Question 1

Objectives:

- Understand dataset with data scientist mindset
- Understand and design computation logic and routines in Python
- Assess use of Python only and Python data structures to perform extract, load, and transformation operations
- Structure code in appropriate Methods (functions), Looping and conditions

In 2020, the COVID-19 is quickly spread out over the whole world. A lot of researchers are dedicated to investigating the breakout trend and its impact on humans. The following URL contains the real COVID-19 related cases in Canada.

https://github.com/LiuFang00/ICT233/blob/master/public-covid-19-cases-canada.csv

(a) Analyse the dataset downloaded from the URL link and answer the following questions:

(i) Load the .csv file into the notebook.

(2 marks)

Can use pandas for this question.

(ii) Summarize the information that can be derived from the dataset, including key features (columns), range of values, and useless values. All the information should be derived by necessary Python codes.

(9 marks)

(iii) Explain **TWO (2)** potential insights that can be derived from the dataset.

(4 marks)

(b) Conduct data pre-processing for the dataset:

(i) Refers to Q1-a)-ii), remove all the useless item like ‘NULL’, ‘Not Report’, etc. You can drop the row or fill it up with a specific value. (3 marks)

(ii) Reformat the Age group as {‘0-19’, ’20-29’, ’30-39’, …, ’90-99’} and update the ‘age’ column.

(6 marks)

When you are summarising the data at aii, for those part that you say useless, then you have to do something about those columns. E.g. if Age cannot interpolate, do you want to drop that row instead of the whole column. (Some are not in the right range, "<18", replace to the correct range, follow the grping, change to "0-19".

(iii) List out the total number of infected person for each age group.

(2 marks)

(c) Save the cleaned and reformatted dataset into a new .csv file.

(3 marks)

Question 2

Objectives:

- Design computation logic and routines in Python
- Assess the Design and use of Database ORM and methods to perform extract, load, transformation and calculation operations

Continue work with the Canada dataset in Question 1. Let's consider the 'load' step in the ETL process. The dataset is required to load in a relational database.

(a) Design and apply a Python ORM(s) (Object Relation Mapping) to store the .csv file obtained in Q1(c). Please specify a table class before inserting the values into the database.

Define ur class, read the data file, then put the data into it, no PK but can use surrogate key that will auto generate a value for you.

(10 marks)

If you have time, use 2 tables.

(b) Compose queries on the database and answer the following questions:

(i) What is the total number of male and female infectors for each month?

(5 marks)

(ii) Sort the age groups with regards to the number of female infectors in descending order.

(5 marks)

(iii) For the person who does not has travel history, what are the top **TWO (2)** months with regards to the number of infectors older than 50?

(5 marks)

Question 3

Objectives:

- Perform simple exploratory data analysis
-
- Design computation logic and routines in Python
- Assess use of REST API, Python only and Python data structures to perform extract, load, transformation, and calculation operations
- Assess use of Pandas and Dataframes to perform extract, load, transformation and calculation operations
- Conduct visualization in an appropriate way

To get a personal-level drill-down of the dataset, both spatial and temporal should be taken into consideration.

(a) Load the .csv file obtained in Q1(c) to Pandas Dataframe and derive the answer for the same 3 question in Q2(b).

(10 marks)

(b) Suppose you are the researcher who would like to discover the infection pattern across the country.

(i) Design a function to find out the top **THREE (3)** provinces of each month with regards to the total number of COVID-19 cases.

(6 marks)

(ii) Draw **ONE (1)** figure to show the total number of COVID-19 cases for each province.

(6 marks)

(iii) For the province with highest number of cases, draw **ONE (1)** figure to describe the age distributions of COVID-19 cases per gender.

can use Pandas to draw.

(6 marks)

(c) Suppose you are the researcher who would like to discover the peak days of COVID-19 cases.

(i) Design a function to compute the days in a week (i.e. Mon, Tue or Wed) for any given date. Then update the column 'report week' in Dataframe by calling the function.

(6 marks)

(ii) List out the top **THREE (3)** days that COVID-19 cases detected.

(6 marks)

(iii) Draw **ONE (1)** figure to describe the number of COVID-19 cases per gender for each month.

(6 marks)