CS412
Tianqi Wu
HW1

Question 1

1) Equal-width binning is more appropriate for this dataset. Our purpose is to divide data into groups with low and high water quality. Equal-width binning divides water quality at 48.35, which separates water with low and high quality quite well, while equal-depth binning divides water quality at 15.8, which does not quite distinguish low and high water well. Groups are divided below:

Group 1

	Experiment	Substance A	Substance B	Water Pollution
2	3.0	7.59	76.7	73.4
4	5.0	7.31	58.4	74.9
7	8.0	7.34	83.4	64.9
9	10.0	7.17	86.9	76.8
13	14.0	7.79	61.4	55.5
14	15.0	0.63	11.3	61.4
16	17.0	7.21	88.0	90.7
17	18.0	6.12	35.4	70.1
18	19.0	4.24	53.3	60.0
21	22.0	0.67	9.2	80.1
23	24.0	1.44	32.9	64.9
28	29.0	2.99	20.6	77.0
30	31.0	8.62	78.9	81.2
32	33.0	7.44	98.8	83.8
35	36.0	1.74	12.7	69.0
38	39.0	2.84	33.6	77.9
39	40.0	1.43	26.2	62.1

Group 2

	Experiment	Substance A	Substance B	Water Pollution
0	1.0	2.84	78.9	11.5
1	2.0	9.34	52.6	17.7
3	4.0	0.21	39.6	11.2
5	6.0	2.77	98.1	6.0
6	7.0	4.41	4.8	15.2
8	9.0	9.94	16.4	14.1
10	11.0	0.66	61.6	11.3
11	12.0	2.73	67.2	9.4
12	13.0	2.14	85.0	13.8
15	16.0	7.36	13.0	21.4
19	20.0	9.79	16.8	6.7
20	21.0	5.07	87.9	13.0
22	23.0	9.83	55.3	8.4
24	25.0	5.76	94.4	9.3
25	26.0	0.17	41.9	9.6
26	27.0	7.90	15.5	11.5
27	28.0	9.29	13.6	7.5
29	30.0	2.00	93.3	11.9
31	32.0	7.21	20.6	14.9
33	34.0	8.70	48.3	15.8
34	35.0	3.05	86.3	16.5
36	37.0	3.41	1.6	7.6
37	38.0	7.50	22.7	10.6

2)

