## HW1

**Due Date: Oct. 21, 2020**

**Submission requirements:**

**Please submit your solutions to our class website.**

1. Suppose that a data warehouse consists of four dimensions, *date*, *spectator*, *location*, and *game*, and two measures, *count and charge*, where charge is the fare that a spectator pays when watching a game on a given date. Spectators may be students, adults, or seniors, with each category having its own charge rate.
2. Draw a *star schema diagram* for the data warehouse.

(b) Starting with the base cuboid [*date, spectator, location, game*]，what specific *OLAP operations* should one perform in order to list the total charge paid by spectators in Chicago in 1999?

(c) Bitmap indexing is a very useful optimization technique. Please present the pros and cons of using bitmap indexing in this given data warehouse.

1. Suppose a hospital tested the age and body fat data for 18 random selected adults with the following result:

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| *age* | 23 | 23 | 27 | 27 | 39 | 41 | 47 | 49 | 50 | 52 | 54 | 54 | 56 | 57 | 58 | 58 | 60 | 61 |
| *%fat* | 9.5 | 26.5 | 7.8 | 17.8 | 31.4 | 25.9 | 27.4 | 27.2 | 31.2 | 34.6 | 42.5 | 28.8 | 33.4 | 30.2 | 34.1 | 32.9 | 41.2 | 35.7 |

1. Calculate the mean, median, and standard deviation of *age* and *%fat*.
2. Draw the boxplots for *age* and *%fat*.
3. Draw a scatter plot based on these two variables.
4. Normalize *age* based on *min-max normalization*.
5. Calculate the *correlation coefficient* (Pearson’s product moment coefficient). Are these two variables positively or negatively correlated?
6. Smooth the fat data by bin means, using a bin depth of 6.
7. Smooth the fat data by bin boundaries, using a bin depth of 6.
8. 某电子邮件数据库中存储了大量的电子邮件。请设计数据仓库的结构，以便用户从多个维度进行查询和挖掘。