~~200~~ $\text{median} = L_1 + \left(\frac{N/2 - (\sum \text{freq})_l}{\text{freq}_{\text{median}}}\right) \text{width}$

2.3 $21 + \left(\frac{1597 - 950}{1500}\right) 50$

$= 35.94$

width = 70

$L_1 = 21$ $n = 3194$ $(\sum_{\delta})_l = 950$, $\delta_{\text{req\_median}} = 1500$

2.6(a) Euclidean distance $= 6.708$

(22,1,42,10)

(20,0,36,8)

$\sqrt{4+1+36+4}$

Manhattan distance

$= 2+1+6+2 = 11$

minkowski distance

$q = 3$

$= \sqrt[3]{8+1+216+8} = 6.6114$ 0.1534

Supreme distance

$\max |x_{i\delta} - x_{j\delta}| = 6$

2.7. The formula in 2.3 in the book

$$\text{median} = L_1 + \left(\frac{N/2 - (\sum \text{freq})}{\text{freq}_{\text{median}}}\right)\text{width}$$

can be used to do median approximation

This method requires us to divide the data into several equal length intervals, ~~It's~~ let's assume it's i intervals

~~Accordin~~

In the formula

L_1 is the lower boundary of the median interval, N is How many value in the entire data set and $\sum \text{freq}$ is the sum of frequencies of all the intervals lower than the median interval, $\text{freq}_{\text{median}}$ is the frequency of median interval., the width is the width of the median interval.

For this formula

The higher the value of i which means more intervals the more accurate the resule will be

The heuristic method I thought is that to find out the outlier first before doing the approximation, and after eliminating the outlier, the median will be more representative

2.7

2.8

a. cosine similarity

$x, x_1 = 1.4 \times 1.5 + 1.6 \times 1.7$

$\frac{1.4 \times 1.5 + 1.6 \times 1.7}{\sqrt{1.4^2 + 1.6^2} \times \sqrt{1.5^2 + 1.7^2}}$

4.84 2.72 2.1

1.96 2.56 2.25

4.52 5.14

2.13 2.27

4.85

4.82004

cos sim $(x, x_1) = 0.9999917$

~~0.99979~~

Cosine similarity

$x, x_2 = \frac{1.4 \cdot 2 + 1.6 \cdot 1.9}{\sqrt{1.4^2 + 1.6^2} \times \sqrt{2^2 + 1.9^2}}$

cos sim (x, x_2) 0.99575

x_4 Cosine similarity

$x, x_4 = \frac{1.4 \cdot 1.2 + 1.6 \cdot 1.5}{\sqrt{1.4^2 + 1.6^2} \times \sqrt{1.2^2 + 1.5^2}}$

cos sim (x, x_4) 0.999028

cosine similarity

$\cos sim(x, x_3)\ x_1, x_3 = \frac{1.4 \times 1.6 + 1.6 \times 1.8}{\sqrt{1.4^2 + 1.6^2} \times \sqrt{1.6^2 + 1.8^2}} = 0.99997$

cosine similarity (x, x_5)

$x_1, x_5 = \frac{1.4 \times 1.5 + 1.6 \times 1}{\sqrt{1.4^2 + 1.6^2} \times \sqrt{1.5^2 + 1}} = 0.96536$

