Question 1

(1)

I choose to use equal-width binning method. Since the question require us to make two groups with high and low pollution, this method can make two groups has significant difference value for water pollution. Besides, there is no obvious outlier. So, I choose this method.

The maximum water pollution is 90.7, the minimum is 6. I take the average of them, that is 48.35, to divide the data into two group as below (Line0-Line3 are Experiment, Substance A, Substance B, Water Pollution) :

Group1:

	0	1	2	3	4	5	6	7	8	9	10
0	3.000	5.000	8.000	10.000	14.000	15.000	17.000	18.000	19.000	22.000	24.000
1	7.590	7.310	7.340	7.170	7.790	0.630	7.210	6.120	4.240	0.670	1.440
2	76.700	58.400	83.400	86.900	61.400	11.300	88.000	35.400	53.300	9.200	32.900
3	73.400	74.900	64.900	76.800	55.500	61.400	90.700	70.100	60.000	80.100	64.900

11	12	13	14	15	16
29.000	31.000	33.000	36.000	39.000	40.000
2.990	8.620	7.440	1.740	2.840	1.430
20.600	78.900	98.800	12.700	33.600	26.200
77.000	81.200	83.800	69.000	77.900	62.100

Group2:

	0	1	2	3	4	5	6	7	8	9	10	11	12
0	1.000	2.000	4.000	6.000	7.000	9.000	11.000	12.000	13.000	16.000	20.000	21.000	23.000
1	2.840	9.340	0.210	2.770	4.410	9.940	0.660	2.730	2.140	7.360	9.790	5.070	9.830
2	78.900	52.600	39.600	98.100	4.800	16.400	61.600	67.200	85.000	13.000	16.800	87.900	55.300
3	11.500	17.700	11.200	6.000	15.200	14.100	11.300	9.400	13.800	21.400	6.700	13.000	8.400

13	14	15	16	17	18	19	20	21	22
25.000	26.000	27.000	28.000	30.000	32.000	34.000	35.000	37.000	38.000
5.760	0.170	7.900	9.290	2.000	7.210	8.700	3.050	3.410	7.500
94.400	41.900	15.500	13.600	93.300	20.600	48.300	86.300	1.600	22.700
9.300	9.600	11.500	7.500	11.900	14.900	15.800	16.500	7.600	10.600

(2)

Min-max normalization:

$$x = \frac{x - min}{max - min}(\max_{new} - \min_{new}) + \min_{new}$$

$\max_{new}$=1, $\min_{new}$=0.

After normalization, I got:

Group1:

	0	1	2	3	4	5	6	7	8	9	10
0	3.000	5.000	8.000	10.000	14.000	15.000	17.000	18.000	19.000	22.000	24.000
1	0.871	0.836	0.840	0.819	0.896	0.000	0.836	0.710	0.492	0.078	0.167
2	0.753	0.588	0.843	0.879	0.619	0.109	0.891	0.358	0.539	0.092	0.332
3	73.400	74.900	64.900	76.800	55.500	61.400	90.700	70.100	60.000	80.100	64.900

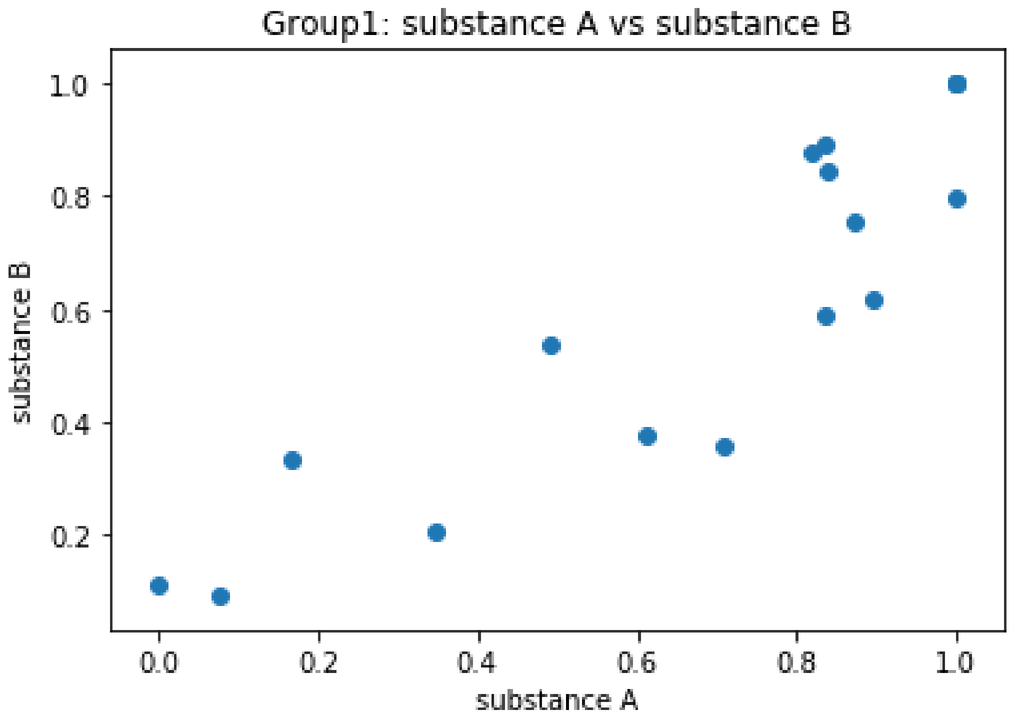
11	12	13	14	15	16
29.000	31.000	33.000	36.000	39.000	40.000
0.347	1.000	1.000	0.613	1.000	1.000
0.208	0.798	1.000	0.376	1.000	1.000
77.000	81.200	83.800	69.000	77.900	62.100

