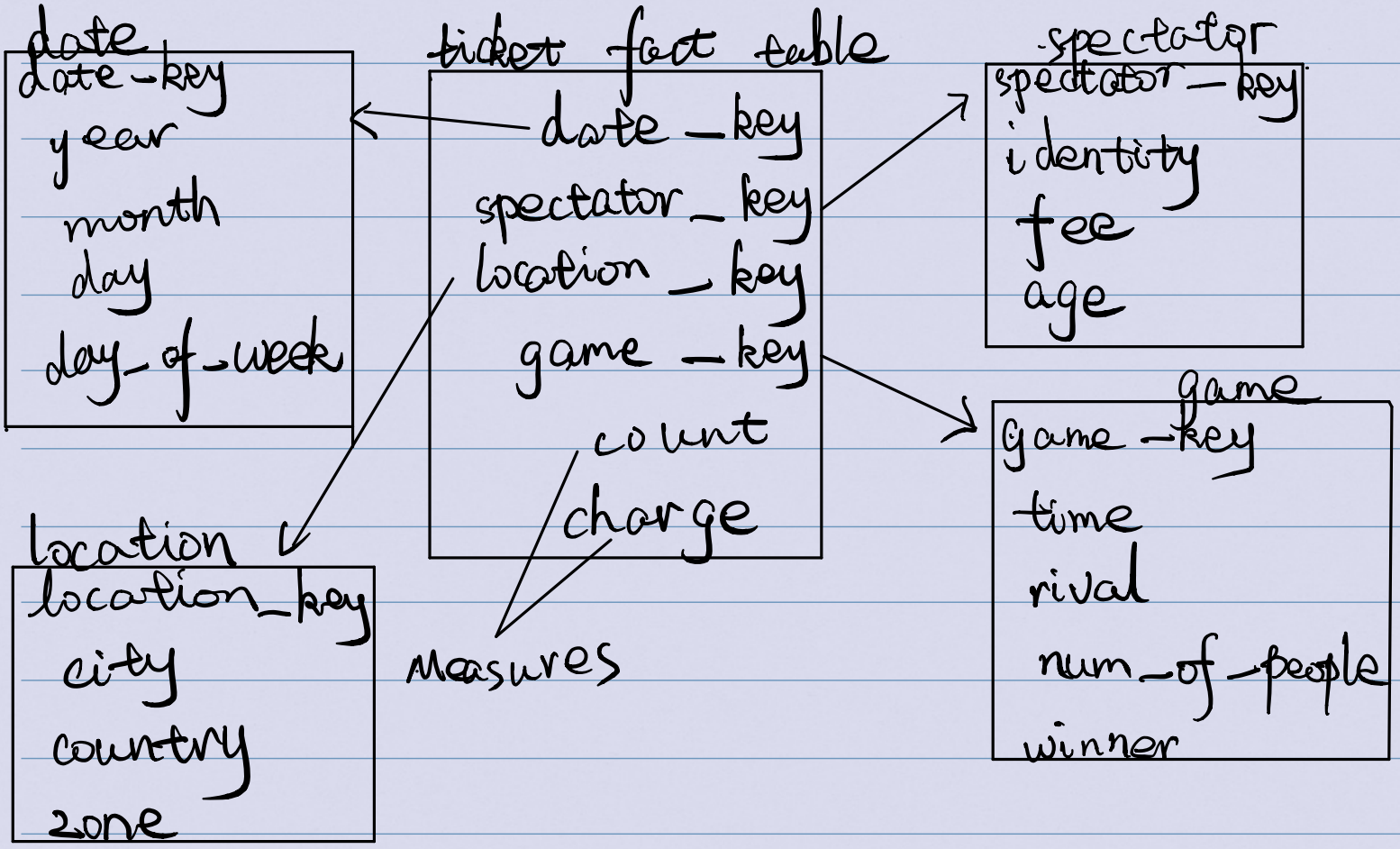
1(a)



(b) ① roll up from date to month to year
② slice for year = "1999"
③ roll up on spectator from individual patient to all
④ slice for spectator = "all"
⑤ roll up from location to city
⑥ slice for city = "chicago"
⑦ get the list of total charge paid by spectators in Chicago in 1999

