**Jackson Romie**

Intelligent Data Analysis – Spring 2017

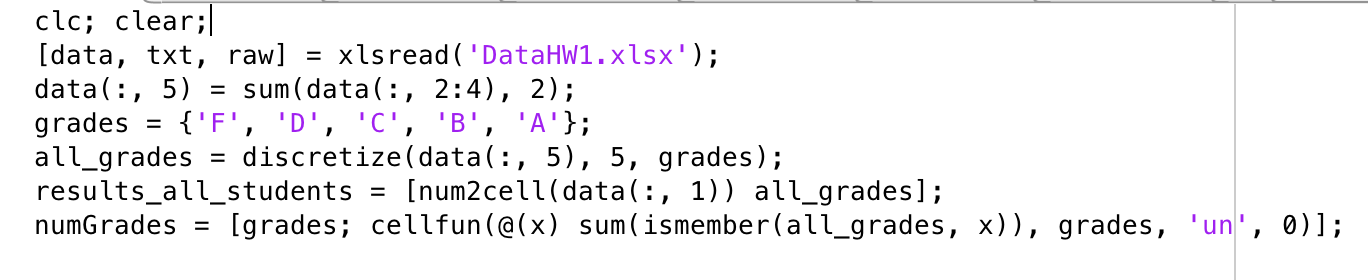
Homework #1

Due Date: Jan 27th, 2017, 7:00PM

The attached excel file contains the data to be used for all the questions listed below. The four columns of the data file are: student id, points obtained in Physics test, points obtained in Maths test, and the grade obtained in English test. For each problem listed below clearly describe and include with submission: (i) a brief list of steps followed by you to solve the problem, (ii) any matlab/python/other code used by you to get the results, and (iii) the results obtained.

1. (10) Add the three scores for each student and write the sum in a new column. Discretize the new sum column into five groups using equal width partitioning and assign one of the five grades (A, B, C, D, and F) to students in the five groups (the highest scores get A and the lowest scores get F). Show the grades assigned to each student in a list sorted according to student id. List the counts of each letter grade awarded.
2. (10) Repeat problem #1 above with the difference that this time use equal frequency partitioning to discretize the sum of points obtained. List the counts of each letter grade awarded. List student ids of those students who would be happier with equal width binning and also of those students’ ids that would be happier with equal frequency binning.
3. (12) Compare the grades assigned in problem #1 and #2 above. Make a list of those student ids whose grades changed when the method of binning changed.
4. (12) Convert the Physics and Maths points to their equivalent z-scores in each column. Sum the three z-scores for each student and use equal frequency binning to create five bins. Assign letter grades to the students and show them in a list sorted by student ids.
5. (12) Compare the grades obtained in #4 and in #2 above. Make a list of students who would be happier with the method in #2 and also a list of those who would be happier with the method in #4. Which method is fairer and why?
6. (12) Consider a student who is happier with the method in #4 compared to the method in #2. Briefly explain why his being happier is justified. Also consider a student who is happier with the method in #2 and briefly explain why he/she is or is not justified.
7. (12) Select two students for each type of grade from the list of grades assigned in #1 above. Create a 10X10 table of all pairwise distances using (i) Euclidean distance, and (ii) Mahalanobis distances.
8. (12) Select the four student pairs with the minimum distance values from the Euclidean distance table and also four student pairs with the minimum distance values from the Mahalanobis distance table. Are the two sets of student pairs the same? If not, point out the differences in the two sets and explain why some students pairs got included in one set but not in the other.
9. (8) These points are for readability and organization of the submitted work.

**Problem1:**

****

results\_all\_students =

40×2 cell array

[ 1] 'B'

[ 2] 'B'

[ 3] 'C'

[ 4] 'A'

[ 5] 'D'

[ 6] 'C'

[ 7] 'C'

[ 8] 'C'

[ 9] 'A'

[10] 'B'

[11] 'D'

[12] 'C'

[13] 'C'

[14] 'A'

[15] 'C'

[16] 'C'

[17] 'D'

[18] 'B'

[19] 'B'

[20] 'C'

[21] 'C'

[22] 'D'

[23] 'C'

[24] 'B'

[25] 'C'

[26] 'C'

[27] 'B'

[28] 'C'

[29] 'B'

[30] 'C'

