CE/CZ4123/SC4023 BIG DATA MANAGEMENT

SEMESTER GROUP PROJECT

College of Computing and Data Science
Nanyang Technological University

1 ASSIGNMENT DESCRIPTION

The goal of this semester's project is to conduct a simple analysis on the resale flats to have a flavor of the data management process. You will be provided with a series of transaction records concerning the resale of HDB flats over the last 11 years (2015 to 2025) in Singapore. These transactions present comprehensive information including approval date, location, flat model, price, etc, enabling us to execute various queries, such as the average resale flat price on a particular street or the price trends for specific types of HDB flats. To facilitate a more manageable workload and evaluation process, we propose to select several specific queries for you to implement.

From the perspective of a potential flat buyer, one may be interested in accessing the *Minimum Price per Square Meter* of the resale HDB flats whose area meets some requirements ($\geq$ y square meters (m^2) in a certain location over x months, considering the range of integers x, y to be $1 \leq x \leq 8$, and $80 \leq y \leq 150$. If for some (x, y) pair, the Minimum Price per Square Meter is at most 4725, we say the (x, y) pair is *valid*. In particular, your program is required to compute **all valid** (x, y) pairs based on the filtering conditions, and the associated record corresponding to the minimum price per square meter for each (x, y) pair.

You are expected to write a program to manage the data in a **column-oriented** manner, including data storage and processing. To be specific, a query is composed of a target time of x months and a list of matched towns, where x is a user input integer with value at least 1 and at most 8. These factors are determined by your matriculation number as follows: a) The last digit of the target year (YYYY) matches the last digit of the matriculation number. Note: the data for 2025 is provided only for query, not as the target year. For example, matriculation number A5656567B corresponds to 2017, while A3245675B corresponds to 2015; b) the commencing month (MM) equals the second last digit of the matriculation number (note that "0" represents October); c) the list of matched towns depends on **all digits** appear in the matriculation number as Table 1 presents; d) the area requirement ($\geq y$ square meters (m^2)) is applicable to all these contents, with y to be an integer at least 80, and at most 150.

Digit	0	1	2	3	4
Town	BEDOK	BUKIT PANJANG	CLEMENTI	CHOA CHU KANG	HOUGANG
Digit	5	6	7	8	9
Town	JURONG WEST	PASIR RIS	TAMPINES	WOODLANDS	YISHUN

Table 1: List of towns corresponding to the digit in matriculation number.

Please use the matriculation numbers of **anyone** of your group members to generate the query condition.

Example**: For $x = 3$, and $y = 85$, a student with matriculation number* **A5656567B** *should scan the resale HDB flats in* **JURONG WEST,* **PASIR RIS***,* **TAMPINES** *from* **Jun. 2017** *to* **Aug. 2017** *to compute the associated Minimum Price per Square Meter of the matched flats not smaller than 85 square meters.*

In this case, an example query should be equivalent to the following SQL query:

Task in SQL

```sql
WITH Tab1 AS (
    SELECT *
    FROM   ResalePricesSingapore
    WHERE  (YEAR(Month) = 2017)
           AND (MONTH(Month) >= 6)
           AND (MONTH(Month) <= 8)
           AND (Town = 'JURONG WEST' or Town = 'PASIR RIS' or Town = '
              TAMPINES')
           AND (Area >= 85)
)
SELECT MIN(Resale_Price*1.0/Floor_Area) FROM Tab1
```

Please note that your task is to find **all** (x, y) that can give at least one record satisfying the filtering conditions, and for each such (x, y) pair, output the corresponding record with the minimum price per square meter that is not higher than 4725.

2 INPUT FORMAT

The input file `ResalePricesSingapore.csv` is the historical transaction records of the resale HDB flat in Singapore from Jan. 2015 to Dec. 2025. Please note that your queries should not start from any month in the year 2025. The data is extracted from an open access dataset published on Singapore's national open data collection website[*] maintained by Data.gov.sg team.

The input data is given in `.csv` format. You can download the data via NTU Learn. The first row is the title row. Each following row contains a line of transaction information as listed, which are separated by a comma ",".

- `Month`: approval date of the resale, in the format `YYYY-MM`.
- `Town`: the town of the associated HDB flat.
- `Block`: the block of the associated HDB flat.
- `Street_Name`: the street of the associated HDB flat.
- `Flat_Type`: the type of flat of the associated HDB flat. In Singapore, there are 1-room flats up to 5-room flats, as well as executive flats.
- `Flat_Model`: it implies the approximate size and the number of rooms for the HDB flat, categorized into types such as Standard, Improved, New Generation, etc.
- `Storey_Range`: In this dataset, the storey range is given in a range of 3 (e.g. 10 to 12, which means the flat is based on the 10th to 12th storey).
- `Floor_Area`: the floor area of the associated HDB flat in square meters.
- `Lease_Commence_Date`: the commence date of the flat lease in (months and) years.
- `Resale_Price`: the resale price of the assocaited HDB flat.

Please note that you should store the entire table and may focus on particular information of interest for your task.

[*] Data from: Data.gov.sg.

3 OUTPUT FORMAT

Your output file `ScanResult_<MatricNum>.csv` should contain all query results associated with the chosen matriculation number. The first row is the title row and the following rows present the query information and results as listed below which are separated by commas “,”. If there is no qualified data in your target range, please take "No result" as the query result.

