Yiduo Feng

CS 699

Assignment 1

Date: 5/14/2022

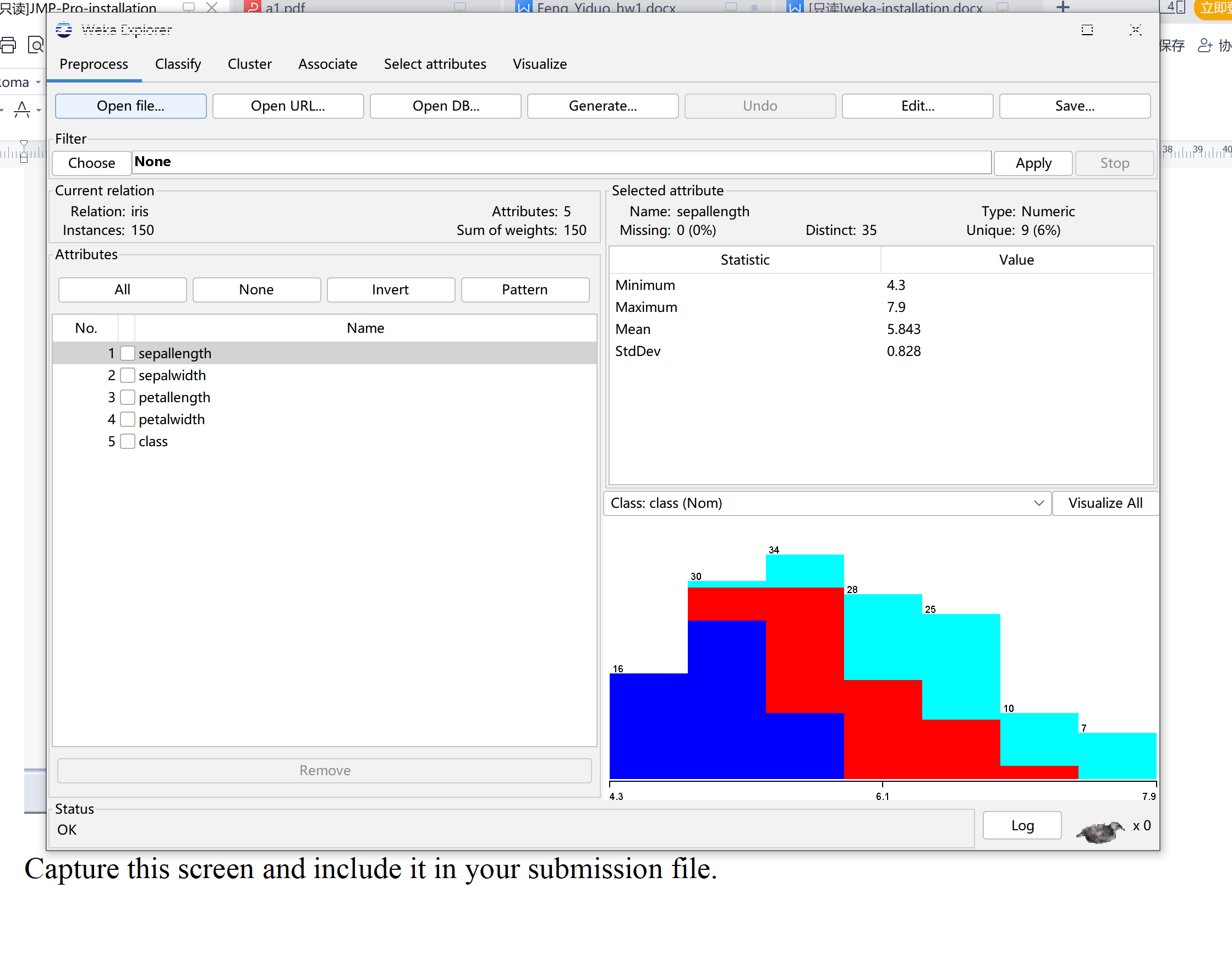
**Problem 1. Install Weka**

 Complete WEKA installation using the attached installation instruction.

