**IDA Assignment 1**

**Name : Sampath Dontharaju**

**Id : M08879603**

Q1) Add the two scores for each student and write the sum in a new column. Discretize the 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.

**Solution :**

1. loaded the data in the excel file using the **dataSet()** function in matlab.
2. Then added the two scores of maths and physics columns and placed the total phy+maths score in a new column in data set called **Totalscore**
3. calculated the min and max of the **Total score** and calculated the width for each bin using (max-min/5).
4. Created five bins for equal width binning and assigned Grades for the students .
5. calculated the no .of students for each grade assigned and printed the values using **fprintf**

and the count is as follows

A grade = 5 students

B grade = 7 students

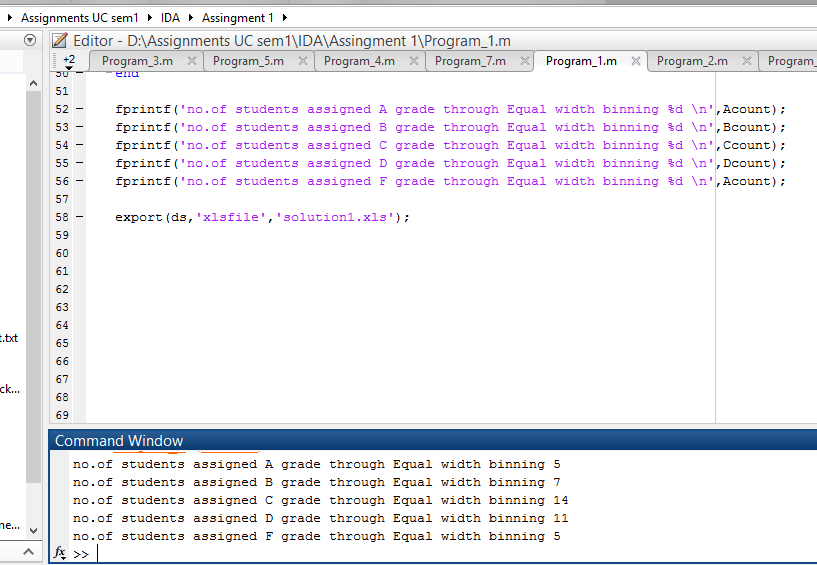
C grade = 14 students

D grade = 11 students

F grade = 3 students

**The above result is displayed in the console when program\_1.m matlab file is executed .**

**see the below screen shot for the output**

****

exported the result to a excel file which shows result of how the grades are assigned to students using Equal width binning.

code for question 1 : 

**Solution 1 excel sheet . This excel sheet will be created in the directory as solution1.xls**

**solution 1 excel sheet** : 

**question 2)**

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 who would be happier with equal frequency binning.

**solution :**

for this solution appended the following logic to the first solution

sorted the dataset based on the **Totalscore(phy+maths score )** using the syntax

**sortedtotal = sortrows(ds,5)**

where 5 is the column number of the total score column.

Assigned the grades using equal frequency binning , divided the 40 student ids into 5 sets of approximate size **8**.

divided the 40 tuples into 5 partitions and assigned **'F'**  grade to least 9 students and '**D'** Grade to **'7'** students and **'C '** grade to **9** students and **'B'** and **'A'** grades to 7 and 8 students respectively.

written the logic such that equal values will be under same bin .

Displayed the list of students happy with **Equal width binning** and **Equal frequency binning** using **fprintf** in command line

