UNIVERSITI MALAYSIA PAHANG

FINAL EXAMINATION

COURSE : DATA WRANGLING

COURSE CODE : BSD2333

COURSE COORDINATOR : MOHD KHAIRUL BAZLI MOHD AZIZ DATE : 23 JUNE 2021

DURATION : 3 HOURS

SESSION/SEMESTER : SESSION 2020/2021 SEMESTER II

INSTRUCTIONS TO CANDIDATES:

1. This examination paper consists of **FIVE** questions. Answer **ALL** questions.
2. All answers should be written on answer sheet.
3. All answers to a new question should starts on a new page.
4. For each answer sheet, write down your ID number and Name.
5. Scan your answer sheet and save in pdf file.
6. Name your pdf file as **<MATRIC NUMBER>** space **<FULL NAME>**. 7. Upload your pdf file in **GOOGLE CLASSROOM** within 15 minutes after the examination ended.

EXAMINATION REQUIREMENTS:

1. None

APPENDIX:

1. None

DO NOT TURN THIS PAGE UNTIL YOU ARE TOLD TO DO SO This examination paper consists of **TEN (10)** printed pages including front page.

QUESTION 1 [19 MARKS]

Data preparation and data wrangling processes involve cleansing and organizing data into a consolidated format. This process is crucial and emphasized because wrong data can lead a business to wrong decisions, conclusions, and poor analysis.

i. Describe the process of data cleansing in data wrangling.

Data cleansing involves detecting and addressing missing values in the dataset. Missing data can be handled through techniques such as replacing missing values with estimated values or deletion such as removing rows or columns with missing data.

[2 Marks]

[CO1, PO1, C1]

ii. Name the typical errors in data wrangling processes based on the statement given.

a.___Missing Data___ : required data is missing.

b. ___Outliers____ : numbers or dates fall outside within a certain range.

c.___Accurate_ : the degree to which the data is close to the true values.

d.___ Sampling Bias_ __ : data not represent the population or phenomenon of the study.

e. _ Inconsistent data : the currency is sometimes in RM and sometimes in YEN.

f.___Duplicate entries_ : the duplicate entries should be removed.

[6 Marks]

[CO1, PO1, C1]

iii. Data wrangling also typically involves transformations and scaling of the cleans data.

Briefly describe all the transformations below.

a. Extraction Extracting relevant data from the dataset based on specific criteria or conditions.

b. Aggregation Combining multiple data points or records into a summarized or aggregated form, such as calculating averages, totals, or counts.

c. Filtration Filtering the dataset to include or exclude data based on certain conditions or criteria.

d. select Selecting specific columns or variables from the dataset for further analysis, while excluding others.

e. Conversion Converting data from one format to another, such as converting categorical data to numerical format or changing the data type of variables.

[5 Marks]

iv. Unstructured data must be transformed into structured data to be processed by computers. Name **THREE** differences of structured and unstructured data.

Structured data and unstructured data differ in several aspects. Structured data is characterized by having a predefined and organized format, often presented in tables with rows and columns. It follows a well-defined schema, where each data element has a specific meaning and purpose. Structured data allows for easy querying and retrieval of information using standard database query languages like SQL, and it can be stored in relational databases, spreadsheets, or other tabular formats. Examples of structured data include data stored in databases or CSV files.

On the other hand, unstructured data lacks a predefined or organized format and does not fit into a traditional tabular structure. It may consist of free-form text, images, videos, social media posts, and other forms of multimedia. Unstructured data does not adhere to a specific schema, making it challenging to analyze. Analyzing unstructured data often requires advanced techniques such as natural language processing, image recognition, or sentiment analysis. Unstructured data is typically stored in documents, text files, multimedia files, or NoSQL databases. Examples of unstructured data include emails, social media posts, audio recordings, images, videos, and web pages.

[6 Marks]

[CO1, PO1, C1]

QUESTION 2 [20 MARKS]

Data structures are the key to organize storage in computers so that we can efficiently access and edit data. Stacks is one of the earliest data structures defined in computer science. A stack also is a very useful data structure that allow us to store and retrieve data sequentially.

i. Name of rule for stack and briefly explain its concept.

The name of the rule for stacks is "Last-In-First-Out" (LIFO) rule. The concept of the LIFO rule is that the last element that is added to the stack will be the first one to be removed. In other words, the element that was most recently pushed onto the stack will be the first one to be popped off the stack.

[1 Mark]

[CO1, PO1, C1]

[1 Mark]

[CO2, PO2, C2]

ii. State **TWO** purposes of using **`append`** method in stacks.

