## 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?
       1. layout lattice / rows=2
       2. lattice columns=2 rows=2

D

* + - 1. lattice layout / columns=2 rows=1
      2. layout lattice / columns=1 rows=2
    1. Which of the following is/are concerning about Agile development principles?
       1. Best architectures, requirements, and designs emerge from self-organizing teams
       2. The term "Agile" means a time effectiveness development approach.
       3. Sustainable development, able to maintain a constant pace.

C

A

* + - 1. Agile development is a series of water-fall lifecycle approaches only.
    1. Which of the following(s) is the correct description of the V-model development approach?
       1. The two life Cycles in the model work in series.
       2. V-model is specific for a large-scale and complex development project.
       3. Requirement Analysis and Module design are essential stages in V-model.
       4. Each stage has its corresponding testing for verification.

D

C

* + 1. Which of the following(s) is/are **NOT** the feature of Big Data?
       1. Visualization
       2. Velocity
       3. Volume
       4. Variety
    2. Which of the following(s) is/are appropriate lifecycle for Data Science Project for achieving client requirements?
       1. Knowledge Discovery from Database
       2. SEMMA
       3. Cross-industry standard process for data mining
       4. Water Fall
    3. Which of the following(s) is/are the correct description of the Data Dictionary?
       1. Data Dictionary is essential information in the Crisp- DM approach.
       2. Data Dictionary can explain words like Jargon or rare vocabulary.
       3. Data Dictionary can introduce a partial impression of the Data set.

C

B

* + - 1. Data Dictionary does not allow data to change its structure and fields.
    1. Which of the following(s) is/are the feature of the R programme?
       1. Recursion
       2. Vectorization
       3. Recycle

B

A

* + - 1. Factorization
    1. 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

* + - 1. 6,8,11,10,10
      2. 6,8,11,10

A

* + - 1. 45
      2. Syntax error
    1. What is the appropriate description of the functionality of

sqldf()?

* + - 1. Translate a MongoDB query to an SQL query.
      2. An API package that connects to SQL database.
      3. It is used to manage data by SQL query and returns a dataframe.

C

* + - 1. Convert Json data to dataframe.
    1. Which of the following R code(s) is/are appropriate translations of the below python code?

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

* + - 1. Random(c("DSA", "No.1", "Sure"))
      2. sample(c("DSA", "No.1", "Sure"))
      3. c("DSA", "No.1",

"Sure")[as.integer(runif(1,1,4))]

C

B

* + - 1. 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

1. 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**.

html <- read\_html(url)

[4 marks]

1. 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]

node <- html\_elements(html, “table”)

1. 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]

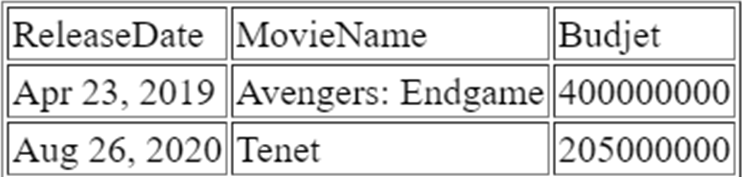
data <- html\_text(node)

budget <- data.frame(data)

5

~Question B1 Continue~

1. 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)



budget$ReleaseData <- as.Date(budget$ReleaseData, format = "%b %d, %Y")

1. Find all movies that were released in April 2022

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

[4 marks]

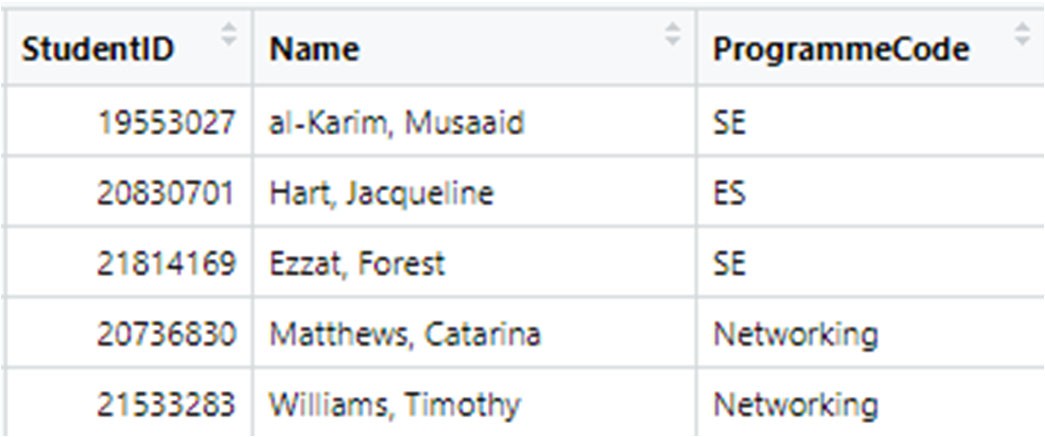
movie <-budget[budget$ReleaseDate >= “2022-04-01” & budget$ReleaseDate < “2022-05-01”]

[4 marks]

~Question B1 End~

6

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**.



