B

Papers

B.1 Paper 1

B.1.1 Task

Working individually and in an entirely reproducible way, please find a dataset of interest on Open Data Toronto – `https://open.toronto.ca` – and write a short paper telling a story about the data.

B.1.2 Guidance

- Find a dataset of interest on Open Data Toronto[1] and download it in a reproducible way using `opendatatoronto` (Gelfand, 2020).
- Create a folder with appropriate sub-folders, add it to GitHub, and then prepare a PDF using `R Markdown` with these sections (you are welcome to use this starter folder: `https://github.com/RohanAlexander/starter_folder`):
  - title,
  - author,
  - date,
  - abstract,
  - introduction,
  - data, and
  - references.
- In the data section thoroughly and precisely discuss the source of the data and the bias this brings (ethical, statistical, and otherwise). Comprehensively describe and summarize the data using text and at least one graph and one table. Graphs must be made in `ggplot2` (Wickham, 2016) and tables must be made using `knitr` (Xie, 2021) (with or without `kableExtra` (Zhu, 2020)). Graphs must show the actual data, or as close to it as possible, not summary statistics. Make sure to cross-reference graphs and tables.
- Add references by using a bib file. Be sure to reference R and any R packages you use, as well as the dataset. Check that you have referenced everything. Strong submissions will draw on related literature and would also reference

[1] `https://open.toronto.ca`

those. There are various options in R Markdown for references style; just pick one that you are used to.
- Go back and write an introduction. This should be two or three paragraphs. The last paragraph should set out the remainder of the paper.
- Add an abstract. This should be three or four sentences. If your abstract is longer than four sentences, then you need to think a lot about whether it is too long. It may be fine (there are always exceptions) but you should probably have a good reason. Your abstract must tell the reader your top-level finding. What is the one thing that we learn about the world because of your paper?
- Then add a descriptive title. 'Paper 1' is not descriptive and there should not be any sign this is a school paper.
- Add a link to your GitHub repo using a footnote.
- Check that your GitHub repo is well-organized, and add an informative README. Comment your code. Make sure that you have got at least one R script in there, in addition, to your R Markdown file.
- Pull this all together as a PDF and check that the paper is well-written and able to be understood by the average reader of, say, FiveThirtyEight. This means that you are allowed to use mathematical notation, but you must explain all of it in plain language. All statistical concepts and terminology must be explained. Your reader is someone with a university education, but not necessarily someone who understands what a p-value is.
- Check there is no evidence that this is a class assignment.
- Via Quercus, submit the PDF.

B.1.3 Checks

- Check you have not included any R code or raw R output in the final PDF.
- Check that although you will probably have most of your code in the R Markdown, make sure that you have at least one R script in the 'scripts' folder.
- Check there is thoroughly commented code that directly creates your PDF. Do not 'knit to html' and then save as a PDF. Do not 'knit to Word' and then save as a PDF
- Check that your graphs, tables, and text are extremely clear, and of comparable quality to those of FiveThirtyEight.
- Check that the date is updated.
- Check your entire workflow is entirely reproducible.
- Check for typos.

B.1.4 FAQ

- Can I use a dataset from Kaggle instead? No, because they have done the hard work for you.
- I cannot use code to download my dataset, can I just manually download it?

No, because your entire workflow needs to be reproducible. Please fix the download problem or pick a different dataset.

- How much should I write? Most students submit something in the two-to-six-page range, but it's really up to you. Be precise and thorough.
- My data is about apartment blocks/NBA/League of Legends so there's no ethical or bias aspect, what do I do? Please re-read the readings to better understand bias and ethics. If you really cannot think of something, then it might be worth picking a different dataset.
- Can I use Python? No. If you already know Python then it doesn't hurt to learn another language.
- Why do I need to cite R, when I don't need to cite Word? R is a free statistical programming language with academic origins so it's appropriate to acknowledge the work of others. It's also important for reproducibility.