Student ids happy with Equal width binning **: 11,34,5,7,17,28,15,21**

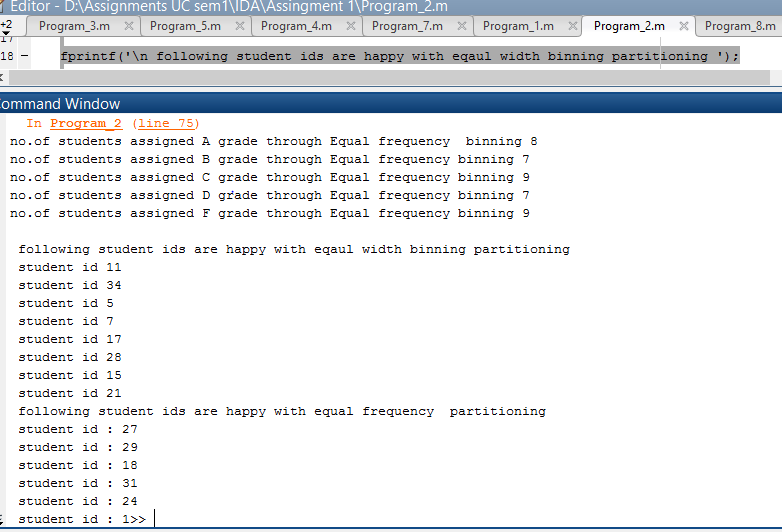
Student ids happy with Equal Frequency binning **: 27,29,18,31,24,1**

**Excel output** generated in the folder showing the grades assigned to students using **Equal freq binning**

code attached for program 2. 

Excel output : 

sample output for solution 2



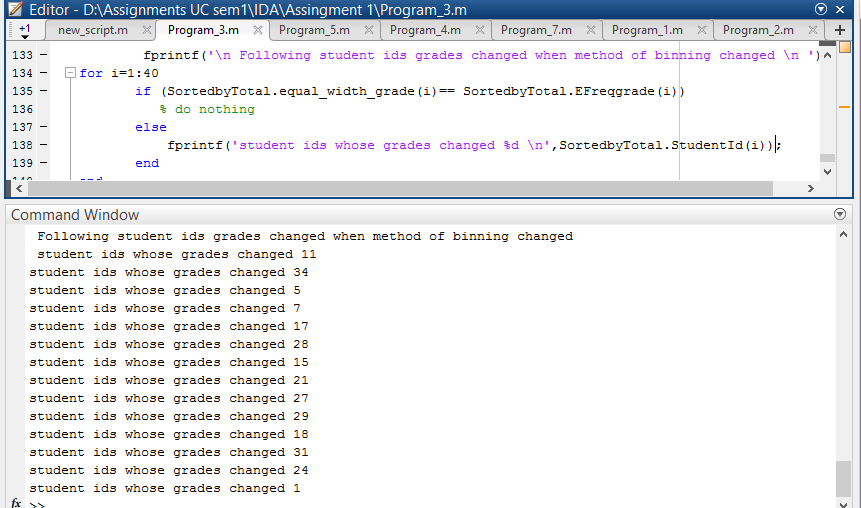
**Question 3**

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.

**solution :** to the code of 2nd problem additional code is added which compares the columns **equal\_width\_grade**  and **EFreqgrade** in the sortedtotal dataset and prints the values of student ids whose grades have changed .

**Students ids whose ids have changed : 11,34,5,7,17,28,15,21,27,29,18,31,24,1**

output screen shot : 

code for Program 3: 

**Question 4 :** **.**  Convert the Physics and Maths points to their equivalent z-scores in each column. Sum the

two z-scores for each student and use equal frequency binning to create five bins. Assign grades

to the students and show them in a list sorted by student ids .

