Project: Data analysis and presentation

DS4200: Information presentation and data visualization

Goal

Pick a dataset and do some analysis and visualization with it. That is your project in a nutshell.

Okay, the actual requirements

The project for this class will consist of analysis and visualize on a dataset of your own choosing. The dataset may already exist, or you may collect your own data using a survey or by conducting an experiment. You can choose the data based on your interests or based on work in other courses or research projects. The goal of this project is for you to demonstrate proficiency in the techniques we have covered in this class (and beyond, if you like) and apply them to a novel dataset in a meaningful way.

To help you success, we divide the project into following parts:

Perperation

The project is designed as a group project. Please checked the group assignment on the Canvas, under the tab People then project group. If you need to adjust your group for any reasons, please email the instructor.

You are welcome to choose any appropriate topics that may not offend your peers. I recommend the topic is neither too broad nor too narrow. For example:

- (Good) How the distribution of campus food is related to the student life?
- (Good) Who are using the Snell library in the evening?
- (Bad) The distribution of restaurants in Boston.
- (Bad) Analysis the major of the students who study in Curry Center.

Based on the topics you have chosen, you can start to collect the data for it. Here is the requirement for data collection.

If you choose to use an exist abstract data (in csv file etc.):

- The number of the observations in the data (after cleaning) should be larger than 2000.
- The number of the features should be larger than 10 (after cleaning and creating additional features).
- Also, it is recommended that the number of the observations is no more than 10000 and the number of attributes is less than 30. You can filter the data if you need. We may have problems to put a really large data on Github.
- The features should have a mix of categorical ones and continuous ones.
- Please prepare a paragraph of writing about why the data is suitable for your topic and what else information maybe missing from the data.

If you choose to use an exist spatial data (map etc.):

- Make sure the data has not only the spatial geometric information but also features.
- The features should have a mix of categorical ones and continuous ones.
- Please prepare a paragraph of writing about why the data is suitable for your topic and what else information maybe missing from the data.

If you choose to collect your own data:

- The observations in the data should be larger than 100. Please contact the instructor for exception.
- The attributes should have a mix of categorical ones and continuous ones.
- Please prepare a paragraph explaining how you collect the data (if there is a survey or something else).

Proposal

The project proposal is due on Feb 14th. The project proposal is a writing report (.txt or .pdf preferred) that need to be submitted as group. In the report, you need to include:

- A title to describe your work (5 pts)
- A paragraph to introduce your topic, and why it is important/interesting? (15 pts)
- Two reference papers that maybe related to your topic (10 pts)
- Two static images from online or other sources that maybe related to your topic.(10 pts)
- An introduction to your data, including the size and source. If you use the online data, please make sure you can load the data. If you decide to collect the data by yourself, just an explanation of how to collect the data will be enough. (20 pts)
- A plan for the later analysis, including any potential pre-processing to the data. What will be the tasks for your analysis and what data visualization (both static and interactive ones are expected) you will include. Please list at least three visualizations you prepare to do. (30 pts)
- Describe each group member's duties. (5 pts)
- The proposal should be no more than 3 pages. Do not include any codes in the proposal. (If you have survey that going to be used in the data collection, please also attached the survey and the survey does not count for the page limit.) (5 pts)

Reusing datasets from class: Do not reuse datasets used in examples/labs/mini-projects in the class.

Project group check point meeting

The week of Feb 18th-21st is the project group check point meeting week. There is no lecture for this week.

Each group should sign up a 15 minutes meeting slot. The link will be posted later.

All meetings are in-person and are in my office at the Meserve Hall 341. All group members should present. Please be on time for your meeting. Please email me if you have special situations that need to meet through zoom.

During the meeting, we will discuss the feedback from your proposal. **Please prepare 2 visualizations for the meeting.**

Final presentation

The final presentation will be in last two weeks. We will post the schedule sign-up sheet when the week approaches. If you decide to present in the earlier weeks, you will get a bonus for the final grade.

- Each group will have 10 minutes for the presentation.
- The attendance for final presentation is required.
- You will present the webpage you made with the data.
- You will need to discuss:
  - Topic and tasks for the project
  - Present your static data visualization and discuss how you design the data visualization
  - Present your interactive data visualization and how to use it to address your tasks. Include a demo to show how to use your interactive data.
  - Conclude what you have learned from the visualization

Final webpage deliverables

The deadline for final project is on Apr 13th. Each group need to present the HTML/CSS/JS file for the webpage you have designed.

You need to setup a Github project webpage for your project work and eventually publish it. On the website, you need to include:

- An overall introduction to your project, including the topic and tasks. (5 pts)
- A paragraph to introduce the data, including the size, source and attributes. (5 pts)
- At least two links to the references. (5 pts)
- You have at least 5 data data visualizations. At least 1 of the visualization is made with Altair and at least 1 of the visualization is made with D3. (5 pts)
- All data visualizations have clear titles, labels, legends (if need). We also need a paragraph to explain the takeaway of the visualization on the webpage, and a paragraph to explain the design idea for the visualization on the separated word document. (40 pts)
- Interactive functions are well-explained and easy to be used. The interaction functions are meaningful and can help people learn more about the data compare to a static plot. The interaction works. (20 pts)
- A summary for your findings, including what you have learned from the data visualization and what can be done in the future. (10 pts)
- Format, including correctly publish the website, no grammar error and keep the webpage tidy and clear.(10 pts)

Peer review

There will be one midterm peer review (after the check point meeting) and one final peer review (after the submission of the final deliverable).

Suggestions

- You should incorporate suggestions from check point meeting into your Final Project.
- Please start as soon as possible and ask for help if you need.
- Show your webpage to more people and ask for any suggestions.
- Review the grading guidelines below and ask questions if any of the expectations are unclear.

Grading

Grading of the project will take into account the following:

- Content - What is the quality of research and/or policy question and relevancy of data to those questions?
- Correctness - Are data visualization carried out and explained correctly? Is there any violations to the design rules?
- Writing - What is the quality of the writing and explanations?
- Creativity and Critical Thought - Is the project carefully thought out? Are the limitations carefully considered? Does it appear that time and effort went into the planning and implementation of the project?

A general breakdown of scoring is as follows:

- 90%-100% - Outstanding effort. Students understand how to apply all data visualization techniques and can put the results into a cogent argument, can identify weaknesses in the argument, and can clearly communicate the results to others.
- 80%-89% - Good effort. Students understand most of the techniques, put together an adequate argument, identify some weaknesses of their argument, and communicate most results clearly to others.
- 70%-79% - Passing effort. Students have some misunderstanding of design concepts in several areas, do not have enough data visualizations, have some trouble putting results together in a cogent argument, and communication of results is sometimes unclear.
- 60%-69% - Struggling effort. Students have making some effort, but have misunderstanding of many concepts and are unable to put together a cogent argument. Communication of results is unclear.
- Below 60% - Students are not making a sufficient effort.