Homework 2

Due: 9/21

- You must show all calculations and important intermediate steps/results. Otherwise, you will lose points even if your answers are correct.
- If you use R, you must submit the R code file.

Problem 1 (10 points). Use the *hw2_p1.csv* file for this problem.

(1). Calculate the mean, standard deviation, minimum, and maximum of *temp*.
(2). Calculate the 25th, 50th, and 75th percentiles of *temp* using the *percentiles* formula that we discussed in the class.
(3). Plot the boxplot of *RH*.
(4). For this problem, consider only the *wind* attribute. Detect outliers using the IQR method that we discussed in the class and show all outlier values.

For problems (2) and (4), you should not use a data analysis or a data mining tool that automatically determines the required values. However, you may use a spreadsheet software just for calculations and you may write a R code that does calculations. In this case, your own code must do the calculations. For example, you should not use the *quantile* function for problem (2) but you can write a R code that does calculations. For problem (1) and (3), you may use any tool.

Problem 2 (10 points). Consider the following two objects with 7 binary attributes:

OID	A1	A2	A3	A4	A5	A6	A7
O1	P	P	N	P	N	N	P
O2	N	P	N	N	P	N	N

(1). Calculate the distance between O1 and O2 assuming all attributes are symmetric attributes.
(2). Calculate the distance O1 and O2 assuming all attributes are asymmetric attributes with P being more important than N.

Problem 3 (10 points)

Consider the following dataset with two objects.

Object	A1	A2	A3	A4
O1	1	second	gold	medium
O2	4	third	bronze	large

Here, all attributes are ordinal attributes and ranks of their values are shown below (lowest rank on the left):

A1: {1, 2, 3, 4, 5}
A2: {first, second, third}
A3: {bronze, silver, gold}
A4: {small, medium, large}

Calculate the distance between O1 and O2 using the method discussed in the class. Use the Euclidean distance measure.

Problem 4 (10 points). Consider the following dataset:

OID	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10
O1	0	1	4	1	3	1	4	1	1	2
O2	2	2	1	5	0	4	0	3	5	2
O3	3	0	5	2	1	3	2	0	3	4

(1). Calculate the cosine similarity between O1 and O2, *cosine*(O1, O2).
(2). Calculate the cosine similarity between O1 and O3, *cosine*(O1, O3).
(3). Is O1 closer to O2 or O3?

Problem 5 (10 points). Consider the following dataset, which has attributes of mixed types.

Object ID	A1	A2	A3	A4	A5	A6	A7
O1	47	1	Yes	Yes	No	Low	hot
O2	49	1	No	Yes	No	Middle	mild
O3	29	0	No	Yes	Yes	Low	cool
O4	35	0	Yes	No	No	Middle	mild
O5	73	1	No	No	No	High	hot
O6	27	0	Yes	No	No	High	mild
O7	52	1	No	Yes	No	Low	cool
O8	36	1	No	No	Yes	High	mild
O9	12	0	Yes	No	Yes	High	hot

- A1 is a numeric attribute.
- A2 A3 are symmetric binary attributes.
- A4 and A5 are asymmetric binary attributes, where Yes is more important than No
- A6 is a categorical (nominal) attribute.
- A7 is an ordinal attribute. The order of values is {cold, cool, mild, hot}, where cold had the lowest rank and high has the highest rank.

Calculate the distance between O7 and O8, *d*(O7, O8), and the distance between O7 and O9, *d*(O7, O9), using the method that we discussed in the class. Is O7 closer to O8 or closer to O9? You must do all calculations yourself.

You must do all calculations yourself.

Include all answers in a single Word or PDF document and upload it to Blackboard. Use *LastName_FirstName_hw2.docx* or *LastName_FirstName_hw2.pdf* as the file name. If you have additional files, such as an Excel file or a R code file, then combine all of them into a single archive file and name it *LastName_FirstName_hw2.EXT*, where *EXT* is an appropriate archive file extension such as *zip* or *rar*.