Midterm Exam Solutions - DSC 80, Summer 2024

Instructions:

- This exam consists of 6 questions. A total of 50 points are available.
- Write name in the top right of each page in the space provided.
- Please write neatly in the provided answer boxes. We will not grade work that appears elsewhere.
- Completely fill in bubbles and square boxes.
  ◯ A bubble means that you should only **select one choice**.
  ☐ A square box means you should **select all that apply**.
- You may refer to one 8.5" × 11" sheet of notes of your own creation. No other resources or technology (including calculators) are permitted.
- Do not turn the page until instructed to do so.

Last name	
First name	
Student ID number	
UCSD email	
All the work on this exam is my own. **(please sign)**	

Name: _____

This page is intentionally left blank, but feel free to use it as scratch paper.

Name: _____

Question 1 ..... ***13 points***

Fill in Python code below so that the last line of each part evaluates to each desired result using the tables `athletes` and `medals` as shown on the Reference Sheet.

(a) (2 points) Compute the number of medals won in July

```
medals[_____medals['date'].dt.month = 7_____].count()[0]
```

(b) (2 points) Compute the number of `teams` (countries) that has won at least one team event medal. (A team event is defined as as an event that involves more than one athlete from the same team)

```
def func(x):
    return (_____x['event'].nunique() != x['event'].shape[0]_____)

medals.groupby(_____'team'_____).filter(func)_____['team'].nunique()_____
```

(c) (2 points) Fill in the blank to calculate the proportion of medalists that competed in a `mixed` event for each `team`. For simplicity, assume that each such event title starts with the word `Mixed`.

```
def foo(x):
    return _____x.str.contains('Mixed')_____.sum() / x.shape[0]

    medals.groupby (____team____)[____event____].____agg____(foo)
```

(d) (2 points) Find the unique names of all gold medalists representing either 'USA' or 'CHN' in 'medalist' as a series:

```
medals.loc[(medals['team'].isin(['USA', 'CHN'])) & (medals['medal'] == 'Gold'), ____'name'____].drop_duplicat
```

(e) (2 points) Find a subset of the `athletes` DataFrame that includes only the rows with the top 5 most *populated* sports. Define a sport's *population* as the number of athletes in `athletes` that compete in that sport.

```
athletes.loc[athletes['sport'].isin(athletes['sport']. __value_counts().index[:5]__)]
```

Name: _____

(f) (3 points) Consider the following DataFrame named `event_medals` derived from the original `medals` DataFrame that contains 1 row for each unique medal. Which of the following snippets correctly produce a Series that display the team with the most number of Gold medals in each sport. It should look like this with `sport` as the index and `team` as values. Assume no ties between teams.

```
Out[10]: sport
         Athletics     SWE
         Badminton     DEN
         Basketball    USA
         Cycling       AUT
         Diving        CHN
         Name: team, dtype: object
```

- ☐ event_medals.groupby(['sport', 'team', 'medal'])['event'].count().idxmax()
- ■ **event_medals[event_medals['medal'] == 'Gold'].groupby('sport')['team'].agg(lambda a: a.value_counts().idxmax())**
- ■ **event_medals[event_medals['medal'] == 'Gold'].groupby(['sport','team']).count().reset_index().sort_ascending=False).groupby('sport')['team'].first()**
- ☐ event_medals[event_medals['medal'] == 'Gold'].pivot_table(index='sport', columns='team', values='event', aggfunc='count').idxmax()
- ■ **event_medals[event_medals['medal'] == 'Gold'].pivot_table(index='team', columns='sport', values='event', aggfunc='count').idxmax()**
- ☐ event_medals[event_medals == 'Gold'].groupby('sport')['team'].transform(lambda a: a.value_counts().idxmax())

Name: _____

Question 2 ..... ***6 points***

(a) (3 points) Write a code snippet using the `merge` method that returns the proportion of medal winners that are women ('sex' == 'F'). Note that there might be multiple people with the same name at the Olympics, but never on the same team.

Solution:

Solution:

```
a = athletes.merge(
    medals,
    on=['team', 'name'],
    how='left'
).drop_duplicates(subset=['name','team'])

a[(a['medal'].isna() == False) & (a['sex'] == 'F')].shape[0] / medals.shape[0]
```

(b) (3 points) Write a code snippet using the `merge` method that returns the proportion of Olympians that did not win a medal for each team. Again, there might be multiple people with the same name at the Olympics, but never on the same team.