 Start Weka and start Explorer.

 Open the *iris.arff* dataset. Provide a snapshot of the resulting screen in your

submission.

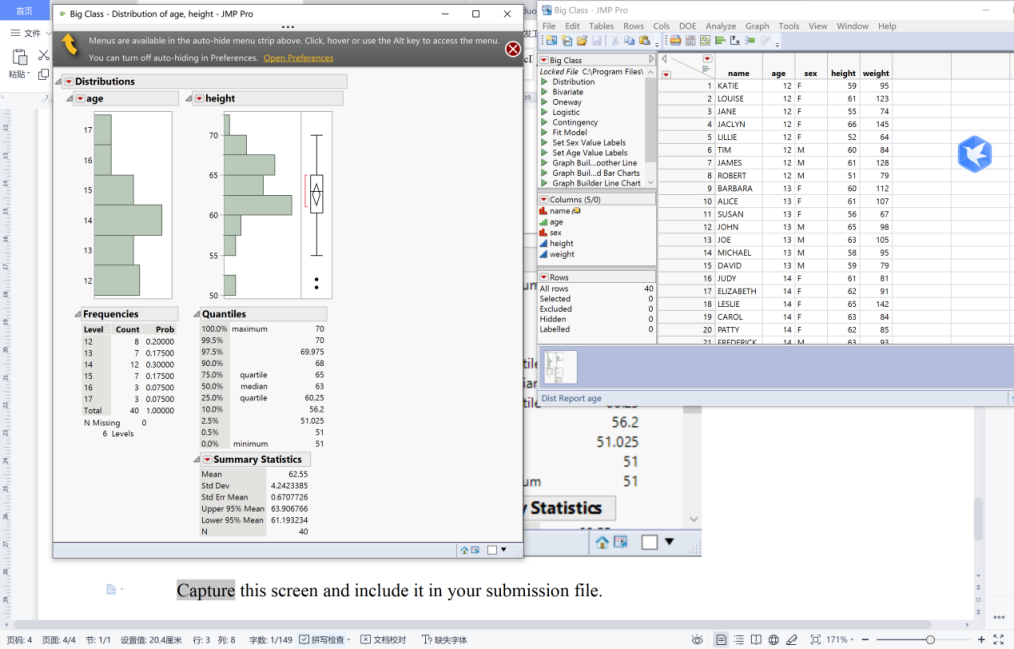


**Problem 2. Install JMP Pro**

 Complete JMP installation using the attached installation instruction.

 Follow the instruction and capture the last screen and include it in your

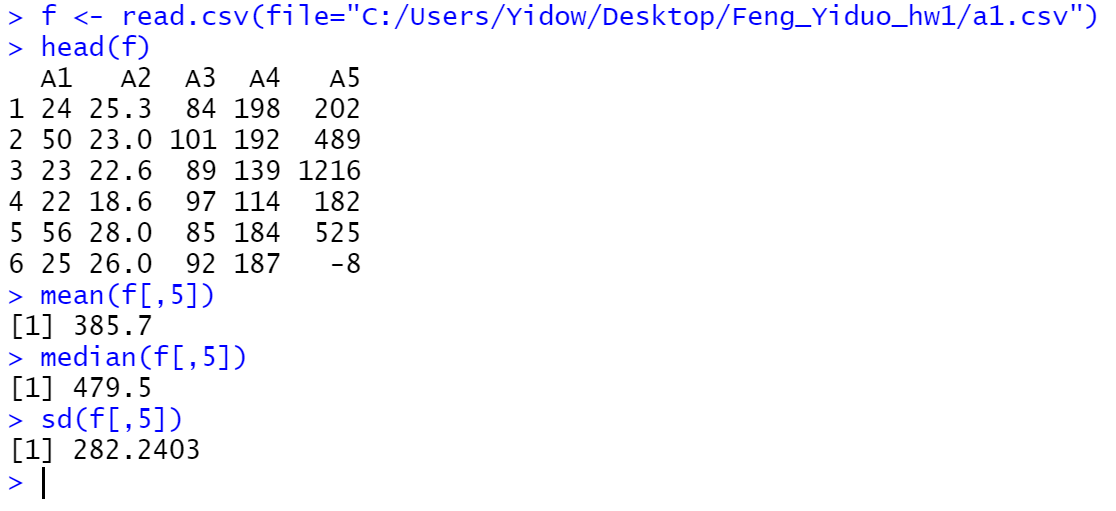
submission file.



