HW1（满分 100）

截至日期: 4 月 21 日

1. 假定数据仓库中包含 4 个维：date, product, vendor, location；和两个度量：sales_volume 和 sales_cost。

(a) 画出该数据仓库的星形模式图. (10 分)

(b) 由基本方体[date, product, vendor, location]开始，列出每年 vendor A 的 sales_volume。(10 分)

2．假设某网络社交平台（例如，抖音，小红书，YouTube 等）数据库中存储了大量信息。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。（25 分）

3. Suppose a hospital tested the age and body fat data for 16 random selected adults with the following result:

age	23	23	27	27	39	41	47	49	50	52	53	53	54	55	56	57
%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	38	36	30.9

(a) Calculate the mean, median, and standard deviation of *age* and *%fat*. (10 分)

(b) Draw the boxplots for *age* and *%fat*. (10 分)

(c) Draw a *scatter plot* based on these two variables. (5 分)

(d) Normalize the two variables based on *Z-score normalization*. (10 分)

(e) Calculate the *correlation coefficient* (Pearson's product moment coefficient). Are these two variables positively or negatively correlated? (5 分)

4. 下面是一个超市某种商品连续 16 个月的销售数据（单位为百元）

23，23，27，27，39，41，47，49，50，52，53，53，54，55，56，57。

(a) 对以上数据进行深度为 4 的 Equal-depth binning，采用 bin median 方法进行平滑；(5 分)

(b) 对以上数据进行深度为 4 的 Equal-depth binning，采用 bin boundaries 方法进行平滑；(5 分)

(c) 对以上数据进行 Equal-width binning，分成 4 个 bin，采用 bin mean 方法进行平滑。(5 分)