Solution:

Solution:

```
a = athletes.merge(medals, on=['team', 'name'], how='left')
.drop_duplicates(subset=['name','team', 'medal'])
(
    a
    .assign(no_medal = a['medal'].isna())
    .groupby(['team'])
    .mean()
    ['no_medal']
)
```

Name: _____

Question 3 ..... ***4 points***

Let's imagine we've done a nice pivot table to look at what proportion of athletes won a medal for a few countries to help us understand how these teams decide how many athletes to send.

(a) (4 points) The below table is the output of that pivot. For the following statements decide if they are Definitely True, Definitely False or Need more information.

team	medals_per_athelete_USA	medals_per_athelete_JPN	medals_per_athelete_CNA	medals_per_athelete_GBR
gymnastics	0.2	0.32	0.2	0.21
athletics	0.3	0.10	0.1	0.35

Assume CHN and USA each sent 10 athletes to each of these two events (for a total of 20 athletes per team). China won more overall medals.

- ◯ Definitely True
- ◯ **Definitely False**
- ◯ Need more information

Solution: This is just math!

Assume JPN sent 20 athletes to these two events and GBR sent 15. These two countries show an example of Simpson's paradox.

- ◯ Definitely True
- ◯ Definitely False
- ◯ **Need more information**

Solution: We don't have a way to calculate the overall medal rate betwee these two groups with just the information given.

If the USA's overall medal rate for these two categories is 0.25 and GBR's overall medal rate is 0.22, this is an example of Simpson's paradox.

- ◯ **Definitely True**
- ◯ Definitely False
- ◯ Need more information

Solution: This is indeed Simpson's paradox. Even though team USA did worse within each individual event, their overall medal rate was higher which must be caused by a difference in how many athletes were sent to each event.

If the USA's overall medal rate for these two categories is 0.25 and GBR's overall medal rate is 0.22, GBR sent more athletes to gymnastics events than USA.

- ◯ Definitely True
- ◯ Definitely False
- ◯ **Need more information**

Solution: This needs more information since some athletes might not have won any medals.

Name: _____

Question 4 ..... ***15 points***

Let's use hypothesis testing to find some patterns in the Olympics!

For the following sets of hypotheses, select the type of test and all valid test statistics:

(a) (3 points) Do athletes over the age of 28 win more medals than all Olympic athletes?

Correct test:

- ◯ **Hypothesis Test**
- ◯ Permutation Test

Test statistic:

- ☐ AverageGold_older - AverageGold_all
- ■ **AverageMedalCount_older - AverageMedalCount_all**
- ☐ |AverageGold_older - AverageGold_all|
- ☐ K-S test

(b) (3 points) Are volleyball players taller on average than basketball players?

Correct test:

- ◯ Hypothesis Test
- ◯ **Permutation Test**

Test statistic:

- ■ **AverageHeight_v - AverageHeight_b**
- ■ **AverageHeight_v / AverageHeight_b**
- ☐ Total variation distance
- ☐ K-S test statistic

(c) (3 points) Did the US female athletes perform differently than their male counterpart?

Correct test:

- ◯ Hypothesis Test
- ◯ **Permutation Test**

Test statistic:

- ☐ AverageGold_F - AverageGold_M
- ■ **|AverageGold_F - AverageGold_M|**
- ■ **Total variation distance**
- ☐ K-S test statistic

(d) (3 points) Did team USA submit a different distribution of players per sport relative to CHN?

Correct test:

- ◯ Hypothesis Test
- ◯ **Permutation Test**

Test statistic:

- ☐ AverageGold_USA - AverageGold_CHN
- ☐ |ProportionGymnasts_USA - ProportionGymnasts_CHN|
- ■ **Total variation distance**
- ☐ K-S test statistic

(e) (3 points) Do athletes from Judo show the same distribution of weights as all athletes?