[31] 'B'

[32] 'C'

[33] 'A'

[34] 'D'

[35] 'F'

[36] 'D'

[37] 'B'

[38] 'C'

[39] 'A'

[40] 'F'

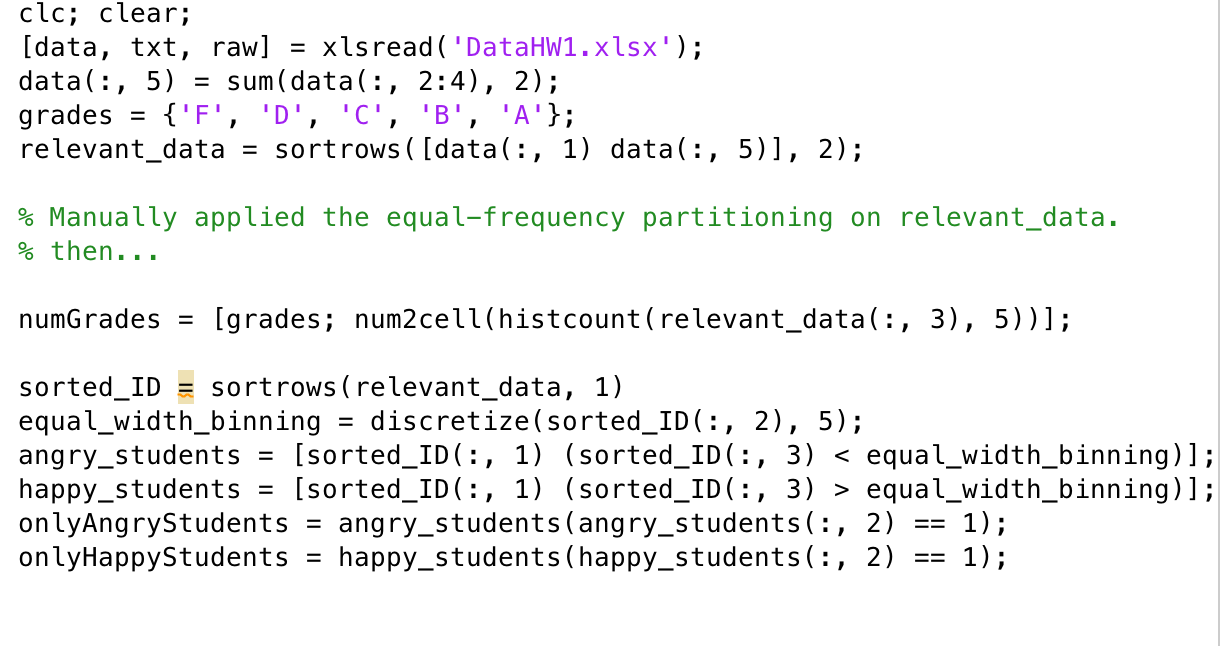
numGrades =

2×5 cell array

'F' 'D' 'C' 'B' 'A'

[2] [6] [17] [10] [5]

**Problem 2**



numGrades =

2×5 cell array

'F' 'D' 'C' 'B' 'A'

[8] [7] [8] [8] [9]

