Assignment 3

Due: 2/12

Note: Show all your work. You can do manual calculations, use R, or use any software (e.g., Weka, Excel, JMP) to answer the questions, unless otherwise noted. In any case, you need to attach the relevant file(s) or screenshot(s) that shows how you obtained your answers.

Problem 1 (20 points) Consider the following dataset (sorted in non-decreasing order):
<18, 22, 36, 36, 40, 48, 50, 60, 65, 71, 93, 105, 124, 128, 130>

(1) Perform the equal width binning on the above data with 3 bins. Note that the bin boundaries are integers in the textbook (to make the discussion simple). But, for this assignment your bin boundaries will include fractions. So, **you must follow the example in the lecture slides**. For each bin, show the bin interval, data values in the bin, and smoothed values using bin means, bin medians, and bin boundaries.
(2) Repeat the same with equal depth binning with 3 bins.
(3) If you transform the dataset into the interval of [0, 100] using Min-max normalization, what is the new value of 48?
(4) If you transform the dataset using z-score normalization using the standard deviation, what is the new value of 48?
(5) If you transform the dataset using z-score normalization using the mean absolute deviation, what is the new value of 48?

Note: For Problem 1-(4) and Problem 1-(5), you need to show the mean, standard deviation, mean absolute deviation, and the new, transformed value as well as all calculation steps.

Problem 2 (10 points) This problem is a practice of calculating correlations between some input attributes (or predictive attributes) and the output attribute (or predictable attribute) in the *machine.csv* dataset. Your task is to calculate the following correlations:

correl(*MYCT*, *ERP*)
correl(*MMAX*, *ERP*)
correl(*CACH*, *ERP*)
correl(*CHMAX*, *ERP*)

Here, *correl*(*X*, *Y*) denotes the Pearson's correlation coefficient between *X* and *Y*.

In your submission, include all four correlations, and indicate the attribute that has the strongest correlation with *ERP*.

The dataset, *machine.csv*, was downloaded from UCI Repository and the description of the dataset can be found in https://archive.ics.uci.edu/ml/datasets/Computer+Hardware.

Problem 3 (10 points) This problem is a practice of determining correlation between two nominal attributes using the chi-square test, which we discussed in the class. Consider the *hw3-p3.arff* dataset. It has 13 attributes and 74 tuples.

(1) Determine whether there is a correlation between attribute *B* and attribute *M*.
(2) Determine whether there is a correlation between attribute *L* and attribute *M*.

Submission:

Name your file *lastName_firstName*_HW3.doc (or *lastName_firstName*_HW3.pdf). If you have multiple files, then combine all files into a single archive file. Name the archive file as *lastName_firstName*_HW3.EXT. Here, "EXT" is an appropriate archive file extension (e.g., zip or rar). Upload this archive file to Blackboard

.