Data Mining Tools and Techniques
Homework 1 – Analyzing Data, Posing Questions
You may work in teams of up to 2 students on this assignment

Summary

For this assignment, you will be working with different types of data sets. Your goal is to understand and characterize one or more of the data sets, pose data mining questions against four data sets and propose additional questions if you had more data.

The Data

Attached to this assignment are four groups of data sets as listed below.

Data Group 1: Ice Cream Data
Data Group 2: 100 Years of Beach Erosions in New Zealand
Data Group 3: Nutrition (There are 2 datasets. McDonald's Link and Nut. values for common foods & products)
Data Group 4: Where to Attend College

If you follow the given links above, you can find out more about each data group.

You may need to do some research to better understand each data set. Quite often when you want to data mine a data set, it is from a domain you are not familiar with. Thus, you need to do some research to understand the attributes within each data group.

Part 1: Ice Cream Data group (20 points)

A) Provide descriptive statistics for each attribute as they apply. This may include minimum, maximum, average, etc.
B) For each attribute, explain whether it is descriptive, discrete, continuous, or discontinuous?
C) Is this a supervised or unsupervised data set? If supervised, what is the class variable(s)?
D) Is this data set time-series, temporal, spatial data, sequence, some combination, or none of these?
E) What measurement scale is the dependent variable (if it exists, it is usually the right-most column)?

Part 2: Formulating questions for the 4 data groups (80 Points Total: 10 points per question per data group)

F) Formulate three specific questions against each of the data sets. Try to vary the question types (e.g. classification, regression, clustering). If this is not possible, explain why. For each question, describe any preprocessing that need to be done to get the data in a format so that you could ask the question. **Make sure to ask data mining, not database type of questons.**
G) If you could collect additional data (other data files), describe what that data would look like. What additional questions would you be able to ask? I am looking for you to be creative and not just repeat what I have done on the next page.

If you work as a group of one do everything above.

If you work as a group of two, do everything above, plus apply the tasks in Part 1 to the *Where to Attend College* data group.

Grading criteria:

1) How well did you describe the data group from Part 1?
2) Do you show a good understanding of the data set?
3) Are your questions reasonable? insightful?
4) How much effort did you put into the assignment?
5) How creative are you to extend a given data group and project additional analysis?

Deliverable: An MS-Word file. Use the *ACM Template for Assignment 1* as a starting point. Your paper length should not exceed 4 pages. Email the assignment to **boetticher@uhcl.edu.** Do not email it to the Teaching Assistant.

Due Date: Monday, June 13th, Noon (Central Time)

Data Mining Tools and Techniques
Homework 1 – Analyzing Data, Posing Questions
You may work in teams of up to 2 students on this assignment

Early Bird Bonus

If you are the **first** person/group to turn the assignment in **early**:	**10 point bonus.**
If you are the **second** person/group to turn the assignment in **early**:	**8 point bonus.**
If you are the **third** person/group to turn the assignment in **early**:	**6 point bonus.**
All other early submissions:	**1 point bonus.**

If you wish to turn it in early, then email me your solution so that I have a time stamp.

Early means that you submit the assignment at least 1 day before the deadline.

If you resubmit the assignment, then I go by the last submission date.

Sample anaylsis of a data set

To help you in your analysis, I have created the following example. I am not claiming this example is complete. However, there is enough details to get you going.

Chicago Crime Data 2001 to present (The data set is included in the Assignment)

The Chicago crime dataset describes criminal activity in the Chicago area.

The data set contains 99 rows of data (The original data set has many more rows.). There are 22 attributes. Each row corresponds to an actual crime and includes a unique Crime ID number (nominal), case number (nominal), timestamp(nominal), type of crime (nominal), location address(nominal), and location coordinates(nominal). The CrimeID, or Case Number, may be used as a key.

It is possible to create various types of problems from this data set.

1) If the types of crime are summed up. Then the total crimes for a given day (or week, month) would be the class variable. This would be viewed as supervised type of problem. The original data set is a temporal data set. It has a time stamp, but does not occur on a regular basis. After the summing the crimes to a daily (or weekly, monthly) basis it now becomes a time-series regression type of problem. The dependent variable would be the day, week, or month and it is nominal.

   We might ask, what is the trend of crime in Chicago?

   If separated out by month, which month has the highest/lowest amount of crimes (this is more of a database question, not a data mining question).

2) If an analysis examined the types of crimes in terms of location coordinates, there is no class variable and this would be an unsupervised type of problem. This would be viewed as supervised type of problem. This would be a spatial type of problem and would be a clustering type of problem.

   Clustering the crimes would lead to the following question: Do resources (e.g. police officers) need to shifted to a hot spot (high rate of crime)?

3) If an analysis groups the data by the type of crime, and sums up each type of crime for a given day (or week, month), the specific date would be a class variable. This would be viewed as supervised type of problem. This would be viewed as supervised type of problem. The original data set is a temporal data set. It has a time stamp, but does not occur on a regular basis. After the summing the crimes to a daily (or weekly, monthly) basis it now becomes a time-series regression type of problem. The dependent variable would be the day, week, or month and it is nominal.

   We might ask, what is the trend of crime in Chicago for a specific type of crime?

   If separated out by month, which month has the highest/lowest amount of a specific type of crime (this is more of a database question, not a data mining question). Is there a correlation between different types of crimes from a time perspective?

4) If an analysis groups the data by the type of crime, then it is separated out by location coordinates, there would not be a class variable and this would be an unsupervised type of problem. This would be a spatial type of problem. This would be a clustering type of problem.

   Clustering the crimes would lead to the following question: Do resources (e.g. police officers) need to shifted to a hot spot (high rate of crime)?

   Is there a correlation between different types of crimes for a given area?

5) Another type of problem combines 1 and 4 with 2 or 3. This would be a spatio-temporal (time-series) type of problem. What are trends of specific types of crimes in a specific area?

Other data sets may be brought in to enhance the analysis. These could include:

- Crime data from other cities. How does Chicago compare to New York, Houston, etc.?
- Temperature data. How does temperature impact crime rates?
- Population data. Divide the crime rate by population size. This would treat crime rates in relative, not absolute terms. So, if the crime rate goes up 10%, but the population doubles, then the crime rate has gone down relative to the population size.
- Real estate values. Are there more crimes in more expensive neighborhoods? Could break out the crimes to types of crimes.
- Moon cycles. Do more crimes happen during a full-moon?
- Budget analysis. How do changes in the budget impact the crime rate if at all?