Data mining Mini-Project – Fall 2018:

The goal for having a project is for you to show that you can do something substantial, and practice the data mining process on some actual data set. The goal of your report is to provide *evidence of learning and understanding with respect to data mining.*

This semester you are asked to do change detection in just one of the five boroughs of New York City.
You will examine other boroughs for comparison, but focus on just one for change detection.
And you will focus on just two months: June and July, of just two years: 2017 and 2018.

The over-arching questions to address are:

- How did the patterns of traffic accidents change from 2017 to 2018?
- How did the patterns of traffic accidents stay the same?

Background of the Data:
In the interest of improving public safety, New York City publishes data on all motor vehicle collisions. The question you need to investigate is: has anything actually changed in the past year? Have any regions or areas gotten much worse, or much better? If you need motivation, imagine that you were hired by New York City to tell them where to improve safety next year.

To avoid issues related to bad weather, we will be just looking at the months of June and July. Remember that July and August are the months when schools are out, and many people visit NY City with their families.

You and your team is to select one of the five boroughs of NY City, and see how the collisions changed between the past years, and the summer of 2018. Did the clusters move around?

Here is a convenient link to the latest database:
https://data.cityofnewyork.us/Public-Safety/NYPD-Motor-Vehicle-Collisions/h9gi-nx95
If the link does not work, go to https://data.cityofnewyork.us and search for "NYPD Motor Vehicle Collisions".

You should be able to find the full database, with over a million records. However, to make things manageable, you only need to examine the difference between 2017 and 2018, and only for the months of June and July. You will need to reduce the data correctly. Select statements are applicable here. You can download the data in several different formats. It took me about two minutes at RIT.

You can use any packages you want for this task. Again, to keep life simple, you can assume that longitude and latitude are locally Euclidean. You do not need to use the Haversine distance for small local distances. (That is the lesson of Taylor's series – you can ignore the high order terms.)

Your assignment:
Pick one of the five boroughs of NY City.
For the months of June and July, and for the years 2017 and 2018, figure out what has changed.

Pick four of the following questions to investigate. I expect you to pick the first four. The others are harder.

1. How was June of 2018 different then June of 2017?
2. How was July of 2018 different then June of 2017?
3. How was July of 2018 different then June of 2018?
4. How was July of 2017 different then June of 2017?
5. Which of the four months is most different from the overall pattern?
   How did you determine this, and how did you justify your decisions.
6. Which locations have the most accidents overall?
7. Which day of the week has the most accidents?
8. Which hour of the day has the most accidents?
9. How do weekends compare to weekdays?
10. Are there any particular aspects pedestrians should know about?
11. Another aspect of your choosing.

Possible Strategies:
Find the average overall clusters with respect to all four months, then see how each of the first four months differ.
OR, find the average for the two Junes, and the average for the two Julys, and see how they differ...

Sections:

Your report should be about 8 to 12 pages double spaced, 12 point font, one inch margins, with figures, and with a final page of references and a bibliography. Getting the paper down to 8 to 12 pages double spaced can be a challenge for some.

It should include the following sections. Here are the sections, with possible questions you might consider.

1. **Title and authors**
2. **No title page – they just waste space**
3. **Abstract – one paragraph**
   Why should I read this paper? What did you find overall? An abstract is an advertisement to get the reader to want to read your paper. It is all one paragraph, and teases the reader to want to read the paper.

4. **Overview / Introduction – about one page.**
   Start with why. Why is this important? Who should care? An overview of what you did and what you found. Which borough did you pick? Is there any reason why you picked that borough? You might select a borough simply because you want to visit there, a borough because you have heard about it, or a borough with the most accidents.

5. **Data Preparation**:
   a. How clean is the data?
   b. Did you quantize the data into regions?
   c. Are there any issues with the data?
   d. Is the data from the two years comparable, or are there any issues between 2017 and 2018?
   e. Are there the same number of weekends in month you are comparing?

6. **Answers to the above four questions you selected.**
   As you answer them, describe the process you used.
   Use data visualizations (probably heat maps or contour maps).
   Possible questions to consider:
   a. What clustering did you do?
   b. What algorithms did you use?
   c. Did you normalize your data somehow?
   d. How did you do any data visualization or create any figures?
   e. How did you compare one month to another?

7. **General Discussion**
   Questions to consider:
   a. What went wrong, or what challenges did you face?
   b. What was interesting about this?
   c. Which algorithm worked best?
   d. What else would you like to share about the project?

8. **Conclusions**
   Questions to consider:
   a. What did you learn overall?
   b. Which algorithms did you finally use?
   c. What went wrong, or what challenges did you face?
   d. What was interesting about this?
   e. What else would you like to share about the project?
   f. What did you learn about data mining by doing this project?