DSC 80 - Final Exam

December 6, 2022

Name:

PID:

Exam Version: A

By signing below, you agree that you will behave honestly and fairly during and after this exam. You should not discuss any part of this exam with anyone who has not yet taken it.

Signature:

Name of student to your **left**:

Name of student to your **right**:

Exam version of student to your **left**:

Exam version of student to your **right**:

(Write "N/A" if a wall/aisle is to your left/right.)

Instructions:

- Your exam version should be different from those to your left/right.
- Write your solutions to the following problems in the spaces provided.
- No calculators are permitted, but a few pages of notes are.
- Write your name or PID at the top of each sheet in the space provided.

(Please do not open your exam until instructed to do so.)

In the next five questions, assume you have access to a dataframe named `pts`, shown below:

	group	color	x	y
0	A	red	3	2
1	B	green	7	1
2	A	blue	2	5
3	A	red	5	3
4	B	blue	10	4
5	A	green	1	1

Problem 1.

What is the **type** of the result of the following line of code?

```
>>> pts.groupby('group')['x'].max()
```

- ◯ `int`
- ◯ `float`
- ◯ `str`
- ◯ `pandas.Series`
- ◯ `pandas.DataFrame`

Problem 2.

What is the result of the following code?

```
>>> pivot = pts.pivot_table(
    values='x',
    index='group',
    columns='color',
    aggfunc='count')
>>> pivot.loc['A', 'red']
```

Your answer should be in the form of a number (or possibly `NaN`). Your answer does **not** need to be exactly what is displayed by Python.

Problem 3.

Suppose `z` is a pandas `Series` containing the data shown below:

```
>>> z
5    20
0     1
2     3
Name: z, dtype: int64
```

Notice that the index of this `Series` does not match the index of `pts`.

Which of the following will be the result of running:

```
>>> pts['z'] = z
>>> pts
```

◯

	group	color	x	y	z
0	A	red	3	2	20.0
1	B	green	7	1	1.0
2	A	blue	2	5	3.0
3	A	red	5	3	NaN
4	B	blue	10	4	NaN
5	A	green	1	1	NaN

◯

	group	color	x	y	z
5	A	green	1	1	NaN
0	A	red	3	2	20.0
2	A	blue	2	5	3.0

◯

	group	color	x	y	z
0	A	red	3	2	1.0
1	B	green	7	1	NaN
2	A	blue	2	5	3.0
3	A	red	5	3	NaN
4	B	blue	10	4	NaN
5	A	green	1	1	20.0

◯ An exception will be raised because `z` is missing some of the rows that are in `pts`.

(The `pts` dataframe is shown again here for convenience.)

	group	color	x	y
0	A	red	3	2
1	B	green	7	1
2	A	blue	2	5
3	A	red	5	3
4	B	blue	10	4
5	A	green	1	1

Problem 4.

Suppose the `costs` dataframe contains the following data:

	color	cost
0	red	5
1	blue	2
2	purple	7

Suppose we run:

```
>>> res = pts.merge(costs, how='left')
```

How many rows will `res` have?

Problem 5.

Suppose we have defined:

```
def foo(ser):
    return (ser - ser.min()).max()
```

What will be the result of:

```
>>> pts.groupby('group')[['x', 'y']].aggregate(foo).loc['A', 'x']
```

Your answer should be in the form of a number.

PID or Name: _____

The Earth Impact Database, curated by the University of New Brunswick's Planetary and Space Science Centre, contains information on almost 200 impact craters caused by meteorites that have crashed into the Earth. In the next five questions, assume you have access to a dataframe named `impacts`, shown below:

	crater_name	state	country	target_rock	diameter_km	drilled
0	Ouarkziz	NaN	Algeria	Sedimentary	3.50	False
1	Glikson	NaN	Australia	Mixed	19.00	False
2	West Hawk	Manitoba	Canada	Crystalline	2.44	True
3	Boxhole	Northern Territory	Australia	Crystalline	0.17	False
4	Hummeln	NaN	Sweden	Crystalline	1.20	True
...	...	...	...	...	...	...
184	Beyenchime-Salaatin	NaN	Russia	Sedimentary	8.00	False
185	Ries	NaN	Germany	Mixed	24.00	True
186	Serpent Mound	Ohio	United States	Sedimentary	8.00	True
187	Clearwater East	Quebec	Canada	Mixed	26.00	True
188	Popigai	NaN	Russia	Mixed	90.00	True

189 rows × 6 columns

A short description of each column follows:

- `crater_name`: the name of the impact crater.
- `state`: if the crater is in the United States, the state containing the crater is listed here; otherwise, it is missing.
- `country`: the country containing the crater.
- `target_rock`: the type of rock that the crater is in.
- `diameter_km`: the diameter of the crater in kilometers.
- `drilled`: whether or not the crater has been drilled to analyze its contents.