onlyAngryStudents =

3

5

7

11

15

17

22

23

28

32

34

36

38

onlyHappyStudents =

1

13

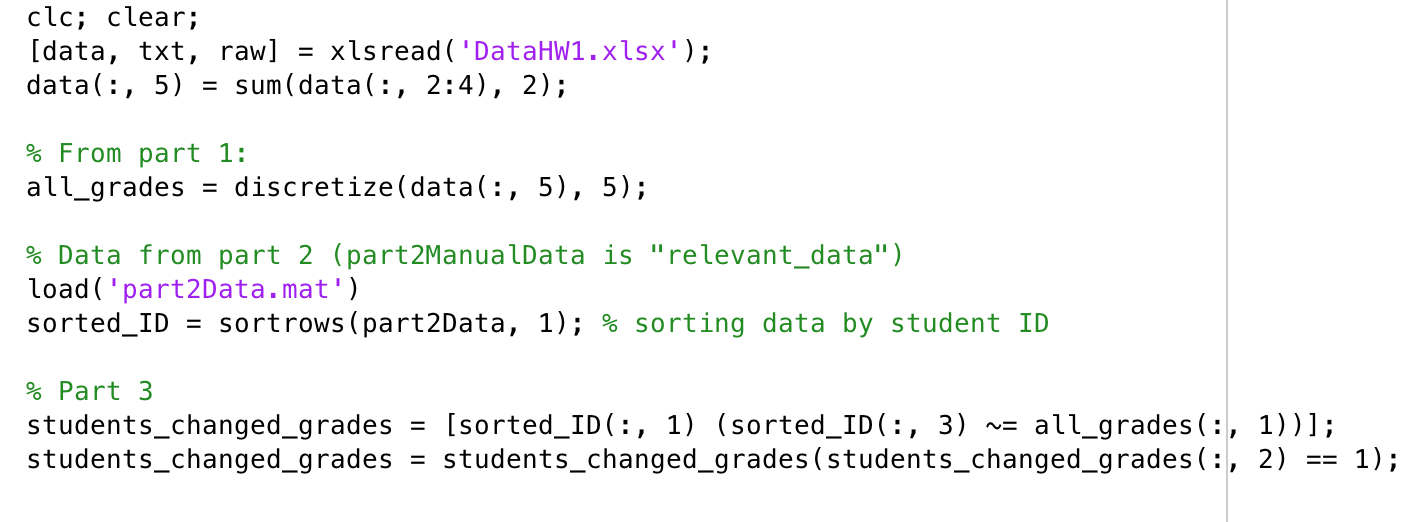
24

30

31

37

**Problem 3**



students\_changed\_grades =

1

3

5

7

11

13

15

17

22

23

24

28

30

31

32

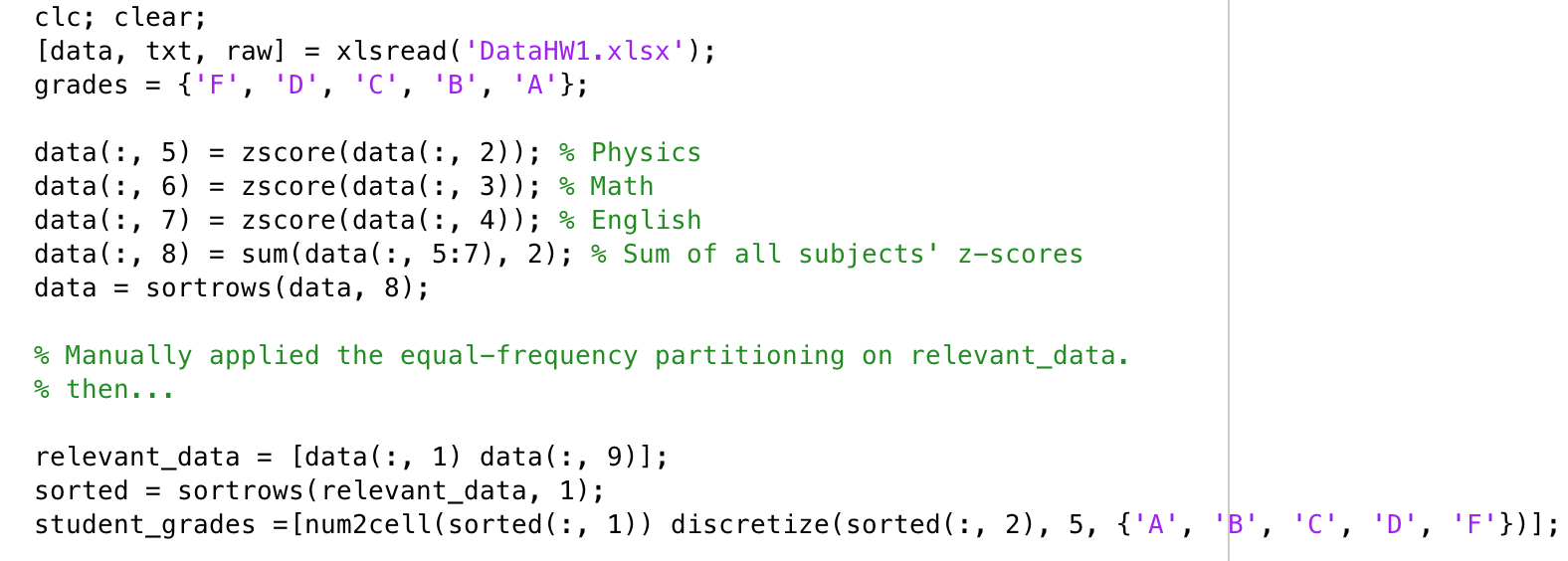
34

36

37

38

**Problem 4**



student\_grades =

40×2 cell array

[ 1] 'C'

