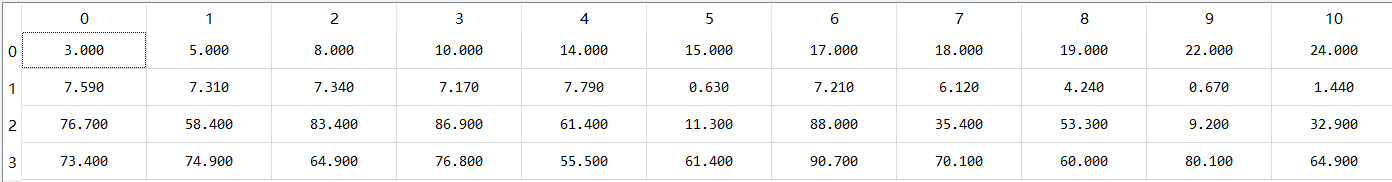
**Question 1**

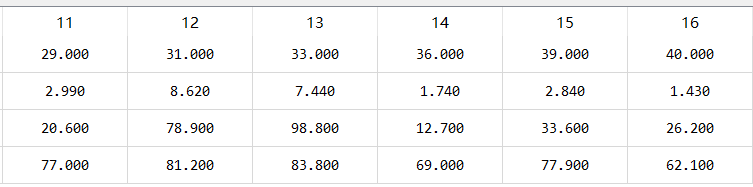
(1)

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.

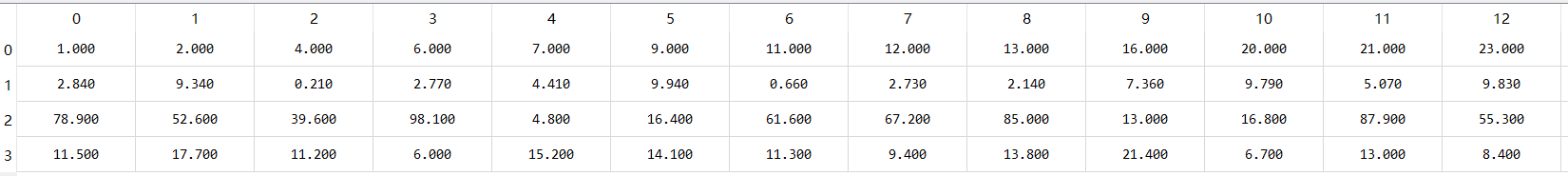
The maximum water pollution is 90.7, the minimum is 6. I take the average of them, that is 48.35, to divide the data into two group as below (Line0-Line3 are Experiment, Substance A, Substance B, Water Pollution) :

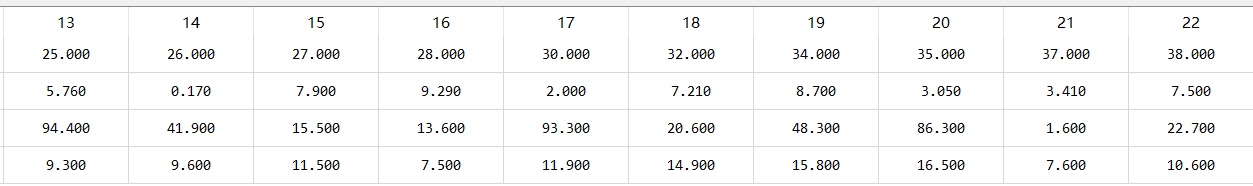
Group1:





Group2:





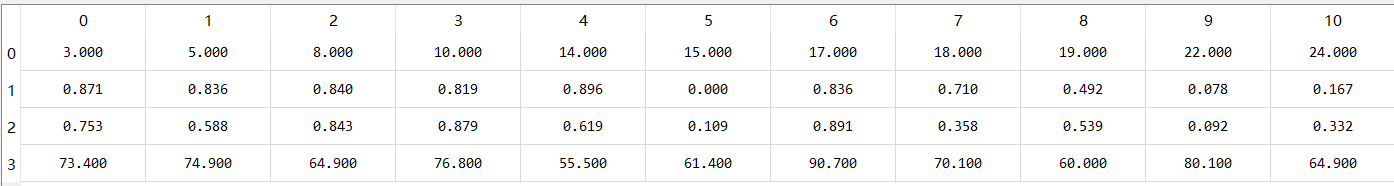
(2)