**Solution :**

to the code from program 3 following logic is appended

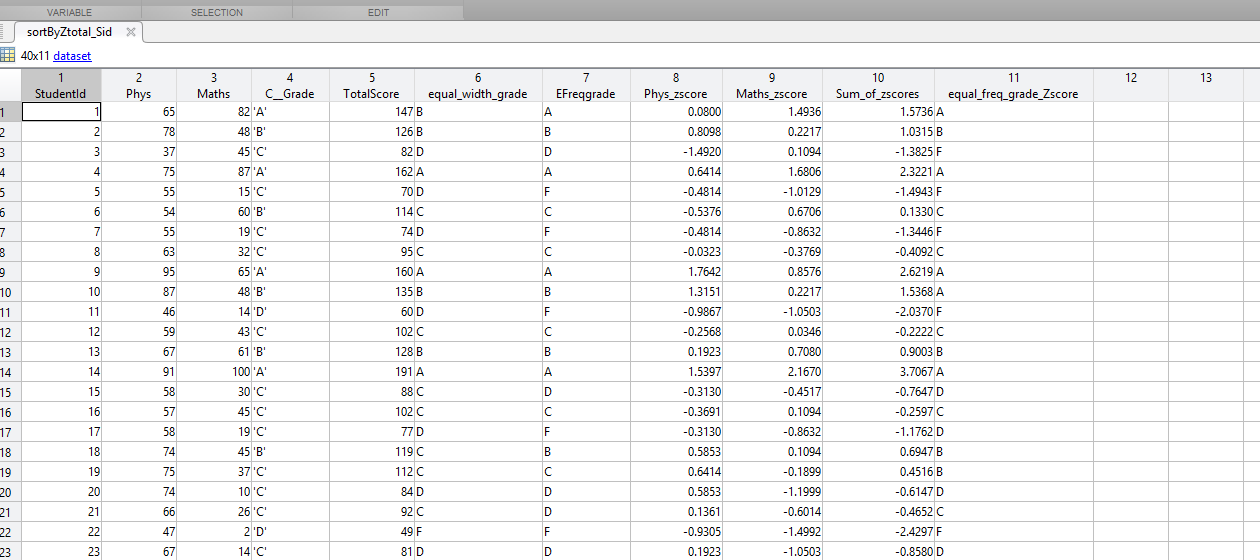
calculated **phy\_zscore** and **Maths\_zcore**  by using the **zscore()** function in matlab. added the two columns in dataset and saved in a new column in dataset by name **Sum\_of\_zscores**

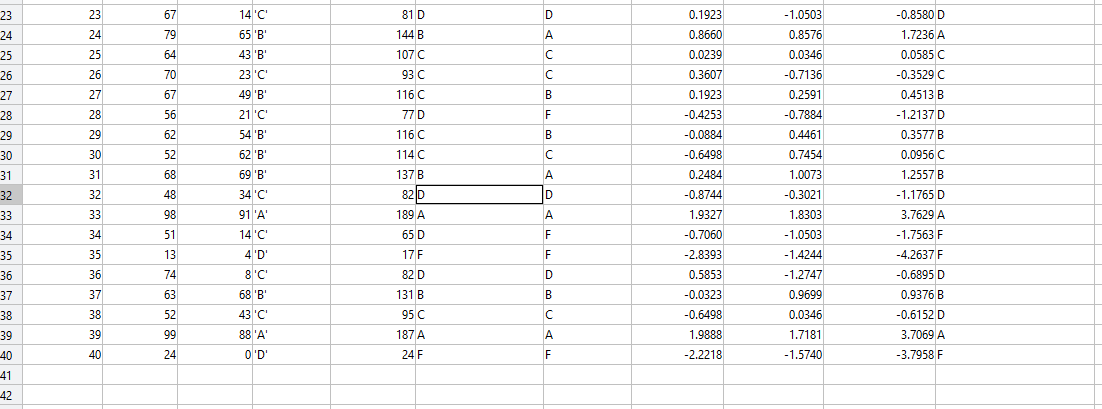
sorted the data set on the column **sum\_of\_zscores** to assign grades to students using **Equal frequency binning.**

Again sorted the dataset according to their student id and displayed in new dataset **sortByZtotal\_Sid**

Exported the result into a excel file named **solution4.xls** . This excel file will be created when the code for 4th program is executed which is attached below.

code for 4th program : 





