Exploring Disparate Data: Part 2

Due November 12th

Overview

This is **Part 2** of the project, in which you'll work with the data that you cleaned in Part 1!

There are several data sets that you cleaned in Part 1. Using at least **three** of these data sets, find something interesting!

Possible ideas:

- Is the intensity of hurricanes related to the sea ice extent?
- Is being aware of climate change associated with being less happy?

Note: You do not have to have fully completed Part 1 in order to do this part. You will need to use the cleaned version of *three* of the data in the other assignments, but you can focus on the important data first. In addition, using the `covid_2020` only for the continents information or using the `pop_data` do *not* count towards the three data sets.

Once you're a member of a group, you will see "Assignment 3 Report" in the Dropbox tool on MyLS. If you do not see this tool, it likely means that you are not yet in a group. Even if you're going solo, you must be a member of a group.

Collaborating across 2 files with 5 people can be hard. Here are som suggestions for collaboration with group members:

- Store the Rmds in a shared OneDrive folder, along with the data and the `tests` folder (everything that comes in the zip file).
- Create an account on posit.cloud, and share the login information amongst yourselves. This makes it easy to avoid issues with package installation and working directory issues, and can be accessed from any device. Uploading data may be a little tricky, unfortunately.
- Use *git*. This is by far the hardest option to learn, but it's also by far the best option for those seeking a career in data science.

Tubric

The word "Tubric" comes from "Two-entry Rubric". For convenience and brevity, I only describe a level 2 and a level 4 rubric item. Level 0 means it's missing/incomplete, level 1 is inadequate, and level 3 means that it was somewhere between the level 2 and level 4 items. In some cases, I include an extra level 0 condition, which means that there are some cases where you'll get an automatic 0 in that category if you meet that condition.

The grader will not read content past page 5 (not including a cover page, if you choose to include that).

There should be a table of summary statistics computed using `dplyr` and three plots made using `ggplot2`.

- **Miscellaneous**
  1. Formatting
     - 0: Any of the cleaned data sets are shown directly in the PDF. You can include subsets of the data to demonstrate something specific about the structure, but do not use `head()`, `glimpse()`, or simply display the data at any point in the PDF.

    - 2: The PDF includes the text and the plots, possibly with some unnecessary code output or odd plot placement.
    - 4: The reader makes it through the whole document without noticing the formatting because it's so natural and error-free.
  2. Structure
    - 2: The document follows the template, possibly with some odd changes.
    - 4: The structure of the document makes navigation and reading easy, including appropriate headers and plots in reasonable places.
  3. Writing Style - you can write with any level of professionalism/casualness and I am lenient with grammar, but the ideas should flow logically from one to another.
    - 2: The written portions of the document have acceptable grammar and provide the correct information.
    - 4: The writing is clear and concise throughout, with appropriate segues between sections. Each idea flows clearly into the next.
- **Abstract**
  4. Abstract
    - 2: There is an abstract that mostly puts the report in context and provides a short summary of the results.
    - 4: The abstract is short and clear, and it accurately represents the entirety of the report.
- **Introduction**
  5. Context - first paragraph or two should place your analysis in context.
    - 2: Paragraph describes the general context.
    - 4: Paragraph concisely describes the context that prepares the reader for the report without the need to reference anything in the report.
  6. Introdutory paragraphs - the next paragraph should provide a general idea of the data and techniques that the reader will encounter in your report.
    - 2: Introductory paragraph mentions most of the techniques used.
    - 4: The introductory paragraphs provide a concise overview of what the report will contain, without repeating any part of the paper verbatim.
  7. The goal of the analysis is described.
    - 2: The goals are adequately explained. The goals can clearly be attained using the data provided.
    - 4: The goals are clearly defined. The goals can be attained using the data provided and limitations are briefly considered.
- **Data Description** - Describe each of the data sets in sufficient (but not excessive) detail.
  8. Data set 1
    - 2: The description describes the columns of the data that are relevant to the analysis.
    - 4: The description describes the information in the data concisely, possibly grouping similar columns into groups for a more concise description.
  9. Data set 2: see data set 1.
  10. Data set 3: see data set 1.
  11. Data cleaning process
    - 2: For each data set, the cleaning process is described.
    - 4: For each data set, the cleaning process is described concisely, but with enough precision that a reader could reproduce their cleaned data without seeing their code.
- **Exploration**
  12. Process description and justification
    - 2: The text describes the how and why of the data exploration
    - 4: The text is concise and fully explains and justifies the process of the exploration. No irrelevant content is included.
  13. Summary statistics table insights
    - 0: The summary statistics are only based on *one* of the data sets provided.
    - 2: The summary statistics provide insights that are relevant to the report.
    - 4: The summary statistics provide clear insights from a quick glance.

14. Summary statistics table formatting
    - 2: The table is formatted appropriately.
    - 4: The table is exceptionally well formatted, using specialized R packages for extra fanciness.
15. Plot 1 insights
    - 2: The plot shows something interesting.
    - 4: The insights of the plot are clear from the data display. Minimal text is required to orient the reader, and otherwise the insight stands out.
16. Plot 2 insights: see plot 1 insights. In addition, this plot *must* have meaningful facets.
    - 0: The second plot does not have meaningful facets.
17. Plot 3 insights: see plot 1 insights. In addition, this plot *must* draw data from two separate data sets and must be a different plot type from both of the first two (if the first two were scatterplots, this one cannot also be a scatterplot; if you hav a violin and a box plot, this one must be something other than a violin or box plot).
    - 0: The third plot is based on only *one* of the data sets and is the same type as the others.
18. Visualization quality
    - 2: The visualizations contain the relevant information, with labels and titles.
    - 4: The visualizations contain only the relevant information (with no irrelevant information), and labels and titles are used effectively.
19. Visualization discussion - discuss the insights in the context of the data (not the general context). For example, "these data demonstrate a moderate correlation between knowledge of climate change and belief that it is happening, as measured by..." rather than "knowing about climate change means you believe that it is happening"
    - 2: For each plot, the insights are discussed within the context of the

- **Conclusion**

20. One-Paragraph Summary
    - 2: Paragraph summarises the report's findings.
    - 4: The paragraph concisely summarises the reports findings and draws reasonable conclusions about the broader context.
21. Future work
    - 2: Ideas for modelling/incorporating other data are discussed.
    - 4: The next steps of the model are concisely detailed, and presented in a way that excites the reader about what you might do next.
22. Limitations
    - 2: The limitations of the data are discussed, with reference to the insights reported.
    - 4: The reader trusts the insights more because of the honesty in which the limitations are discussed.

- **References** (Note that links to websites are fine, but make sure that the links are relevant.)

23. Insight-Related
    - 2: Less than 5 external references are used to support your insights.
    - 4: There are at least 5 references from different sources used to provide further evidence for the stated insights. The formatting makes it easy to see which reference applies to which insight. See template for example.
24. Other
    - 2: There are at approximately 5 references that direct the reader to further information about the data or the techniques used.
    - 4: There are more than 5 highly relevant references that direct the reader to further information about the data and the techniques used. There are no irrelevant references, or references to help files/general documentation.

- **Bonus Marks**
  - Up to 4 marks for an especially interesting use of the provided data.
  - Up to 2 marks for using data beyond what was provided in an interesting way.
  - Up to 4 marks for proving that you used git for collaboration, with all group members contributng via commits (must be a group size of 3 or more to recieve these marks; email me a link to the repo whn you submit).