## Institution of Vocational Education

**DEPARTMENT OF INFORMATION TECHNOLOGY**

**(Tsing Yi)**

**Higher Diploma in Data Science and Analytics ITP4887 Big Data Management Semester 4 2021-2022**

**Written Test**

## Question and Answer Booklet

**Instruction:**

1. **Students are forbidden to have any communications during the test, otherwise, we will judged as cheating.**
2. **This test is a close book test.**
3. **All answers should write in this question and answer booklet.**
4. **Any Drafts on this question booklet is allowed.**
5. **Full mark of this test: 80 marks Resource Required:**
   1. **Your Soul and your brain**
   2. **Test Question booklet**

**Class:**

**ID:**

**Name:**

Date: /11/2020 Time Allowed: 90 Mins

# Part A Multiple Choice(s) 20%, TEN Questions 2% e.a.

## Please write your answer(s) in the box .

**(Hint: Each question may contain one or two answers or even no suitable answer, please cross out ** **the box if there is no proper answer. Marks only be awarded for choosing ALL appropriate answer(s).)**

* + 1. Which of the following(s) is/are an essential element of the lifecycle of a Big Data Project?
       1. Financial Planning
       2. Human Resource Management
       3. Data Visualization

C

D

* + - 1. Business Case Evaluation
    1. Which of the following(s) is/are the correct description of Data Dictionary.
       1. Data Dictionary is essential information in the Crisp- DM approach only.
       2. Data Dictionary can explain words like Jargon or rare vocabularies.
       3. Data Dictionary can introduce the partial impression of the Data set.
       4. There is no such item as called Data Dictionary in any Big Data Project.

C

B

* + 1. Which of the following package(s) is/are used for web crawling purposes in R?
       1. Beautiful Soup 3
       2. Rvest

D

B

* + - 1. Request
      2. Selenium
    1. Data.Frame() is included in the package.
       1. Pandas
       2. DataFrame
       3. Base

C

D. Table

* + 1. Which of the following is/are an appropriate **web data acquisition** approach(es) in Big Data Project Lifecycle?
       1. Web Scrapping from a website.
       2. Get data by sample() in R.

D

A

* + - 1. Scan a paper data into the computer.
      2. Retrieve Data from the authorised API package.
    1. Which of the following is/are concerning about Agile development principles?
       1. The term "Agile" means a fast and time effectiveness development approach.
       2. A face-to-face conversation is the best form of communication
       3. Customer satisfaction by early and continuous delivery of valuable software

C

B

* + - 1. Agile development is a series of water-fall lifecycle approaches.
    1. Which of the following is the main role(s) of Scrum?
       1. Scrum Master
       2. Product Owner
       3. Scrum Team

B

A

* + - 1. Product Analyst
    1. Each stage of the sprint should last a length period in Scrum.
       1. Flexible
       2. Continuous

C

* + - 1. Fixed
      2. Seven days
    1. Who is/are responsible for listing the backlog item in the Scrum approach?
       1. Scrum Master
       2. Client

D

* + - 1. Development Team
      2. Product Owner
    1. proc can manipulate data by a query language.
       1. QBE
       2. SQL

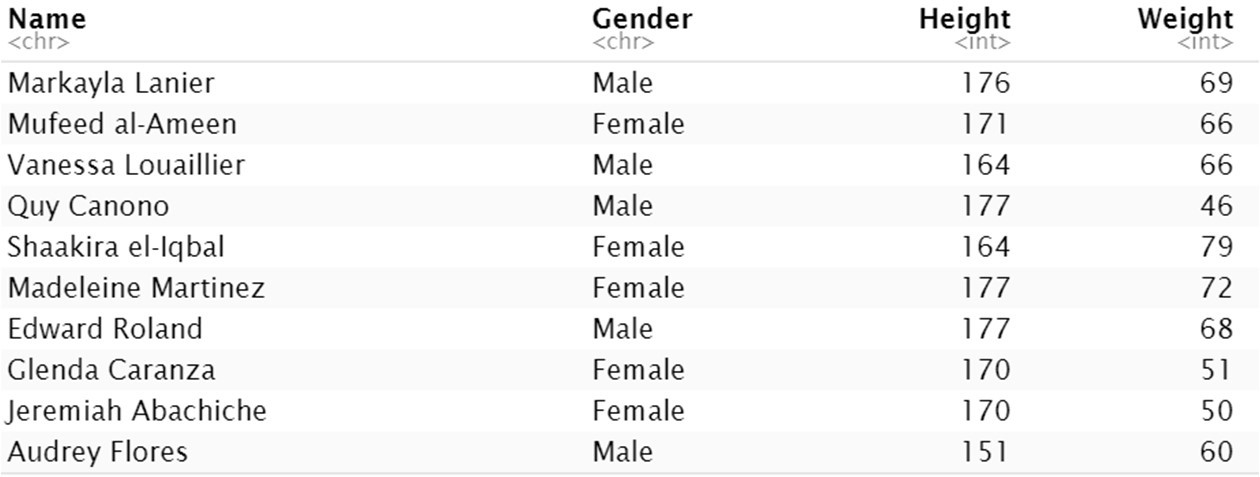
C

* + - 1. Print
      2. Template

# Part B Short Questions 60%, FOUR questions 15% e.a.

## Question B1

There is a dataset stored in an object BodyMass:



Write R code to

1. Update the Name "Vanessa Louailler" to "Benson Lau". [5 marks]

BodyMass$Name[BodyMass$Name == “Vanessa Louailler”] <- “Benson Lau”

1. find all the Male that their Height is taller than 170.