Pushing Elements: The append method is used to add elements to the top of the stack.
Growing the Stack: As new elements are added to a stack, the size of the stack grows.

[2 Marks]

[CO1, PO1, C1]

iii. Give **TWO** examples of any systems in real life that use stacks concept.

Browser History: Web browsers often maintain a history of visited web pages, allowing users to navigate back and forth between previously visited pages.

Undo-Redo Functionality in Software Applications: Many software applications, such as text editors, image editors, and drawing tools, implement undo-redo functionality using a stack.

[2 Marks]

[CO1, PO1, C1]

iv. Apart from stacks, another important high-level data structure type is queue. What is the rule for queue and briefly explain its concept.

The rule for a queue is known as the "First-In-First-Out" (FIFO) rule. The concept of the FIFO rule in a queue is that the first element that is added to the queue will be the first one to be removed. In other words, the element that has been in the queue for the longest time will be the first one to be dequeued.

[1 Mark]

[CO1, PO1, C1]

[1 Mark]

[CO2, PO2, C2]

v. Give **TWO** examples of any systems in real life that use queue concept.

Supermarket Checkout Lines: Customers join the queue at the back, and the cashier serves the customers one by one from the front of the queue. The first customer to join the queue is the first one to be served.

Ticketing Systems: Customers join a queue to purchase or collect their tickets, and they are served by the ticketing agent at the front of the queue. Again, the FIFO rule applies, and the customer who has been waiting in the queue the longest will be the first one to be served.

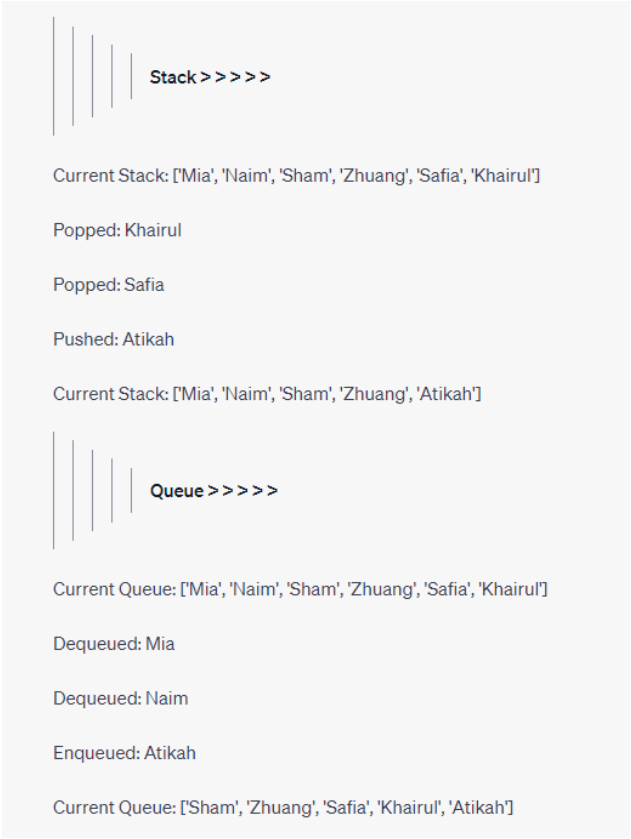
[2 Marks]

[CO2, PO2, C2]

vi. Write the detail outputs of the following Python codes including the 'print' command.

```
from time import sleep
from stack import Stack
from queue import Queue

print("> > > > > Stack > > > > >")
print("")
myStack = Stack(["Mia","Naim","Sham","Zhuang","Safia","Khairul"])

sleep(0.2)
myStack.output()
print("")
sleep(0.2)
myStack.pop()
sleep(0.2)
myStack.pop()
sleep(0.2)
myStack.push("Atikah")
```

```
sleep(0.2)
print("")
myStack.output()
print("")
print("")
sleep(0.2)
print("> > > > > Queue > > > > >")
print("")

myQueue = Queue(["Mia","Naim","Sham","Zhuang","Safia","Khairul"])

sleep(0.2)
myQueue.output()
print("")
sleep(0.2)
myQueue.dequeue()
sleep(0.2)
myQueue.dequeue()
sleep(0.2)
myQueue.enqueue("Atikah")
sleep(0.2)
print("")
myQueue.output()
print("")
```



QUESTION 3 [18 MARKS]

As a data wrangling engineer, we have to filter and group data based on the characteristics of the data before processing them and producing separate datasets as the final output for separate machine learning models.

i. Briefly explain what is subsetting, filtering and grouping.

1. Subsetting:
Subsetting refers to the process of selecting a specific subset of data from a larger dataset based on certain conditions or criteria. It involves extracting a portion of the data that meets specific requirements or matches certain characteristics. Subsetting allows you to focus on a smaller, more relevant portion of the data for analysis or further processing.

