TE WHARE WĀNANGA O TE ŪPOKO O TE IKA A MĀUI

VICTORIA
UNIVERSITY OF WELLINGTON

EXAMINATIONS – 2021

TRIMESTER 2

COMP 132

PROGRAMMING FOR THE NATURAL AND SOCIAL SCIENCES

Time Allowed: 120 minutes for questions; 10 minutes for loading data and submitting answers

OPEN BOOK

Permitted materials:

- Notes, calculators and dictionaries are permitted.
- Do not use phone, online chat, email etc to communicate with anyone.
- Do not post any questions or any answers.

Instructions: Before the test starts:

- Click the `Test2` link from the green menu on the left hand side of our course home page.
- Download the template file called `TermTest2.ipynb`, run all the cells and make sure there are no error messages.
- **Do not open this test paper yet.**

During the test:

- Save your file frequently to avoid data loss.
- The questions started with a * are harder and you can do them later.

At the end of the test:

- Save your file and do not modify your file after the cut-off time. Submit your file immediately after the test using our online submission system.

The test will be marked out of 100 and is worth 50% of your grade.

Questions

1. Programming basics [15]
2. Numpy and line plot [15]
3. Pandas [55]
4. Machine learning [15]

Question 1. Programming basics **[15 marks]**

(a) **[7 marks]** Define a function `calculate_fee(num)` to calculate the cost of the tickets for a group of people. The parameter `num` is the number of people in the group and you may assume it is always a valid positive integer. The returned value should be the cost.

- If the number of people is less than 3, then the ticket is $10 per person,
- If the number of people is in the range of [3,6] (3 to 6, inclusive), it is $8 per person.
- If the group has 7 people or more, it is $7 per person

You may test your function as follows:

`print(calculate_fee(2))` should print `20`
`print(calculate_fee(4))` should print `32`
`print(calculate_fee(8))` should print `56`

(b) **[8 marks]** Write a `for` loop to print the following table. The first column is an integer (1 to 10, inclusive) and the second column is the square root of the integer number rounded to two decimal places. Feel free to use any built-in libraries or functions.

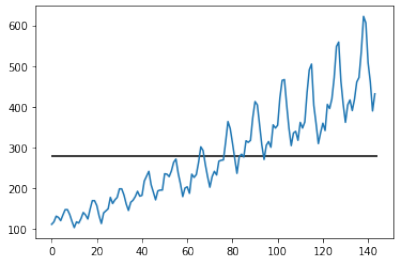
```
1 1.0
2 1.41
3 1.73
4 2.0
5 2.24
6 2.45
7 2.65
8 2.83
9 3.0
10 3.16
```

Question 2. Numpy and line plot **[15 marks]**

Run the following code in the template file to load data into a Numpy array.

```
flights = load_dataset('flights')
data = np.array(flights['passengers'])
data
```

(a) **[8 marks]** Make a line plot to show the data in the array and draw a horizontal line to show the average of the values. Your figure should look like this:



(b) **[7 marks]** Find all the elements that are larger than the average, and count them. Print these elements, their indexes and the number (count) of these elements.

Question 3. Pandas **[55 marks]**

Run the following code in the template file. It loads a built-in data set into a dataframe, and then deletes all the rows with missing data.

```
df_raw = load_dataset('penguins')
df = df_raw.dropna()
df
```

You should see a table with 7 columns and 333 rows .

The data contains size measurements for three penguin species observed on three islands in the Palmer Archipelago, Antarctica. Write code for the following tasks:

(a) **[5 marks]** Find and print the names of the three different species from the values in the **species** column.

(b) **[5 marks]** Show all the records of the `Adelie` penguins who live on the `Dream` island.

(c) **[5 marks]** Find out which species has the biggest average **body_mass_g**. (Show relevant data as evidence)

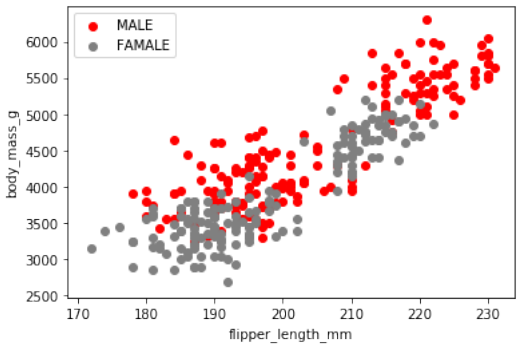
(d) **[5 marks]** Show the top 8 records with the biggest **flipper_length_mm**.

(e) **[5 marks]** Show the number of penguins in each species on each of the three islands.

(f) **[10 marks]** Show evidence for the following questions, and write answers (`True/False`) using comments.

- Is it true that the number of `MALE` and `FEMALE` penguins are equal in the dataset? Is this still true if we only consider the `Adelie` species?
- Is it true that at least half of the `MALE` penguins have bigger **body_mass_g** than the heaviest `FEMALE` one?

(g) **[10 marks]** * Create a scatter plot of **flipper_length_mm** as `x` and **body_mass_g** as `y`. Use different colours for different sex (`MALE, FEMALE`). Add `x, y` labels and a legend. Your figure should look as follows. The figure is printed black and white. You may use any two colours for the dots.



(h) **[10 marks]** * Add a new column called **size** and it should have just two values: `BIG` or `SMALL`. For each record, the value is `BIG` if the sum of its three measures **bill_length_mm, bill_depth_mm**, and **flipper_length_mm** is bigger than or equal to 250; otherwise, the value is `SMALL`.

Show the first 10 records of the dataframe with the new column.

(You can ignore the warning message or do this step on the original data `df_raw` to avoid it)

Question 4. Machine learning **[15 marks]**

Create a linear regression model to predict the **body_mass_g** values using two features: **bill_length_mm**, and **flipper_length_mm**.

Use the data collected on the `Biscoe` island as the training data, and the data on the `Dream` island as the test data.

Show the MSE (Mean Squared Error) on the test data. Use comments to explain why the error is big.

* * * * * * * * * * * * * * *