[ 2] 'D'

[ 3] 'B'

[ 4] 'F'

[ 5] 'A'

[ 6] 'B'

[ 7] 'C'

[ 8] 'D'

[ 9] 'F'

[10] 'C'

[11] 'A'

[12] 'B'

[13] 'C'

[14] 'F'

[15] 'B'

[16] 'B'

[17] 'B'

[18] 'F'

[19] 'D'

[20] 'D'

[21] 'C'

[22] 'A'

[23] 'B'

[24] 'F'

[25] 'C'

[26] 'D'

[27] 'D'

[28] 'C'

[29] 'D'

[30] 'C'

[31] 'D'

[32] 'B'

[33] 'F'

[34] 'A'

[35] 'A'

[36] 'A'

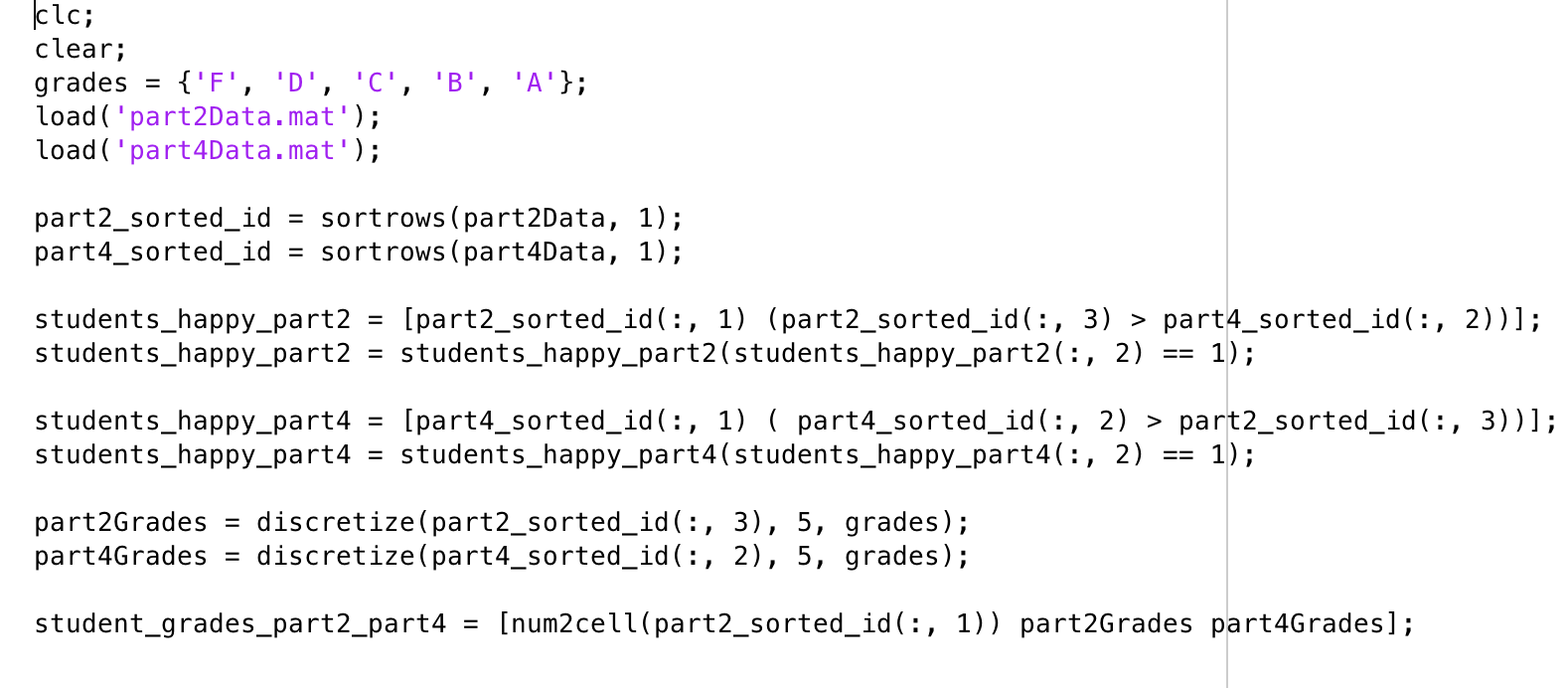
[37] 'F'

