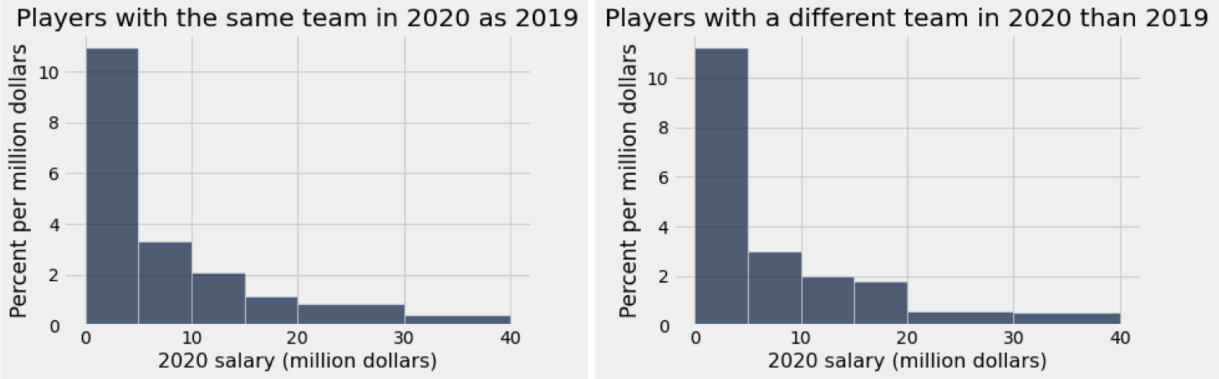
- ■ The number of divisions that had at least 5 players paid more than $20,000,000 in 2020
- ■ The name of the team that paid the most player salary per seat in its arena in 2020 (Note: The number of seats in an arena is its capacity.)
- ☐ The number of players who retired after the 2019 season.
- ☐ The name of the player that made the most additional salary by changing teams in 2020 compared to the amount they would have made staying at their 2019 team
- ☐ None of these

(e) (16.0 points)

The two histograms below displaying 2020 salaries were generated from data in the `players` table. The first histogram only includes players who had the same `19team` and `20team`. The second only includes players who had a different `19team` from their `20team` and played in both 2019 and 2020.

The bins are `make_array(0, 5, 10, 15, 20, 30, 40)`.

Additional information: Among the 440 players who played in both 2019 and 2020, 60% played on the same team and 40% played on different teams.



i. (2.0 pt) About what percentage of the players who had the **same** `19team` and `20team` had a salary between \$10 million and \$20 million in 2020?

- (●) 15%
- ( ) 20%
- ( ) 25%
- ( ) 30%
- ( ) 35%

ii. (2.0 pt) About what percentage of the players who had the **same** `19team` and `20team` had a salary of \$10 million or more in 2020?

- ( ) 10%
- (●) 30%
- ( ) 50%
- ( ) 70%
- ( ) 90%

iii. (4.0 pt) About how many players played on **different** teams in 2019 and 2020 and made between \$5 million and \$10 million in 2020?

Please express your answer as a Python expression (e.g., `0.1 * 0.2 + 0.3`) rather than simplifying it to a single number.

```
440 * 0.4 * 3 * 5 / 100
```

iv. **(4.0 pt)** Select **all** of the quantities below that can be determined from **only** these two histograms and the **additional information** that appears just above the histograms.

Reminder: The additional information was that among the 440 players who played in both 2019 and 2020, 60% played on the same team and 40% played on different teams.

- ☐ The total number of players who played in 2019 and had a 2020 salary below $2 million
- ■ The total number of players who played in 2019 and had a 2020 salary below $20 million
- ■ Among all players who played in both 2019 and 2020, the proportion who had a salary of $20 million or more
- ■ Among all players who played in 2019 and had a 2020 salary of $20 million or more, the proportion who played on the same team in 2019 and 2020
- ☐ None of these.

v. **(2.0 pt)** How would you use these histograms to determine whether the 2020 salary distribution was different for players with a different team than for players with the same team?

- ● Compare the two histograms visually and look for differences.
- ○ Use the two histograms to perform an A/B test.
- ○ Use the histograms to compute the average salary for both groups and compare those averages.
- ○ Use the histograms to compute the total salary for both groups and compare those totals.

vi. **(2.0 pt)** The $30-$40 million bin is slightly taller for players with a different team (right histogram) than for players with the same team (left histogram). What can we conclude from this difference?

- ○ Players who switch teams are paid more.
- ○ Players who switch teams are more likely to end up with a salary of $30-$40 million.
- ● Within that bin, the density among players with a different team is higher than the density among players with the same team.
- ○ Within that bin, the number of players with a different team is higher than the number of players with the same team.

