Institution of Vocational Education

DEPARTMENT OF INFORMATION TECHNOLOGY
(Tsing Yi)

Higher Diploma in Data Science and Analytics

ITP4887 Big Data Managment

Semester 4 2022-2023

Written Test for absentees

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: 100 marks**

Resource Required:

1. **Your Soul and your brain.**
2. **Test Question and answer booklet.**

Class:	**ID:**

Name:

Date: 3/1/2023

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 SAS code able to arrange two plotting in a **vertically** sequenced display?

A. `layout lattice / rows=2`
B. `lattice columns=2 rows=2`
C. `lattice layout / columns=2 rows=1`
D. `layout lattice / columns=1 rows=2`

☐☐

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

A. Best architectures, requirements, and designs emerge from self-organizing teams
B. The term "Agile" means a time effectiveness development approach.
C. Sustainable development, able to maintain a constant pace.
D. Agile development is a series of water-fall lifecycle approaches only.

☐☐

3 Which of the following(s) is the correct description of the V-model development approach?

A. The two life Cycles in the model work in series.
B. V-model is specific for a large-scale and complex development project.
C. Requirement Analysis and Module design are essential stages in V-model.
D. Each stage has its corresponding testing for verification.

☐☐

4 Which of the following(s) is/are **NOT** the feature of Big Data?

A. Visualization
B. Velocity
C. Volume
D. Variety

5 Which of the following(s) is/are appropriate lifecycle for Data Science Project for achieving client requirements?

A. Knowledge Discovery from Database
B. SEMMA
C. Cross-industry standard process for data mining
D. Water Fall

6 Which of the following(s) is/are the correct description of the Data Dictionary?

A. Data Dictionary is essential information in the Crisp-DM approach.
B. Data Dictionary can explain words like Jargon or rare vocabulary.
C. Data Dictionary can introduce a partial impression of the Data set.
D. Data Dictionary does not allow data to change its structure and fields.

7 Which of the following(s) is/are the feature of the R programme?

A. Recursion
B. Vectorization
C. Recycle
D. Factorization

8 What is the value of the variable **final** after the following R codes was executed?

```
Vector_A <- c(2,3,4,5,6)
Vector_B <- c(4,5,7,5)
final <- Vector_A + Vector_B
```

A. `6,8,11,10,10`
B. `6,8,11,10`
C. `45`
D. `Syntax error`

9 What is the appropriate description of the functionality of `sqldf()`?

A. Translate a MongoDB query to an SQL query.
B. An API package that connects to SQL database.
C. It is used to manage data by SQL query and returns a dataframe.
D. Convert Json data to dataframe.

10 Which of the following R code(s) is/are appropriate translations of the below python code?

```
random.choice(["DSA", "No.1", "Sure"])
```

A. `Random(c("DSA", "No.1", "Sure"))`
B. `sample(c("DSA", "No.1", "Sure"))`
C. `c("DSA", "No.1", "Sure")[as.integer(runif(1,1,4))]`
D. `c("DSA", "No.1", "Sure")%>%choice()`

Part B Short Questions 80%, FOUR questions 20% e.a.

Question B1

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

```
<body>
    <table border="1">
        <tr>
            <td>ReleaseDate</td>
            <td>Movie Name</td>
            <td>Budget</td>
        </tr>
        <tr>
            <td>Apr 23, 2019</td>
            <td>Avengers: Endgame</td>
            <td>$400,000,000</td>
        </tr>
        <tr>
            <td>Aug 26, 2020</td>
            <td>Tenet</td>
            <td>$205,000,000</td>
        </tr>
    </table>
</body>
```

Write R code to

a. Request the server to return an HTML document and store the result to the object html. The address of the server is given to the object **url**. [4 marks]

b. Extract the content of table tag and store it into the object "node". (Hint: you can use the value from the previous answer) [4 marks]

c. Convert the extracted value in node to Dataframe and store it into object **budget**. (Hint: you can use the value from the previous answer) [4 marks]

~Question B1 Continue~

d. Assume that the data type of the release date is a string. Convert the release date to date type. (Hint: you can use the value from the previous answer) [4 marks]

ReleaseDate	MovieName	Budjet
Apr 23, 2019	Avengers: Endgame	400000000
Aug 26, 2020	Tenet	205000000

e. Find all movies that were released in April 2022 [4 marks]
(Hint: you can use the value from the previous answer)

~Question B1 End~

Question B2

There is a student dataset in a common separate format named **student.csv**. Below is the data sample of the dataset. Suppose all variables are shareable within question **B2**.

StudentID	Name	ProgrammeCode
19553027	al-Karim, Musaaid	SE
20830701	Hart, Jacqueline	ES
21814169	Ezzat, Forest	SE
20736830	Matthews, Catarina	Networking
21533283	Williams, Timothy	Networking

Write R code to

a. import the dataset to the working environment and store it to the object **student**. Suppose the user will select the dataset file through the file directory. [4 marks]

b. extract each student's registration year from the studentID using the stringr package and add the data to the dataframe as a new column named **registrationyear**. Suppose the first two digits of studentID are the registration year. [6 marks]

StudentID	Name	ProgrammeCode	registrationyear
20972776	al-Karim, Musaaid	SE	20
21166560	Hart, Jacqueline	ES	21
19634557	Ezzat, Forest	SE	19

c. amend the inappropriate ProgrammeCode with respect to the below reference table, named **ref**, using `merge()`. [10 marks]

Wrong	Correct
DSANo.1,D SA	DSA
ES, Software	SE
GDS, GG	GSD

~Question B2 End~

Question B3

This is Climate Data recorded from 1st July 2021 to 31st August 2021. The data contains the date recorded, temperature (°C) , humidity(%), and WindSpeed(m/sec), then stored it into the dataframe **Record**. The following is some data sample:

date <date>	temperature <int>	humidity <int>	WindSpeed <int>
2021-07-01	32	31	18
2021-07-02	33	43	16
2021-07-03	28	35	15
2021-07-04	30	37	17
2021-07-05	33	47	19
2021-07-06	33	37	18
2021-07-07	31	32	17
2021-07-08	33	47	15
2021-07-09	29	30	20
2021-07-10	29	43	18

Write R code to

a. calculate the water vapour pressure value and add it to the dataframe as a new column named **WVP.** This is the equation of water vapour pressure: [5 marks]

$$\text{WVP} = \left(\frac{\text{humidity}}{100}\right) * 6.105 * \exp\left(17.27 * \text{temperature}/(237.7 + \text{temperature})\right)$$

b. calculate the apparent temperature value and add it to the dataframe as a new column named AT. This is the equation of apparent temperature. [5 marks]

$$\text{AT} = 1.04 * temperature + 0.2 * WVP + 0.65 * windspeed - 2.7$$

c. calculate the average apparent temperature by weekday. [5 marks]

```
  Friday    Monday  Saturday    Sunday  Thursday   Tuesday Wednesday
21.00000  20.33333  19.77778  20.77778  20.88889  19.77778  20.62500
```

d. delete the columns humidity and WVP. [5 marks]

~Question B3 End~

Question B4

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

`"/home/u44771062/Data/Ballot.csv"`

The following is the data dictionary of the Ballot.csv

Column	Data Type	Description	Sample
Age	Integer	The age of the voter	17
Gender	String	Gender of the voter	M
Candidate	String	The target candidate	Houshou Marin

Write SAS code to

a. import dat.csv from the provided path and store it in variable **ballot**. [5 marks]

b. Find all Female vote that younger than 30. [5 marks]

c. Plot proportion of voter gender by `template` and `sgrender` proc. [10 marks]

~End of Question B4 and paper~