**Question 5) 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.**

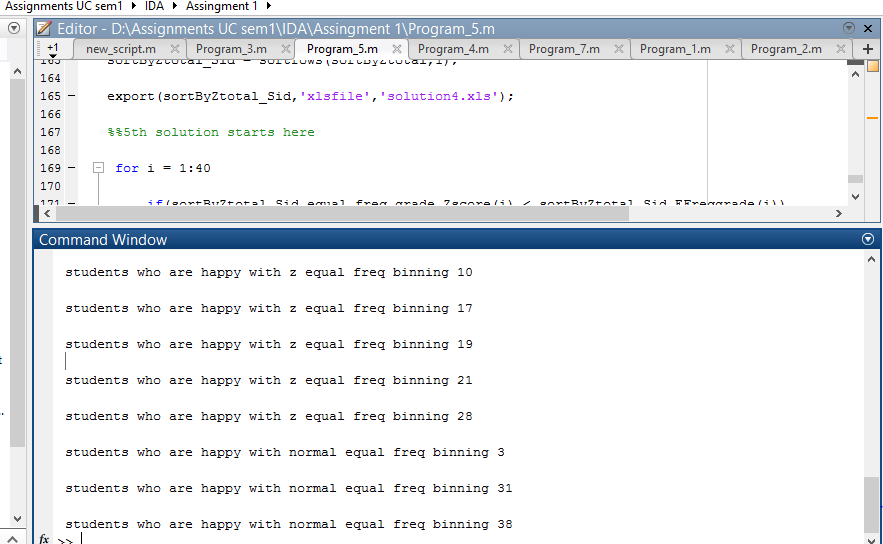
**solution)** compared the **equal frequency grades** obtained in solution 2 with the **equal\_freq\_grade\_Zscore**  grades assigned in solution 4 and the list of students who are happy with grades obtained by equal freq binning are displayed in command window and list of students who are happy with equal frequency grading based on Zscore.

list of student ids who are happy with grades obtained in **solution 2** .i.e., the list of student who are happy with grades obtained with Normal **Equal\_Freq\_binning : 3,31,38**

list of student ids who are happy with grades obtained in **solution 4** .i.e., the list of student who are happy with grades obtained with **Zscore -Equal\_Freq\_binning : 10,17,19,21,28**

**code for question 5: **

**Output screen shot which are displayed in command window after executing the program code given above**

****

**Question 6) 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.**

solution 6) From the above solution we can see that student ids **10,17,19,21,28** are happy with the grades which they received using **ZScore\_Equal\_frequency** than with **normal equal freq distribution grading** .

Zscore grading is generally preferred as it normalises the scores .

for instance consider **student id 10** he gets a grade "**B**" by **equal frequency grading** and he gets a grade "**A**" by **zscore equal frequency Grading .**

Getting "A" grade is more logical because the score distribution of Maths is more different compared to physics and the range of marks for maths is more and we can understand that the toughness level of maths is greater as many students have scored less i.e. avg score or mean of maths is less whereas in physics the avg score or mean is high.

Hence **student 10**  who scored a 87 in physics and 48 in maths should get grade **"A". Zscore** normalizes the maths scores and assigns grades based on the range of the distribution and it does **Data Normalization**  whereas equal frequency binning does not perform normalization.

**Question 7)** For Physics and Maths scores individually, perform the following. Use z-scores to assign

label “Low” to those students whose z-score is strictly below -0.3, the label “Mid” to those

whose z-score is between -0.3 and +0.3 (both values inclusive), and the label “High” to those

students whose z-score is strictly greater than 0.3. Show the data table sorted according to

student ids.

**solution )**  assigned the values **H** to values whose zscores are greater than 0.3 and value label **M** to values whose zscores are between -0.3 and 0.3(inclusive) and  **L** to values whose zscores are strictly less than -0.3.