- `(x,y)`: the (x, y) pair; follow the increasing order of x; for two pairs with the same x, follow the increasing order of y.
- `Year`: the year of the matched record, in the format of `YYYY`.
- `Month`: the month of the matched record, in the format of `MM`.
- `Town`: the town where the matched HDB flats locate.
- `Block`: the block of the associated HDB flat.
- `Floor_Area`: the floor area of the associated HDB flat in square meters.
- `Flat_Model`: profile of HDB flat with types Standard, Improved, New Generation etc.
- `Lease_Commence_Date`: the commence date of the flat lease in years.
- `Price_Per_Square_Meter`: the minimum price per square meter rounded to integers.

Example*: For matriculation number* **A6626226B** *the task is to scan resale HDB records in* **CLEMENTI** *and* **PASIR RIS** *from Feb. 2016 with different* (x, y) *pairs. For each* (x, y) *pair, identify the record with the Minimum Price per Square Meter and print it as follows. Please ensure that your output file includes all fields shown in the demo.*

ScanResult_A6626226B.csv (example)

```
(x, y),Year,Month,Town,Block,Floor_Area,Flat_Model,Lease_Commence_Date,Price_Per_Square_Meter
(1, 80),2016,02,PASIR RIS,149,126,Improved,1995,3413
...
(4, 80),2016,04,PASIR RIS,140,126,Improved,1994,3175
...
(1, 120),2016,02,PASIR RIS,149,126,Improved,1995,3413
...
```

4 SUBMISSION

Time: During Week 14 (By April 24 unless otherwise specified)

Method: Via NTULearn

The required files include the output file, the source code of your program, and an assignment report. They should be compressed and submitted in a `.zip` file. Name the `.zip` file with your **group number**. The requirements of each files are as follows:

- Output Files `ScanResult_<MatricNum>.csv`: the scan results following the requirements in **Output Format** Section. Do not include any raw or intermediate data files.

- Source Code `source`: the file or folder containing the source codes that input the file `ResalePricesSingapore.csv` and the matriculation number, and output the corresponding `ScanResult_<MatricNum>.csv`. Source codes should be well-commented and contains essential documentations to help understand the functionalities.
- Report `Report.pdf`: the report exported in `.pdf` format. Your report sections and contents should follow the requirements in **Report Format** in Appendix. The report should be at most 5 pages (single column, font size 11pt, excluding cover page and contribution form). If your really want to place big figures such as screenshots (or design figures) that you cannot squeeze into the 5 main pages after your best trial, you may optionally add an Appendix within two pages. This section is fully optional which should only contain figures and corresponding captions, and the assessment would still be mainly based on the 5-page main content.

5 FORMING GROUPS

The expected group size is 3. We encourage you to form groups autonomously by editing the **Online Form** before **February 8** (Week 4 Sunday). Students who are not involved in any group will be randomly assigned to a group by the TA.

6 ASSESSMENT

This is a **group project**. Your submission will be evaluated in multiple aspects, including design sophistication (e.g., whether the program meets basic requirements and whether there are additional optimizations), output accuracy, code quality (e.g., whether you can reuse some functions for conciseness), and report quality. Late submission will be penalized. The evaluation of an individual is based on the contribution form.

7 GENERAL GUIDELINES

1. If you are not familiar with the `.csv` format input file, you can regard it as a plain text file (just like `.txt` format).

2. Please note assuming that the `Month` column is monotonically increasing may lead to inaccurate query outcomes. Additionally, the resale HDB flats locating in the same town may not be strictly clustered together in the `Town` column.

3. While we recommend Java, you are free to choose any programming language in case you are not familiar with Java.

4. Ensure that your program is implemented in the column-store manner. Avoid high-level data tools when storing and processing the data. Example 1: Python pandas is not column-oriented. Example 2: simple SQL implementation is not column-oriented.

5. Please validate the accuracy of your query results with supplementary tools, such as Microsoft Excel or Google Sheets, and include supporting evidence in your report. For instance, you can select a few output (x, y) pairs, and for the corresponding query, attach the associated Office Scripts in Excel (or Excel formulas) and screenshots of the results to compare with your outputs.

6. The computation time of your program will not be evaluated as the working environments and hardware configurations vary widely. However, we still encourage you to implement optimizations to improve efficiency and present the enhancements in your report. You may consider dealing with various scenarios, such as handling data too large to fit in main memory, possibility of reusing intermediate results, speeding up data scanning with additional index or specialized data layouts, and boosting computational efficiency through data compression techniques, etc. The design sophistication will be counted into assessment.

7. The code and report should be developed on your own, and using AI tools to generate codes and reports is not allowed.

Appendix: Report Format

REPORT FORMAT

Name and Matriculation Number

1 Data Storage

In this section, explain how your program handles and stores the data. You may present your design and experience (whether success or failure) related to:

- How to store the data in the column-store approach[‡];
- How to design data columns for efficient processing[‡];
- How to read and write the input/output files;
- How to handle possible exceptions such as empty qualified entries in the code.

2 Data Processing

In this section, explain how your program scans the data and finds the values. You may present contents related to:

- How to scan columns according to task conditions;
- How to determine and retrieve the matched data;
- How to improve the efficiency in scanning columns and reuse intermediate results [‡];

3 Experiment Result

In this section, present the experimental results that your program successfully complete the tasks. The following contents are compulsory:

- Screenshots that your program executes and outputs results successfully;
- Evaluations that the output results are correct (You may put screenshots in the report comparing your output results with the correct results output by other tools such as Excel; given the examination for a few (x, y) pairs is sufficient).

[‡] Exploration and improvements on these aspects are encouraged. You are expected to come up with the ideas on your own.

CONTRIBUTION FORM

Group Number

Name	Matriculation Number	Detailed Individual Contribution	Percentage (100% in total)

Name and Signature from all group members:

Name and Signature of Member 1

Name and Signature of Member 2

Name and Signature of Member 3

Name and Signature of Member 4