Student id: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Name: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

COMP 120: Test 2 (Part B)

# Overview

The goal of this test is to evaluate your ability to:

1. perform data manipulation operations
2. perform data aggregation operations
3. perform data reshaping operations
4. perform web scraping operations
5. handle relational data using relational operations
6. perform data modelling
7. read and write files
8. create visualisation using ggplot
9. utilize functions such as *is.na* and *ifelse* appropriately

This part of the test (part B) is worth 25% of your final mark. There are seven questions, with some containing sub-parts. **You must answer all the questions.** The marks for each question are indicated beside the question. If you need clarification about a question, you are welcome to contact the coordinator using the Zoom link provided. Only clarification questions are allowed. Please don’t call for hints about how to solve a problem (this is a test and not a lab!). You must complete the test by 5pm of 2nd October and submit the R file (test-2.R) through Blackboard.

## Documents and Instructions

For this test, you have been given *ten* documents:

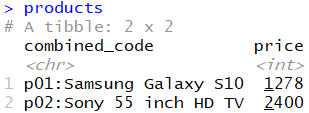
1. The instructions and questions file (this document you are reading) which is 5 pages long.
2. The other nine files are described below. Of these four are optional.
3. test-2.R is the file in which you will be recording your answers (code). Write your code in the placeholders for each question in this file. **You must upload the file on Blackboard by 5pm**.
4. A html file called products.html from which the data about products must be scraped.
5. Data files called products-raw.csv and products-processed.csv that contain details about products that may be used (optionally) as a part of questions 2 and 3 respectively.
6. Data files named customers.csv and sales.csv which you require for answering question 3.
7. Data file named output1.csv which may be used (optionally) as a part of questions 4 and 5.
8. Data file named house\_prices.csv which must be used as a part of question 6.
9. Data file named money\_and\_trophy.csv which may be used (optionally) as a part of question 7.

If you haven’t done so already, before you start answering the questions, save the above mentioned files in an appropriate folder under COMP120 folder (e.g., *practical\_test\_2* folder) and make this folder your working directory. Also, remember to set the .libPaths() correctly! Keep saving your test-2.R file at regular intervals of time so that you do not lose your work.

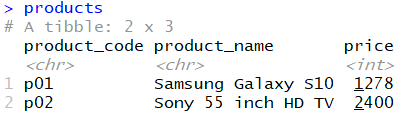
(QUESTIONS START IN THE NEXT PAGE - PLEASE TURN OVER)

# Questions

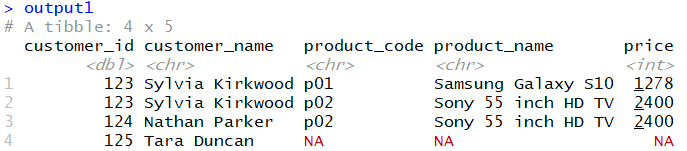
1. The file named products.html contains the details of two products in a table. Extract the data of these products into a *tibble* called products. Your code must use the pipe operator to connect the functions together into a single command. The expected output is shown below. [1.5 marks]



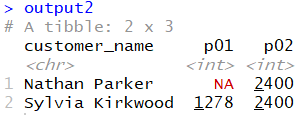
1. Each entry in the *combined\_code* column in the products dataset (shown above) contains two values combined using the “:” symbol. This column must be split into two columns. For example, the value “p01:Samsung Galaxy S10” has details about two aspects of a product, the product code (i.e., p01) and the product name (i.e., Samsung Galaxy S10). Split the data in the *combined\_code* column into two columns: *product\_code* and *product\_name*. After this separation, the output should resemble the snapshot given below. If you are not sure about the correctness of the *products* tibble you created for question 1, you can load the data from products\_raw.csv file. [1.5 marks]



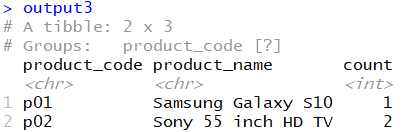
1. Inspect two files that have been given to you – customers.csv and sales.csv. These contain details about the customers and the sales they are involved with. The data in customers.csv contains two columns: *customer\_id* and *customer\_name*. The data in the sales.csv file contains two columns, *customer\_id* and *product\_code*. This file links the customers with the products they have bought. Using the data from these two files and the data in the *products* tibble you created in question 2 above (or the data stored in products\_processed.csv), write code to produce the outcome shown in the snapshot below. The resulting output should be stored in a variable called output1. [2 marks]



1. Starting with the output from the previous question (output1), use appropriate functions to produce the transformed data given in the snapshot below. The output below shows for each customer, their names and the prices corresponding to each product (given by its product code) they had purchased. Note that the results for customers who have not bought any products have been omitted. If you are unsure about your output for the previous question, load the data from output1.csv given to you. Store the result to a variable called output2. [3 marks]



1. Now, using the data obtained from output1 write code to produce the output shown below. (Note output1 was obtained as a part of result for question 3. Alternatively, you can also load the data from output1.csv). The output below shows the product code, product name and the count of each product that has been sold. Store the result in a variable called output3. [3 marks]