2. (11.0 points) Sus

In the mobile game Among Us, Crewmates on a spaceship work together to complete tasks while a few randomly-selected Imposters secretly try to eliminate crewmates. If all Crewmates complete their tasks, the Crewmates win; if the Imposters eliminate all but one of the crewmates, the Imposters win.

Matty made a `games` table listing each game they played in 2021, ordered chronologically. The first three rows:

team	outcome	length	completed
Crewmate	Win	981	7
Imposter	Loss	840	8
Crewmate	Loss	520	3

The columns include:

- `team`: which team Matty was on in the game.
- `outcome`: whether Matty's team won or lost.
- `length`: the duration of the game in seconds.
- `completed`: the number of tasks completed by all crewmates before the game ended.

(a) (3.0 points)

Choose which type of visualization would be most useful for investigating each of the following.

i. (1.0 pt) The distribution of game lengths.

- ○ Bar Chart
- ● Histogram
- ○ Line Plot
- ○ Scatter Plot

ii. (1.0 pt) The association between game length and number of tasks completed.

- ○ Bar Chart
- ○ Histogram
- ○ Line Plot
- ● Scatter Plot

iii. (1.0 pt) The average game length for each outcome.

- ● Bar Chart
- ○ Histogram
- ○ Line Plot
- ○ Scatter Plot

(b) (8.0 points)

i. (4.0 pt) The result of which of the following expressions contains in **one** of its cells the total number of games in which Matty won? **Select all that apply.**

- ☐ `games.pivot('outcome', 'team')`
- ■ `games.group('outcome')`
- ☐ `games.group('team')`
- ☐ `games.group(['team', 'outcome'])`
- ☐ None of these

ii. (4.0 pt) The result of which of the following expressions contains in **one** of its cells the total number of tasks completed in all games for which Matty was a Crewmate and lost? **Select all that apply.**

- ☐ `games.pivot('completed', 'team', 'outcome', collect=sum)`
- ■ `games.pivot('team', 'outcome', 'completed', collect=sum)`
- ☐ `games.group('team').group('outcome').group('completed', collect=sum)`
- ■ `games.group(['team', 'outcome'], collect=sum)`
- ☐ None of these

3. (24.0 points) Chances

Each pet photo at the end of a lab is chosen from a collection of 20 pets with 10 cats, 9 dogs, and 1 bird.

For each event below, choose the Python expression that evaluates to the probability of that event.

(a) (8.0 points)

i. (2.0 pt) When **one** pet is chosen at random, the probability that it is either a cat or a bird.

- ○ `(9 / 20) ** 2`
- ○ `(10 / 20) * (1 / 20)`
- ● `(10 / 20) + (1 / 20)`
- ○ `1 - (9 / 20) ** 2`
- ○ `1 - (10 / 20) * (1 / 20)`
- ○ `1 - ((10 / 20) + (1 / 20))`

ii. (2.0 pt) When **two** pets are chosen at random with replacement, the probability that they are both dogs.

- ● `(9 / 20) ** 2`
- ○ `(10 / 20) * (1 / 20)`
- ○ `(10 / 20) + (1 / 20)`
- ○ `1 - (9 / 20) ** 2`
- ○ `1 - (10 / 20) * (1 / 20)`
- ○ `1 - (10 / 20) + (1 / 20)`

iii. (2.0 pt) When **two** pets are chosen at random with replacement, the probability that the first is a cat and the second is not.

- ○ `10 / 20 + 10 / 20`
- ● `(10 / 20) * (10 / 20)`
- ○ `(10 / 20) * (9 / 20) * (1 / 20)`
- ○ `1 - (10 / 20) * (10 / 20)`
- ○ `1 - (10 / 20 + 10 / 20)`
- ○ `1 - (10 / 20) * (9 / 20) * (1 / 20)`

iv. (2.0 pt) When **two** pets are chosen at random with replacement, the probability that the first chases the second. Assume dogs only chase cats, cats only chase birds, and birds don't chase.

- ● `(10 / 20) * (10 / 20)`
- ○ `(19 / 20) * (10 / 20)`
- ● `(10 / 20) * (1 / 20) + (9 / 20) * (10 / 20)`
- ○ `1 - ((9 / 20) * (1 / 20) + (10 / 20) * (9 / 20))`
- ○ `1 - ((10 / 20) ** 2 + (9 / 20) ** 2 + (1 / 20) ** 2)`
- ○ `1 - ((10 / 20) ** 2 + (9 / 20) ** 2 + (1 / 20))`

(b) (8.0 points)

The pygmy hippo is a small, reclusive (and cute) hippopotamid type that is native to the forests and swamps of West Africa. Two teams of zoologists set out to estimate the proportion that are male by sampling at random from the population. The first team samples 100 hippos and finds the proportion of males in their sample to be A. The second team samples 40 hippos and finds the proportion of males in their sample to be B. The full population has all 2,500 wild pygmy hippos; the proportion P of males in the population is 50% (but unknown to the zoologists).

i. (4.0 pt) Which of the following are more likely than not? Select **all** that apply.

