Data Mining
HW 1

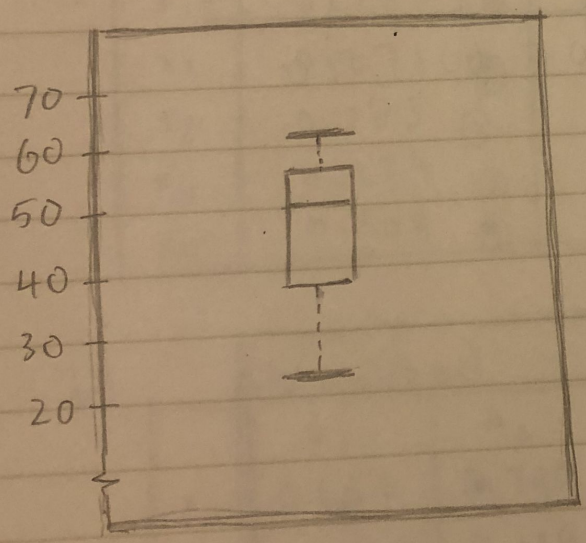
2.4

a)

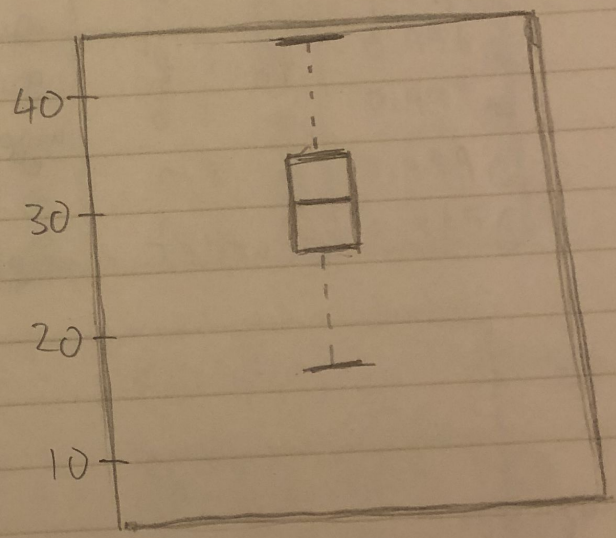
	Age	% fat
Mean	46.44	28.78
Median	51	30.7
SD	12.85	8.99