2. Filtering:
Filtering is a technique used to extract specific observations or rows from a dataset based on certain conditions. It involves setting criteria or conditions that the data must meet to be included in the filtered subset. By applying filters, you can remove unwanted or irrelevant data and retain only the observations that meet the specified conditions. Filtering is commonly used to remove noise, outliers, or to isolate specific subsets of data for analysis.

3. Grouping:
Grouping involves categorizing or dividing data into distinct groups or categories based on one or more variables. It is typically used to analyze data based on different attributes or characteristics and understand the patterns and trends within each group. By grouping data, you can aggregate and summarize information within each group separately, allowing for comparisons and insights. Grouping is often used in conjunction with summary statistics or calculations to analyze data at different levels of granularity.

[3 Marks]

[CO2, PO2, C2]

ii. Tidy data is a standard way of mapping the meaning of a dataset to its structure.
Describe the characteristics of a tidy data.

Each variable forms a column: In tidy data, each variable is represented by a separate column. This means that each attribute or characteristic being measured or observed should have its own column in the dataset. For example, if you have a dataset about students that includes columns for "Name," "Age," and "Grade," each of these attributes represents a separate variable.

Each observation forms a row: Each row in the dataset represents a unique observation or instance. In other words, each row contains the values for all the variables related to a single observation. For example, if you have a dataset about students, each row would contain the information about a specific student, including their name, age,

grade, and any other relevant variables.

Each type of observational unit forms a table: Tidy data should be organized into separate tables, where each table represents a specific type of observational unit. This means that data that belongs together should be kept together in the same table. For example, if you have data about students and teachers, you would have a separate table for students and a separate table for teachers, rather than combining them into a single table.

[3 Marks]

[CO2, PO2, C2]

iii. Visualize the output for the following Python code.

```
# Import module
import pandas as pd

# Creating Data
car_selling_data = {'Brand': ['Mazda', 'Mazda', 'Mazda', 'Mazda',
                              'Hyundai', 'Hyundai',
                              'Toyota', 'Honda', 'Honda',
                              'Ford', 'Toyota', 'Ford'],
                    'Year': [2010, 2011, 2009, 2013,
                             2010, 2011, 2011, 2010,
                             2013, 2010, 2010, 2011],
                    'Sold': [6, 7, 9, 8, 3, 5,
                             2, 8, 7, 2, 4, 2]}

# Creating Dataframe of car_selling_data
df = pd.DataFrame(car_selling_data)

# printing Dataframe
print(df)
```

[5 Marks]

[CO2, PO2, C2]

iv. Rewrite the Python code given in **(iii)** for the year 2010 output. Then, visualize the output.

```
import pandas as pd

# Creating Data
car_selling_data = {'Brand': ['Mazda', 'Mazda', 'Mazda', 'Mazda',
                              'Hyundai', 'Hyundai',
                              'Toyota', 'Honda', 'Honda',
                              'Ford', 'Toyota', 'Ford'],
                    'Year': [2010, 2011, 2009, 2013,
                             2010, 2011, 2011, 2010,
                             2013, 2010, 2010, 2011],
                    'Sold': [6, 7, 9, 8, 3, 5,
                             2, 8, 7, 2, 4, 2]}

# Creating Dataframe of car_selling_data
df = pd.DataFrame(car_selling_data)


# Filter the dataframe for the year 2010


filtered_df = df[df['Year'] == 2010]


# Visualize the output


print(filtered_df)
```

```
    Brand  Year  Sold

   Mazda  2010     6

  Hyundai  2010     3

    Honda  2010     8

    Ford  2010     2

  Toyota  2010     4
```

[7 Marks]

[CO2, PO2, C2]

QUESTION 4 [21 MARKS]

A join is basically a method to retrieve linked rows from two tables using any kind of primary key and foreign key relation. There are several types of join, such as inner, left outer, right outer and full outer join.

Table 1: Prices Table 2: Quantities

Product	**Price**
Potatoes	RM3
Avocados	RM4
Kiwis	RM2
Onions	RM1
Melons	RM5
Oranges	RM5
Tomatoes	RM6

Based on **Table 1** and **Table 2**,

Product	**Quantity**
Potatoes	45
Avocados	63
Kiwis	19
Onions	20
Melons	66
Broccoli	27
Squash	92

i. define inner join and show the output for inner join of the tables.

	Product	Price	Quantity
0	Potatoes	RM3	45
1	Avocados	RM4	63
2	Kiwis	RM2	19
3	Onions	RM1	20
4	Melons	RM5	66

[5 Marks]

[CO2, PO2, C2]

iii. define left outer join and show the output for left outer join of the tables.