[38] 'A'

[39] 'F'

[40] 'A'

**Problem 5:**



students\_happy\_part2 =

1

6

10

12

13

16

30

31

38

students\_happy\_part4 =

7

8

17

18

20

26

28

*The following output is a matrix where the first column is the student ID, second column is grades according to part2, and third column is grades according to part 3:*

student\_grades\_part2\_part4 =

40×3 cell array

[ 1] 'A' 'C'

[ 2] 'B' 'B'

[ 3] 'D' 'D'

[ 4] 'A' 'A'

[ 5] 'F' 'F'

[ 6] 'C' 'D'

[ 7] 'D' 'C'

[ 8] 'C' 'B'

[ 9] 'A' 'A'

[10] 'B' 'C'

[11] 'F' 'F'

[12] 'C' 'D'

[13] 'B' 'C'

[14] 'A' 'A'

[15] 'D' 'D'

[16] 'C' 'D'

[17] 'F' 'D'

[18] 'B' 'A'

[19] 'B' 'B'

[20] 'C' 'B'

[21] 'C' 'C'

[22] 'F' 'F'

[23] 'D' 'D'

[24] 'A' 'A'

[25] 'C' 'C'

[26] 'C' 'B'

[27] 'B' 'B'

[28] 'D' 'C'

[29] 'B' 'B'

[30] 'B' 'C'

[31] 'A' 'B'

[32] 'D' 'D'

[33] 'A' 'A'

[34] 'F' 'F'

[35] 'F' 'F'

[36] 'F' 'F'

[37] 'A' 'A'

[38] 'D' 'F'

[39] 'A' 'A'

[40] 'F' 'F'

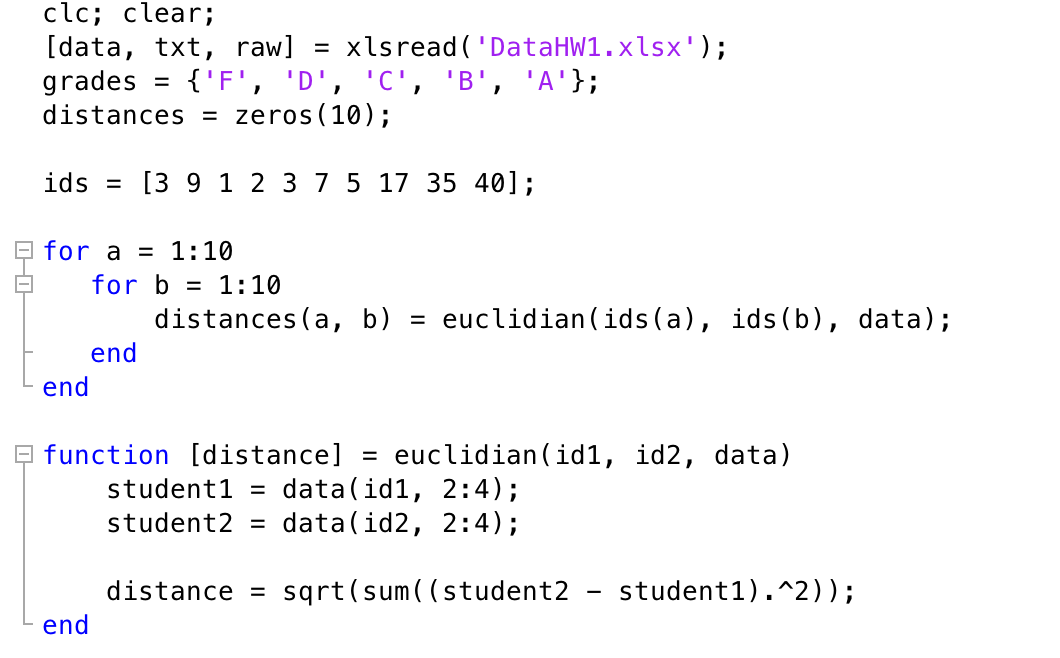
Part 4 is a fairer approach because in part 2, it is not considering the average score. The average score could be very poor, and thus the majority of the class gets that score, where using z-score you’re measuring how many standard deviations from average each student gets.