rank for cosine similarity
x_1, x_3, x_4, x_2, x_5

Manhatten distance

$d(x, x_1) = |1.4 - 1.5| + |1.6 - 1.7| = 0.2$

Manhatten distance

$d(x, x_2) = |1.4 - 2| + |1.6 - 1.9| = 0.9$

Manhatten distance

$d(x, x_3) = |1.4 - 1.6| + |1.6 - 1.8| = 0.4$

Manhatten distance

$d(x, x_4) = |1.4 - 1.2| + |1.6 - 1.5| = 0.3$

Manhatten distance

$d(x, x_5) = |1.4 - 1.5| + |1.6 - 1.0| = 0.7$

Rank for Manhatten rank: x_1, x_4, x_3, x_5, x_2

Euclidean distance

$$d(x, x_1) = \sqrt{(1.4-1.5)^2 + (1.6-1.7)^2} = 0.14142$$

$$d(x, x_2) = \sqrt{(1.4-2)^2 + (1.6-1.4)^2} = 0.67082$$

$$d(x, x_3) = \sqrt{(1.4-1.6)^2 + (1.6-1.8)^2} = 0.28284$$

$$d(x, x_4) = \sqrt{(1.4-1.2)^2 + (1.6-1.5)^2} = 0.22360$$

$$d(x, x_5) = \sqrt{(1.4-1.5)^2 + (1.6-1.0)^2} = 0.60828$$

rank, x_1, x_4, x_3, x_5, x_2.

Supreme distance

$d(x, x_1) = 0.1$
$d(x, x_2) = 0.6$
$d(x, x_3) = 0.2$
$d(x, x_4) = 0.2$
$d(x, x_5) = 0.6$

Rank $x_1, (x_3, x_4), (x_2, x_5)$

Normalize
(1.4, 1.6).

(b) $1.5 \times 1.5 = 2.25$
2.80.

$\frac{1.96 + 2.56}{\sqrt{1.4^2 + 1.6^2}} = 2.1260$
(0.6588, 0.7526)

2.26715681

Euclidean norm.

$\sqrt{1.5^2 + 1.7^2} = 2.26715681$

(1.5/2.26715681, 1.7/2.26715681)
= (0.6616, 0.7498)

$\sqrt{2^2 + 1.9^2} = 2.75862$
(2/2.75862, 1.9/2.75862)
(0.72500, 0.68878)

$\frac{2.56 \quad 3.24}{\sqrt{1.6^2 + 1.8^2}} = 2.4083.$
(0.6644, 0.7474)

$\frac{1.44 \quad 2.25}{\sqrt{1.2^2 + 1.5^2}} = 1.9209.$
(0.6247, 0.7809)

$\frac{2.25 + 1}{\sqrt{1.5^2 + 1.0^2}} = 1.8027$
(0.8321, 0.5547)

New Euclidean distance

$$C(X, X_1) = \sqrt{(0.6585 - 0.6616)^2 + (0.7526 - 0.7448)^2}$$
$$\approx 0.0042$$

$$(X, X_2) = \sqrt{(0.6585 - 0.725)^2 + (0.7526 - 0.6875)^2}$$
$$= 0.0922$$

$$(X, X_3) = \sqrt{(0.6585 - 0.6644)^2 + (0.7526 - 0.7474)^2}$$
$$= 0.00786$$

$$(X_1, X_4) = \sqrt{(0.6585 - 0.6247)^2 + (0.7526 - 0.7809)^2}$$
$$= 0.0441$$

$$(X_1, X_5) = \sqrt{(0.6585 - 0.8321)^2 + (0.7526 - 0.5597)^2}$$
$$= 0.2635$$

rank. X_1, X_3, X_4, X_2, X_5

3.1

From the perspective of quality, different intended use can have different requirement for ~~data quality~~ accuracy

For example a marketing analyst may find ~~an~~ a database ~~with~~ of customer address with 80% accurate data good enough for ~~his~~ him ~~analyze~~ to analyze while the sales manager think it's not good enough.

another example military may need higher accuracy ~~for~~ of gps location database than normal citizen who use ~~gps~~ gps for navigation

From the perspective of completeness ~~some of~~ ~~some~~ the data may be incomplete beaus they were not necessary ~~when~~ when the database was created. But later were considered important.

For example, a database created to record ~~how many cars~~ the information of cars of each household may not record the color of car but later it becomes importane when we ~~can't~~ want to study what's the most popular car color.

From the perspective of consistency the data

The data might be inconsistent if there are requirement about ~~consistency~~ completeness of data, because the data might be incomplete if inconsistent data are deleted.

For example, if a customer has two home address record in database, it's difficult to determine which one to use, but it's complete

Two other dimension
timeliness, believability
interpretability

3.3

13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70.

