National University of Singapore
School of Computing

CS2102: Database Systems
Semester 2, 2018/19

Assignment 1 (1 mark)

Due: January 31, 2019 (Thursday, 6pm)

1 Introduction

A database management system (DBMS) is a specialized piece of software for managing data. Before DBMSs were invented, applications manage data using what is known as *file processing techniques*, where the data is stored on the operating system file system and is processed by application programs written in some programming language. The following illustrates the typical processing steps on a single data file.

```
initialize some book-keeping information I
open data file F
while (F is not empty)
        read next record r from F
        if (r satisfies some condition) then
              do something with r
        update I if necessary
do something with I if necessary
close file F
```

The objective of this assignment is to let you get a sense of what it takes to manage data without using a DBMS.

2 Database

In this assignment, we have a very simple database that consists of a single comma-separated values (CSV) file that maintains the resale flat prices in Singapore from January 1990 to July 2017[1]. The file contains a total of 757,024 records, where each record consists of the following sequence of 10 columns:

1. Resale month (YYYY-MM)
2. Town (e.g., ANG MO KIO, YISHUN)
3. Flat type (e.g., 1 ROOM, 2 ROOM, EXECUTIVE, MULTI GENERATION)
4. Block number

[1]The CSV file is derived from `https://data.gov.sg/dataset/7a339d20-3c57-4b11-a695-9348adfd7614`.

5. Street name
6. Storey range (e.g., 01 TO 03, 49 TO 51)
7. Floor area in square metres
8. Flat model (e.g., DBSS, MAISONETTE, MODEL A)
9. Lease commencement date (YYYY)
10. Resale price

3 What to do

You are to attempt AT LEAST ONE of the following five questions. Each question requires writing a program (using your favourite programming/scripting language) to perform a data management task.

Question 1: Find all the records with flat model = 'ADJOINED FLAT' and flat type = '3 ROOM'. Output these records (in any order) to a CSV file named `q1.csv`.

Question 2: Compute the following four statistics for each town: (1) the number of resale flat transactions, (2) the maximum price per square metre (psm) for the town, (3) the average psm for the town, and (4) the minimum psm for the town. The *psm* metric is defined as the ratio of *resale price* to *floor area*. All psm values should be rounded up to the nearest integer values. Output each town and its four statistics as a single record to a CSV file named `q2.csv`. The records are to be sorted in descending order of average psm values.

Question 3: Resale transactions can be classified into either *good* or *bad* as follows. A resale transaction x is defined to be *bad* if there is another resale transaction y where both x and y are for the same town and they satisfy one of the following conditions:

(a) The resale price of y is higher than that of x, and the floor area of y is lower than that of x;
(b) The resale price of y is the same as that of x, and the floor area of y is lower than that of x; or
(c) The resale price of y is higher than that of x, and the floor area of y is the same as that of x.

A resale transaction that is not classified as bad is considered to be a *good* resale transaction. Find all the good resale transactions for the town named 'BISHAN' and output them to a CSV file named `q3.csv` sorted in ascending order of resale month values.

Question 4: Compute the cumulative number of resale transactions for the town named 'LIM CHU KANG' as follows. For each resale month m where there is some transaction for the town named 'LIM CHU KANG', count the total number of resale transactions for 'LIM CHU KANG' up till month m. Output the distinct pairs of resale month and total to a CSV file named `q4.csv`. The records are to be sorted in ascending order of resale month values.

Question 5: Update all the records located in the town 'YISHUN' by increasing their floor area by 10%. The updated values should be rounded up to the nearest integer values.

4 How to Start

Download the following file: http://www.comp.nus.edu.sg/~cs2102/cs2102-assign1.zip

The unzipped directory contains the following files:

- A CSV database file resale-flat-prices.csv.
- A directory output-files/ which contains an output CSV file for each of the questions. You may use these files to compare against your program outputs.

5 How to Submit

Submit your source code for each of the attempted questions. Name each source file using an appropriate question number (e.g., q1.java, q2.py). Do not submit any CSV files.

If you are submitting only a single file, simply upload that file to the IVLE folder named Assignment 1 Submissions. Otherwise, upload a zip file containing all your source files to that folder.

This assignment is due on January 31 (Thursday, 6pm).