Final Project

Up until now, we have given you fairly detailed instructions for how to design data analyses to answer specific questions about data -- in particular, how to set up a particular analysis and what steps to take to run it. In this project, you will put that knowledge to use!

Put yourself in the shoes of a data scientist being given a data set and asked to draw conclusions from it. Your job will be to understand what the data is showing you, design the analyses you need, justify those choices, draw conclusions from running the analyses, and explain why they do (or do not) make sense.

We are deliberately not giving you detailed directions on how to solve these problems, but feel free to come to office hours and lab hours to brainstorm.

The final project is due December 6th (the last day of class).

Objectives

There are three possible paths through this project:

1. You may use data set #1, which captures information about bike usage in New York City. See below for the analysis questions we want you to answer.
2. You may use data set #2, which captures information about student behavior and performance in an online course. See below for the analysis questions we want you to answer.
3. You may propose a new analysis problem to us that you would like to solve. See below for the steps for doing that.

Partners

On this project **you may work with a partner**. Working with a partner is optional, and working with a partner will not impact how the project is graded. When cloning your repository on GitHub classroom, you will have an opportunity to join an existing team (i.e., your partner has already created a group) or create a new team (i.e., your partner hasn't set up their repository yet). By doing so, you and your partner will *share a repository*.

Common material for projects

For each project, in addition to any code you write to perform your analyses, we would like you to turn in two files:

- `info.txt` : This text file should first list which of the three project paths you are taking. It should then include the list of team members in your group. List each person's full name and Purdue username. If you are working by yourself, please still submit this file, with just your information in it.
- `report.doc` or `report.pdf` : The report you write for this project. Each report should consist of:
    - A section describing the data set you are working with.
    - A section describing the analyses you chose to use for each analysis question (with a paragraph or two justifying why you chose that analysis and what you expect the analysis to tell you)
    - A section (or more) describing the results of each analysis and what your answers to the analysis questions are as a result of performing these analyses. Visual aids are helpful here, if necessary to back up your conclusions. Note that it is OK if you do not get "positive" answers from your analysis, but you must explain why that might be.

Path 1: Bike traffic

The data set here gives information on bike traffic across a number of bridges in New York City. In this path, the analysis questions we would like you to answer are as follows:

1. You want to install sensors on the bridges to estimate overall traffic across all the bridges. But you only have enough budget to install sensors on three of the four bridges. Which bridges should you install the sensors on to get the best prediction of overall traffic?
2. The city administration is cracking down on helmet laws, and wants to deploy police officers on days with high traffic to hand out citations. Can they use the next day's weather forecast to predict the number of bicyclists that day?
3. Can you use this data to predict whether it is raining based on the number of bicyclists on the bridges?

Path 2: Student performance

`behavior-performance.txt` contains data on how students watched videos and performed on in-video quizzes in an online course. `readme.txt` details the information contained in the data fields. In this path, the analysis questions we would like you to answer are as follows:

1. The instructor for this course wants to know whether students are naturally grouped into different categories of video-watching behavior. How well can the students be clustered by the given behavioral features (`fracSpent` , `fracComp` , `fracPaused` , `numPauses` , `avgPBR` , `numRWs` , and `numFFs`)? You should use all students that complete at least five of the videos in your analysis.
2. It would save people a lot of time if course grades could be inferred based on a student's behavior in studying course material as opposed to sitting for tests. So, how well can you predict a student's average

performance (i.e., average score `s` across all quizzes) in this course by the behaviors they exhibited while watching the course videos? You should use all students that complete at least half of the quizzes in your analysis.
3. Taking this a step further, how well can you predict a student's performance on a particular in-video quiz question (i.e., whether they will be correct or incorrect) in this course by the behaviors they exhibited while watching the corresponding video? You should use all student-video pairs in your analysis.

Path 3: Choose your own adventure

You are also welcome to design your own project. The project must be at least as difficult as one of the pre-set projects!

If you want to choose this path, you will need to submit a project writeup. This writeup should include:

1. A list of who will be in your project team.
2. A description of the data set(s) you are going to analyze.
3. A set of analysis questions you want to answer
4. A sketch of what analyses you will use to answer these questions (you do not need to have performed the analysis yet!) -- this is basically the equivalent of section 2 of your report.

Please submit your project description (by email) to the professors by November 22nd.

What to turn in

See "Common material for projects" above.