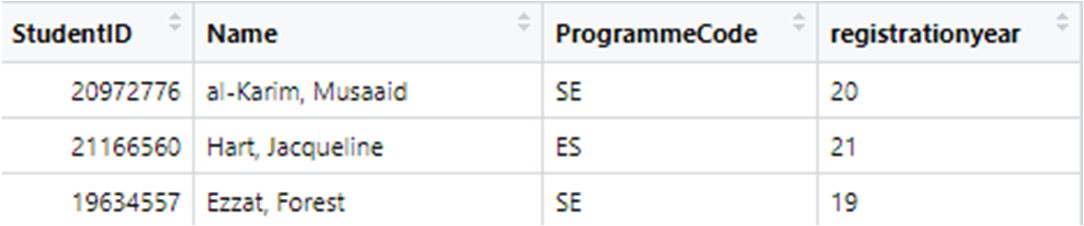
Write R code to

1. 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.

path <- file.choose()

student <-read.csv(path)

[4 marks]

1. 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.

student$registrationyear <- str\_sub(student$StudentID, 1, 2)

1. amend the inappropriate ProgrammeCode with respect to the below reference table, named **ref**, using merge().

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

[6 marks]

ref <- data.frame(

Wrong = c("DSANo.1", "D SA", "ES", "GDS", "GG"),

Correct = c("DSA", "DSA", "SE", "GSD", "GSD")

)

df\_merged <- merge(student, ref, by.x = "ProgrammeCode", by.y = "Wrong", all.x = TRUE)

df\_merged$ProgrammeCode <- ifelse(is.na(df\_merged$Correct), df\_merged$ProgrammeCode, df\_merged$Correct)

# 刪除不需要的列

df\_merged <- df\_merged[, c("ProgrammeCode")]

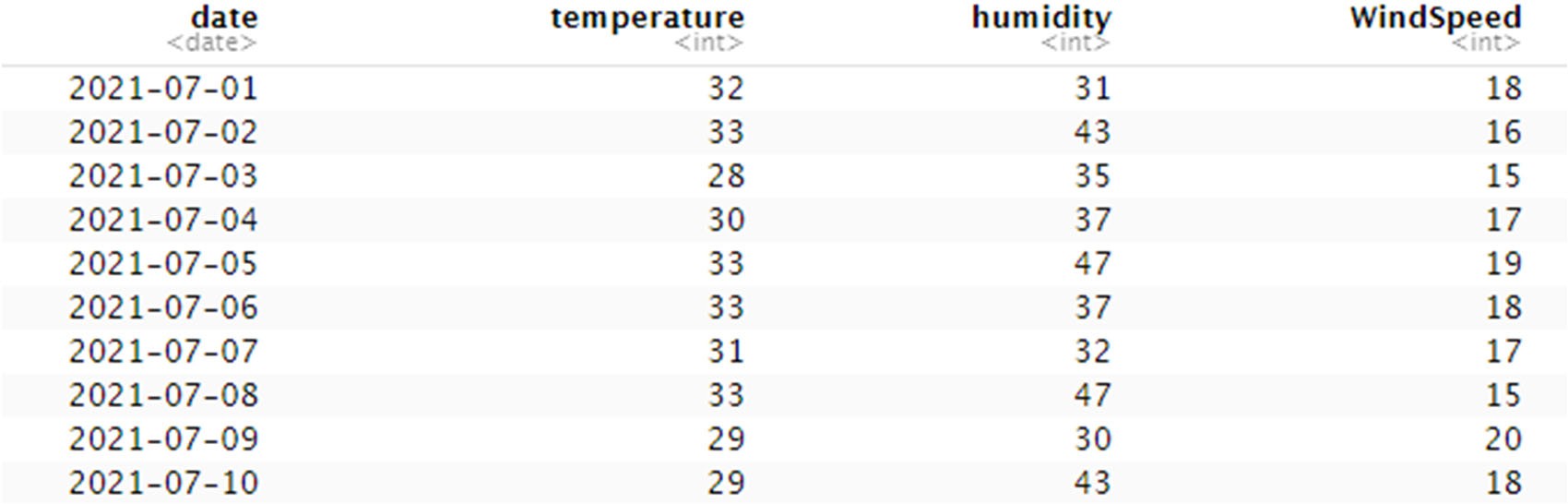
# 顯示結果

print(df\_merged)

[10 marks]

7

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:



Write R code to

1. 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:

WVP = (humidity)∗ 6.105 ∗ exp (17.27 ∗ temperature/(237.7 + temperature))

100

Record$WVP <- Record$humidity / 100 \* 6.105 \* exp(17.27 \* Record$temperature / (237.7 + Record$temperature))

[5 marks]

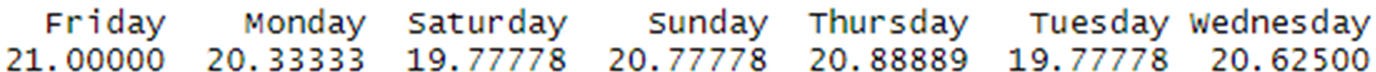
1. calculate the apparent temperature value and add it to the dataframe as a new column named AT. This is the equation of apparent temperature.

AT = 1.04 ∗ 𝑡𝑒𝑚𝑝𝑒𝑟𝑎𝑡𝑢𝑟𝑒 + 0.2 ∗ 𝑊𝑉𝑃 + 0.65 ∗ 𝑤𝑖𝑛𝑑𝑠𝑝𝑒𝑒𝑑 — 2.7

[5 marks]

Record$AT <- 1.04 \* Record$temperature + 0.2 \* Record$WVP + 0.65 \* Record$WindSpeed – 2.7

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



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

Record <- [,-c(3, 5)]

8

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

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

proc import

out ballot

datafile = “/home/u44771062/Data/Ballot.csv”

dbms = csv

replace;

run;

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

proc print data = ballot

where Gender == ‘F’ & Age < ‘30’

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

proc template;

define statgraph dashboard;

begingraph;

layout lattice / column = 1 rows = 1;

piechart category=ballot /stat=percent;

endlayout;

endgraph;

end;

run;

proc sgrender data = ballot template = dashboard;

run;

~End of Question B4 and paper~

9