Stat 133
Group Project

Students working in groups of 3-5 are responsible for an initial plan, a final written report as well as either a 3 picture poster presentation or a 6 minute 3 slide oral presentation to the class. We will accomodate as many oral presentations as possible (around 15) but sign up soon as possible on the discussion board soon since space is limitted. Students doing the oral presentation will get a .25 out of 10 points extra credit.

Schedule:

1 Project info announced Monday, April 11

2 Initial plan due Monday, April 18 at 8pm

3 Presentation slides due Monday, May 2 at 8pm for those doing class presentation

4 Monday May 2 (RRR week) poster presentations (1-3pm in classroom)

5 Wednesday May 4 (RRR week) slide presentations (1-3pm in classroom)

6 Written report due Monday, May 9 at 8 pm

All materials should submitted to bCourses. Please include the names of all your group members.

Overview

Working in groups of 3-5, your task is to visualize (and perhaps statistically analyze) a large dataset. To give you a rough sense of what I mean by "large," I list some examples at the end of this handout, each of which consists of many thousands of observations. You should also identify a set of substantive questions that your visualizations/analyses will answer. The steps in completing the project will be very similar to what you have already done in the homework assignments, with the addition of deciding on a topic and obtaining the data yourself. These are

- Deciding on a topic, questions, and plan with your group
- Obtaining and cleaning the data (may involve web scraping)
- Creating graphical displays of the data
- Performing any appropriate statistical analyses of the data
- Writing up your results in a report
- Giving a short presentation (poster or slide)

Scope, or "How Much is Enough?"

I am leaving the precise definition of the project purposefully vague, and I encourage you to aim high. I'm expecting the scope of your project to be roughly equivalent to two full-length homework assignments. There are a variety of areas in which you can invest your time, and I expect some groups will put more effort into some and less into others. For example, if obtaining the data requires quite a bit of coding, then you can spend less time on the visualization/analysis and perhaps address only one or two questions. If the data are readily available, then you should plan on addressing multiple questions in depth. You may or may not wish to include formal statistical analysis in your project, depending on the knowledge and experience of your group members. If you do not do a formal analysis, you should put more effort into visualization. I will give you feedback based on your initial plan about whether your proposed project has the appropriate scope.

Initial Plan

Meet at least one more time with your group over the upcoming week, and prepare a one-page (or shorter) document with answers to the following questions.

- Choose a name for your group.
- What is your overall topic? What question(s) will you address?
- Where will you get the data? What is involved in obtaining it? What variables will you use?
- Is there any initial data processing must be done to put the data in a form that is suitable for visualization or analysis?
- What plots will you make of the data? Will you perform any analyses, and if so, what?
- List the responsibilities of each person in the group in completing the rest of the assignment.
- So that I may help you resolve any issues, what do you think will be the most difficult part of doing this project?

Note: It's ok if the questions you address change at some point after you turn in your initial plan, for example if you find something unexpected in the data. My main goal is that you start mapping out your plan early and make expectations clear to all group members.

Written Report

Write a report summarizing your findings. Be sure to include an introduction section motivating your visualizations/analyses, with a description of the substantive context and why it is interesting. You may wish to cite a few references. You may use whatever bibliographic style you like, as long as you're consistent. Put your code in an appendix and include comments. There is no page

requirement. Successful groups in previous years have averaged about 10-15 pages with figures and tables, but not counting appendices. I regret that I am not able to read drafts of the report in their entirety, but I am happy to look them over in office hours and give general feedback about structure. When you send the report, please do NOT include extra separate files such as data files or R script files. Incorporate any code you want me to see into the appendix of the report itself.

Presentation

Your group will present EITHER a five minute overview of the project or a 3 picture poster. You should prepare *exactly three slides* to turn in. My suggestion is that you prepare one slide with an overview of the project and questions, one describing the data, and one with a plot showing some results, but you may change this if another format seems like a better fit. It is up to you whether you wish to have one or multiple presenters. Be prepared for questions from me and/or your classmates. *Note: We will use one (Mac) laptop for all slide presentations. The pdf format is safest for transferring slides. Powerpoint files may not render as you expected.*

Grading

Grades will be assigned based on the following breakdown:

- 10% - Initial plan, based on completeness of answers
- 80% - Written report, based on
  - appropriateness of any visualizations/analyses (e.g. right plot to address the question)
  - adequate descriptions of all steps and of the plots/results
  - clarity and organization and reproducibility of the report (not grammar unless it interferes with my ability to understand your report – if in doubt get someone to proofread)
  - creativity and originality
- 10% - Presentation, based on clarity of description of what you did

All group members will be assigned the same grade. However, each group member will also complete an anonymous survey asking what percentage of effort was made by each person in the group. If a member receives consistently low percentages, he/she may receive a lower grade.

Common Problems to Avoid

Here are a few issues that have shown up in past projects. Avoid these!

- Not enough detail about the data. I, as the reader of your report, should be able to figure out exactly how you obtained and processed the data, well enough that I could recreate your steps if needed. Don't just give the overall website; tell me what file you used and how to get it. If you had to write code to get the data, describe what it does and then note where it is in the appendix.
- Plots that are hard to read, poorly labeled, or not to referred to appropriately in the text.
- Names of functions or actual code in the text. Put it in the appendix!
- Overly strong conclusions. Be careful about any claims you make based on your analysis. Consider whether there may be alternative explanations or problems making general conclusions based on the data you have.

Some Possible Topics and Data Sources

- Airline Flight Information
  `http://stat-computing.org/dataexpo/2009/` (also see datasets for other years)
- Crime
  `http://www.crimemapping.com`
- US Climate
  `http://cdiac.ornl.gov/epubs/ndp/ushcn/ushcn.html`
  `http://www.ncdc.noaa.gov/oa/ncdc.html`
- US Social Trends
  `http://www.norc.org/GSS+Website/`
- World Development
  `http://www.gapminder.org/`
- US Health and Disease
  `http://statecancerprofiles.cancer.gov/`
  `http://www.statehealthfacts.org/`
- US Unemployment
  `http://www.bls.gov/data/`
- California Traffic
  `http://traffic-counts.dot.ca.gov/`
- Stock Data
  `http://finance.yahoo.com/q/hp?s=GE` (substitute stock name; also see `R` package `quantmod`)
- Radiohead House of Cards Video
  `https://code.google.com/p/radiohead/`

- Baseball
  `http://www.baseball-databank.org/`
- Some interesting blogs about data visualization (for inspiration)
  `http://flowingdata.com/`
  `http://junkcharts.typepad.com/junk_charts/`
  `http://www.floatingsheep.org/`

Some websites explicitly prohibit web-scraping. This information is usually contained in a link labeled "Terms of Use" or "Terms of Service." While one can argue about whether any website has the legal right to prohibit this, you should err on the side of caution and avoid these sites.