Homework #1

Q1. (15 points) Given the following dataset:
10,11,13,17,17,17,19,21,21,25,27,28,29,29,31,33,33,35,36,36,36,39,39,40,40,,43,52,72.

1) Show a boxplot of the data.
2) Show the histogram of the data with bin width = 10.
3) Show the data after the following normalization:
    a. min-max normalization to the range [0, 1]
    b. z-score normalization
    c. decimal scaling normalization

Q2. (20 points)

TID	List of Items_IDs
1	I1, I2, I5
2	I2, I4
3	I2, I3
4	I1, I2, I4
5	I1, I3
6	I2, I3
7	I1, I3
8	I1, I2, I3, I5
9	I1, I2, I3

Suppose MinSup= 2, MinConf = 75%,
(a) List all frequent itemsets.
(b) List all maximal frequent itemsets and closed frequent itemsets.
(c) List all association rules (with support *s* and confidence *c*).

Q3. (15 points)
Consider a clinical trial of an experimental drug. 90 patients with a certain disease were randomly assigned to one of two groups. The treatment group (45 patients) received the new drug and the control group (45 patients) did not. Outcome was categorized as 1 year survival or non-survival. The outcome of patients is presented by group in the table below.

Outcome	Treatment Group	Control Group
Survived 1 year	40	30
Did not survive	5	15

Compare the four correlation measures: lift, X2(Chi-Square), all-conf, and cosine . Determine if survival and treatment is positively correlated.

Q4 (50 points. Implementation Project) Implement frequent itemset mining methods Apriori (or ECLAT or FP-Tree) using C++ or Python. You program should be able to read in a transaction data file of the format:

T1: I1, I2, I5
T2: I2, I4
T3: I1, I6, I8, I12, I15
T4: I3, I7
….

The output should be (1) All Frequent Itemsets and their supports (2) All association rules and their confidences (3) All Maximal Frequent Itemsets

The output should be in the following format:

Frequent Itemsets:
{I1}; Sup=10
{I1, I2}; Sup=9
{I1, I2, I3}; Sup=9
{I2, I3}; Sup=8
…
Association Rules:
{I1 -> I2, I3}; Confident: 75.0%
{I1, I2 -> I3}; Confident: 82.1%
…

Maximal Frequent Itemsets
{I1, I2, I3}; Sup=9
{I2, I3}; Sup=8
…

These result should be write into a *.txt file with file name (1) “Frequent Itemsets.txt” (2) “Association Rules.txt” and (3) “Maximal Frequent Itemsets.txt”

Please note:

(1) A readme file is required to describe how to run your code. At most 10% points will be cut if it is not clear. You should also provide a makefile if you use C++.
(2) The input arguments should be exactly (1) full path of the data file, (2) Minsup, (3) MinConf namely.
(3) 80% percent will be graded for correctness, and 20% percent will be graded for running time of your code. You have to tic-toc your running time (including time for reading data) and print it on the screen. The speed part is considered regardless of which programming language you chose.
(4) If your code is not giving right answer (including situations like cannot be compiled on any machine, segmentation fault, or giving the incorrect answer), you have one chance to correct it. But the final score cannot be higher than 80% of the total score.
(5) The implementation should be your work and only your work. The penalty of copying, plagiarism and other forms of cheating is a grade of zero on the entire homework.