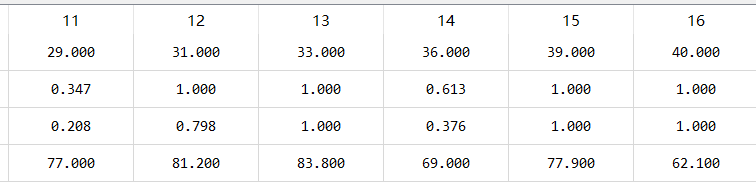
Min-max normalization:

=1, =0.

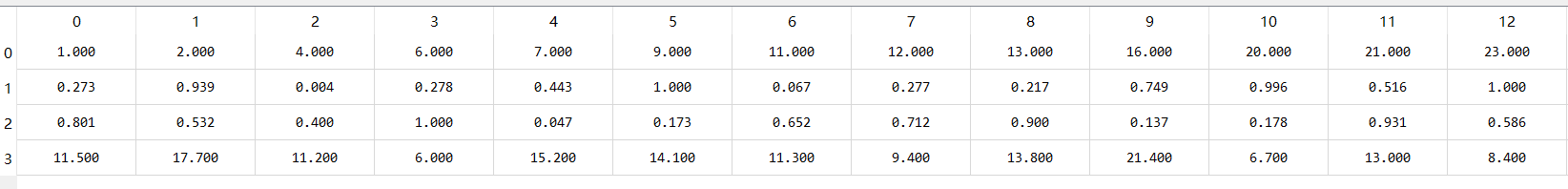
After normalization, I got:

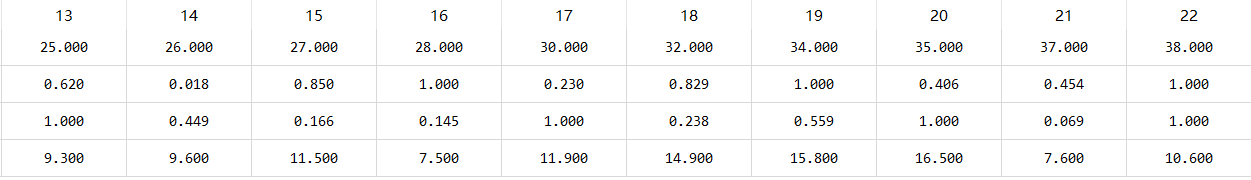
Group1:



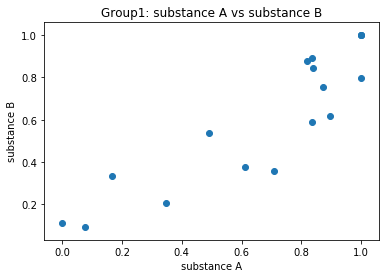


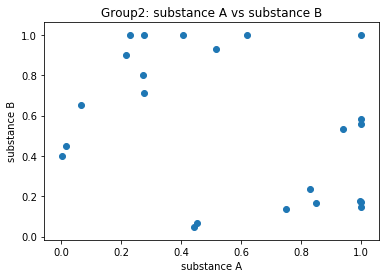
Group2:





(3)





(4)

Pearson’s correlation coefficient:

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

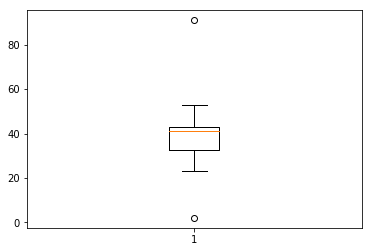
Q1=32.75

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:

B:

C:

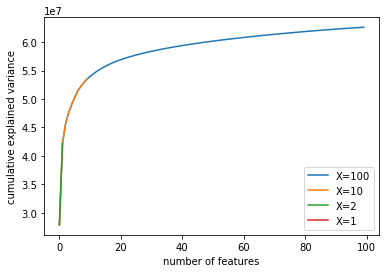
D:

E:

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.