	Product	**Price**	**Quantity**
0	**Potatoes**	**RM3**	**45.0**
1	**Avocados**	**RM4**	**63.0**
2	**Kiwis**	**RM2**	**19.0**

3 Onions RM1 20.0

4 Melons RM5 66.0

5 Oranges RM5 NaN

6 Tomatoes RM6 NaN

[5 Marks]

[CO2, PO2, C2]

iv. define right outer join and show the output for right outer join of the tables.

	Product	Price	Quantity
0	Potatoes	RM3	45
1	Avocados	RM4	63
2	Kiwis	RM2	19
3	Onions	RM1	20
4	Melons	RM5	66
5	Broccoli	NaN	27
6	Squash	NaN	92

[5 Marks]

[CO2, PO2, C2]

v. define the full outer join and show the output for full outer join of the tables.

	Product	Price	Quantity
0	Potatoes	RM3	45.0
1	Avocados	RM4	63.0
2	Kiwis	RM2	19.0
3	Onions	RM1	20.0
4	Melons	RM5	66.0
5	Oranges	RM5	NaN
6	Tomatoes	RM6	NaN
7	Broccoli	NaN	27.0
8	Squash	NaN	92.0

[6 Marks]

[CO2, PO2, C2]

QUESTION 5 [22 MARKS]

Web scraping is an essential part of data wrangling in today's world, as we can find nearly everything on the Web. Python libraries can be used to explore web pages, search for information, and collect it for your reporting with web scraping.

i. Explain what is web scraping and give **TWO** reasons why it is important in data wrangling.

Web scraping refers to the process of extracting data from websites by automatically fetching and parsing the HTML content of web pages.

Accessing Unstructured Data: The web contains vast amounts of unstructured data, such as articles, reviews, social media posts, product information, and more. Web scraping allows you to convert this unstructured data into structured data that can be easily analyzed and processed.

Real-time and Updated Data: Web scraping allows data wranglers to fetch the latest data at regular intervals, ensuring that the analysis and insights are based on the most current information available.

[4 Marks]

[CO2, PO2, C2]

ii. In your opinion, is web scraping legal and ethical? Give a reason.

The legality and ethics of web scraping can vary depending on the specific context and the terms and conditions set by website owners. In general, web scraping can be legal and ethical when performed within legal boundaries and with proper consent. It is important to consider the legality of web scraping in relation to factors such as the website's terms of service, copyright restrictions, and applicable laws regarding data protection and privacy. It is advisable to consult legal and ethical guidelines and obtain permission when scraping data from websites to ensure compliance and ethical practices.

[2 Marks]

[CO2, PO2, C2]

iii. Name **TWO** applications of web scraping in business.

Market Research and Competitive Analysis: Web scraping allows businesses to gather data from competitor websites, industry directories, e-commerce platforms, social media, and other sources. By scraping data from these sources, businesses can analyze and compare their performance, identify market opportunities, and make informed decisions to stay competitive.

[2 Marks]

[CO1, PO1, C1]

iv. Name Python library that allows us to read HTML tables directly from a URL.

pandas

[1 Mark]

[CO1, PO1, C1]

v. Name Python command that should be used to read from the website page containing tabular data.

```
import pandas as pd
url = "https://en.wikipedia.org/wiki/Lee_Zii_Jia"
table = pd.read_html(url)
table1 = table[0]
table2 = table[1]
table
```

[1 Mark]

[CO1, PO1, C1]

vi. Use the Python command in **(iii)** to write a detailed Python code to read from the Wikipedia page containing 2010–11 Premier League data. The webpage URL is **https://en.wikipedia.org/wiki/2010-11_Premier_League**.

```
import pandas as pd
url = "https://en.wikipedia.org/wiki/2010-11_Premier_League"

# Read the HTML tables from the URL
tables = pd.read_html(url)

# Find the desired table index or inspect the tables list to identify the table of interest
desired_table_index = 2
premier_league_table = tables[desired_table_index]

# Print the extracted table
print(premier_league_table)
```

[2 Marks]

[CO2, PO2, C2]

vii. A simple loop needs to run for searching of the particular table on a webpage. What is the Python command that we should use to examine the number of rows and the number of columns of each of the tables on the webpage in **(vi)**? Then, write the Python code to get the output as in **Figure 1**:

```
(21, 2)
(20, 4)
(20, 5)
(9, 7)
(22, 11)
(1, 1)
(20, 21)
(11, 4)
(17, 5)
(10, 6)
(4, 13)
(9, 2)
(1, 2)
(23, 2)
(0, 2)
(4, 2)
(1, 2)
(0, 2)
(5, 2)
(5, 2)
```

Figure 1

```
import pandas as pd

url = "https://en.wikipedia.org/wiki/2010-
11_Premier_League"

# Read the HTML tables from the URL
tables = pd.read_html(url)

# Loop through each table and print its shape
for table in tables:
    print(table.shape)
```

[3 Marks]

viii. Based on the output in **Figure 1**, write a detailed Python code to extract the table as in **Figure 2** from the webpage.

