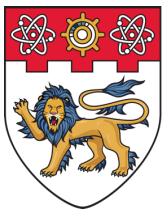

NANYANG TECHNOLOGICAL UNIVERSITY
SINGAPORE

CE/CZ4123/SC4023 BIG DATA MANAGEMENT

SEMESTER GROUP PROJECT

SCHOOL OF COMPUTER SCIENCE AND ENGINEERING
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 10 years (2014 to 2023) in Singapore. Each transaction presents comprehensive information including approval date, location, flat model, price, etc. Given the provided data, you can execute diverse queries to extract various information, 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 statistics of the resale HDB flats in a certain location over several months. Therefore, your program is required to compute the minimum value, average value, and standard deviation of floor area and price for resale HDB flats within a specific town over 3 consecutive months in a given year.

You are expected to write a program to manage the data in a **column-oriented** manner, including data storage and processing. Your program should first receive queries, then scan the data columns to find out the satisfactory lines, and compute the query results according to the task requirements.

The query described above is determined by several conditions which are derived according to your matriculation number. Specifically, the last digit of the required years equals to the last digit of the matriculation number; the commencing month equals the second last digit of the matriculation number (note that "0" represents October); the town depends on the third last digit of the matriculation number as Table 1 presents. Please use the matriculation numbers of anyone of your group members to generate queries.

Digit	0	1	2	3	4
Town	ANG MO KIO	BEDOK	BUKIT BATOK	CLEMENTI	CHOA CHU KANG
Digit	5	6	7	8	9
Town	HOUGANG	JURONG WEST	PUNGGOL	WOODLANDS	YISHUN

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

Please note that each query produces one and only one statistic presented in Table 2. Furthermore, your report should contain an assessment of all 6 statistic types in arbitrary sequence. All results should be **rounded to the hundredth place**.

Minimum Area	Average Area	Standard Deviation of Area
Minimum Price	Average Price	Standard Deviation of Price

Table 2: List of statistics for testing your program.

***Example**: For querying the average price, a student with matriculation number* **A1234567B** *should scan the resale HDB flats in* **HOUGANG** *from* **Jun. 2017** *to* **Aug. 2017** *to compute the average value of the associated prices.*

In this case, an example query for the average price of the resale HDB flats in HOUGANG during Jun. 2017 to Aug. 2017 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 = 'HOUGANG')
)
SELECT AVG(Resale_Price) FROM Tab1
```

2 INPUT FORMAT

The input file `ResalePricesSingapore.csv` is the historical transaction records of the resale HDB flat in Singapore during the past 10 years (Jan. 2014 – Dec. 2023). 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 may focus on particular types of information for your task.

3 OUTPUT FORMAT

Each output file `ScanResult_<MatricNum>.csv` should contain the results from one matriculation number query. The first row is the title row. The following rows present the query results including the average price or standard deviation, along with the corresponding querying period and location information, separated by a comma ",". If there is no data in your target range, please take "No result" as the query result.

*Data from: Data.gov.sg.

We would like to remind you that all of the 6 types of statistics presented in Table 2 should be tested with your program. The columns are listed as follows:

- `Year`: the year in the query, in the format of `YYYY`.
- `Month`: the month you **start** to collect statistics, in the format of `MM`.
- `Town`: the town where the queried HDB flats locate.
- `Category`: type of statistic associated with the query. Please use the expression provided in Table 2 in output, such as "`Average Price`" or "`Standard Deviation of Price`".
- `Value`: the value of the query outcome.

***Example**: Suppose two consecutive queries accessing the average price and standard deviation of floor area for HDB flats in `ANG MO KIO` during `January 2023` to `March 2023`, respectively, which are 530000.05 and 100.00. Then the corresponding result rows in the output file should be:*

ScanResult_A1234013B.csv (example)

```
Year,Month,town,Category,Value
2023,01,ANG MO KIO,Average Price,530000.05
2023,01,ANG MO KIO,Standard Deviation of Area,100.00
...
```

4 SUBMISSION

Time: During Week 14 (By April 19 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).

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 Thursday). 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 on the comprehensive basis including design sophistication (e.g., what if data go big and cannot be stored in main memory), 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 attach the associated Office Scripts in Excel (or Excel formulas) and screenshots of the results to compare with your outputs.

6. 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 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 decide and record the statistics (e.g. the minimum values);
- How to improve the efficiency in scanning columns‡;

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).

‡Exploration and improvements on these aspects are encouraged.

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