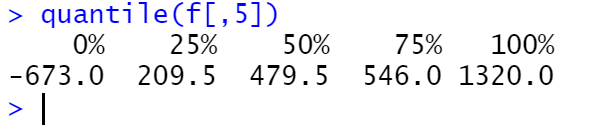
**Problem 3** Consider the dataset *a*1*.csv* which is posted along with this

assignment. It has 100 instances and 5 attributes.

1. . Calculate the mean, median, and standard deviation (sample) of the attribute *A*5.

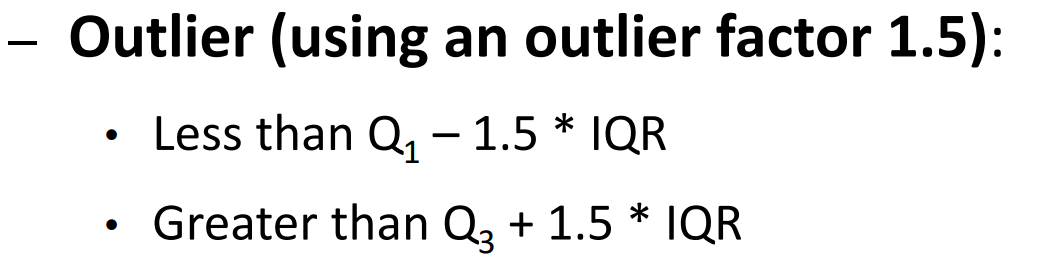


1. . Determine Q1, Q2, and Q3 of *A*5.



1. . Detect outliers using the IQR method, which we discussed in the class, and show the *A*5 values of the detected outliers. When detecting outliers, use only the *A*5 values.

According to the notes



IQR = Q3 - Q1 = 546 - 209.5 = 336.5

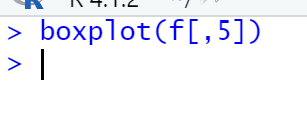
Q1 - 1.5\*IQR = 209.5 - 1.5\*336.5 = -295.25

