Yurim Park HW1

1. a) $\frac{809}{27} \approx 30$, 25

b) The mode is the value in the data set that appear more frequently. In this data 25,35 is the mode.

Since two values appear more frequently, the data set is bimodal.

c) $\frac{83}{2} = 41.5$

d) 20, 35

e) Minimum : 13

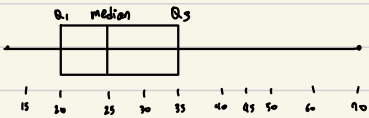
Q1 : 20

Median : 25

Q3 : 35

Maximum : 90

f.



g. The key difference is that a quantile plot focuses on the distribution of a single dataset, while a Q-Q plot compares the distribution of two datasets to see how well they align.

2. a) mean(age) ≈ 46.44

mean(%fat) ≈ 28.78

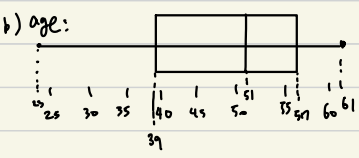
median(age) = 51

median(%fat) = 30.7

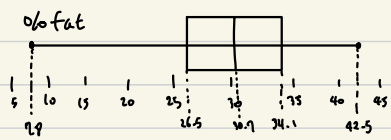
6age ≈ 12.85

6%fat ≈ 8.99

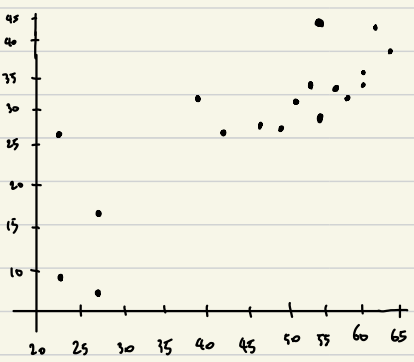
b) age:



%fat



c)



Scatter plot

3. a) $\sqrt{45} \approx 6.71$

b) 11

c) $\sqrt[3]{233} \approx 6.1534$

d) 6

4.

	cosine similarity	Euclidean distance	Manhattan distance
a)	1. X_1	1. X_1	1. X_2
	2. X_3	2. X_4	2. X_1
	3. X_4	3. X_3	3. X_3
	4. X_2	4. X_5	4. X_4
	5. X_5	5. X_2	5. X_5

b) X_1: 0.0041

X_2: 0.0922

X_3: 0.0098

X_4: 0.0441

X_5: 0.2632

5. a) smoothing by bin means involves replacing values in each group with the mean of the group

b) outliers can be identified using statistical measures like z-scores or the interquartile range.

c) Other methods for data smoothing include moving averages, exponential smoothing, polynomial regression, and Savitzky-Golay filtering.

6. a) 0.39

b) 0.39

c) 0.35

d) Min-Max and Z-score are commonly used, but Z-score might be more appropriate if the data is normally distributed.

7. a)

age	23	23	27	27	39	41	47	49	50
Z-age	-1.83	-1.83	-1.31	-1.31	-0.58	-0.38	0.04	0.20	0.28
% fat	9.5	26.5	17.8	18.8	31.4	25.9	27.4	27.2	31.2
Z-% fat	-2.14	-0.25	-2.33	-1.22	0.29	-0.32	-0.13	-0.18	0.27

a)

52	54	54	56	57	58	58	60	61
0.43	0.59	0.59	0.74	0.82	0.90	0.90	1.06	1.13
34.6	42.5	28.8	33.4	30.2	34.1	32.9	41.2	35.7
0.65	1.53	0.0	0.51	0.16	0.59	0.46	1.38	0.00

b) 0.82, since the correlation coefficient is positive, the two attributes are positively related.