Final Project Outline

Computer Science 21800
Fall 2020

You will work in groups of four to analyze a dataset of your choice. This project is open-ended assignment where you will: formulate a question, acquire some dataset that can evaluate/answer/explore this question, and write a "longform" article on your findings. This project works best if the domain has some significance to you and your partners.

(Step 1). Selecting a Group

Due. October 9th, 2020

We ask that you work in groups of four people. You will have to coordinate with your partners outside of class to complete the project. For this project, you will receive both a group grade and an individual grade. The group grade will be given to you based on the TAs and my judgment about the overall quality of your final article. For, the individual grade, you will be asked to grade each of your group members. We will aggregate both the group grade and your individual scores to give each of you a final grade for the project. I encourage you to form your own groups. And, once you have done so please submit the names to:
`https://forms.gle/LjrUDdcoPKF8h9iv7`

After the groups are settled, each group will be assigned to one of our TAs. They will help you throughout the project and periodically check-in to see how you are doing.

Step 1 FAQs

1. *What if I can't find a group of four by October 9th?* Fill out the form with as many group members as you can (even if it's just you!), and we will try to match you up with others.
2. *What if I add the class after October 9th?* We will try our best to find a group willing to take you. We anticipate some students to drop the class so there will be openings.
3. *What if I want to work alone?* We want you to work in groups—the only reason that we **may** allow you to work alone is if you are remote in an inconvenient time-zone.
4. *If do all of the work of the project, will I get more credit?* No. Please learn how to delegate responsibility.

(Step 2) Selecting a Dataset and Question

Due. October 16th, 2020

While we won't micromanage you on what dataset you use or what question you explore, there are a few general guidelines. First, this project is supposed to make use of the techniques in this class. We ask that the dataset be sufficiently large ($>$ 10,000 rows of data) to be interesting and statistically significant. Second, the questions that you ask of this dataset shouldn't be "obvious", and your findings should be generally informative. It also goes without saying that these questions/data should be respectful of your peers coming from different backgrounds and be professional in nature.
To help get you started, here are some example questions and available data to answer them:

- *How do coaching changes affect play calling in American Football games?* [1]
- *How much money does a landlord actually make?* [2]
- *Does an artist's color palette predict an early demise?* [3]

[1] https://github.com/ryurko/nflscrapR-data
[2] https://www.kaggle.com/paultimothymooney/zillow-house-price-data
[3] https://www.kaggle.com/ikarus777/best-artworks-of-all-time

By October 16th, you will email your group name, research question, and dataset to your assigned TA. The TA will give you feedback (e.g., make it more specific) and pointers (e.g., consider using X library) either over email or a Zoom call.

Step 2 FAQs

1. *Can I use data from another class or research project?* Absolutely!
2. *Can I collect my own data through web-scraping, social media, etc?* As long as you don't break any laws, you can collect any dataset that you want.
3. *What if the answer to my research question turns out to be inconclusive?* That's part of the process! An inconclusive answer is sometimes just as valuable as a definitive one.
4. *What if I want to build an app based on data rather than test a hypothesis?* This is fine too! We just ask that you have some way of evaluating the performance, accuracy, efficacy of the final system.

(Step 3) Model Selection

Due. Nov 13th, 2020

After you find a dataset and a question, we will ask you to describe the mathematical models that you will use to evaluate your question. You must include: (1) background knowledge about your dataset and what it represents, (2) a full mathematical description of the model(s) you will use, (3) assumptions that need to be true about the data to apply such a technique, and (4) any biases or artifacts of data collection that might affect your conclusions. These four points must be written in scientific prose with enough background information so that one of your peers could understand what you did.

By November 13th, you will email a writeup of a minimum of 2 pages (11pt font, single-spaced with formulas) to your assigned TA. The TA will setup a Zoom call with you and your group to go over your submission.

Step 3 FAQs

1. *Can I use mathematical models not discussed in class?* Yes, but we ask that you have a thorough understanding of every thing that you use for your project.
2. *Will we be graded on writing?* Yes. Use a spell-checker, use a grammar checker, and proofread! All of our TAs are strong writers who have written multiple research articles, so you will be graded on this.

(Step 4) Initial Results

Due. November 25th, 2020 You will have two weeks to produce initial findings for your dataset and question. By now, you should know: (1) what is an initial answer to your research question, (2) is the mathematical model you used accurate and how do you measure accuracy, (3) what are the confounding factors/explanations, and (4) what else could you do with more time. These four points must be written in scientific prose with the aid of charts, tables, and visualizations when appropriate. For visualizations, we ask you to give a minimum of 1 paragraph explanation of what the visualization is showing. A simple way to think about this is WALTER: (W)hat is the chart trying to show, (A)xes and their units, (L)ines or marks in the chart and how they relate to the visualized data series, (T)rends in the lines or marks that are shown, (E)xceptions to those trends if any, and (R)ecap of the major insight from the chart.

By November 25th, you will email a writeup explaining key results to your assigned TA. The TA will setup a Zoom call with you and your group to go over your submission.

Step 4 FAQs

1. *What if I need additional computing resources (like servers) to get results?* Email us and we can make it happen!

2. *What if we can't get it done in two weeks?* Try your best to submit "initial" results. Results that are informative enough that the TA can give you feedback.

(Step 5) Final Submission

Due. December 10th, 2020 Based on the feedback from your assigned TA, you will have one month to submit a final writeup. This writeup should contain:

1. An introduction motivating the research question and describing why it is interesting.
2. A literature review of others that have looked into such research questions before.
3. A description of your models and methods.
4. Results

While we won't ask for a minimum overall length, successful articles should aim for 8-10 pages single-spaced with formulas and charts.
By December 10th, you will submit your final writeups to:
`https://forms.gle/Ghh2fHS8E8EqUGGX8`
All students must submit on this form since you will grade your peers.