- ☐ A is smaller than B.
- ☐ A is larger than B.
- ■ P is closer to A than B.
- ☐ P is closer to B than A.
- ☐ None of these.

ii. (2.0 pt) Which of the following is largest?

- ○ The chance that A is above 55%
- ● The chance that B is above 55%
- ○ The chance that A is above 60%
- ○ The chance that B is above 60%

iii. (2.0 pt) Which Python expression evaluates to the probability that B is not 0 and not 1, but instead a proportion between 0 and 1?

- ○ 0
- ○ 1
- ○ 0.5 ** 40
- ○ 1 - (0.5 ** 40)
- ○ 0.5 ** 40 + 0.5 ** 40
- ● 1 - (0.5 ** 40 + 0.5 ** 40)

(c) (8.0 points)

Complete the code below that uses a simulation repeated 10,000 times to estimate the chance that the average dice outcome when rolling 5 fair 6-sided dice is within 0.5 of 3.5. (That is, larger than 3 and smaller than 4.) For example, the average dice outcome of rolling (3, 2, 2, 6, 4) from the 5 dice is (3+2+2+6+4)/5 = 3.2, which is within 0.5 of 3.5.

```
def within(x, y, z):
    "Return whether z is strictly within x of y."
    return _________
              (a)
count = 0
for i in np.arange(10000):
    if within(0.5, 3.5, np.average(_________(_________))):
                                        (b)       (c)
        count = count + 1

estimate = count / _________
                     (d)
```

i. (3.0 pt) Fill in blank (a). You may call the built-in function `abs` to compute the absolute value of a number.

```
abs(z-y) < x or y-x < z < y + x or y-x < z and z < y + x
```

ii. (1.0 pt) Which of these could fill in blank (b)?

- ○ `sample_proportions`
- ● `np.random.choice`
- ○ `Table.sample`
- ○ `within`
- ○ `max`
- ○ `min`

iii. (3.0 pt) Fill in blank (c). You may include one or more commas.

```
np.arange(1, 7), 5
```

iv. (1.0 pt) Which of these could fill in blank (d)?

- ○ `counts`
- ○ `trials`
- ○ `len(count)`
- ● `10000`

4. (24.0 points) Wordle

In the game of Wordle, a player guesses up to 6 words until they correctly guess the secret word of the day or run out of guesses. Their *guess count* is either the guess number that was correct, 1 through 6, or X if all 6 guesses were incorrect.

For all 1,000 UC Berkeley students who played Wordle yesterday, we have collected the *proportion* of students with each guess count. These proportions appear in the table below and an array called `berkeley`.

1	2	3	4	5	6	X
0.0	0.17	0.33	0.27	0.20	0.02	0.01

```
berkeley = make_array(0.00, 0.17, 0.33, 0.27, 0.20, 0.02, 0.01)
```

Wordle's creator, Josh Wardle, sent us the proportion of guess counts for all players who tried to guess yesterday's word in an array called `everyone`.

1	2	3	4	5	6	X
0.0	0.09	0.25	0.32	0.28	0.03	0.03

```
everyone = make_array(0.00, 0.09, 0.25, 0.32, 0.28, 0.03, 0.03)
```

(a) (12.0 points)

Let's investigate whether the distribution of guess counts for UC Berkeley students differs from the distribution for all players on yesterday's Wordle. Describe a hypothesis test that would aid this investigation.

i. (2.0 pt) Complete the null hypothesis: *The distribution of guess counts for UC Berkeley students is ...*

- ◯ uniform with a 1/7 chance for each possible guess count.
- ◯ like a random sample from a uniform distribution with a 1/7 chance for each possible guess count.
- ◯ different from a uniform distribution with a 1/7 chance for each possible guess count.
- ◯ the population of guess counts for all Wordle players.
- ● like a random sample from the population of guess counts for all Wordle players.
- ◯ different from the population of guess counts for all Wordle players.

ii. (2.0 pt) Complete the alternative hypothesis: *The distribution of guess counts for UC Berkeley students is ...*

- ◯ the same as the distribution of guess counts for all Wordle players.
- ● different from the distribution of guess counts for all Wordle players.
- ◯ the same as the uniform distribution.
- ◯ different from the uniform distribution.

iii. **(2.0 pt)** Which test statistic is best for choosing between the null and alternative hypotheses?