1- sore the data

2. Bin 1: 13, 15, 16,
Bin 2: 16, 19, 20,
Bin 3, 20, 21, 22,
Bin 4, 22, 25, 25,
Bin 5: 25, 25, 30,
Bin 6: 33, 33, 35,
Bin 7. 35, 35, 35
Bin 8: 36, 40, 45,
Bin 9: 46, 52, 70.

3. replace each value in a bin by the mean value of the bean.

Bin 0: 44/3, 44/3, 44/3 (14.67)
Bin 2: 55/3, 55/3, 55/3 (18.33)
Bin 3: 21, 21, 21
Bin 4: 24, 24, 24
Bin 5: 80/3, 80/3, 80/3 (26.67)
Bin 6: 101/3, 101/3, 101/3 (33.67)
Bin 7: 35, 35, 35
Bin 8: 121/3 121/3 121/3 (40.33)
Bin 9: 56, 56, 56,

~~We may not be a~~
we can't spot the biggest and smallest data and it's hard to identify outliers

(b) We can ~~cluter~~ use clustering to identify out liers, put similar data into clusters, and data that are not in clusters will be considered as outlier.

(C) some other methods for data smoothing are smoothing by bin medians
smoothing by bin boundaries
Regression

3.5

(a) Min-Max Normalization

from the formula

$$\frac{v_i - min_A}{max_A - min_A}$$

we get that the maximum value will be transformed to 1
the minimum value will be transformed to 0, other value will be transformed to decimal between 0 and 1.

(b) Z-Score normalization

$$v_i' = \frac{v_i - \bar{A}}{\sigma_A}$$

$\bar{A}$ — mean

σ_A — standard deviation

if value above mean: positive
value below mean, negative
range decide by standard deviation

$[-\infty, \infty]$ $\left[\frac{min-\bar{A}}{\sigma_A}, \frac{max-\bar{A}}{\sigma_A}\right]$

[c] Z-score normalization using mean absolute deviation

$$v_i' = \frac{v_i - \bar{A}}{s_A}$$

range decide by mean absolute deviation

$[-\infty, \infty]$ $\left[\frac{min-\bar{A}}{s_A}, \frac{max-\bar{A}}{s_A}\right]$

