DS2500
Spring 2025
Practice for Exam 2

This practice exam contains some sample questions on the same topics that will appear on Exam #2. Please complete the questions on your own, and we'll publish the solutions on Piazza and review in lab on March 31

Exam #2 takes place on April 4th, 2025. It will be administered on paper, during our usual lecture time.

The exam is designed to be shorter than our 100-minute lecture time, but you may use the entire class time to complete it. Don't rush, take your time with each question, and double-check your solutions before handing it in.

For the exam, you may bring one 8.5x11-inch piece of paper, with anything written or typed on it (one side only). You will submit this cheat sheet along with your exam, and you will not be permitted to use any other materials or notes during the exam.

Part One -- Python Programming

Practice Question 1
Suppose you're scraping a website that contains the source code below. Which of the following calls to a BeautifulSoup method would retrieve all of the descriptions?

```
<table>
<tr scope="row">
<td class="description" style="text-align:center">
      Growing up, McKay is coached...
</td>
<td class="description" style="text-align:center">
      Rue and Jules reconcile as Rue recovers...
</td>
</tr>
</table>
```

A. `bs.find_all("td", {"class" : "description"})`
B. `bs.find_all("class", {"class" : "description"})`
C. `bs.find_all("tr", {"scope" : "class"})`
D. `bs.find_all("Growing up", "Rue and Jules")`
E. `bs.find_all("<td></td>")`

Practice Question 2

Which Pandas method would you use to change the data type of a column (e.g., converting a column from string to integer)?

A) `df.astype()`
B) `df.convert_dtype()`
C) `df.change_type()`
D) `df.to_type()`

Practice Question 3

Which of the following file types are considered structured data formats? Select all that apply.

A. XML
B. HTML
C. Plaintext
D. JSON
E. Markdown

Practice Question 4

Suppose you make an API request for the weather forecast and get back JSON data that looks like the snippet below, and it is saved in a Python variable named *data*. Write a Python code snippet that would print out all of the *temp* values (no matter how many there are!):

```
{
  "cod": "200",
  "message": 0,
  "list": [
    {
      "dt": 1741284000,
      "main": { "temp": 285.13, "feels_like": 284.68 },
      "weather": [
        {"main": "Clouds"}
      ],
    }
    {
      "dt": 1741294800,
      "main": {"temp": 284.26, "feels_like": 283.72},
      "weather": [
        { "main": "Rain" }
            ],
      }
   ]
}
```

Practice Question 5

You have a CSV file named **`sales_data.csv`** that contains sales records with the following columns:

- `"Date"` (string)
- `"Product"` (string)
- `"Quantity"` (integer)
- `"Price"` (float)

The file is stored at the highest level on your personal Google Drive (`/content/drive/MyDrive`). Your job is to write separate cells that would accomplish the following tasks (comments are not required, and you may assume that all required libraries have been imported):

A. Mount your google drive and load the CSV file into a Pandas dataframe.

B. Add a new column called "Total Sales" that contains the total revenue for each row (`Quantity * Price`).

C. Replace any `NaN` values with 0

D. Created a new DataFrame that filters the original one so that only Total Sales greater than $500 are included.

Practice Question 6

What does the following code snippet print?

```
import numpy as np

ray1 = np.array([10, 20, 30])
ray2 = np.array([5, 6, 7])
print(ray1 * ray2 * 10)
```

Practice Question 7

Draw the image generated by the code below.

```
grid = np.zeros((2, 2, 3), dtype = int)
grid[0, 1] = [255, 255, 255]
plt.imshow(grid)
```

Part Two -- Data Science Algorithms

Practice Question 8

Given a dataset with range [0, 200], what is the min-max scaled value of 100?

A. 0.5
B. 0.75
C. 1.0
D. 0.25

Practice Question 9

For each of the following metrics, identify the correct range of possible values:

The *polarity* score in sentiment analysis?

Correlation coefficient (r-value)?

The scaled value in min-max normalization?

Practice Question 10

Which statement below best describes the Sigmoid Function used in logistic regression?

A. A linear activation function that maps any real number to a specific range of values
B. Outputs probabilities by mapping input values to a range between 0 and 1, making it useful for binary classification.
C. Computes the residual error in logistic regression by measuring the difference between predicted and actual values.
D. Transforms categorical variables into numerical values for use in regression models.

Practice Question 11
What are some possible downsides of KNN Classification? Select all that apply.

A. A poor choice of *k* can lead to overfitting or underfitting
B. The algorithm is computationally inefficient
C. The lack of labels on training data means accuracy is hard to measure
D. KNN never performs better than other classification algorithms such as decision trees and support vector machines.

Practice Question 12

Consider the following testing set, the result of running a KNN Classification algorithm on labelled data.

A	**B**	**Actual Label**	**Predicted Label**
10	5	1	0
8	6	1	0
7	7	0	1
1	12	1	1

A. Compute and fill in the following values:

   True Positives: _____

   True Negatives: _____

   False Positives: _____

   False Negatives: _____

B. What is the *accuracy* of the model?

Practice Question 13

Suppose you have a list of *actual* labels and a list of *predicted* labels for a testing set used in a KNN Classifier. You can assume that the two lists are the same length and non-empty. Complete the Python function below that computes the *precision* for the given label. Your function should return a float.

```
def compute_precision(actual, predicted, label):
```

What would your function return if I call it with `compute_precision([1, 1, 0, 1, 1], [0, 0, 1, 1, 1], 1)`?