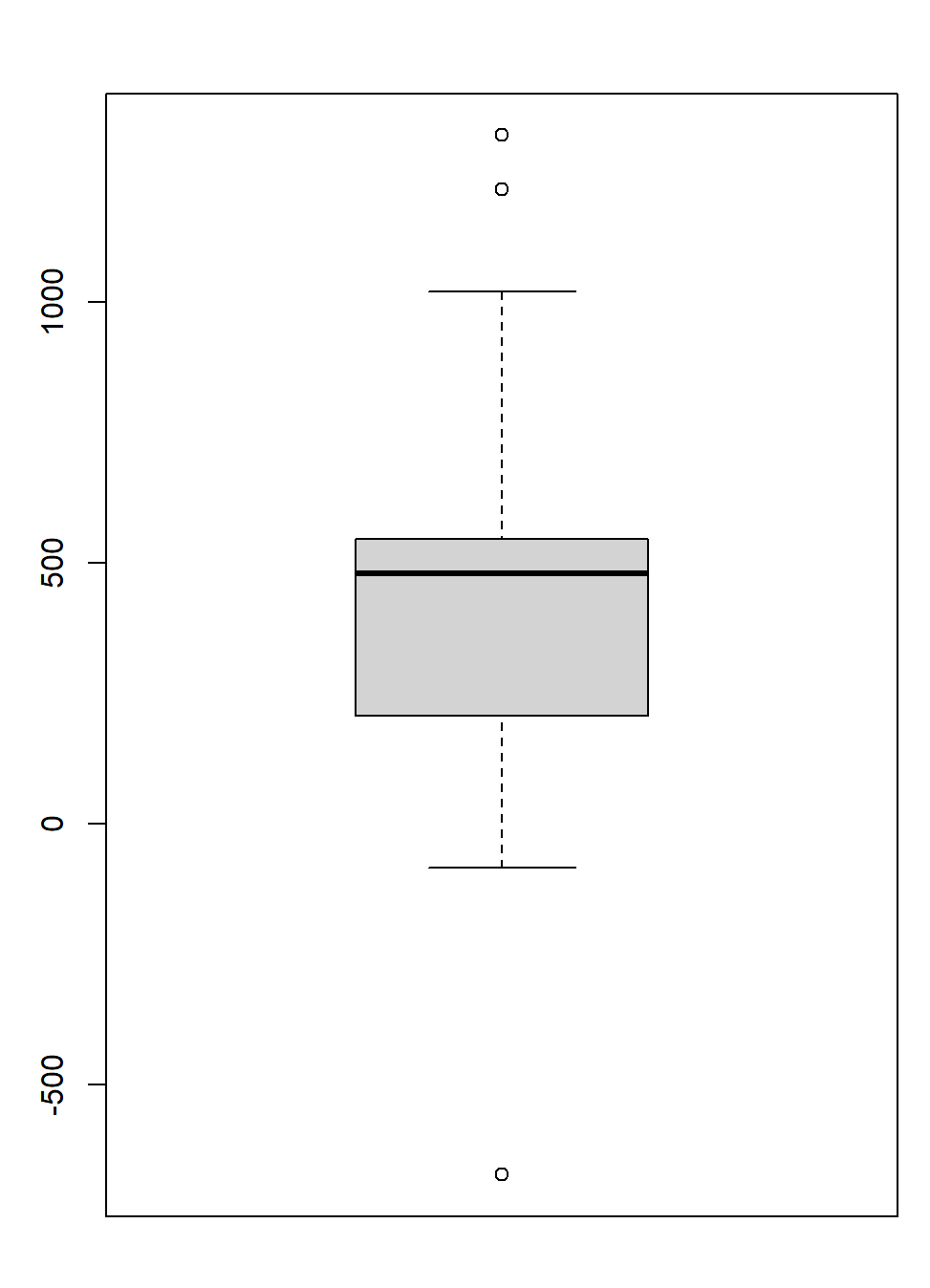
Q3 - 1.5\*IQR = 546 - 1.5\*336.5 = 41.25

Less than -295.25 greater than 41.25

(4). Plot the boxplot of the attribute *A*5. In your boxplot, you need to show outliers

separately.





**Problem 4**. This problem also uses *a*1*.csv* dataset.

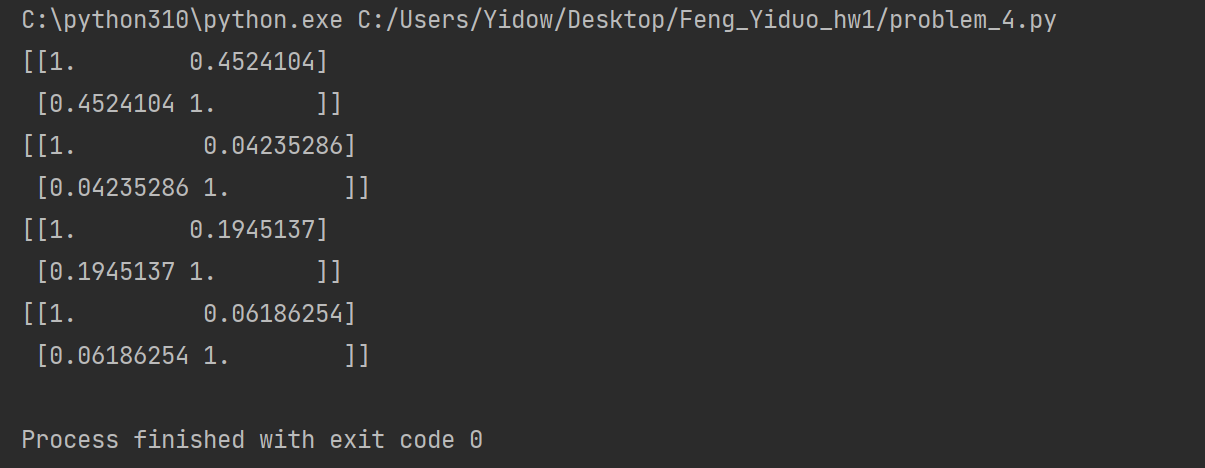
(1). Calculate the following four Pearson's correlation coefficients:

