LIS 545 / Mayernik **World Happiness Report** **Joseph F. Saultz**
WI 2024

World Happiness Report: https://www.kaggle.com/datasets/unsdsn/world-happiness
Data and Metadata Profile

Data

The data are happiness scores and rankings by country (approximately 160 ranked) based on the Gallup World Poll surveys, for the years 2015 through 2019 (total 5). The surveys are based on answers to the main life evaluation question being asked in the poll – the Cantril Ladder – with the best possible life being a “10” to the worst possible life being a “0” and then using 6 factors (levels of Economy (GDP per capita), Family, Health (Life Expectancy), Freedom, Trust (Government Corruption) and Generosity) as estimates that contribute or correlate to making life evaluations higher (i.e. happier) in each country. These results are then compared to a hypothetical country benchmark, Dystopia, with its values equal to the world’s lowest national averages for each of the 6 factors, as a residual or *unexplained component*. The use of the Dystopia benchmark allows that each of the 6 key variables are either a positive or “0” measure.

Key stakeholders in the data are the participants from the countries surveyed, Gallup, the UN Sustainable Development Solutions Network, the Oxford Wellbeing Research Center, the World Happiness Report’s (WHRs) Editorial Board (John F. Helliwell, Richard Layard, and Jeffrey Sachs), and Abigail Larion (the data administrator from Kaggle).

There are a total of 5 data files, in .CSV (comma-separated values) format, that uses commas to separate values and stores tabular data in numbers and plain text. The columns contain headers limited to Country, Region, Happiness Rank, Happiness Score, Standard Error, Economy (GDP per Capita), Family, Health (Life Expectancy), Freedom, Trust (Government Corruption), Generosity, and Dystopia Residual. The CSV format has wide interoperability and is supported by almost all spreadsheet and database management systems (e.g., Apache OpenOffice, Apple Numbers, Google Sheets, LibreOffice, and Microsoft Excel).

The data is in the Public Domain or Open Source, with original crediting to the World Happiness Report 2015-2019, New York: Sustainable Development Solutions Network.

Metadata

There is tabular metadata included in the .CSV files, themselves, containing the characteristics of the contents (i.e., Country, Region, Happiness Rank, etc.), which is likely based on the CSV on the Web (CSVW) Metadata Standard from W3C [link]. There is also unstructured, free-form metadata in the *About Dataset* fields on Kaggle with sub-sections for *Context*, *Content*, *Inspiration*, *What is Dystopia?*, *What are the residuals?*, and *What do the columns succeeding the Happiness Score(like Family, Generosity, etc.) describe?* [sic]. Other prescribed metadata fields on Kaggle are largely left empty besides *Collaborators* (Sustainable Development Solutions Network (Owner) and Abigail Larion (Admin)). There are also additional Tags including Arts and Entertainment, News, Social Science, Religion and Belief Systems, and Economics.

Data Enrichment

When comparing data from World Happiness Report [link], the .CSV files appear to be different. If data has been processed (simplified, for example, for clarity) and this is a derived data product, the workflow should be documented and links to the original dataset provided. The data could be enriched by including unedited .CSV files from the original publisher or at least links to these .CSV files as they contain additional statistics and measures as well as cleaner explanations for categories (for example, what is meant by Family? Freedom?) and additional tables. Whenever possible, keep raw data raw or, at least, keep it separate.

The file names are indicated by year, but it is unclear what the year signifies (i.e., is this when the survey data was collected or is this the time when the data was analyzed? The file names should be more descriptive and changing the file name to better clarify this would aid in discovery. Further, additional notes or a separate metadata file could help to clarify this – perhaps one that even includes overall organization of the dataset.

Provide a citation and document provenance for the dataset. The metadata could be enriched by completing all available fields in the Kaggle repository and including extra kernels such as proper citations, documentation of provenance, and acknowledgements (which would help make the data more "discoverable"). It would also be recommended to include links to additional appendices, available data collections summaries, ISBN of the original report, and links to Frequently Asked Questions (FAQs) from the original publisher site. With the inclusion of the additional tables from the original publisher as well as the links to the comprehensive report, it would contribute to the ability to use the data in new ways. This data/metadata enhancement would provide for better interpretation of the data, which would deepen the knowledge gained by the reader for further use.

Publications

On Kaggle, when looking at the *Code* tab, there are approximately 1076 products that use World Happiness Report from this dataset (this is also linked in the *Related Notebooks* selection near the bottom of the homepage). Because the dataset is part of the public domain, performing a general web search outside of Kaggle using a search engine (e.g., Google Search) presents challenges in understanding provenance, especially because the Collaborator / Owner, Sustainable Development Solutions Network, provides the data in other repositories.

Chosen Repository: ResearchWorks Archive – University of Washington [link]
Repository Profile (n.b. under review)

I primarily chose this repository because I am a University of Washington (UW) graduate student, so familiarity with this repository would provide a benefit to me as I continue with my graduate studies. Also, for faculty or peers in the UW community who would need access to my data files, membership would be available for them, too. Additionally, the repository accepts a wide array of content and formats, including data sets like mine that include standard office documents.

