**Assignment 1: Data Description**

In this assignment, you are asked to answer the following questions using tools of your choice. Example tools may include MS Excel, any database, R, Matlab, etc. You may also choose to write a program (in C/C++, Java, Php, Python, etc.) to answer the questions. Feel free to discuss with your classmates in doing the homework. However, you should answer these questions entirely by yourself.

Put the questions, your answers, and explanation on how you get the answers in a PDF file and submit to e-learning. If you have source codes or other related files, pack all of them in a single zip file and submit to e-learning.

**Question 1:**

The file **data.online.scores.txt** contains 1000 records of students’ exam scores. The first column is students’ id, the second column is the mid-term scores, and the third column is the final scores. The data are split by tab. Calculate the following statistical description **about mid-term scores**.

1. Max, min
2. First quartile Q1, median, third quartile Q3.
3. The mean score.
4. The mode score.
5. Empirical variance.

Repeat the above **on the final scores**.

1. Max, min
2. First quartile Q1, median, third quartile Q3.
3. The mean score.
4. The mode score.
5. Empirical variance.

**Question 2:**

Using data in the file **data.online.scores.txt**, normalize the mid-term score using z-score normalization.

1. Compare the empirical variance before and after normalization.
2. Given original score of 90, what is the corresponding score after normalization?
3. Pearson’s correlation coefficient between midterm scores and final scores is:
4. Covariance between midterm scores and final scores is:

**Question 3:**

Given the inventories of two libraries, Citadel's Maester Library (CML) and Castle Black's Library(CBL), you are asked to compare the similarity between the two libraries by using different proximity measures.

1. Using 200 books, Table 1 summarizes how many books are served by corresponding library. For example, the table cell for CBL=0 and CML=0 is 20, which means there are 20 books that are served neither by CBL nor CML. The table cell for CBL=1 and CML=0 is 2, which means that there are 2 books that are served by CBL but not CML. Using Table 1, calculate the Jaccard coefficient of CBL and CML.

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | Citadel's Maester Library (CML) | |
|  |  | 0 | 1 |
| Castle Black's Library(CBL) | 0 | 20 | 120 |
| 1 | 2 | 58 |

Table 1: Book supplement summary

1. File **data.libraries.inventories.txt** has two data records, one record for each library. Each record shows numbers of copies of 100 selected books the library owns. Calculate the Minkowski distance of the two records with regard to different h values:
   1. h = 1
   2. h = 2
   3. h = ∞
2. Using file **data.libraries.inventories.txt**, calculate Cosine similarity between CML and CBL.
3. Using file **data.libraries.inventories.txt**, calculate Kullbac-Leibler divergence between CML and CBL. Let be the number of copies of book\_i in a library. Assume a person picks up a book randomly, the probability that he/she picks up book\_i is then .

Based on this probability, calculate the Kullback–Leibler divergence of CML and CBL.

**Question 4:**

Table 2 is a summary about customers’ purchase history of diapers and beer. Calculate the chi-square correlation value.

|  |  |  |
| --- | --- | --- |
|  | Buy diaper | Do not buy diaper |
| Buy beer | 150 | 40 |
| Do not buy beer | 15 | 3300 |

Table 2: Purchase history.

**Submit your answers to e-Learning Dropbox.**