NATIONAL TECHNICAL UNIVERSITY OF ATHENS

SCHOOL OF ELECTRICAL AND COMPUTER ENGINEERING
DATA AND KNOWLEDGE MANAGEMENT LAB

Data Science and Machine Learning Master's Programme

Εξόρυξη Γνώσης από Δεδομένα (Data-Driven Knowledge Extraction)

Fall 2022
Instructors: V. Kantere, D. Tsoumakos

Assignment #2

Part I:

Suppose that the data for analysis includes the attribute `age`. The age values for the data tuples are: 21, 25, 16, 25, 33, 19, 45, 25, 22, 35, 52, 36, 70, 20, 35, 22, 35, 25, 15, 35, 20, 30, 33, 13, 40, 46, 16.
Use smoothing by:
(a) bin means (15%) and
(b) bin boundaries (15%),
to smooth the above data, using a bin depth of 3. Illustrate your steps.

Part II:

A data warehouse consists of three dimensions: `time`, `doctor`, and `patient`, and two measures: `count` and `charge` (charge is the fee that a doctor charges a patient for a visit).
(a) Draw a star schema diagram for the above data warehouse. Assume some useful attributes for each dimension table besides the necessary ones. (30%)
(b) Starting with the base cuboid `[day, doctor, patient]`, what specific OLAP operations should be performed in order to list the total fee collected by each doctor for 2021? (30%)
(c) Write a SQL query that answers the query of (b) assuming the data is stored in a single relational DB table with the schema:

`warehouse (day, month, year, doctor, hospital, patient, count, charge)`. (10%)

Deliverable:

- This is an individual assignment.
- Your solutions must be uploaded to the helios class page in pdf format by the deadline. No late submissions or different file formats will be accepted.