Problem 6.

There are many missing values in the `state` column. Upon inspection, you find that a crater is missing a state if and only if the crater is not located in the United States.

What is the most likely type of the missingness in the `state` column?

◯ Not Missing at Random (NMAR)

◯ Missing at Random (MAR)

◯ Missing Completely at Random (MCAR)

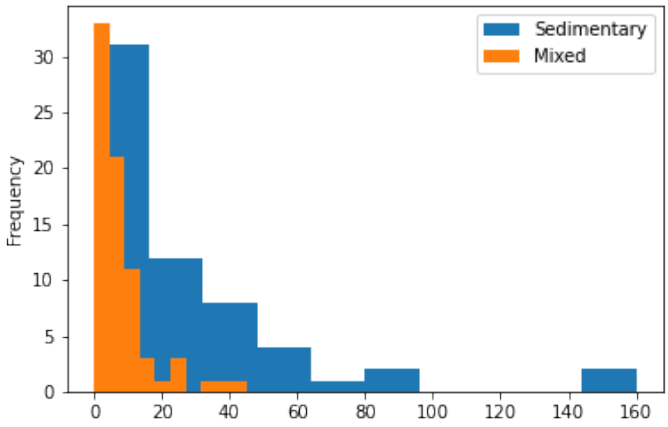
◯ Missing by Design (MD)

Problem 7.

Suppose we want to test the following hypotheses:

- **Null Hypothesis**: the crater diameter of impacts in sedimentary rock comes from the same distribution as the crater diameter of impacts in mixed rock.
- **Alternative Hypothesis**: the crater diameter of impacts in sedimentary rock is significantly larger on average.

When the distributions of crater diameter are plotted, we see the following:



Which one of the following is the best test statistic in this case?

◯ Total Variation Distance (TVD) between the distributions

◯ Kolmogorov-Smirnov (K-S) distance between the distributions

◯ the **signed** difference between the mean crater diameter of impacts in sedimentary rock, minus the mean crater diameter of impacts in mixed rock

◯ the **unsigned** (absolute) difference between the mean crater diameter of impacts in sedimentary rock, minus the mean crater diameter of impacts in mixed rock

Problem 8.

Suppose it is observed that some values in the `diameter_km` column are missing. To determine if there is an association between this missingness and the values in the `country` column, a permutation test will be performed with the null hypothesis that the distribution of countries when the diameter is missing is the same as the distribution of countries when the diameter is not missing.

Which of the following test statistics should be used?

◯ the Total Variation Distance (TVD) between the distribution of countries when the diameter is missing and the distribution of countries when the diameter is not missing

◯ the Kolmogorov-Smirnov statistic between the distribution of countries when the diameter is missing and the distribution of countries when the diameter is not missing

◯ the **signed** difference between the mean crater diameter of impacts where the country is missing, minus the mean crater diameter of impacts where the country is not missing

◯ the **unsigned** (absolute) difference between the mean crater diameter of impacts where the country is missing, minus the mean crater diameter of impacts where the country is not missing

Problem 9.

Suppose the permutation test described in the previous problem fails to reject the null hypothesis. Assuming that NMAR and MD have been ruled out already, what can be said about the missingness in `diameter_km`?

◯ it is MCAR

◯ it is MAR

◯ We cannot say for sure without first testing for an association between the missingness and the other columns besides `country`.

Problem 10.

Suppose we fill in the missing values in the `diameter_km` column by random sampling. That is, for each missing diameter, we randomly sample from the the set of observed diameters. You may assume that these samples are drawn from the uniform distribution on observed diameters, and that they are independent.

