CS 422: Data Mining

Department of Computer Science
Illinois Institute of Technology
Vijay K. Gurbani, Ph.D.

Fall 2022: Homework 1 (10 points)

Due date: Sun, Sep 11 2022, 11:59:59 PM Chicago Time

Please read all of the parts of the homework carefully before attempting any question. If you detect any ambiguities in the instructions, please let me know right away instead of waiting until after the homework has been graded.

1. Exercises (4 points)

1.1 Tan, Chapter 1 (2 points divided evenly among the questions)

Besides the lecture, make sure you read Chapter 1. After doing so, answer the following questions at the end of the chapter: 1, 3.

1.2 Tan, Chapter 2 (2 points divided evenly among the questions)

Besides the lecture, make sure you read Chapter 2, sections 2.1 – 2.3. After doing so, answer the following questions at the end of the chapter: 2, 3, 7, 12.

2. Programming Problem

Please label your answers clearly, see Template.Rmd R notebook for an example (Template.Rmd R notebook is available in "Blackboard → Assignment and Projects → Homework 0". Rename Template.Rmd to firstname.lastname.Rmd.) Each answer must be preceded by the R markdown as shown in the Homework 0 R notebook (### Part 2.1-A-ii, for example). Failure to clearly label the answers in the submitted R notebook will lead to a loss of 2 points.

2.1 Exploratory data analysis (6 points divided evenly among the constituent parts)

You are provided a dataset of the United States Covid-19 deaths and testing in each of the 50 US states, Puerto Rico, and the District of Columbia (please see us-covid.csv). This dataset has the following nine attributes:

state	*State or Territory*
level	*Level of community transmission*
total_cases	*Total cases = confirmed + probable*
confirmed	*Confirmed cases*
probable	*Probable cases (not confirmed)*
cases_last_7_days	*Last 7 days case*
case_rate_per_100K	*Case rate per 100,000 residents*
total_deaths	*Total deaths recorded*
confirmed_deaths	*Confirmed deaths*

Feel free to use any R package that you think will make benefit you. For instance, the dplyr package can be used to answer many of the questions below. See https://courses.cs.ut.ee/MTAT.03.183/2017_fall/uploads/Main/dplyr.html for some examples of using dplyr.

(i) Read the dataset into a R dataframe; call the dataframe 'data.df'. Pay attention to the first seven lines of the dataset. These contain comments as indicated by the '#' character in dataset file. When reading the dataset, ignore the lines that contain comments. (See the manual page for read.csv() and find out the parameter that will allow you to ignore comments.)

(ii) You will notice that the last column of the dataset has many "N/A" values (Not Applicable). Drop this column so that the data.df dataframe now contains only eight columns, or attributes. The remaining analysis will be done on the dataframe with eight columns.

(iii) Sort the dataframe:

(a) Sort the dataframe by **descending** order of total cases, then print the **top** six entries.

(b) Sort the dataframe by **descending** order of total cases, then print the **bottom** six entries.

(iv) Create a correlation visualization using the pairs.panels() API from the R library psych. Note that the first two columns can be excluded from correlation analysis. Plot the correlation and use it to answer the following questions:

(a) Why should we exclude the first two columns from correlation analysis?

(b) Which pair of columns have the highest correlation?

(c) Which pair of columns has the lowest correlation?

(v) Focus on confirmed cases, probable cases, and total deaths.

(a) Draw a plot of confirmed+probable cases (on the X-axis) against the total deaths (on the Y-axis). Label the plot appropriately.

(b) As you see the plot, there appears to be an anomaly. Looking at the data, briefly describe the anomaly.

(c) Print out the state name, total deaths, confirmed cases, probable cases and total cases of all such states that show this anomaly.