correl(A1, A5)

correl(A2, A5)

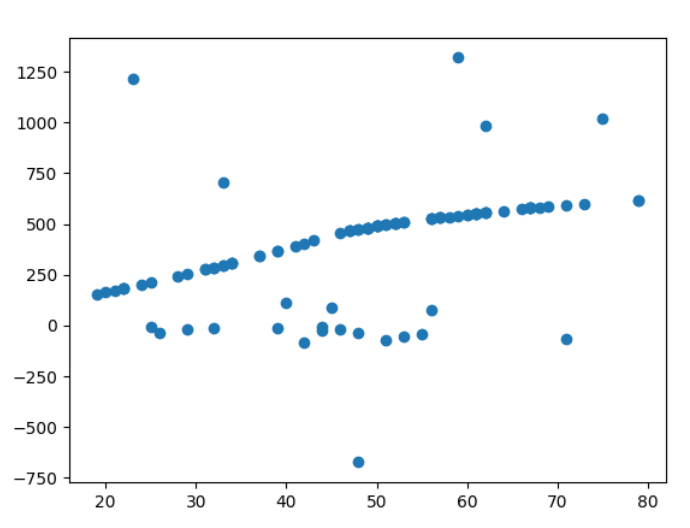
correl(A3, A5)

correl(A4, A5)

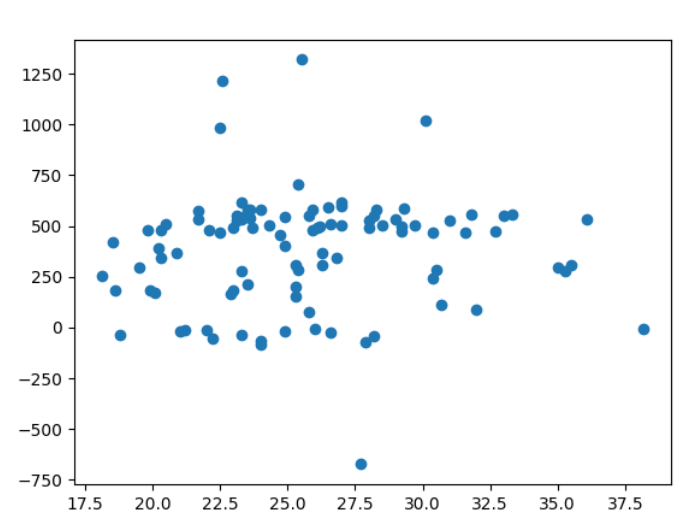


(2). Show the four scatterplots.

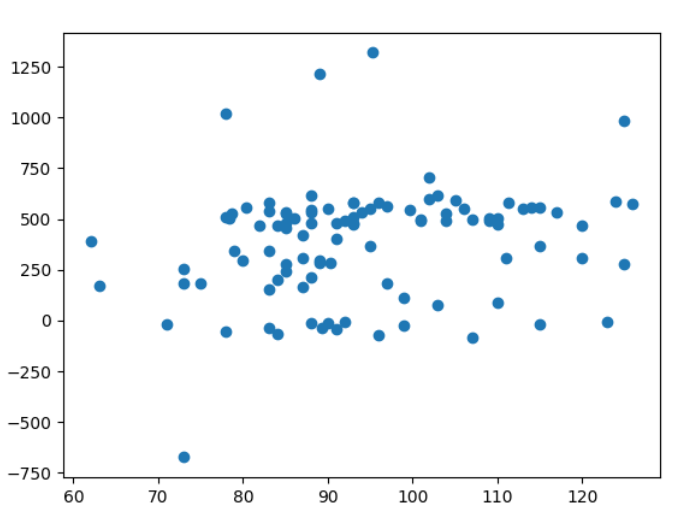
correl(A1, A5)



