TE WHARE WĀNANGA O TE ŪPOKO O TE IKA A MĀUI

VICTORIA
UNIVERSITY OF WELLINGTON

EXAMINATIONS – 2018

TRIMESTER 2

COMP 132

PROGRAMMING FOR THE NATURAL AND SOCIAL SCIENCES

Time Allowed: TWO HOURS

OPEN BOOK

Permitted materials: Notes, calculators and dictionaries are permitted.
You are not allowed to use phone, online chat, or email.
You are not allowed to post any questions or any answers.

Instructions: Before the examination starts, you may login to a lab computer, create a file called `exam.ipynb`, and write your name and ID on top of the file as comments. Do not open the examination paper.

During the examination, save your file frequently to avoid data loss. Do not spend too much time on a question if you get stuck.

At the end of the examination, save your file and do not modify your file after the cut-off time. Submit your file immediately after the examination using our online submission system. You can access the submission system by going to our course home page, clicking on “exam” from the green menu on the left hand side and then clicking the “submit” link.

The examination will be marked out of 120.

Questions

1.	Programming Basics	[20]
2.	Numpy	[10]
3.	Matplotlib	[15]
4.	Pandas	[55]
5.	Machine Learning	[20]

Question 1. Programming Basics **[20 marks]**

(a) **[10 marks]** Define a function that accepts a string parameter, calculates and returns the number of words in the string parameter. You may assume that every pair of words is separated by one space, there is no space at the end and the string is not empty.

Test your function using different arguments, for example,

`count_words("This is an example")` should return 4
`count_words("Carry on")` should return 2
`count_words("Done")` should return 1

(b) **[10 marks]** Write a loop to calculate the sum of the integer numbers from `1` to `100` except for the numbers which are the multiples of three. In other words, your code should calculate:
1 + 2 + 4 + 5 + 7 + 8 + 10 + 11 + 13 + 14 + 16 + ... + 100.

Question 2. Numpy **[10 marks]**

- Create a numpy array named `a` containing the following numbers: `8, 5, 3, 7, 2, 9, 1, 0, 4`
- Print the second and the fifth element of the array.
- Print the last three elements.
- Find the average of all the elements.
- Use array slicing to put the 3rd to 8th (both inclusive) elements in a new variable named `b` and make sure that `b` is a copy (not a view). Change the last element of `b` to `110`. Print both `a` and `b`.

Question 3. Matplotlib **[15 marks]**

Suppose the following numbers are the yearly average temperatures from the year 2001 to 2005 for City-A and City-B respectively.

City-A: `16.5, 15.5, 16.8, 15.6, 15.4`
City-B: `12.1, 10.1, 12.6, 13.3, 13.2`

- Create two numpy arrays to store the data.
- Plot the temperatures for the two cities in two steps:
  - Plot the temperatures of the two cities in the same figure, using different line styles, colors and markers, and different labels "City A" and "City B". Show these line-styles/colors/markers/labels in the figure legend.
  - Make the x-axis display 2001, 2002, 2003, 2004 and 2005.

Question 4. Pandas **[55 marks]**

The following online DataFrame (tsv file) contains a list of food orders in a restaurant.

https://bit.ly/chiporders

(a) **[5 marks]** Load the DataFrame and print the first 10 rows.

(b) **[5 marks]** Extract the last column of rows 25-35 (inclusive) into a variable and print it. Then print the data type of the result.

(c) **[5 marks]** Print the subset of the DateFrame in which “Side of Chips” has been ordered and “order_id” is bigger than 1000. Then print the shape of the result.

(d) **[5 marks]** Considering that NaN represents a missing value in this DataFrame, calculate the number of cells in which “choice_description” is missing.

(e) **[7 marks]** Find and print the number of times that “Chicken Burrito” has been ordered with “Tomatillo Red Chili Salsa”.

Note that food details such as “Salsa” are in the “choice_description” column.

(f) **[8 marks]** Calculate the total number of Chicken Burritos (total quantity) that have been ordered.

Note that the quantity of some orders is more than one.

(g) **[10 marks]** Calculate the total purchase price for the order with “order_id”= 1483.

Note that order 1483 occupies several rows in the DataFrame.

(h) **[10 marks]** Suppose we want to do some more data analysis on food ordered with Hot Salsa.

- Find the food items that are ordered with Hot Salsa.
  Note that it is possible that the two words may be separated in the description.
- Among the food items with hot salsa, find the food items for which the total number of this food item (total quantity) is greater than 10.
- Print the results of the second step, consisting of the food item names and the total quantities ordered.

Question 5. Machine Learning [20 marks]

The following online DataFrame (csv file) is a medical dataset that contains a list of **Head Size** and **Brain Weight** measurements. You will be using the DataFrame to predict the **Brain Weight** based on **Head Size**.

https://bit.ly/2PUdmgo

(a) **[5 marks]** Load the DataFrame. Split the DataFrame based on the default setting into a train set and a test set, assuming **Head Size** as the feature and **Brain Weight** as the outcome, and ignoring other columns.

Note that we only have one feature, but you will still need to use a list to represent the feature set.

(b) **[5 marks]** Use the training set to train a Linear Regression model with the default setting.

(c) **[5 marks]** Use the model to make predictions on the test set. Show a scatter plot with the **Brain Weight** true values in the x-axis and the predicted values on the y-axis.

(d) **[5 marks]** Evaluate your regression model by calculating the Root Mean Squared Error (RMSE) metric.

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