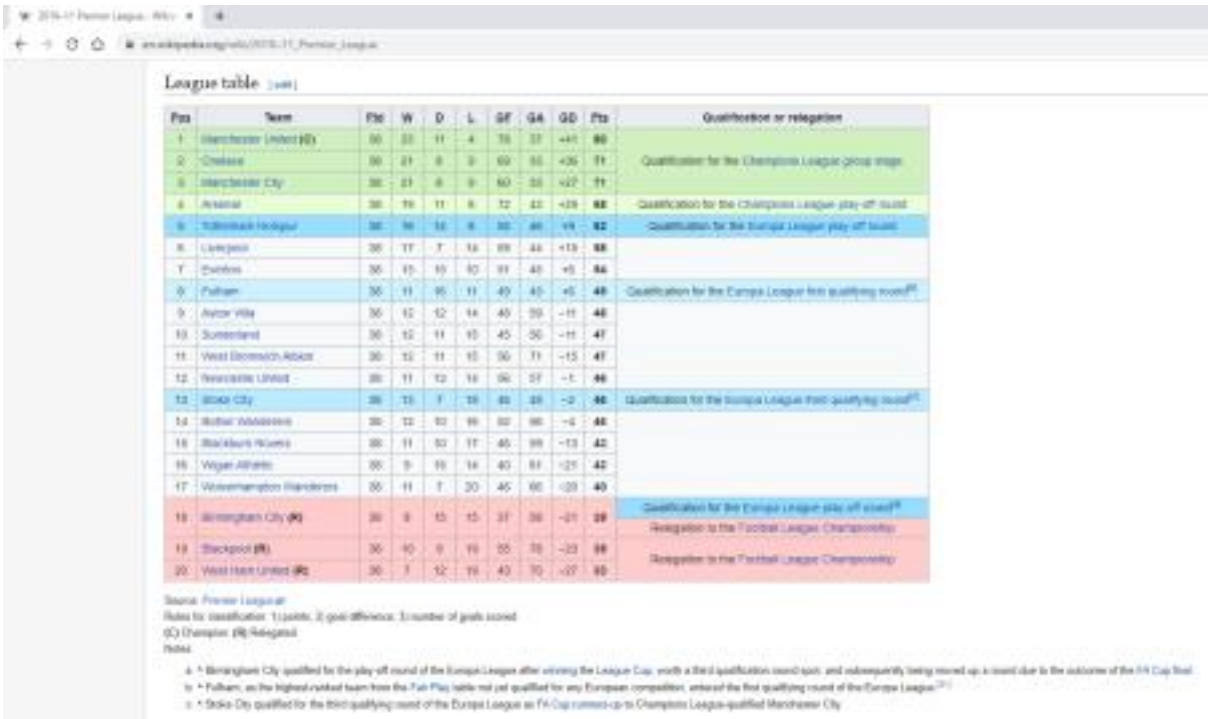


Figure 2: Wikipedia webpage

```
import pandas as pd

url = "https://en.wikipedia.org/wiki/2010-11_Premier_League"
```

 # Read the HTML tables from the URL
tables = pd.read_html(url)

 # Find the index of the table with shape (20, 5)
desired_table_index = 2

 # Extract the desired table
desired_table = tables[desired_table_index]

 # Print the extracted table
print(desired_table)

[2 Marks]

[CO2, PO2, C2]

ix. Name a Python library used to save a DataFrame as an Excel (xlsx) file. Then, write a detail Python code to save the table in Excel file and name the file as 'Manchester United Champions 2010-11'.

```
import pandas as pd

url = "https://en.wikipedia.org/wiki/2010-11_Premier_League"

desired_table_index = 2

# Read the HTML tables from the URL

tables = pd.read_html(url)

# Extract the desired table

desired_table = tables[desired_table_index]

# Save the table to an Excel file
```

```
file_name = "Manchester United Champions 2010-11.xlsx"

desired_table.to_excel(file_name, index=False)

print("Table saved successfully as", file_name)
```

[1 Mark]

[CO1, PO1, C1]

[4 Marks]

[CO2, PO2, C2]

END OF QUESTION PAPER