(c) pros: ① 具有良好的 join 查询性能
② 适合大型数据仓库.

cons: ① 每个字段都需要一个，成本太高
② 对于不同观众收费不同，需要设置更多

2 (a)

age:

mean: 46.44

median: $\frac{50+52}{2} = 51$

standard deviation: $\sqrt{\Sigma(age(i)-mean)^2} \approx 12.846$

%fat

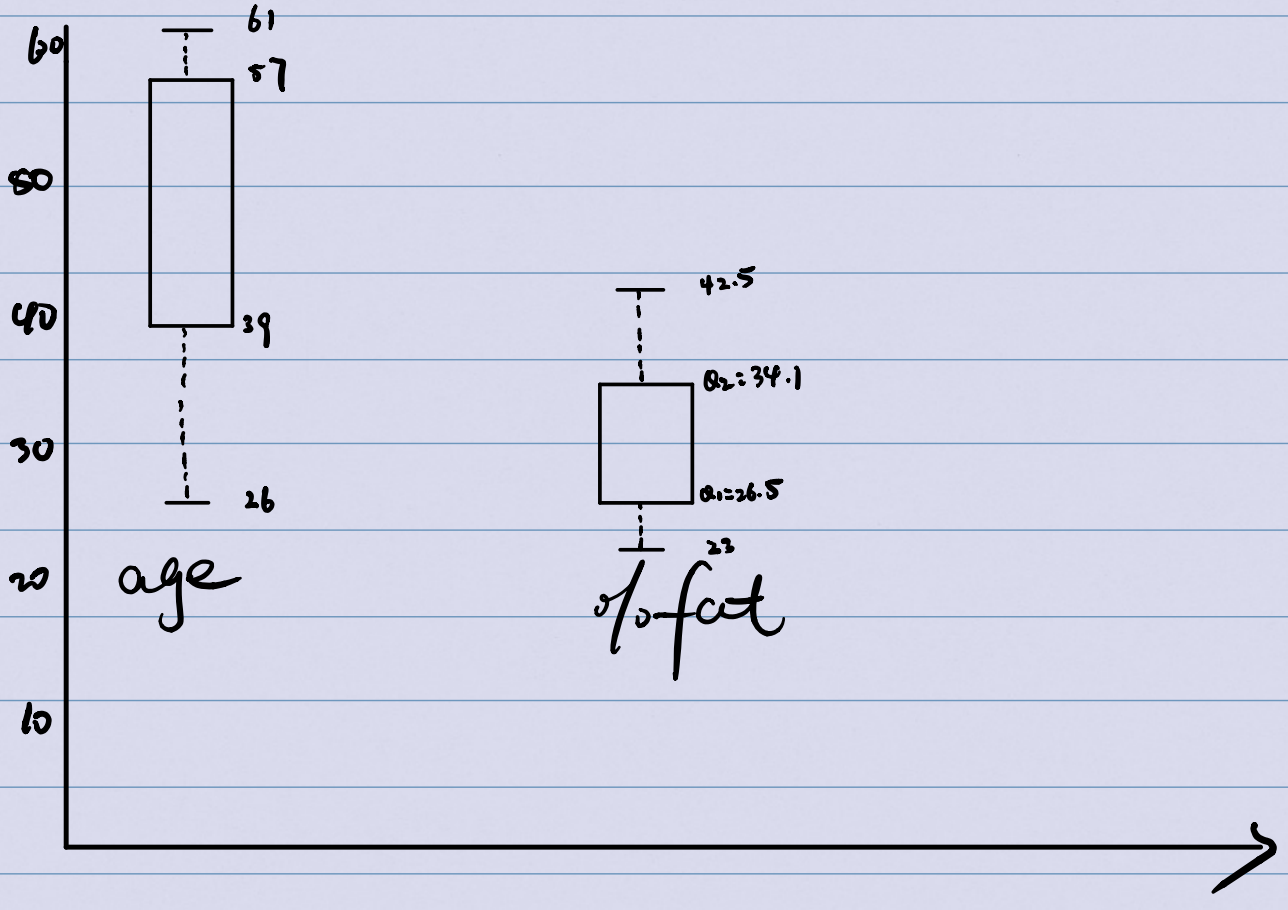
mean: 28.78

median: 30.7

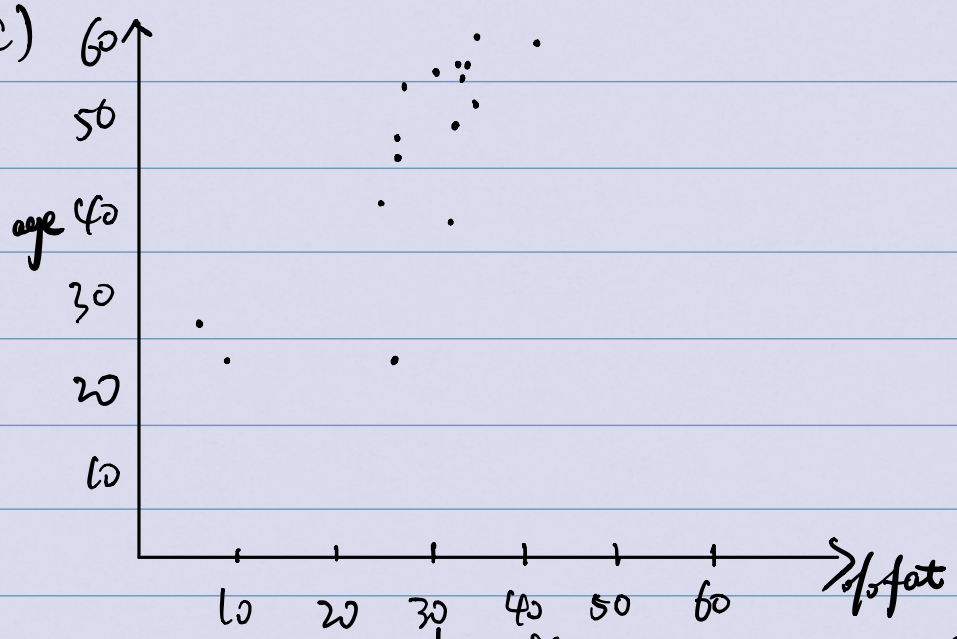
standard deviation: 8.994

(b) age: $Q_1 = 39$ %fat: $Q_1 = 26.5$

$Q_2 = 57$ $Q_2 = 34.1$



(c)



(d) min-max normalization on age = $\frac{age(i) - min\_age}{max\_age - min\_age}$

after min-max normalization:

age: [0, 0, 0.11, 0.11, 0.42, 0.47, 0.63, 0.68,
0.71, 0.76, 0.81, 0.81, 0.87, 0.89, 0.92,
0.92, 0.97, 1]