Assume that it is known that the missingness in the `diameter_km` column is MAR. Which of the following is true about the overall mean of the `diameter_km` column after imputation?

◯ It is likely to be an **unbiased** estimate of the true mean.

◯ It is likely to be a **biased** estimate of the true mean.

Questions 11-13 below reference the following HTML.

```
<html>
    <head>
        <title>ZOMBO</title>
    </head>

    <body>
        <h1>Welcome to Zombo.com</h1>
        <div id="greeting">
            <ul>
                <li>This is Zombo.com, welcome!</li>
                <li>This is Zombo.com</li>
                <li>Welcome to Zombo.com</li>
                <li>You can do anything at Zombo.com -- anything at all!</li>
                <li>The only limit is yourself.</li>
            </ul>
        </div>
        <div id="footnotes" class="faded">
            <h3>Footnotes</h2>

            <ol id="footnotes">
                <li>Please consider <a href="paypal.html">donating!</a></li>
                <li>Made in California with <a href="https://reactj.org">React.</a></li>
            </ol>
        </div>
    </body>
</html>
```

Problem 11.

Consider the node representing the `body` tag in the Document Object Model (DOM) tree of the above HTML. How many children does this node have?

Problem 12.

The page shown above contains five "greetings", each one a list item in an unordered list. The first "greeting" is "This is Zombo.com, welcome!", and the last is "The only limit is yourself.".

Suppose we have parsed the HTML into a `BeatifulSoup` object stored in the variable named `soup`.

Which of the following pieces of code will produce a list of `BeautifulSoup` objects, each one representing a single greeting list item? **Mark all which apply.**

☐ `soup.find('div').find_all('li')`

☐ `soup.find_all('li', id='greeting')`

☐ `soup.find('div', id='greeting').find_all('li')`

☐ `soup.find_all('ul/li')`

Problem 13.

Suppose you perform an HTTP request to a web API using the `requests` module. The response succeeds, and you get the following (truncated) content back:

```
>>> resp = requests.get("https://pokeapi.co/api/v2/pokemon/squirtle")
>>> resp.content
b'{"abilities":[{"ability":{"name":"torrent","url":"https://pokeapi.co"...
```

What type of data has been returned?

◯ HTML

◯ JSON

◯ XML

◯ PNG

Problem 14.

You are scraping a web page using `requests`. Your code has been working fine and returns the desired result, but suddenly you find that your code takes much longer to finish (if it finishes at all). It does not raise an exception.

What is the most likely cause of the issue?

◯ The page has a very large GIF that hasn't stopped playing.

◯ You have made too many requests to the server in a short amount of time

◯ The page contains a Unicode character that `requests` cannot parse

◯ The page has suddenly changed and has caused `requests` to enter an infinite loop.

In the next five questions, you will be asked to determine which strings are matched by various regular expression patterns when `re.search` is used. For these questions, remember that `re.search(pattern, s)` matches `s` if the pattern can be found anywhere in `s` (not necessarily at the beginning). For example, `re.search("name", "my name is justin")` matches, while `re.search("foo", "my name is justin")` does not.

Problem 15.

Which of the below strings are matched by `re.search` using the pattern `'a+'`? Select all that apply.

☐ `"aa bb cc"`

☐ `"aaa bbb ccc"`

☐ `"abaaba"`

☐ `"abacaba"`

Problem 16.

Which of the below strings are matched by `re.search` using the pattern `'a+ b+'`? Select all that apply.

☐ `"aa bb cc"`

☐ `"aaa bbb ccc"`

☐ `"abaaba"`

☐ `"abacaba"`

Problem 17.

Which of the below strings are matched by `re.search` using the pattern `r'\baa\b'`

Recall that the `r` at the front of the pattern string above makes it a "raw" string; this is used so that `\b` is not interpreted by Python as a special backspace character.

Select all that apply.

☐ `"aa bb cc"`

☐ `"aaa bbb ccc"`

☐ `"abaaba"`

☐ `"abacaba"`

Problem 18.

Which of the below strings are matched by `re.search` using the pattern `'(aba){2,}'`? Select all that apply.

☐ `"aa bb cc"`

☐ `"aaa bbb ccc"`

☐ `"abaaba"`

☐ `"abacaba"`

Problem 19.

