CS5750: Methods of Visualisation and Exploratory Analysis

(some) Project Guidelines

Dr. Matteo Fontana
Dept. of Computer Science
Royal Holloway, University of London

The Projects

- As you know already… the project will be part of the evaluation! (20%)
- Solo projects.

Deliverable – Moodle Submission (21st of July) with

- Written report (.pdf, .docx, .html …)
- Relevant Code (.py files, python notebooks …)
- The data source (we will need to run the code)

My suggestion: use notebooks, it’s better

The Project

For the project you are expected to apply the techniques you have seen in the course to tackle a real-world problem.

The data should be:

- Multivariate
- "Very interesting" for you.
- Should come up with interesting and meaningful research questions that can be possibly solved using the methods presented during the course.
- Publicly available.

The Project - Evaluation

The projects are normally evaluated looking at:

- How meaningful your question is
- How well you structure your «process» to answer the research question
- How well you decide the methods to use during your «process», and how well you implement those.
- How compelling your report is.
- How tidy and well structured your code is (YES, you will be required to hand it in)

To deem a project as EXCELLENT (and to impress us… ;)) we expect you to go the extra mile… e.g.

- Nonstandard visuals
- Nonstandard methods (Dimensionality reduction, statistical tests… use material from other courses too!)
- …but don't get carried away too much! We still want you to use methods we teach in CS5750

The Report – Expected Outline

(This is actually useful also for learning how to develop a proper statistical analysis of a dataset, either for academic or professional purposes)

1. Problem statement – Why this problem? Why is this problem relevant? What are the open themes in the area of knowledge you are exploring? How is the data you're using able to solve this kind of questions (here a quick description of the dataset is useful, and some considerations in terms of data cleaning and wrangling, as well as a codebook for the dataset)?
2. Research questions – Starting from 1., what are the questions you're asking to your dataset? What kind of visualization methods do you want to use for these questions?
3. Analysis – What do your visualization show? What kind of insights can you gather from those? Are they reliable?
4. Conclusions – How you will proceed in your analysis to tackle the questions 2.? How will the insights gathered in 3. guide your analysis?

Some Suggestions on the Choice of the Dataset

Choose something…

- …that you are passionate about! Doing a project on a dataset that you find boring is… well, boring.
- …on which you have (or really would like to have) some background knowledge. As we have seen multiple times, having a good grasp of what's going on under the hood of your data is fundamental…
- …that triggers interesting questions! They not need to be necessarily real, but at least realistic (and justifiable…)

We will give you some possible sources where you can find interesting data. Don't feel forced to use one of those, but they are a good starting point

Some Socio-Economic Sources

This is a (VERY biased towards socioeconomic and remote sensing applications) list of sources to gather your data:

- https://ec.europa.eu/eurostat/home?
- https://www.ons.gov.uk/
- https://tfl.gov.uk/info-for/open-data-users/our-open-data
- https://data.jrc.ec.europa.eu/
- https://data.oecd.org/
- https://www.ecb.europa.eu/stats/html/index.en.html
- https://www.bankofengland.co.uk/statistics
- https://www.data.gov.uk/
- https://data.un.org/
- https://ukdataservice.ac.uk/

Some Biosciences Sources

If socioeconomic applications are not your thing, you can try also these repos, that have a more Biosciences flavour:

- https://biolincc.nhlbi.nih.gov/teaching
- http://featureselection.asu.edu/datasets.php
- https://mimic.physionet.org/
- https://higgi13425.github.io/medicaldata/

Other resources

And, more general purpose repo…

- http://mmds.org/
- https://www.kaggle.com/
- https://archive.ics.uci.edu/ml/datasets.php
- https://github.com/rfordatascience/tidytuesday

If you want something else, feel free to propose anything you think it could be interesting. (Just be careful about data ownership…)

Questions?