Lab of Data Science

Exercises for the first mid-term

CORRECTION ON October 25, 2018

Exercise 1 (8 pts). Consider the `foodmart` database. The *dissimilarity index* of year 1998 is defined as:

$$D = \frac{1}{2} \sum_{i=1}^{n} \left| \frac{f_i}{F} - \frac{m_i}{M} \right|$$

where n is the number of stores, and, for a store id i:

- f_i is the number of distinct female customers who made at least one purchase in the store i during 1998;
- m_i is the number of distinct male customers who made at least one purchase in the store i during 1998;
- $F = \sum_{i=1}^{n} f_i$ is the sum of the f_i's;
- $M = \sum_{i=1}^{n} m_i$ is the sum of the m_i's.

Write a Python program Dissimilarity.py which outputs such the value D. The Python program can submit only SQL queries of the form "SELECT * FROM table".

What to deliver: Dissimilarity.py.

Exercise 2 (8 pts). Develop a SSIS package that outputs on a CSV file the result of Ex. 1. The usage of GROUP BY / WHERE / ORDER BY clauses in SQL queries to perform computation at server side is not permitted. All the work must be done by the SSIS package.

What to deliver: SSDT solution.