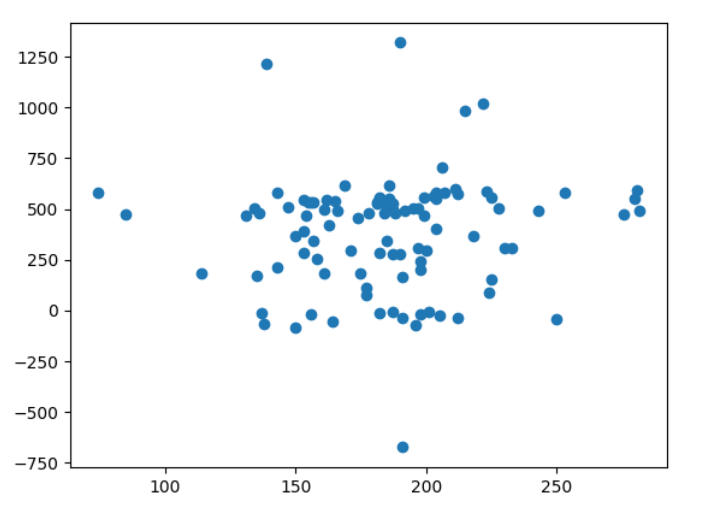
correl(A2, A5)



correl(A3, A5)



correl(A4, A5)



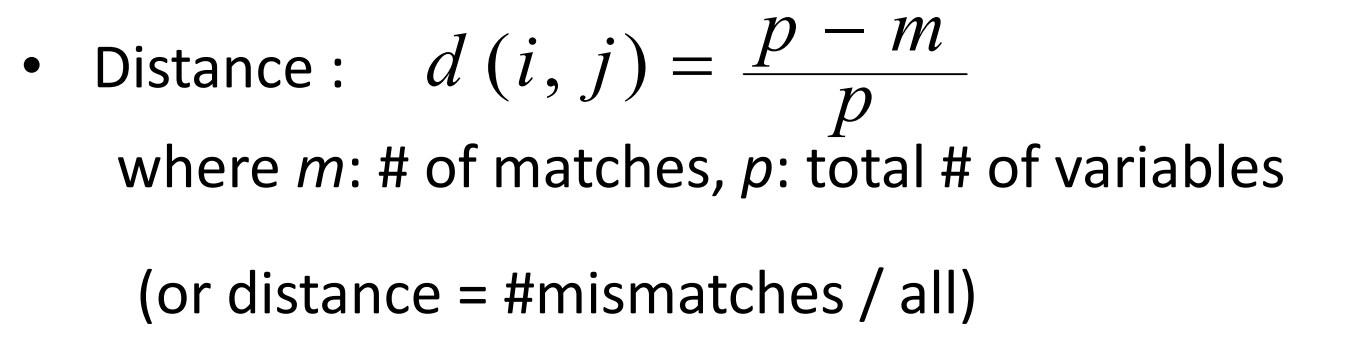
(3). Among A1, A2, A3, and A4, which one has the strongest correlation with A5?

According to the graph and data above, A1 has the strongest correlation with A5.

**Problem 5** . Consider the following dataset that has some information about 10 people.

Calculate the distance between P4 and P5, *d*(P4, P5), and the distance between P4 and P6, *d*(P4, P6). Is P4 closer to P5 or P6? Here, all attributes are nominal attributes.