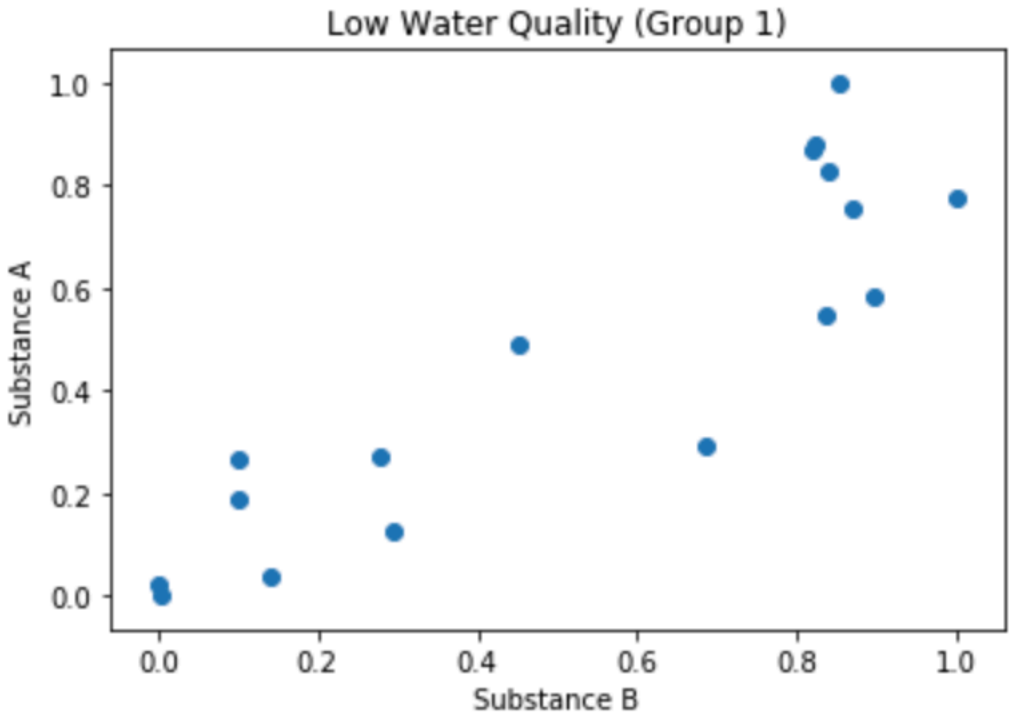
Group 1 Normalized

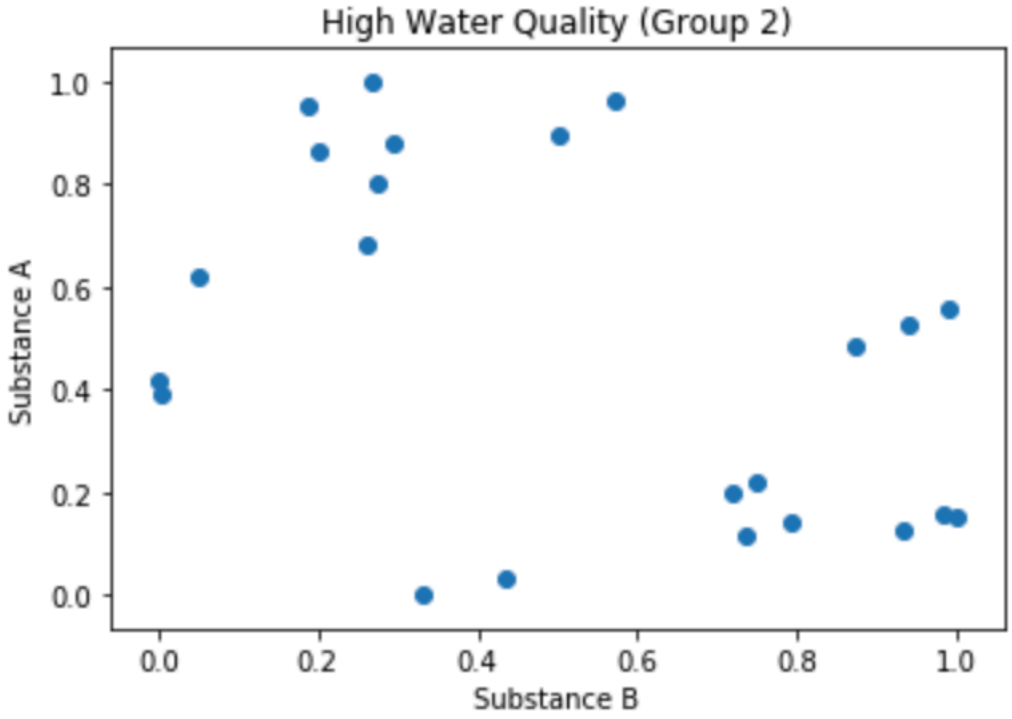
```
 Substance A    Substance B
 [[ 0.87108886  0.75334821]
 [ 0.83604506  0.54910714]
 [ 0.83979975  0.828125  ]
 [ 0.81852315  0.8671875 ]
 [ 0.89612015  0.58258929]
 [ 0.          0.0234375 ]
 [ 0.82352941  0.87946429]
 [ 0.68710889  0.29241071]
 [ 0.45181477  0.4921875 ]
 [ 0.00500626  0.        ]
 [ 0.10137672  0.26450893]
 [ 0.29536921  0.12723214]
 [ 1.          0.77790179]
 [ 0.85231539  1.        ]
 [ 0.13892365  0.0390625 ]
 [ 0.27659574  0.27232143]
 [ 0.10012516  0.18973214]]
```

Group 2 Normalized

```
Substance A    Substance B
[[ 0.27328557  0.80103627]
 [ 0.93858751  0.52849741]
 [ 0.00409417  0.39378238]
 [ 0.26612078  1.        ]
 [ 0.43398158  0.03316062]
 [ 1.          0.15336788]
 [ 0.05015353  0.62176166]
 [ 0.26202661  0.67979275]
 [ 0.20163767  0.8642487 ]
 [ 0.73592631  0.11813472]
 [ 0.98464688  0.15751295]
 [ 0.50153531  0.89430052]
 [ 0.98874104  0.55647668]
 [ 0.57215967  0.96165803]
 [ 0.          0.41761658]
 [ 0.79119754  0.14404145]
 [ 0.93346981  0.12435233]
 [ 0.18730809  0.95025907]
 [ 0.72057318  0.19689119]
 [ 0.87308086  0.48393782]
 [ 0.29477994  0.87772021]
 [ 0.33162743  0.        ]
 [ 0.75025589  0.21865285]]
```

3)


Low Water Quality (Group 1)
Substance A
Substance B
0.0
0.2
0.4
0.6
0.8
1.0


High Water Quality (Group 2)
Substance A
Substance B
0.0
0.2
0.4
0.6
0.8
1.0

4)

```
Group 1 Pearson correlation coefficient:
 (0.89382425337583649, 1.3317228787170414e-06)
Group 2 Pearson correlation coefficient:
 (-0.44962747254599289, 0.031353743604282829)
```

5) From the scatter plot and Pearson correlation coefficient, it can be seen that substance A and substance B are nearly negatively correlated for the high water quality group(2). Hence, adding the two substance in reverse amount(one high and one low) would decrease water pollution.

Question 2:

1) The method used here is the function pd.describe( ) from python data frame.

	age
count	20.000000
mean	39.750000
std	16.577966
min	2.000000
25%	32.750000
50%	41.000000
75%	43.000000
max	91.000000

2) From the plot, age of 2 and 91 are the outliers.