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.

The following URL contains the real HDB flat transactions from Jan 2017 onwards:
https://github.com/DataTensor/hdb/blob/main/resale-flat-prices.csv

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

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

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

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

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

(i) Refer to Q1(a)(ii), remove all the data rows with missing data values.

(ii) In Singapore, HDB flats have a 99 years' leasehold. Compute the remaining lease in years for each transacted flat on its transacted date.

(iii) List out the top ten of remaining lease in years (computed by Q1(b)(ii) having the greatest number of transacted flats.

(c) Save the cleaned and transformed dataset into a new CSV file.

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 the work with the HDB flat resale dataset saved from the Q1(c). Let's consider the "load" step in the ETL process to load the dataset into a relational database.

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

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

(i) What is the total number of transactions for each month?

(ii) Sort the data by town and the number of resale transactions in descending order.

(iii) For resales transacted on and after Jan 2019 and storey range being level 10 and above, what are the top three towns having the greatest number of transactions?

Question 3

Objectives:

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

(a) Load the CSV file obtained from Q1(c) to a Pandas dataframe and derive the answers for the same three questions in Q2(b)

(b) Suppose you are a researcher who would like to discover data patterns in HDB resale transaction after 2017:

(i) Design a function to find the top three towns of each month with the greatest number of transactions.

(ii) Draw **ONE (1)** figure to show the median resale price in 2020 of each town.

(iii) For the town with the greatest number of transactions in 2020, draw **ONE (1)** figure to visualize the median resale price per flat type for each month.

(iv) Is there any correlation between the storey range and resale price? Draw **ONE (1)** figure to visualize the correlation.

(v) Is there any correlation between the remaining lease in years and resale price? Draw **ONE (1)** figure to visualize the correlation.

(vi) For Yishun, draw **ONE (1)** figure to visualize how was the median resale price being changed over time per flat type?