The result only includes the column "Name" and "Height".

[5 marks]

result <- BodyMass[BodyMass$Gender == “Male” & BodyMass$Height > 170,c(“Name”, “Height”)]

~Question B1 Continue~

1. Delete the column Gender. [2 marks]

BodyMass$Gender <- NULL

1. Suppose the unit of the Height and the Weight is centimetre and kilogramme, respectively. Then, add a new column "BMI" to the "BodyMass", and the value should base on the below formula:

𝑊𝑒𝑖𝑔ℎ𝑡 𝑖𝑛 𝑘𝑖𝑙𝑜𝑔𝑟𝑎𝑚𝑚𝑒

[3 marks]

𝐵𝑀𝐼 =

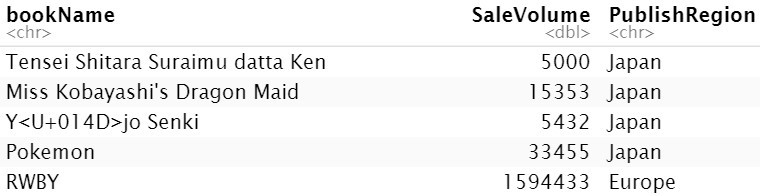
(𝐻𝑒𝑖𝑔ℎ𝑡 𝑖𝑛 𝑚𝑒𝑡𝑟𝑒)2

BodyMass$BMI <- BodyMass$Weight / (BodyMass$Height / 100)^2

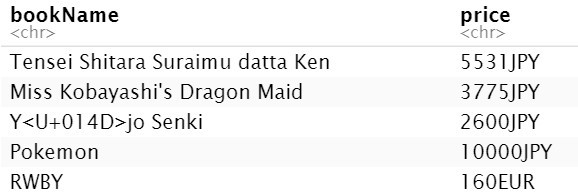
~Question B1 End~

## Question B2

There are datasets, and one of the datasets is stored in an object BookSale:



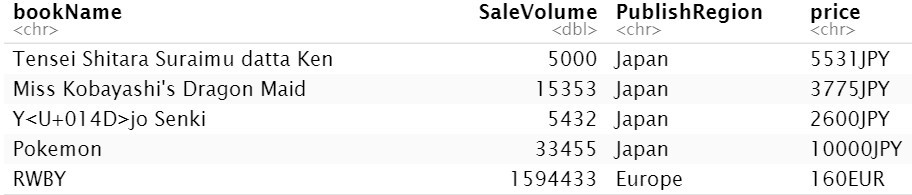
Another one of the datasets is stored into an object BookInformation:



Write R code to

1. combine the above datasets to a single dataframe object Bookstat by sqldf(). Assume the sqldf package is imported. And, below is the sample for reference.

[5 marks]



~Question B2 Continue~

1. Exchange all the currency of the price to HKD. The following is the exchange rate:
   1. JPY : HKD = 1: 0.070755877
   2. EUR : HKD = 1 : 9.1951366

[8 marks]

1. calculate the revenue by Publishregion by tapply(). (hint: Revenue = SaleVolume x price)

[2 marks]

~Question B3 Continue~

You are required to scape the data from the following website structure:

<body>

<a *href*="/notice/12378.html">This is the result notice...</a>

<a *href*="/news/17377.html">The latest news </a>

</body>

Write R code to

1. read the HTML document from the webpage "[http://IT114116.edu.hk](http://IT114116.edu.hk/)" and store the result to the object html.

html <- read\_html(“http://IT114116.edu.hk”)

[5 marks]

1. Extract elements from all anchor tags and store into the object "node".

(Hint: you can use the value from previous answer)

[5 marks]

node <- html\_elements(html, “a”)

1. extract the URL link from the anchor tag and store it in full URL format into the object URL.

Full URL example: "<http://IT114116.edu.hk/notice/12378.html>" (Hint: you can use the value from previous answer)

URL <- html\_attr(node,"href")

[5 marks]

## Question B4

Suppose the membership dataset HoroData.csv is uploaded into SAS OnDemand Folder with the following path:

"/home/u44771062/Data/HoroData.csv"

The following is the data dictionary of the HoroData.csv

|  |  |  |  |
| --- | --- | --- | --- |
| **Column** | **Data Type** | **Description** | **Sample** |
| **Name** | String | Name of the member | Kim, Dennis |
| **Horo** | String | Horoscope of the member | Gemini |
| **Height** | Integer | Height of the member | 156 |
| **Weight** | Integer | Weight of the member | 84 |

Write SAS code to

1. import data HoroData.csv from the provided path and store it to variable Horo [3 marks]

proc import

out = Horo

datafile = ‘/home/u44771062/Data/HoroData.csv’

dbms = csv

replace;

run;

1. display SEVEN rows of data that the Horoscope is Gemini. [2 marks]

proc print data = Horo(obs=7);

WHERE Horo = ‘Gemini’;

run;

~Question B4 Continue~

1. Plot a bar chart about the average Weight of members by their Horoscope by

template and sgrender proc.

[10

proc template;

define statgraph barchart;

begingraph;

entrytitle “bar chart”;

layout lattice / columns = 1 rows = 1;

barchart category=Horo response = height / stat = mean;

endlayout;

endgraph;

end;

run;

proc sgrender data = Horo template = barchart;

run;

marks]

~End of Question B4 and paper~