According to the notes



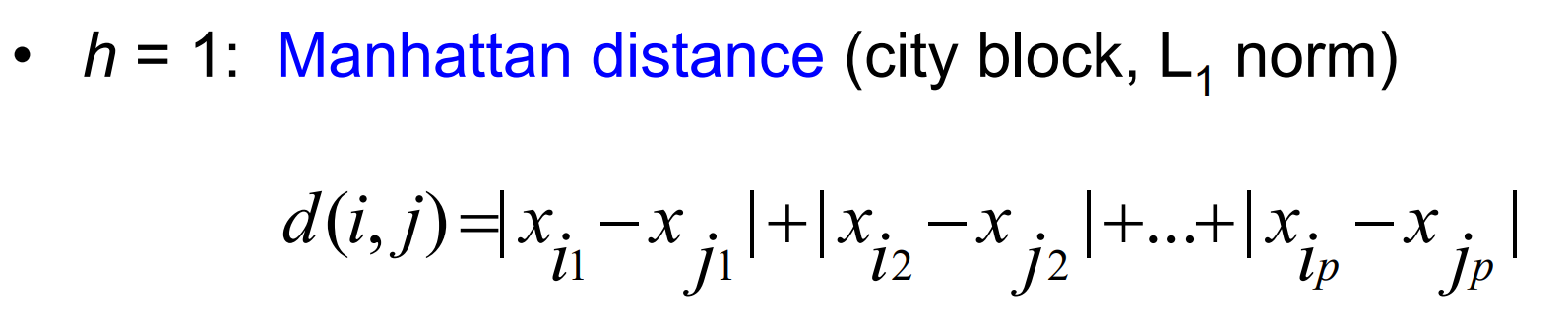
d(P4, P5) = (7-3)/7 = 4/7

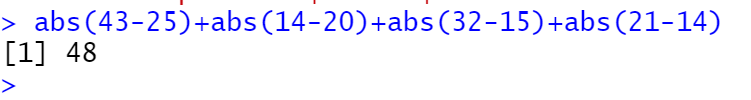
d(P4, P6) = (7-4)/7 = 3/7

So P4 is closer to P6.

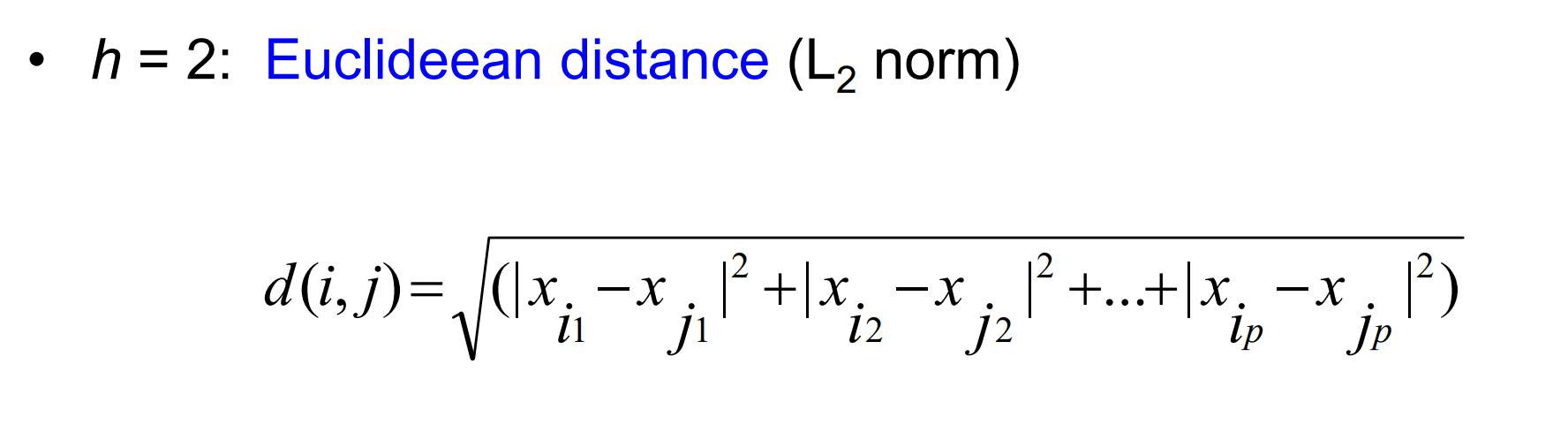
**Problem 6** . Consider the following dataset with two objects.

