Data mining Project – Fall 2023:

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:
Staten Island, Manhattan, the Bronx, Queens, or Brooklyn. You pick! If you don't know anything else, pick Queens.

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: 2019 and 2020.

The over-arching questions to address are:

- How did the patterns of traffic accidents change?
- 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: did anything actually change between the summer of 2019 versus 2020? 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 partner are to select one of the five boroughs of NY City, and see how the collisions changed over recent years.
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 the two years. 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. And, to keep life simple, 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. Load the data into a database. Find the answers to the following seven questions:

1. For the two years given, figure out what has changed in the summer from one year to the next.
   Figure out how to visualize the difference, in some way.

2. How was June of 2019 different then June of 2020?
   Figure out how to show or demonstrate the difference.

3. How was July of 2019 different then July of 2020?
   Figure out how to show or demonstrate the difference.

4. For the year of January 2019 to October of 2020, which 100 consecutive days had the most accidents?

5. Which day of the week has the most accidents?

6. Which hour of the day has the most accidents?

7. In the year 2020, which 12 days had the most accidents?
   Can you speculate about why this is?

Sections:
Your report should be about 7 to 12 pages double spaced, 12 point font, one inch margins, including figures.

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

1. **Title and Teammates**

2. **No title page – they just waste space**

3. **Data Preparation**
   Discussion.
   a. How clean is the data?
   b. Which data did you ignore?
   c. What data did you focus on?
   d. Did you quantize the data into regions?
   e. Are there any issues with the data?
   f. Is the data from the two years comparable?
   g. Are there any other issues you found?

4. **Answers to the above questions.**
   As you answer them, describe the process you used. Use data visualizations (probably heat maps or contour maps).
   Possible questions to consider: What clustering did you do? What processing did you do? What algorithms did you use? Did you need to normalize your data somehow? How did you do any data visualization or create figures?

5. **Conclusions**
   Questions to ponder:
   a. What did you learn overall?
   b. What went wrong, or what challenges did you face?
   c. What was interesting about this?
   d. Which algorithm worked best?
   e. What else would you like to share about the project?
   f. Which algorithms did you finally use?
   g. What went wrong, or what challenges did you face?
   h. What was interesting about this?
   i. What else would you like to share about the project?
   j. What did you learn about data mining by doing this project?