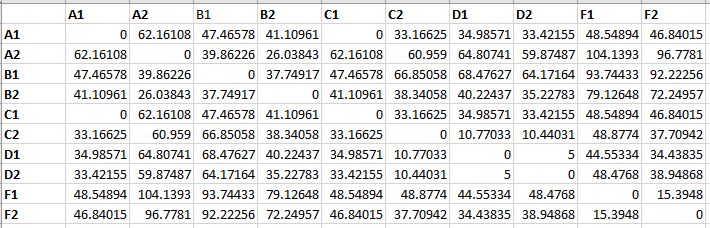
**Problem 6**

A student who is happier with method 4 is justified because the grading is done based on average. If he/she ended up with a 59% on the test, but the average was a 60%, then the student is going to get a better grade because he/she is probably only a fraction of a standard deviation from average. Equal partitioning will likely end up unpredictably assigning grades to a student who got a common score.

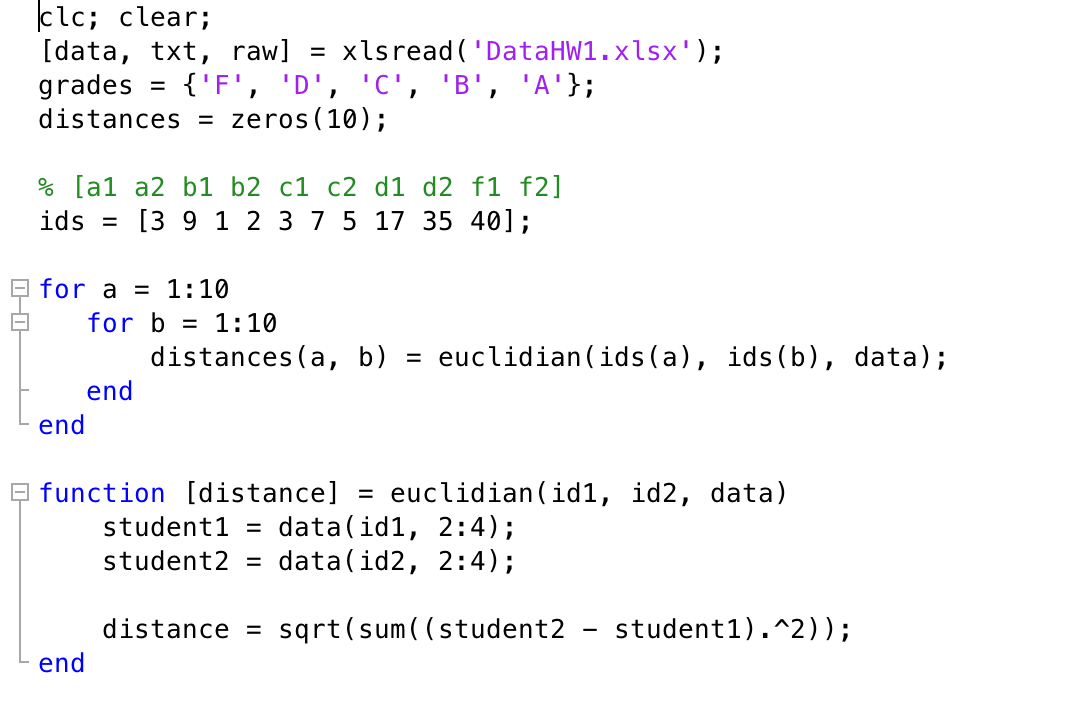
A student is not justified to be happy with method 2 because he/she might have gotten a relatively high score, but not using z-score could dissatisfy a lot of other students. Take Student 6. They got a C overall using method 2, and a D overall using method 4. Student 21 got a C both ways, but was closer to the average.

**Problem 7**

****



**Problem 8**



* Student D1 – D2 distance of 5
* Student D2 – C2 distance of 10.44031
* Student D1 – C2 distance of 10.77033
* Student F1 – F2 distance of 15.3948