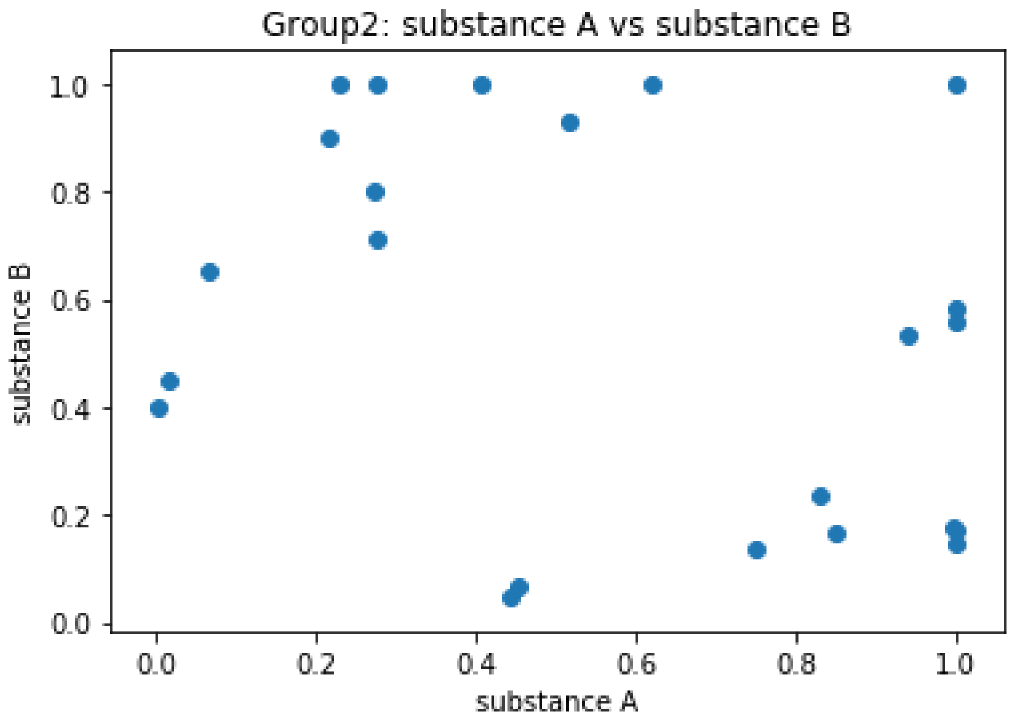
Group2:

	0	1	2	3	4	5	6	7	8	9	10	11	12
0	1.000	2.000	4.000	6.000	7.000	9.000	11.000	12.000	13.000	16.000	20.000	21.000	23.000
1	0.273	0.939	0.004	0.278	0.443	1.000	0.067	0.277	0.217	0.749	0.996	0.516	1.000
2	0.801	0.532	0.400	1.000	0.047	0.173	0.652	0.712	0.900	0.137	0.178	0.931	0.586
3	11.500	17.700	11.200	6.000	15.200	14.100	11.300	9.400	13.800	21.400	6.700	13.000	8.400

