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?

A. Financial Planning
B. Human Resource Management
C. Data Visualization
D. Business Case Evaluation

□□

2 Which of the following(s) is/are the correct description of Data Dictionary.

A. Data Dictionary is essential information in the Crisp-DM approach only.
B. Data Dictionary can explain words like Jargon or rare vocabularies.
C. Data Dictionary can introduce the partial impression of the Data set.
D. There is no such item as called Data Dictionary in any Big Data Project.

□□

3 Which of the following package(s) is/are used for web crawling purposes in R?

A. Beautiful Soup 3
B. Rvest
C. Request
D. Selenium

□□

4 `Data.Frame()` is included in the ________ package.

A. Pandas
B. DataFrame
C. Base
D. Table

□□

5 Which of the following is/are an appropriate **web data acquisition** approach(es) in Big Data Project Lifecycle?

A. Web Scrapping from a website.
B. Get data by `sample()` in R.
C. Scan a paper data into the computer.
D. Retrieve Data from the authorised API package.

□ □

6 Which of the following is/are concerning about Agile development principles?

A. The term "Agile" means a fast and time effectiveness development approach.
B. A face-to-face conversation is the best form of communication
C. Customer satisfaction by early and continuous delivery of valuable software
D. Agile development is a series of water-fall lifecycle approaches.

□ □

7 Which of the following is the main role(s) of Scrum?

A. Scrum Master
B. Product Owner
C. Scrum Team
D. Product Analyst

□ □

8 Each stage of the sprint should last a _______ length period in Scrum.

A. Flexible
B. Continuous
C. Fixed
D. Seven days

□ □

9 Who is/are responsible for listing the backlog item in the Scrum approach?

A. Scrum Master
B. Client
C. Development Team
D. Product Owner

□ □

10 ______ proc can manipulate data by a query language.

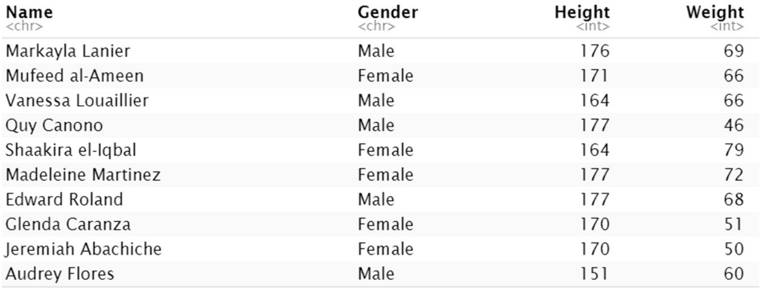
A. QBE
B. SQL
C. Print
D. Template

□ □

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

Question B1

There is a dataset stored in an object BodyMass:



Name	Gender	Height	Weight
<chr>	<chr>	<int>	<int>
Markayla Lanier	Male	176	69
Mufeed al-Ameen	Female	171	66
Vanessa Louaillier	Male	164	66
Quy Canono	Male	177	46
Shaakira el-Iqbal	Female	164	79
Madeleine Martinez	Female	177	72
Edward Roland	Male	177	68
Glenda Caranza	Female	170	51
Jeremiah Abachiche	Female	170	50
Audrey Flores	Male	151	60

Write R code to

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

b. find all the Male that their Height is taller than 170. [5 marks]
The result only includes the column "Name" and "Height" .

~Question B1 Continue~

c. Delete the column Gender. [2 marks]