note : **H - High**

**M- Medium**

**L - Low**

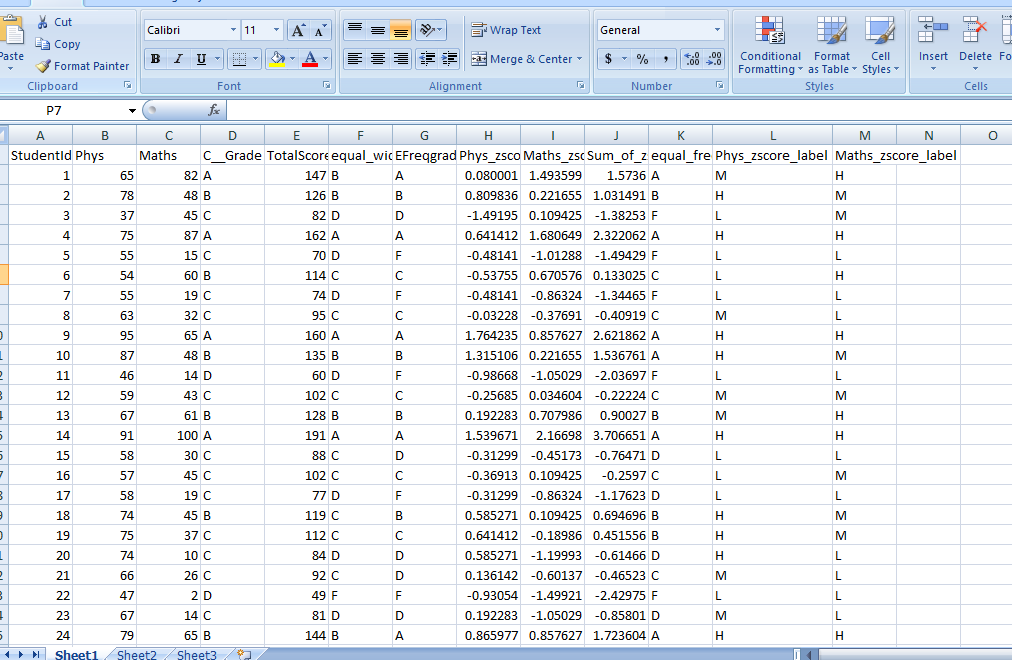
Assigned the labels **H,M,L to** physics and maths Zscores and created two columns for physics and maths zscores as **Phys\_zscore\_label**  and **Maths\_zscore\_label.**

Travesed the phys z-score values and assigned the labels **H,M,L**  and placed them under **Phys\_zscore\_label**  and traversed maths Z scores values and assigned **H,M,L**  Value and placed them under **Maths\_zscore\_label.**

Sorted the entire data set on student id and exported it to a excel file **solution7.xls** which is generated on execution of the code below attached

code for problem 7: 

**output :** output of the excel generated on execution looks like as shown below .



**Question 8)** We want to use the labels assigned in #7 above for Physics and Maths scores to predict the

letter grade a student would obtain his/her C++ class. Use the entropy method to determine

which course’s label (Physics or Maths) is a better predictor of a student’s grade in C++ class.

Show all your work to arrive at your answer.

**Solution 8)**

Calculated the weighted entropy of physics and maths using the formula

Weighted entropy :(|D1|/|D|).Entropy(D1) + (|D2|/|D|).Entropy(D2)+(|D3|/|D|).Entropy(D3)

Here D =40 i.e., no of tuples

and D1,D2,D3 determines how the data is split. In the previous problem we have split the physics and maths Zscores as **H**(High) **, M(**medium)**, L(low)**  based on whether their zscore is above 0.3 or between -0.3 and 0.3 or below -0.3 respectively.

here D1, D2, D3 denote **H,M,L of the physics and maths zscores.**

**entropys of maths medium**, **low** and **high** are calculated and these values are used to calculate the weighted entropy of Maths .

weighted entropy of maths =

**(|D1|/|D|).Entropy(D1) + (|D2|/|D|).Entropy(D2)+(|D3|/|D|).Entropy(D3)**

= ( **Maths** low/|D|).Entropy(**Maths** low)+ ( **Maths** medium/|D|).Entropy(**Maths**medium) + ( **Maths**  high/|D|).Entropy(**Maths** high)

Entropy\_Maths Low = -(13/17\*log10(13/17)+(4/17)\*log10(4/17));

where 13 is the no.of the corresponding '**C**' grades and '**D**' grades of C++ received by the 17 students in maths low classififcation.

Similarly entropy for Maths  **medium** and Maths **high** are claculated by calcuating the c++ grade counts in their classification and perform entropy.

Once Entropy for Maths low ,Maths medium and Maths High are calculated , these are used in the below formula to calculate weighted entropy for maths

( **Maths** low/|D|).Entropy(**Maths** low)+ ( **Maths** medium/|D|).Entropy(**Maths**medium) + ( **Maths**  high/|D|).Entropy(**Maths** high)

**similarly entropys of physics medium**, **low** and **high** are calculated and these values are used to calculate the weighted entropy of Physics.

After calculating the weighted entropys of both maths and physics

**Weighted Entropy of Maths = 0.9**

**Weighted Entropy of Physics =1.3**

since weighted entropy of maths is less.

**Math's is a better predictor of a student's grade in C++ class.**

**code of program 8: **