(e) $r_{A,B} = \frac{\sum(a_i - \bar{A})(b_i - \bar{B})}{(n-1)\sigma_A \sigma_B}$ $\bar{A} = 46.44$ $\bar{B} = 28.78$

$$r_{age,\%fat} = \frac{(23-46.44)(9.5-28.78)+(23-46.44)(26.5-28.78)+\cdots+(61-46.44)(28.78-35.7)}{(18-1)\times 12.846 \times 8.994}$$

$$= \frac{1700.33}{1964.127708} > 0$$

因此, age 和 fat 呈正相关

(f) fat data:

7.8, 9.5, 17.8, 25.9, 26.5, 27.2, 27.4, 28.8, 30.2, 31.2, 31.4, 32.9, 33.4, 34.1, 34.6, 35.7, 41.2, 42.5

Bin 1: 19.1, 19.1, 19.1, 19.1, 19.1, 19.1

Bin 2: 30.3, 30.3, 30.3, 30.3, 30.3, 30.3

Bin 3: 29.8, 29.8, 29.8, 29.8, 29.8, 29.8

(g) Bin 1: 7.8, 7.8, 27.2, 27.2, 27.2, 27.2

Bin 2: 27.4, 27.4, 32.9, 32.9, 32.9, 32.9

Bin 3: 33.4, 33.4, 33.4, 33.4, 42.5, 42.5

3. time 维, classification 维, location 维