Data 221 - Spring 2022 Midterm group project

Intro Data Science II - Trimble

Due: Friday, Feb 17, 2023, 11:59pm

Midterm group project - Modeling

1. For this assignment, you are asked to select a dataset and explore it using some of the tools we talked about this quarter. You can choose groups of 2 or 3; you are encouraged to find a dataset that is interesting, even if you are only able to explain a tiny part of it.

   - You should investigate and briefly explain the origin and the meaning of the dataset. When, where is the data from and why might we care about it?
   - A footnote or an entry in the bibliography must indicate the origin of the dataset in sufficient detail to permit it to be found, with dataset name, author, revision number in addition to the URL where it can be found today.
   - You should report something about what other people have found looking at this data or data of this sort. This requires research outside the dataset, and requires a footnote or bibliographic entry.
   - You should apply a data modeling / data reduction technique from the class to summarize / compress / understand / predict something about the dataset. Some kind of evaluation of your model is appropriate.
   - Present summaries of the data and your model of the data as visualizations. Visualizations should be appropriate and informative. Can you tell which terms in the model are most important?
   - Visualizations should be free of correctable flaws; everything that needs a label must be labeled, fonts must be no smaller than half the font size of the text in the report; included images must not be grainy or illegible. Best practice is to write a caption for each figure that succinctly explains the figure.
   - Most of your grade will be on the quality of your data reporting. Explaining what is in a dataset is difficult, and doing it in a way that is easy to read is even more so. Does the dataset have obvious utility or immediately raise questions? Can you answer at least some of those for someone unfamiliar with the dataset?
   - Page limit: 8 pages including visualizations, not including bibliography or submitted code. Some people could do a good job in 5 pages.
   - You do not have to use any specific tools to produce the visualization, but you must submit the substantial code you used to perform modeling and generate visualizations.
   - Group projects require a statement as to who did what, though I assure you it will not be scrutinized the way the visualization and bibliography will be.

   Data producers and publishers have their reasons for collecting and publishing; the datasets at UCIrvine and kaggle are many, though they have probably been well scrutinized.

   - Awesome public datasets (github awesomedata)
     `https://github.com/awesomedata/awesome-public-datasets`
   - The Guttmacher institute publishes data related to women's health, historical trends and geographical trends in fertility. `https://www.guttmacher.org/public-use-datasets`
   - Centers for Medicare and Medicaid Open Payments publishes a database of payments that pharmacutical manufacturers make to prescribers in the US, (dubbed by Propublica "Dollars for Docs") `https://openpaymentsdata.cms.gov/`
   - Chicago City Data portal. Municipal datasets of various sorts; city finances, communication, and enforcement. `https://data.cityofchicago.org/`

- Google Dataset search: `https://datasetsearch.research.google.com/`
- `https://blog.google/products/search/discovering-millions-datasets-web/`
- CDC: `https://data.cdc.gov/browse`
- 500 cities: `https://www.cdc.gov/500cities/index.htm`
- UN: `http://data.un.org/`
- Kaggle: `https://www.kaggle.com/datasets`
- AWS: `https://registry.opendata.aws/`
- FEC: `https://www.fec.gov/`
- FiveThirtyEight: `https://github.com/fivethirtyeight/data`