b) Boxplot

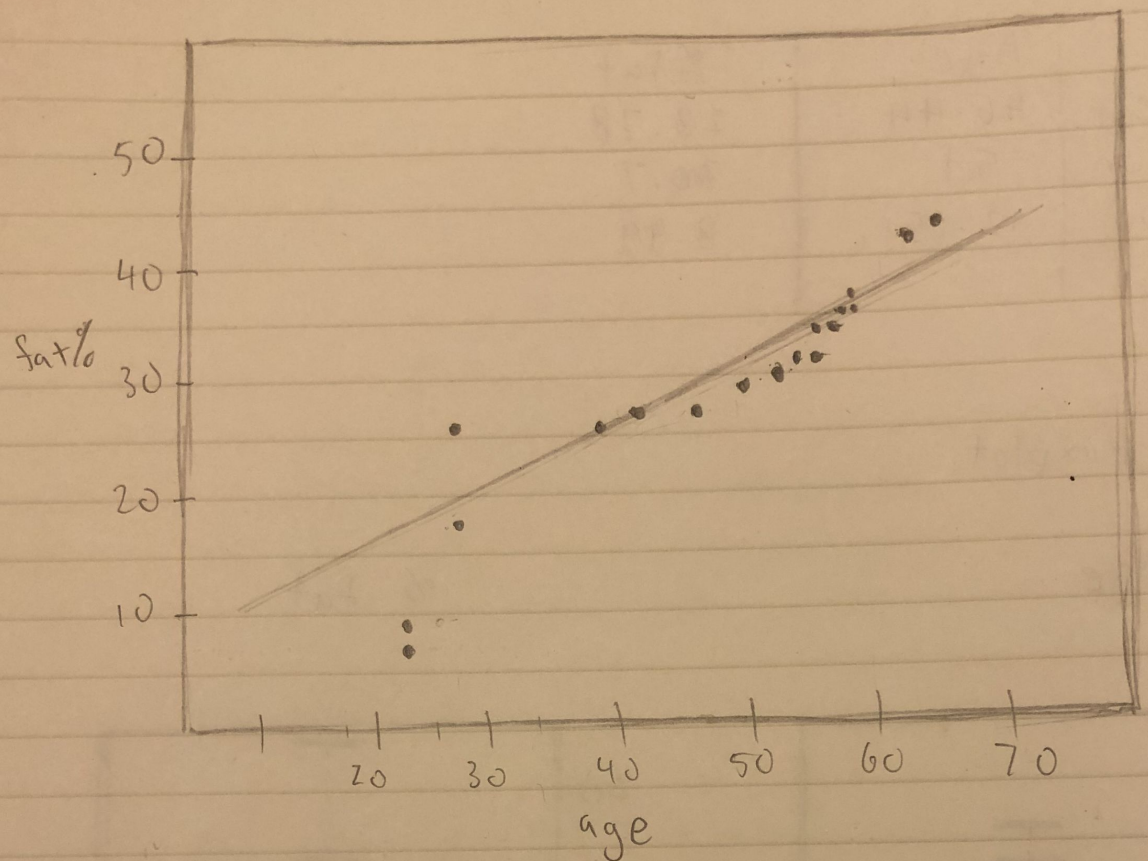
Age



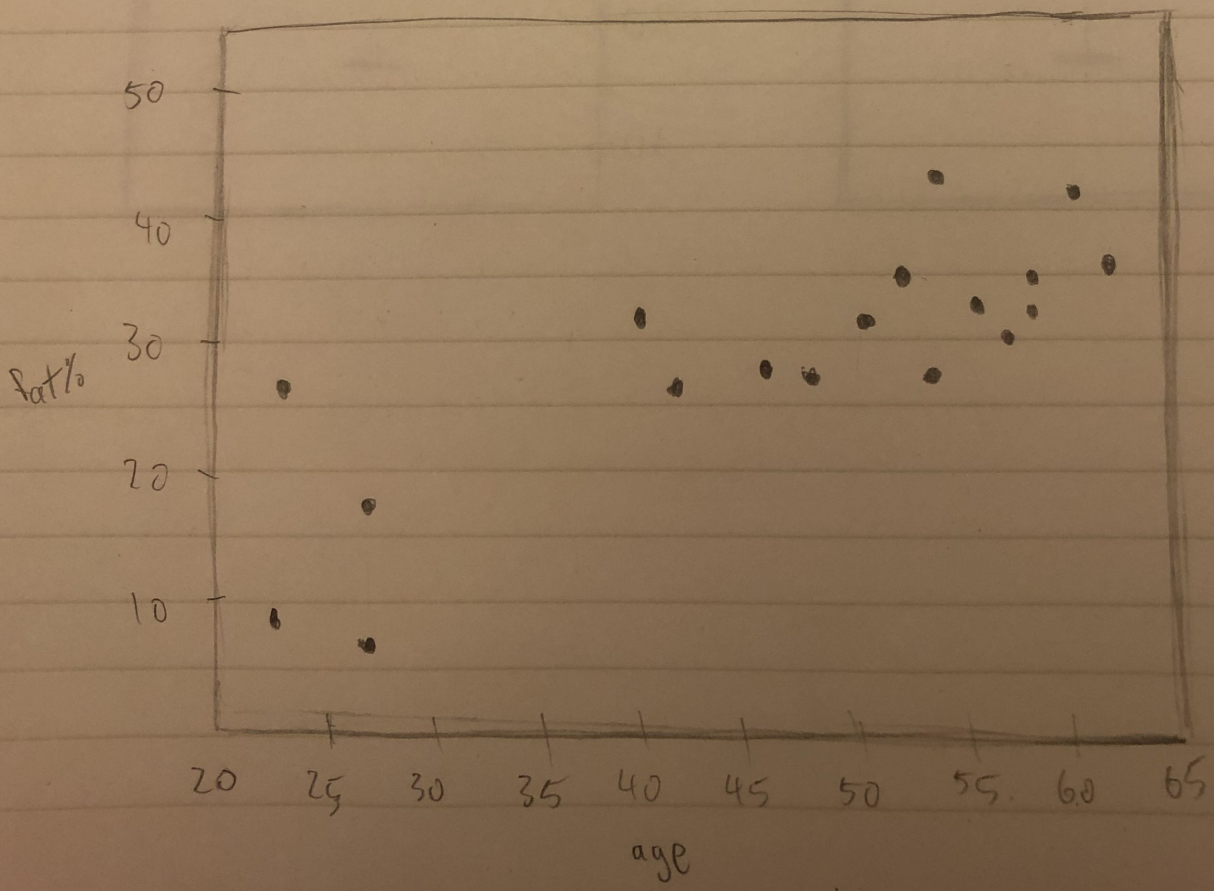
% fat



c) qq plot



scatter plot



2.8.

a) Euclidean Distance :

$$d(i,j) = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \dots + (x_{in} - x_{jn})^2}$$

Manhattan Distance :

$$d(i,j) = |x_{i1} - x_{j1}| + |x_{i2} - x_{j2}| + \dots + (x_{in} - x_{jn}|$$

Supreme Distance :

$$d(i,j) = \lim_{h \to \infty} \left(\sum_{f=1}^{p} |x_{if} - x_{jf}|^h \right)^{\frac{1}{h}} = \max_{f}^{p} |x_{if} - x_{jf}|$$

* Calculate using excel and online calculator

	Euclidean		Manhat		Supreme		Cosine	
x_1	0.141	①	0.2	①	0.1	①	0.999	①
x_2	0.671	⑤	0.9	⑤	0.6	④	0.996	④
x_3	0.283	③	0.4	③	0.2	②	0.999	②
x_4	0.224	②	0.3	②	0.2	③	0.999	③
x_5	0.608	④	0.7	④	0.6	⑤	0.965	⑤

b) Normalized

	A_1	A_2
x_1	0.662	0.749
x_2	0.725	0.689
x_3	0.664	0.747
x_4	0.625	0.781
x_5	0.832	0.555

Euclidean Distance

x_1	0.004	①
x_2	0.092	④
x_3	0.008	②
x_4	0.044	③
x_5	0.263	⑤

3.7

a) Age 35 → [0.0, 1.0]

$= \frac{(35-13) \times 1}{57} + 0 = \frac{22}{57} = 0.386$

b) $= \frac{value - M}{\sigma}$

$= \frac{35 - \frac{809}{27}}{12.94} = 0.389$

c) $= \frac{value}{10^j}$

$= \frac{35}{10^2} = \frac{35}{100} = 0.35$

d) Decimal Scaling would be a better method for normalization because it would still result in the same distribution, but in a simpler fashion, allowing for further mining / analyzing. Min-max wouldn't be very appropriate due to outliers such as 70.

3.9 a) 1 : 5, 10, 11, 13
2 : 15, 35, 50, 55
3 : 72, 92, 204, 215

b) 3 equi-width $\frac{215-5}{3} = 70$ 5 — 75 — 145 - 215
1 : 5, 10, 11, 13, 15, 35, 50, 55, 72
2 : 92
3 : 204, 215