Correct test:

- ◯ **Hypothesis Test**
- ◯ Permutation Test

Test statistic:

- ■ **|AverageWeight_Judo - AverageWeight_all]**
- ☐ |AverageWeight_Judo - AverageWeight_Tennis|
- ☐ Total variation distance
- ■ **K-S test statistic**

Name: _____

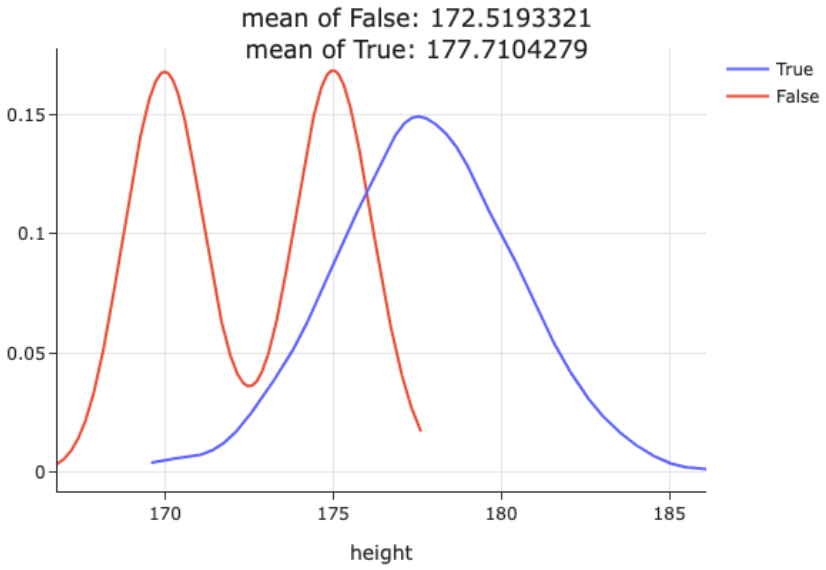
Question 5 ..... ***12 points***

Brendan got curious about some data patterns among Olympic athletes but must have pulled the data from an unreputable website because some of the data is missing from the `weight` column.

(a) (3 points) He wants to determine the missingness mechanism. Which of the following is a correct pairing of missingness Mechanisms and logical reasoning?

- ■ **MCAR, Brendan's internet connection is spotty and dropped some random packets when downloading the data**
- ☐ Missing by design, weight is irrelevant to the analysis the data creators cared about
- ■ **MAR, weights were only collected for American athletes**
- ■ **MAR, weights are more likely to be missing when `‘sport == Judo‘`**
- ☐ NMAR, weights are more likely to be missing when `‘sport == Judo‘`

(b) (3 points) Imagine we suspect that there is a MAR relationship between `‘height‘` and `‘weight‘` and we show the below plot of heights GIVEN whether the weight is missing (`weight_missing = True` or `weight_missing = False`. Which of the following are valid hypothesis pairs?



- ■ **NULL: Weights are not missing due to height.**
  ALTERNATIVE: Weights are missing from taller individuals.
- ■ **NULL: Weights that are missing and weights that are not missing come from the same distribution.**
  ALTERNATIVE: Weights that are missing and weights that are not missing come from different distributions.
- ■ **NULL: Weights that are missing and weights that are not missing come from the same distribution.**
  ALTERNATIVE: Weights that are missing come from a distribution with a higher average height than weights that are not missing.
- ☐ NULL: Weights are MCAR
  ALTERNATIVE: Weights are not MCAR

(c) (2 points) Which is the *most* appropriate test statistic for this test among methods discussed in class?

- ◯ **comparison of means**
- ◯ K-S test
- ◯ TDS

Name: _____

(d) (2 points) Having found a significant test result and rejecting the null, Brendan decided to impute missing weight. Which imputation method did he use based on the three lines of code below?

```
def impute(s):
    return s.fillna(s.mean())

weights_new = athletes.groupby('sport')['weight'].transform(impute).to_frame()
```

- ◯ Listwise deletion
- ◯ Mean imputation
- ◯ **Conditional mean imputation**
- ◯ Probabilistic
- ◯ Multiple Imputation

(e) (2 points) In 1-2 sentences, which method would you use? what are some benefits to it? There are multiple justifiable answers, so the important thing is to justify the benefits!

Solution:

Solution: Students could argue for a number of solutions, as long as the justification is appropriate, e.g. the conditional mean is good because we want to preserve the mean. Or conditional probabilistic is better because it doesn't reduce variance etc.

Name: _____

Question 6 ..... ***0 points***

Optional: Draw a Picture About UCSD Data Science (or just use this page for scratch work)