Which of the below strings are matched by `re.search` using the pattern `'a..a'`? Select all that apply.

☐ `"aa bb cc"`

☐ `"aaa bbb ccc"`

☐ `"abaaba"`

☐ `"abacaba"`

Problem 20.

Which of the below strings are matched by `re.search` using the pattern `'.*'`? Select all that apply.

☐ `"aa bb cc"`

☐ `"aaa bbb ccc"`

☐ `"abaaba"`

☐ `"abacaba"`

Problem 21.

Consider the following four sentences:

- "this is one"
- "this is two"
- "this is the third"
- "and this is the fourth"

Suppose these sentences are encoded into a "bag of words" feature representation. The result is a dataframe with four rows (one for each sentence). How many columns are in this dataframe? Your answer should be in the form of a number.

Problem 22.

Again consider the same four sentences shown above.

What is the TF-IDF score for the word "this" in the first sentence? Use base-2 logarithm.

Your answer should be in the form of a number.

Problem 23.

Again consider the same four sentences shown above.

What is the TF-IDF score for the word "and" in the last sentence? Use base-2 logarithm.

Your answer should be in the form of a number.

PID or Name: _____

Problem 24.

Consider the dataframe shown below:

	group	color	x	y
0	A	red	3	2
1	B	green	7	1
2	A	blue	2	5
3	A	red	5	3
4	B	blue	10	4
5	A	green	1	1

Suppose you wish to use this data in a linear regression model. To do so, the `color` column must be encoded numerically.

True or False: a meaningful way to numerically encode the `color` column is to replace each string by its index in the alphabetic ordering of the colors. That is, to replace "blue" by 1, "green" by 2, and "red" by 3.

◯ True

◯ False

Problem 25.

Suppose you perform a one-hot encoding of a Series containing the following strings:

```
["red", "blue", "red", "green", "green", "purple", "orange", "blue"]
```

Assume that the encoding is created using the `OneHotEncoder(drop='first')` from `sklearn`. Note the `drop='first'` keyword argument: `skearn`'s documentation says that this will "drop the first category in each feature."

How many columns will the resulting one-hot encoding table have?

Problem 26.

Suppose you split a data set into a training set and a test set. You train a classifier on the training set and test it on the test set.

True or False: the training accuracy must be higher than the test accuracy.

◯ True

◯ False

Problem 27.

Suppose you are using `sklearn` to train a decision tree model to predict whether a data science student is enrolled in DSC 80 or not based on several pieces of information, including their hours spent coding per week and whether or not they have heard of the Kolmogorov-Smirnov test statistic.

Suppose you train your model, but achieve much lower training and test accuracies than you expect. When you look at the data and make predictions yourself, you are easily able to achieve higher train and test accuracies.

What should be done to improve the performance of the model?

◯ Decrease the `max_depth` hyperparameter; the model is "overfitting".

◯ Increase the `max_depth` hyperparameter; the model is "underfitting".

The next four questions concern the following example.

A Silicon Valley startup candy company buys a factory from a lightbulb company and repurposes some of the existing equipment to make chocolate bars with the goal of disrupting the candy industry. Their IPO is delayed, however, when they discover that some of their chocolate bars contain broken glass.

The company's engineers quickly build an AI which looks at the chocolate bars coming off of the conveyor belt and predicts which bars contain broken glass ("yes") and which don't ("no"). The results are shown in the following confusion matrix:

	Actually No	Actually Yes
Predicted No	87	6
Predicted Yes	3	4

Problem 28.

What is the accuracy of their model as a percentage (between 0% and 100%)?

Problem 29.

What is the recall of their model as a percentage (between 0% and 100%)?

Problem 30.

From a safety perspective, which metric should be maximized in this situation?

◯ Precision

◯ Recall

Problem 31.

Suppose the company's investors wish to improve the model's **precision**. Which should they do (besides hire better data scientists)?

◯ Lower their model's threshold for predicting that a bar contains glass, thus throwing out more candy.

◯ Raise their model's threshold for predicting that a bar contains glass, thus throwing out less candy.

Before turning in your exam, please check that your name is on every page.

After turning in your exam, have a good winter break!

You may detach and use this page for scratch work. You do not need to turn it in.

You may detach and use this page for scratch work. You do not need to turn it in.