- ○ total guess count
- ○ most common guess count
- ○ guess count
- ● total variation distance
- ○ observed average

iv. **(2.0 pt)** Which line of code simulates a distribution of proportions for 1000 Berkeley students **under the null hypothesis**?

- ○ `sample_proportions(1000, berkeley)`
- ● `sample_proportions(1000, everyone)`
- ○ `sample_proportions(1000, make_array('1', '2', '3', '4', '5', '6', 'X'))`
- ○ `sample_proportions(1000, make_array(1/7, 1/7, 1/7, 1/7, 1/7, 1/7, 1/7))`

v. **(2.0 pt)** How does **increasing** the number of times a distribution is simulated under the null hypothesis affect the outcome of the hypothesis test?

- ○ The probability that the null hypothesis is false will increase.
- ○ The probability that the null hypothesis is true will increase.
- ○ The observed distribution of guess counts for Berkeley students will be more similar to the distribution for all players.
- ○ The observed test statistic for Berkeley students will be more similar to the test statistic for all players.
- ● The empirical distribution of the test statistic under the null hypothesis will be more similar to its theoretical distribution.

vi. **(2.0 pt)** If the null hypothesis is rejected because the p-value of this hypothesis test is very small, what can we conclude? Select **all** that apply.

- ☐ Attending Berkeley improves most people's Wordle performance.
- ☐ Attending Berkeley changes most people's Wordle performance.
- ☐ Attending Berkeley does not improve most people's Wordle performance.
- ☐ Attending Berkeley does not change most people's Wordle performance.
- ■ None of these.

(b) (4.0 points)

Assume the observed test statistic is assigned to `obs`. We simulate under the null hypothesis 10,000 times and append each simulated test statistics to an array `sim`. Complete this Python expression that computes the p-value for this hypothesis test.

```
_________ ( _________ >= _________ ) / len( _________ )
   (a)          (b)           (c)                (d)
```

i. (1.0 pt) Fill in blank (a).

`sum` **or** `np.count_nonzero`

ii. (1.0 pt) Fill in blank (b).

`sim`

iii. (1.0 pt) Fill in blank (c).

`obs`

iv. (1.0 pt) Fill in blank (d).

`sim`

(c) Define *reading more* as spending an extra two hours a day reading The New York Times, and a *good game* of Wordle as one in which the player guesses the word in 3 or fewer tries. We want to test if reading more leads to a higher proportion of good games.

Among the 1000 Berkeley students who played Wordle yesterday, 500 were selected at random (without replacement) one month ago and asked to read more. All 1000 played yesterday's Wordle, and the number of guesses each student took was recorded.

i. (2.0 pt) How would a permutation test be used to investigate whether reading more leads to a higher proportion of good games?

- ● Repeatedly, all 1000 students would be partitioned at random without replacement into two groups of 500, and the proportion of good games in those two groups would be compared for simulating a null distribution.
- ○ Repeatedly, all 1000 students would be partitioned at random without replacement into two groups of 500, and within each group the proportion of good games for students who read more would be compared to that of the students who didn't.
- ○ Repeatedly, the proportion of good games for students who read more would be compared with the proportion of good games of a random permutation of those who didn't.
- ○ Repeatedly, the proportion of good games for students who read more would be compared with the proportion of good games of a random permutation of all 1000 students.

ii. (2.0 pt) Suppose we consider the following alternative hypothesis: Among the 1000 students, the proportion of good games would be higher if they all read more than if none of them read more.

Complete this null hypothesis: *Among the 1000 students, the proportion of good games ...*

- ○ would be lower for students who read more than for those who didn't.
- ○ for the 500 students who were selected to read more is the same as for the other 500 students.
- ○ for students who read more would be 50%.
- ● would be the same whether they all read more or none of them read more.

iii. (2.0 pt) Which of the following test statistics is best for choosing between the null and alternative hypotheses above?

- ● The difference between the proportion of good games in each group.
- ○ The absolute difference between the proportion of good games in each group.
- ○ The difference between the proportion of good games in the "read more" group and 0.5.
- ○ The difference between the proportion of good games in the "didn't read more" group and 0.5.

iv. (2.0 pt) When we conduct this permutation test, we compute a p-value of 0.002. Assume we had chosen a p-value cut-off of 0.05. Which of the following can we conclude about the 1000 Berkeley students based on this result? Select **all** that apply.

- ■ Reading more increases the proportion of good games.
- ■ There is an association between reading more and the proportion of good games.
- ☐ Being a Berkeley student is a confounding factor for the association between reading more and the proportion of good games.
- ☐ None of these