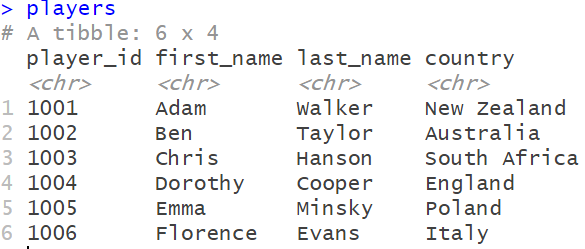
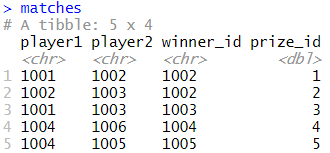
1. The data about house prices in a city is given in the file house\_prices.csv. Load that data into R as a tibble and inspect the columns. There are six columns: *price*, *bedrooms*, *bathrooms*, *sqft\_living*, *sqft\_lot* and *floors*. These depict the selling price, the number of bedrooms, the number of bathrooms, the living area in square feet, the land area in square feet and the number of floors in the house, respectively. [6 marks]
   1. Write code to create a **linear model** called mod1 that can be used to predict the price based on three explanatory variables, bedrooms, bathrooms and sqft\_living. Use all the data (100% of the data) in the dataset to create the model. Once the model has been constructed, write code to display the R-squared value.
   2. Next, write code to create a **randomForest model** called mod2 with four explanatory variables: three indicated in question 6a (see above) and sqft\_lot. Use all the data (100% of the data) in the dataset to create the model. Once the model has been constructed, write code to display the R-squared value.
   3. Using a comment discuss which of the two models might be a better model just based on comparing the R-squared values. You are not expected to conduct additional analyses (e.g. create and visualise residual plots).
   4. Finally, using the best model write code to predict the price of the house which has:
      * 4 bedrooms
      * 2 bathrooms
      * living area of 1000 square feet
      * land area of 5000 square feet

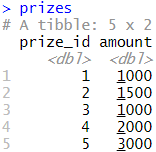
7. You have been given data pertaining to six international tennis players in four tibbles (players, matches, prizes, and trophies) whose snapshots are given below. The code to create these tibbles are already included in the test-2.R file given to you. The players tibble contains the *player\_id* that uniquely identifies the player. It also has details about their first\_name (*first\_name*), last name (*last\_name*) and the *country* they are from. The matches tibble contains the details about the matches played in the Dunedin Meadows event. Each row contains the first and second player’s id (given by *player1* and *player2* columns respectively), the id of the player who won the match (*winner\_id*) and also the id of the prize (*prize\_id*). The prizes tibble contains the *prize\_id* and the corresponding prize *amount* in New Zealand dollars. The trophies tibble contains the full names of the players as the column names, and the number of trophies each player has won in their career (i.e., trophy count) as values (i.e., these are the entries in the first (and only) row of the tibble).

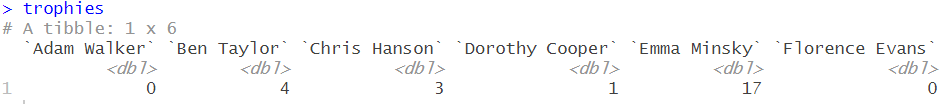
a) Using the information in the four tibbles and appropriate functions you have learnt in this course, write code that computes the sum of prize money won by each player, and shows the trophies won, and stores it in a tibble called money\_and\_trophy. This tibble when printed produces the result shown in the snapshot below in the box highlighted in green (containing six rows and four columns). You must use pipes to create the money\_and\_trophy tibble. Not using pipes throughout will attract a penalty of 0.5 mark.

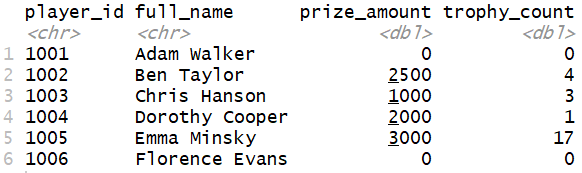
b) Using the money\_and\_trophy tibble created for question 7a (see above), write code to create the visualisation given in the next page. If you are unsure about your output for the money\_and\_trophy tibble you have created, load the data from money\_and\_trophy.csv given to you.

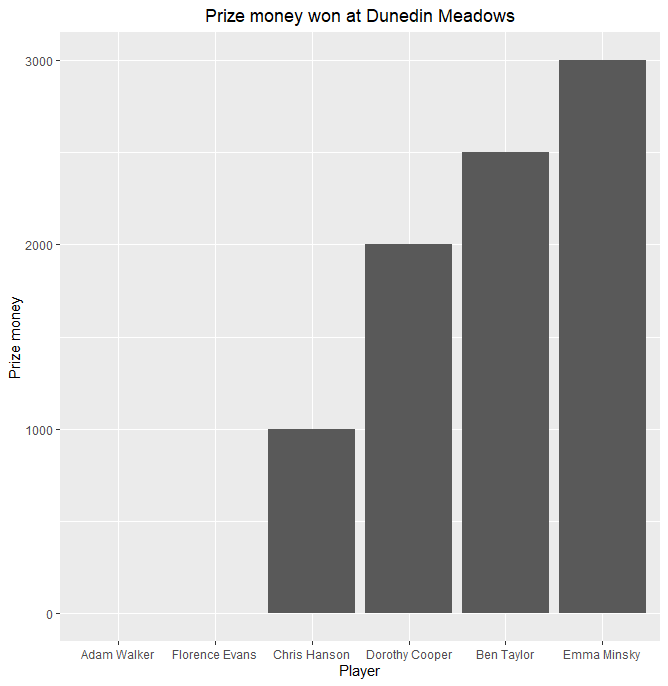
[**8 marks** (6 for creating *money\_and\_trophy* tibble and 2 for visualisation)]









**IMPORTANT: After completing the test, upload test-2.R on Blackboard by 5pm!**