CPSC 368: Databases in Data Science
Finding datasets and choosing a topic

Assignment Description

In this assignment, you and your team will brainstorm ideas for the term-long data science project. As a first step, we want you to identify datasets and brainstorm ideas for what you want to do for your project. The class collectively voted on a project topic so when thinking of ideas, be sure to think of ideas and datasets that are related to the topic.

We will ask you to brainstorm **three** distinctly **different** ideas that use different datasets. This is to maximize the likelihood of finding a project that will meet the requirements we are looking for.

While you can use the same dataset in more than one proposed idea, no two ideas should share identical datasets. That is, imagine if you had datasets A, B, C, D, E, and F. Idea 1 can use datasets A, B, and C, idea 2 can use datasets D, E, and F, and idea 3 can use datasets A, C, and E. We would consider the three ideas to be using different datasets. If idea 3 used datasets A, B, and C, then that would be a problem.

For each idea, you must:

1. Explain the proposed idea.
2. Identify how the idea relates to the topic chosen by the in-class vote.
3. List three datasets you wish to work with. The idea is that you will combine these three datasets to get the data that you need.
    a. For each dataset:
        i. Provide a link. If there is more than one dataset listed on the linked page, be specific about which one you want to work with.
        ii. Comment on the trustworthiness and reliability of the data. Do not choose data that is of questionable quality.
    b. Your datasets do not need to be in CSV form. If you have data that needs to be fetched via API call, that is fine. You will need to fetch the data to eventually create a CSV file but that can be done in a future assignment.
    c. The datasets that you choose should have at least 4 columns and 50 rows.
4. Describe the data analysis you wish to perform. Specifically, explain:
    a. What you are trying to analyze and how you wish to go about analyzing it. Why is this the best method to analyze the data? List two other possible ways to approach the problem and explain why you are not using these approaches?
    b. The purpose of the data analysis (i.e., why is this important?)
    c. The statistical rigour of the methodology you are choosing. If you are not using a statistical approach, explain how your approach ensures that the results you obtain are meaningful.
        i. Keep in mind that you will be able to use existing libraries when implementing your project. The focus of this part of the assignment is for you to explore possibilities and pick something suitable for your use case.

5. Identify a visualization (or a series of visualizations) that you will use to communicate your results. Provide a sketch.

Deliverables

Please submit your work as a single PDF file.

Each group only needs one submission (i.e., a representative from the group can submit for everyone; there is no need for each individual group member to submit).

The teaching team warmly recommends that you review the rubric on Canvas prior to submission.