13	14	15	16	17	18	19	20	21	22
25.000	26.000	27.000	28.000	30.000	32.000	34.000	35.000	37.000	38.000
0.620	0.018	0.850	1.000	0.230	0.829	1.000	0.406	0.454	1.000
1.000	0.449	0.166	0.145	1.000	0.238	0.559	1.000	0.069	1.000
9.300	9.600	11.500	7.500	11.900	14.900	15.800	16.500	7.600	10.600

(3)


Group1: substance A vs substance B
substance B
substance A
1.0
0.8
0.6
0.4
0.2
0.0
0.2
0.4
0.6
0.8
1.0


Group2: substance A vs substance B
substance B
substance A
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0

(4)

Pearson's correlation coefficient: $\mathrm{r(x,y)} = \frac{\sum(x_i - \bar{X})(y_i - \bar{Y})}{n\sigma_x\sigma_y}$

Group1: r=0.89

Group2: r=-0.34

(5)

From Pearson correlation coefficient and scatter plot, when pollution is low, the correlation between A and B is negative. And, when pollution is high, the correlation between A and B is positive. Thus, to efficiently decrease pollution, we need add A and B in a reverse amount.

Question 2

(1)

Mean=39.75

Median=41.0

Standard deviation=16.16

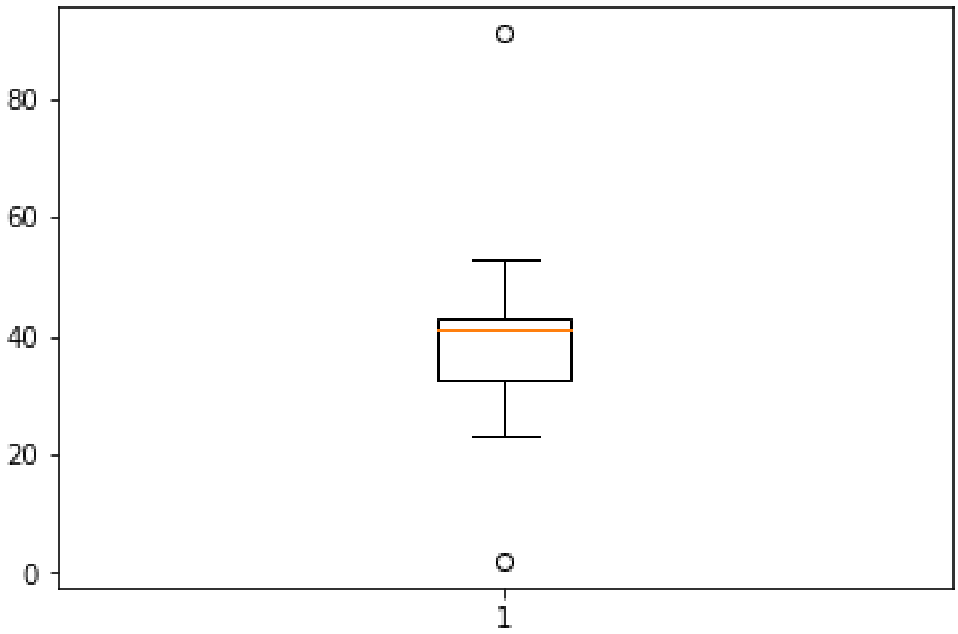
Q1=32.75

Q2=43.0

I calculate the quantile by np.percentage(x,percent).

(2)

Part(a):



From the boxplot, there are outliers: 2 and 91. Drop them and I got:

New dataset A = [34, 32, 53, 33, 43, 43, 38, 41, 42, 49, 25, 41, 36, 42, 52, 32, 23, 43]

Part(b):

Smooth by bin means with bin depth of 5:

New dataset B = [22.8,22.8,22.8,22.8,22.8,36.4,36.4,36.4,36.4,36.4,42.2,42.2,42.2,42.2,42.2,57.6,57.6,57.6,57.6,57.6]

Part(c):

Smooth by bin boundaries with bin depth of 5:

New dataset C= [2,32,32,32,32,33,33,33,41,41,41,41,41,43,43,43,43,43,43,91]

(3)

Data	Original data	New data A	New data B	New data C
Mean	39.75	39.0	39.75	39.15
Median	41.0	41.0	39.3	41.0
Standard deviation	16.16	8.03	12.48	14.99
Q1	32.75	33.25	33.0	32.75
Q3	43.0	43.0	46.05	43.0

If we need the lowest variance, we should drop the outlier, which will decrease the variance in the greatest degree.