ResearchWorks at the University of Washington does have a defined collection scope that includes faculty, academic staff, and students at UW. The work must be scholarly or research-oriented and must be submitted with the intent that it be made available on a permanent basis. Further, the author/owner should be willing to grant UW the right to preserve and distribute the work via ResearchWorks Archive (RWA) and, further, if the work is part of a series, other works in that series should also be contributed so that RWA can offer as full a set as possible.

The repository is open for submissions and will accept a wide array of data content including: data sets, published articles where copyright allows, working papers and technical reports, manuscripts, and any other form of research output that can be technically loaded to the repository. These submission types include all the parameters that qualify my dataset. There appears to be no limits towards data types, domain, or format type, though they do indicate a caveat that, "the proprietary or executable nature of some file formats, however, may make it impossible to guarantee persistent access to all deposited works as digital technologies evolve." A list of preferred file format may be found at this [link]. The RWA Collection Policy can be found online by visiting this [link].

The repository provides that if/when you are adding to ResearchWorks for the first time that you contact them via email at: libanswers@uw.edu. It does not provide guidance on *what to include* in your email, including any accompanying data or metadata beyond your dataset for submission (such as in a submission information package). To that end, if you are logged into the ResearchWorks Homepage [link], you may click on Submissions under the My Account section and review existing unfinished submissions by Title/Collection/Submitter or choose to *Start Another Submission*, which asks you to Choose a Collection before continuing (which can be a bit dismaying if you are just getting started). You also have the option to complete a *Mediated Deposit Form* (limited to one file, 100MB or smaller, per submission). This deposit form does solicit a small amount of information from you e.g., depositor's name, email address, Author/Creator's Name, Title of Research You are Depositing, and the option to enter the Internet location where the file can be found and downloaded by a ResearchWorks administrator for submission to the repository. The presumption is that the information you provide will contribute to the initial metadata. Regarding metadata specifically, on the *Collection Policy* website, it is stated that "… some of this metadata will be automatically generated by the software used by ResearchWorks Archive; other pieces, including author, title, and date, will be provided by the depositor." With that said, under *Quality Control*, it does state that "… quality control will be handled at the community or collection level …" and "If a community/collection is not administered by a specific organizational unit, a small group of staff within the Libraries will take responsibility for vetting the submissions." The metadata standard name/scheme is Dublin Core/DDC per re3data.org [link].

By default, items submitted in RWA have no access restrictions, which is to say that they are openly available via the World Wide Web. There is, however, the ability to create access restrictions on either an item (3 tiers: level 1 – restricted to UW faculty, staff, and students with a UWNetID and password only, level 2 – restricted to a specific group defined and maintained within RWA, or level 3 – embargoed or closed access for a specific period of time) or a collection (matching levels 1 and 2 for item restrictions) by either the individual depositor or the RWA Community. You may complete a registration on the website which will be authenticated using your UWNetID and password and [this] would allow you to bypass level 1 restrictions. When downloading, you are not surveyed for explanation for your purpose in the download, including any guidance for dissemination methods for how to share the resource or what to do with the downloaded item. Downloads, themselves, are direct downloads (for the individual data, itself, for example, a dissertation as a standalone document) completed either by clicking the "View/Open" option on the left-anchored menu or when viewing the full item record and clicking the "View/Open" option in the "Files" section below the metadata. After sampling several dozen downloads using different document types (e.g., .pdf, .docx,, .txt, .zip, etc.), it appears that the data/item is simply the raw data provided by the submitter. On the item's respective homepage, which one can identify through URI, you have the option to view the *full item record*, which has an extensive metadata listing using Dublin Core schema (e.g., dc.contributor.author, et. al.,). It does not readily seem obvious if any statistics are provided (for example, downloads, citations, etc.) to the contributor from the RWA.

Data Citation and Preservation

For citation style, I recommend American Psychological Association (APA) as the data set relates to both Psychology and Sciences. I do not recommend establishing a DOI for the World Happiness Report simplified data set as it is a derivative data set and the original is from the same principal author. Moreover, the original data set is linked to the author's website where they own the domain, making the data very discoverable, and better takes into account long-term preservation potential. The data file formats are .CSV, which does not require proprietary software to save, open, or edit and can even be opened with a simple text editor. There is little risk of this format becoming obsolete in the near future.

Copyright and Human Subject Considerations

The data set is part of the public domain, so the authors and publishers have waived their rights to the works worldwide under copyright law. By being part of the public domain, the content of the work is able to be used and reused without concerns for liability. Although the content of the data set is a "measure" of human happiness country-by-country, the data does not contain personally identifiable information about specific people. There is, however, some ethical risk in misrepresentation of the data to be predictive instead of descriptive (i.e., to use the "scores" to be prejudicial), though this is not the "fault" of the data or the authors, but the understanding of the relationship between correlation and causality.