B.1.5 Rubric

- Go/no-go #1: R is cited - [1 'Yes', 0 'No']
  - Both referred to in the main content and included in the reference list.
  - If not, no need to continue marking, just give paper 0 overall.
- Title - [2 'Exceptional', 1 'Yes', 0 'Poor or not done']
  - An informative title is included.
  - Tell the reader what your story is, don't waste their time.
  - Ideally tell them what happens at the end of the story.
  - 'Problem Set X' is not an informative title. There should be no evidence this is a school paper.
- Author, date, and repo - [2 'Yes', 0 'Poor or not done']
  - The author, date of submission, and a link to a GitHub repo are clearly included. (The later likely, but not necessarily, through a statement such as: 'Code and data supporting this analysis is available at: LINK').
- Abstract - [4 'Exceptional', 3 'Great', 2 'Fine', 1 'Gets job done', 0 'Poor or not done']
  - An abstract is included and appropriately pitched to a general audience.
  - The abstract answers: 1) what was done, 2) what was found, and 3) why this matters (all at a high level).
- Introduction - [4 'Exceptional', 3 'Great', 2 'Fine', 1 'Gets job done', 0 'Poor or not done']
  - The introduction is self-contained and tells a reader everything they need to know, including putting it into a broader context.
  - Your introduction should provide a bit of broader context to motivate the reader, as well as providing a bit more detail about what you're interested in, what you did, what you found, why it's important, etc.
  - A reader should be able to read only your introduction and have a good idea about the research that you carried out and what you found.
  - It would be rare that you would have tables or figures in your introduc-

    tion (again there are always exceptions but think deeply about whether yours is one).
  - It must outline the structure of the paper.
  - For instance (and this is just a rough guide) an introduction for a 10 page paper, should probably be about 3 or 4 paragraphs, or 10 per cent, but it depends on specifics.
- Data - [10 'Exceptional', 8 'Great', 6 'Good', 4 'Some issues', 2 'Many issues', 0 'Poor or not done']
  - When you discuss the dataset (in the data section) you should make sure to discuss at least:

    1) The source of the data.
    2) The methodology and approach that is used to collect and process the data.
    3) The population, the frame, and the sample (as appropriate).
    4) Information about how respondents were found. What happened to non-response?
    5) What are the key features, strengths, and weaknesses about the source generally.

  - You should thoroughly discuss the variables in the dataset that you use. Are there any that are very similar that you nonetheless don't use? Did you construct any variables by combining various ones?
  - What do the data look like?
  - Plot the actual data that you're using (or as close as you can get to it).
  - Discuss these plots and the other features of these data.
    - These are just some of the issues strong submissions will consider. Show off your knowledge. If this becomes too detailed, then you should push some of this to footnotes or an appendix.
    - 'Exceptional' means that when I read your submission I learn something about the dataset that I don't learn from any other submission (within a reasonable measure of course).
- Numbering - [2 'Yes', 0 'Poor or not done']
  - All figures, tables, equations, etc are numbered and referred to in the text.
- Proofreading - [2 'Yes', 0 'Poor or not done']
  - All aspects of submission are free of noticeable typos.
- Graphs/tables/etc - [4 'Exceptional', 3 'Great', 2 'Fine', 1 'Gets job done', 0 'Poor or not done']
  - You must include graphs and tables in your paper and they must be to a high standard.
  - They must be well formatted and camera-ready. They should be clear and digestible.
  - They must: 1) serve a clear purpose; 2) be fully self-contained through appropriate use of labels/explanations, etc; and 3) appropriately sized and colored (or appropriate significant figures in the case of stats).

- References - [4 ‘Perfect’, 3 ‘One minor issue’, 0 ‘Poor or not done’]
  - All data/software/literature/etc are appropriately noted and cited.
  - You must cite the software and software packages that you use.
  - You must cite the datasets that you use.
  - You must cite literature that you refer to (and you should refer to literature).
  - If you take a small chunk of code from Stack Overflow then add the page in a comment next to the code.
  - If you take a large chunk of code then cite it fully.
  - 3 means one minor issue. More than one minor issue receives 0.
- Reproducibility - [4 ‘Exceptional’, 3 ‘Great’, 2 ‘Fine’, 1 ‘Gets job done’, 0 ‘Poor or not done’]
  - The paper and analysis must be fully reproducible.
  - A detailed README is included.
  - All code should be thoroughly documented.
  - An R project is used. Do not use `setwd()`.
  - The code must appropriately read data, prepare it, create plots, conduct analysis, and generate documents. Seeds are used where needed.
  - Code must have a preamble etc.
  - You must appropriately document your scripts such that someone coming in could follow them.
  - Your repo must be thoroughly organized.
- General excellence - [3 ‘Exceptional’, 2 ‘Wow’, 1 ‘Huh, that’s interesting’, 0 ‘None’]
  - There are always students that excel in a way that is not anticipated in the rubric. This item accounts for that.

B.1.6 Previous examples

Some examples of papers that well in the past include those by: Amy Farrow[2], Morgaine Westin[3], and Rachel Lam[4].

[2] `inputs/pdfs/Mandatory_minimums-Amy_Farrow.pdf`

[3] `inputs/pdfs/Mandatory_minimums-Morgaine_Westin.pdf`

[4] `inputs/pdfs/Mandatory_minimums-Rachel_Lam.pdf`