TE WHARE WĀNANGA O TE ŪPOKO O TE IKA A MĀUI

VICTORIA
UNIVERSITY OF WELLINGTON

EXAMINATIONS – 2020

TRIMESTER 2

COMP 132

PROGRAMMING FOR THE NATURAL AND SOCIAL SCIENCES

Time Allowed: 50 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:

- Create a file called `exam.ipynb`, and write your name and ID on top of the file as comments.
- **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.
- You can access the submission system by clicking the **Term Test 2** link from the green menu on the left hand side of our course home page.

The test will be marked out of 50 and is worth 15% of your grade.

Questions

1.	Data filtering and visualization	[22]
2.	Data analysis	[22]
3.	Machine learning	[6]

Use the following code to load a built-in data set into a dataframe data.

```
from seaborn import load_dataset
data = load_dataset('tips')
data
```

Run the code and you should see a table with 7 columns. Ask for help if you have any problems in this step.

The description of the data is as follows:

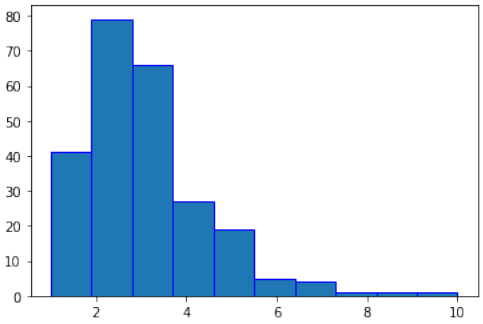
total_bill	Total bill (cost of the meal), including tax, in US dollars
tip	Tip (gratuity) in US dollars
sex	Sex of person paying for the meal (Male, Female)
smoker	Smoker in party? (No, Yes)
day	Thur, Fri, Sat, Sun
time	Lunch, Dinner
size	Size of the party (Number of customers)

Question 1. Data filtering and visualization **[22 marks]**

(a) **[2 marks]** Create a new dataframe with only the records for Saturday Dinner time.

(b) **[2 marks]** Create a new dataframe for the first 10 records with only two columns: **total_bill** and **tip**.

(c) **[4 marks]** Show the distribution of **tip**, and your figure should look like this:



(d) **[6 marks]** Create a bar plot to show the total number of customers at Lunch time and Dinner time. Please note that the number of customers in each party (e.g. each table) may be different.

(e) **[8 marks]** * Create a scatter plot of **total_bill** (as x) and **tip** (as y). Use different colors for different days (Thur, Fri, Sat, Sun). Add a legend. Also use the size of the party to decide the size of the dots, and draw the dots with reasonable sizes.

Question 2. Data analysis [22 marks]

(a) **[2 marks]** Show the top 10 records with the highest **total_bill**.

(b) **[2 marks]** Among the top 10, print the one where the maximum amount of tips is paid.

(c) **[10 marks]** * Write code to investigate gender differences in the data.

- Is it true that `Male` customers often pay for the meal?
- Is it true that a `Male` customer is more generous when paying tips than a `Female` customer?

  Hint: You need to consider **total_bill** for the second question. A person who paid $7 for a $100 bill was more generous than a person who paid $10 for a $200 bill. Please also note that the numbers of payments made by `Male` and `Female` are different.

  You code should print some numbers to show the evidence for each question.

(d) **[8 marks]** * Suppose the restaurant wants to encourage more people to come for `Lunch`, so it gives a 10% discount on the **total_bill** during `Lunch` time. Add a column to show the total amount to pay which includes the **total_bill** after the discount and the tips. Show the dataframe.

Question 3. Machine learning [6 marks]

(a) **[4 marks]** Use Simple Linear Regression to train a model to predict the **tip** value based on the **total_bill** value. To simplify the problem, please use ALL the data to train the model, so there is not a test dataset.

(b) **[2 marks]** Use the model to predict the tip value for an unseen **total_bill** value: e.g. 100.

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