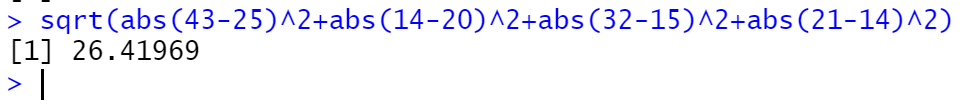
1. . Calculate the distance between O1 and O2 using the Manhattan distance.





(2). Calculate the distance between O1 and O2 using the Euclidean distance.





**Problem 7 (10 points).** Consider the following dataset with two objects.

Here, all attributes are ordinal attributes and ranks of their values are shown below (lowest rank on the left):

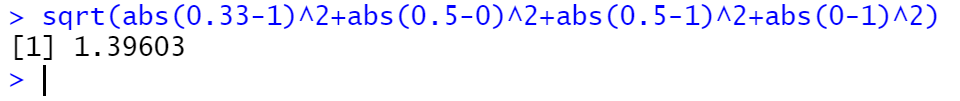
A1: {1, 2, 3, 4}

A2: {first, second, third}

A3: {bronze, silver, gold}

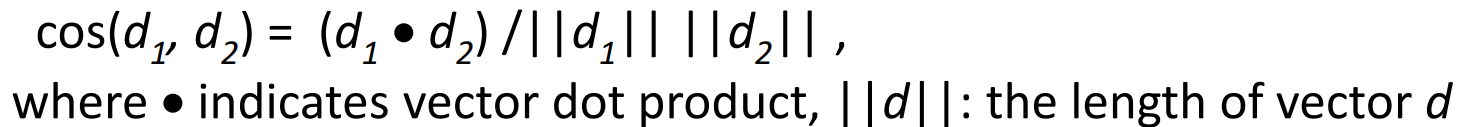
A4: {small, medium, large}

Calculate the distance between O1 and O2 using the method that we discussed in the class. Use the Euclidean distance measure.



**Problem 8**. Consider the following dataset.

Calculate the similarity between D1 and D2, *cos*(D1, D2), and the similarity between D1 and D3, *cos*(D1, D3), using the cosine similarity measure. Is D1 closer to D2 or D3? You must calculate the cosine similarity yourself (i.e., you must not use a built-in function of a software).



D1=( 5 1 2 1 3 2 4 2 3 1 )

D2=( 2 1 0 0 4 2 3 3 1 1 )

D3=( 2 0 1 0 3 1 0 2 3 2 )

D4=( 1 3 4 3 0 3 5 0 4 0 )

||D1|| = (5\*5+1\*1+2\*2+1\*1+3\*3+2\*2+4\*4+2\*2+3\*3+1\*1)^0.5 = 8.602325

||D2|| = (2\*2+1\*1+0\*0+0\*0+4\*4+2\*2+3\*3+3\*3+1\*1+1\*1)^0.5 = 6.708204

||D3|| = (2\*2+0\*0+1\*1+0\*0+3\*3+1\*1+0\*0+2\*2+3\*3+2\*2)^0.5 = 5.656854

D1\*D2 = 5\*2+1\*1+2\*0+1\*0+3\*4+2\*2+4\*3+2\*3+3\*1+1\*1 = 49

D1\*D3 = 5\*2+1\*0+2\*1+1\*0+3\*3+2\*1+4\*0+2\*2+3\*3+1\*2 = 38

*cos*(D1, D2) = 49/(8.602325\*6.708204) = 0.8491296

*cos*(D1, D3) = 38/(8.602325\*5.656854) = 0.7808952

So D1 is closer to D2.