Final Project

Due: 30th April, 2021 11:59pm ET

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 recitation hours to brainstorm.

Objectives

There are two 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.

Partners

On this project **you may work with one partner** (except for Honors contracting students who must work individually). Working with a partner is optional, and working with a partner will not impact how the project is graded. If you want to work with a partner, it is your responsibility to pair up; feel free to use Piazza's "Search for Teammates" feature (https://piazza.com/class/kjn0wpn1k2v5ju?cid=5) to facilitate this.

If you are working with a partner, *you must share a repository*. This means that one of you will clone your repository on GitHub classroom, and the other will join this team (rather than cloning a second repository). When you go through the normal steps for cloning your repository 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).

If you plan to team up with someone who has already cloned their repository, it is very important that you follow these instructions to join their team, rather than creating a new repository.

Path 1: Bike traffic

The `NYC_Bicycle_Counts_2016_Corrected.csv` 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 related to video-watching behavior

`behavior-performance.txt` contains data for an online course on how students watched videos (e.g., how much time they spent watching, how often they paused the video, etc.) and how they performed on in-video quizzes. `readme.pdf` details the information contained in the data fields. There might be some extra data fields present than the ones mentioned here. Feel free to ignore/include them in your analysis. In this path, the analysis questions we would like you to answer are as follows:

1. How well can the students be naturally grouped or clustered by their video-watching behavior (`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. Can student's video-watching behavior be used to predict a student's performance (i.e., average score `s` across all quizzes)? This type of analysis could ultimately save significant time by avoiding the need for tests. 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) based on their video-watching behaviors while watching the corresponding video? You should use all student-video pairs in your analysis.

What to turn in

You must turn in two sets of files, by pushing them to your team's Github repository:

- `report.pdf` : A project report, which should consist of:
  - A section with the names of the team members (maximum of two), your Purdue username(s), and the path (1 or 2) you have taken.
  - A section describing the dataset 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 questions are based on your results. 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.
- All Python `.py` code files you wrote to complete the analysis steps.

Note: You may encounter instances in the data where you would need to make a decision on including extra information, changing non number values in files to appropriate numbers. How you choose to handle this is up to you. You may visit office hours to seek guidance.