CIS 635 Data Mining

Homework 2

Part A – concepts

2. which rows had missing data: 57 and 65

3. which rows had the noise: 24 89 133

4. what was the value of the outlier: 106

5. a) Rating of 1 is more similar to 2

   b) They are both the same

Part B – document tables

	you	at	have	are	be	am	study	meet	class
doc 1	1	1	0	1	1	0	0	0	0
doc 2	1	0	1	0	0	0	0	0	0
doc 3	1	1	0	0	0	1	0	1	1
doc 4	1	0	0	1	1	0	1	0	0
doc 5	0	0	0	0	0	0	0	0	0

Part C transactions

when you see milk, you often see cereal

Part D sampling

result from 25 samples 2.494819, 50 samples 1.442, 100 samples 0.6203624 paste the R function below:

```r
avgDiffSample = function(numOfSamples, meanNum) {
  result = c()
  for (item in 1:10) {
    # Each time it will sample the number of records using the parameter
    newSample = sample(tchrAsstData[,6], numOfSamples)

    # Each time you sample, calculate the mean value of classSize
    newMeanNum = mean(newSample)

    # calculate the absolute value of the difference between this mean from
    # the parameter
    diff = abs(newMeanNum - meanNum)

    # append to result vector
    result = c(result, diff)

  }
  # return the average of the differences calculated in the loop.
  return( mean(result) )
}
```