[d] normalization by decimal scaling

$[-1, 1]$

3. [a] min-max normalization

min = 13
max = 70

$\frac{35-13}{70-13}(1-0)+0 \approx 0.386$

22
57

[b] z-score normalization

$v_i' = \frac{v_i - \bar{A}}{\sigma A}$

sum of all data

809

mean = 809/27 = 29.96

z-score normalization $\frac{35-29.96}{12.94} \approx 0.389$

[c] data scaling $v_i' = \frac{35}{10^j}$

j is smallest integer such that $\max(|v_i'|) < 1$

$j = 2$

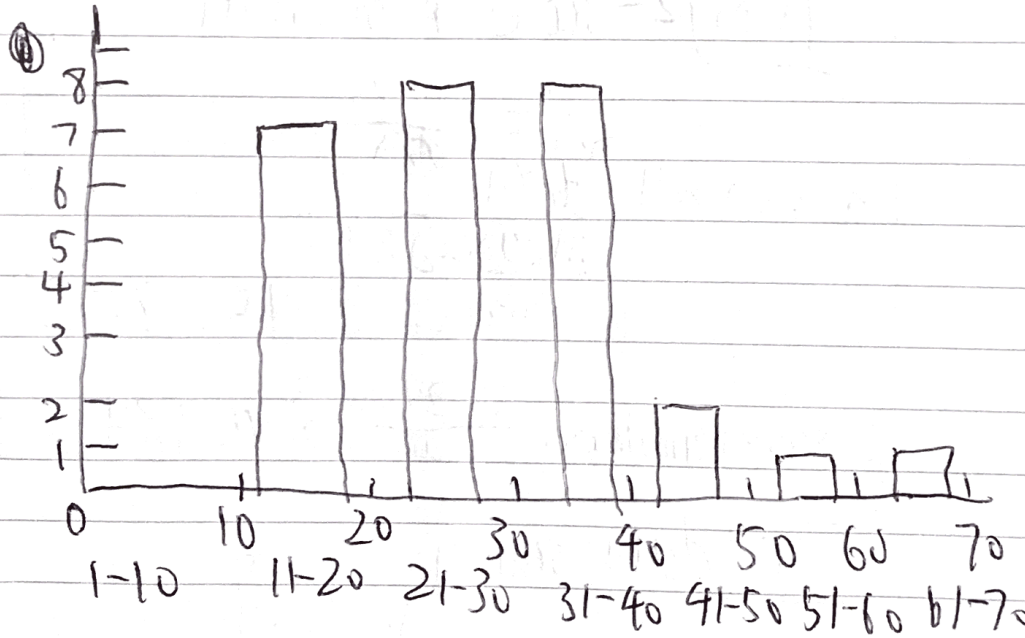
$v_i' = 0.35$

(d) I think decimal scaling is the bese in this case because it preserves the distribution of the data and intuitive for mining into.

Min-Max normalization prevent future data to fall outside current max and min value

and z score normalization. interpret the data in terms of their distance from mean which might not be all that useful

3.11



~~SRSWOR~~

$T_1 = 13$
$T_2 = 15$
$T_3 = 16$
$T_4 = 16$
$T_5 = 19$
$T_6 = 20$
$T_7 = 20$
$T_8 = 21$
$T_9 = 22$
$T_{10} = 22$
$T_{11} = 25$
$T_{12} = 25$
$T_{13} = 25$
$T_{14} = 25$
$T_{15} = 30$
$T_{16} = 33$
$T_{17} = 33$
$T_{18} = 35$
$T_{19} = 35$
$T_{20} = 35$
$T_{21} = 35$
$T_{22} = 36$
$T_{23} = 40$
$T_{24} = 45$
$T_{25} = 46$
$T_{26} = 52$
$T_{27} = 70$

SRSWOR $n = 5$

T_1	13
T_7	20
T_{12}	25
T_{24}	45
T_{27}	70

SRSWR $n = 5$

T_4	16
T_9	22
T_9	22
T_{21}	35
T_{26}	52

clusters

$T_1 = 13$
$T_2 = 15$
$T_3 = 16$
$T_4 = 16$
$T_5 = 19$

$T_6 = 20$
$T_7 = 20$
$T_8 = 21$
$T_9 = 22$
$T_{10} = 22$

$T_{11} = 25$
$T_{12} = 25$
$T_{13} = 25$
$T_{14} = 25$
$T_{15} = 30$

$T_{16} = 33$
$T_{17} = 33$
$T_{18} = 35$
$T_{19} = 35$
$T_{20} = 35$

$T_{21} = 35$
$T_{22} = 36$
$T_{23} = 40$
$T_{24} = 45$
$T_{25} = 46$

$T_{28} = 52$
$T_{27} = 70$

SRSWOR

$S = 2$

$T_{16} = 33$
$T_{17} = 33$
$T_{18} = 35$
$T_{19} = 35$
$T_{20} = 35$

$T_6 = 20$
$T_{27} = 20$
$T_8 = 21$
$T_9 = 22$
$T_{10} = 22$

~~Strai~~ Stratified sampling

T1	13	young
T2	15	young
T3	16	young
T4	16	young
T5	19	young
T6	20	young
T7	20	young

T8	21	young
T9	22	young
T10	22	young
T11	25	young
T12	25	young
T13	25	young
T14	25	young

T15	30	mid-age
T16	33	mid-age
T17	33	mid-age
T18	33	mid-age
T19	33	midage
T20	35	mid-age
T21	35	midage
T22	36	mid-age

T23	40	midage
T24	45	mid-age
T25	46	midage
T26	52	mid-age
T27	70	senior

T7	20	young
T9	22	young
T15	30	mid
T19	33	mid
T27	70	senior