If we need the least skewness, we should use smoothing by bin means, which has the closest mean and median.

If we need closest Q1 and Q3, we should use smoothing by bin boundaries.

Question 3

A:$Z = \frac{x-u}{s} = \frac{3.5-3.2}{0.5} = 0.6$

B:$Z = \frac{x-u}{s} = \frac{3.7-3.4}{0.4} = 0.75$

C:$Z = \frac{x-u}{s} = \frac{3.4-3.2}{0.35} = 0.57$

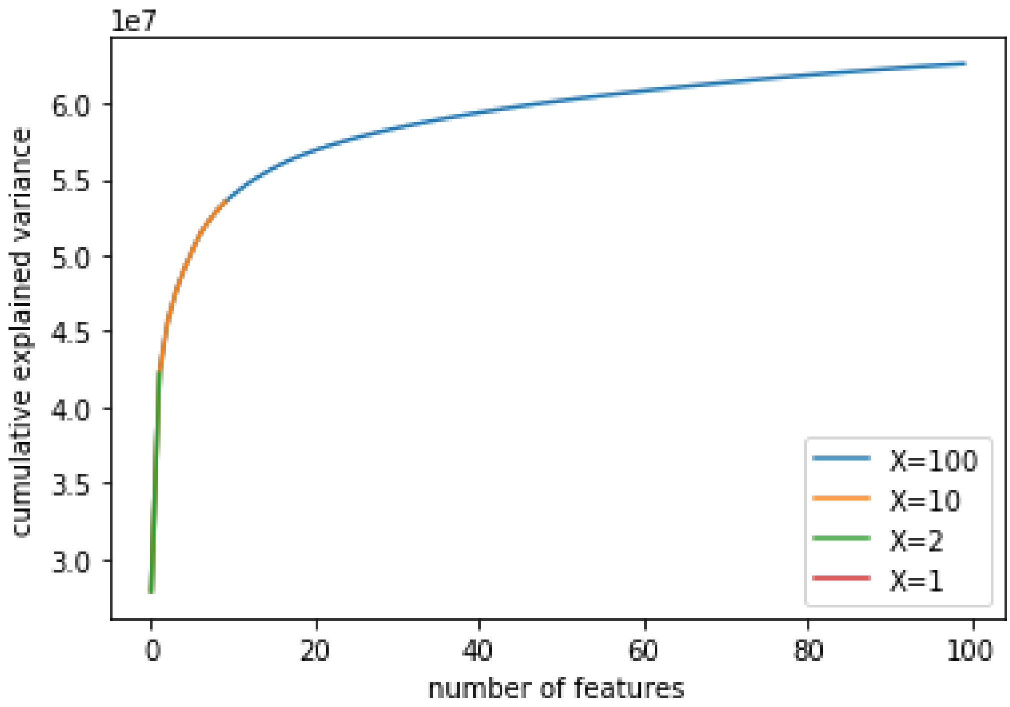
D:$Z = \frac{x-u}{s} = \frac{3.8-3.9}{0.5} = -0.2$

E:$Z = \frac{x-u}{s} = \frac{3.9-3.8}{0.2} = 0.5$

Normalized GPA: B>A>C>E>D

Question 4

(1)



The more features, cumulative explained variance become higher. However, when features reach 10, the increasing rate of cumulative explained variance become much slower. If we keep increasing features after 10, the variance will not increase too much.

(2)

	Manhattan	Euclidean	Supremum	Cosine
Original Dataset (X=-1)	54 19 94 46 56	54 19 56 94 46	86 16 75 79 29	54 19 56 94 46
X=1000				
X=100	94 19 54 56 95	54 19 56 94 46	46 56 19 54 5	54 19 56 94 46
X=10	94 19 54 46 56	94 19 54 46 56	94 19 54 46 44	94 19 54 46 56
X=2	94 44 5 62 70	94 44 70 62 5	94 44 70 62 91	94 16 70 44 62
X=1	94 16 70 44 62	94 16 70 44 62	94 16 70 44 62	5 6 10 12 14

Ps: Since there are only 100 instances, sklearn library cannot support that number of features is greater than number of instances. So, I choose features up to 100.

(3)

From the table, when X=100 or 10, the most of similar patients are same with the original dataset. Also, from the chart in (1), the accumulative explained variance grows slowly since X=10. Thus, I'd like to choose X=10. Because, compared with X=100, its accuracy does not decrease much and it save much computational cost.