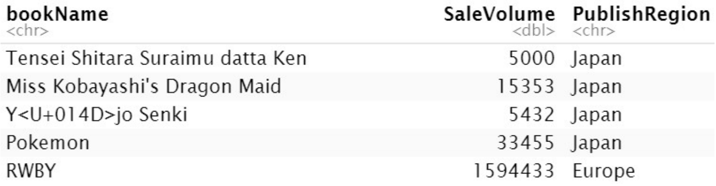
d. Suppose the unit of the Height and the Weight is centimetre and kilogramme, respectively. Then, add a new column "BMI" to the "BodyMass", and the value should base on the below formula: [3 marks]

$$BMI = \frac{Weight\ in\ kilogramme}{(Height\ in\ metre)^2}$$

~Question B1 End~

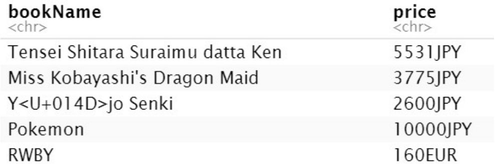
Question B2

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



bookName	SaleVolume	PublishRegion
<chr>	<dbl>	<chr>
Tensei Shitara Suraimu datta Ken	5000	Japan
Miss Kobayashi's Dragon Maid	15353	Japan
Y<U+014D>jo Senki	5432	Japan
Pokemon	33455	Japan
RWBY	1594433	Europe

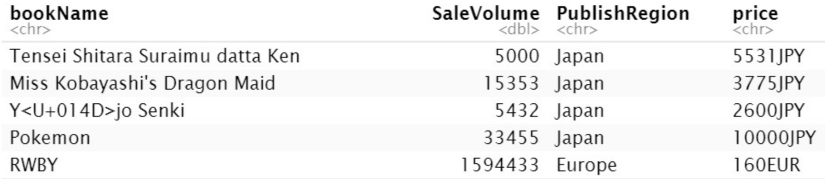
Another one of the datasets is stored into an object `BookInformation`:



bookName	price
<chr>	<chr>
Tensei Shitara Suraimu datta Ken	5531JPY
Miss Kobayashi's Dragon Maid	3775JPY
Y<U+014D>jo Senki	2600JPY
Pokemon	10000JPY
RWBY	160EUR

Write R code to

a. 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]



bookName	SaleVolume	PublishRegion	price
<chr>	<dbl>	<chr>	<chr>
Tensei Shitara Suraimu datta Ken	5000	Japan	5531JPY
Miss Kobayashi's Dragon Maid	15353	Japan	3775JPY
Y<U+014D>jo Senki	5432	Japan	2600JPY
Pokemon	33455	Japan	10000JPY
RWBY	1594433	Europe	160EUR

~Question B2 Continue~

b. Exchange all the currency of the price to HKD. The following is the exchange rate: [8 marks]

i. JPY : HKD = 1: 0.070755877

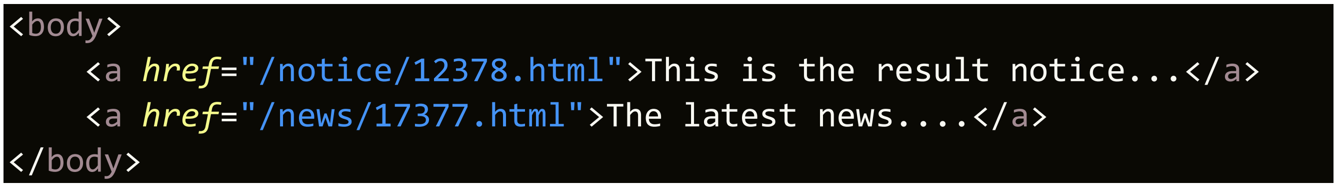
ii. EUR : HKD = 1 : 9.1951366

c. calculate the revenue by Publishregion by `tapply()`. [2 marks]
(hint: Revenue = SaleVolume x price)

~Question B3 Continue~

Question B3

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

a. read the HTML document from the webpage `"http://IT114116.edu.hk"` and store the result to the object `html`. [5 marks]

b. Extract elements from all anchor tags and store into the object `"node"`. [5 marks]
(Hint: you can use the value from previous answer)

c. extract the URL link from the anchor tag and store it in full URL format into the object URL. [5 marks]
Full URL example:
`"http://IT114116.edu.hk/notice/12378.html"`
(Hint: you can use the value from previous answer)

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

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

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

~Question B4 Continue~

c. Plot a bar chart about the average Weight of members by their Horoscope by `template` and `sgrender` proc. [10 marks]

~End of Question B4 and paper~