STAT 5000 Final Project Dataset Sources

STAT 5000

Statistical Methods and Applications I

How Do I Find a Good Dataset?

Please read through this and ask some of these questions when looking for a dataset to use for the final project.

- **Description**: The website contains a story on what the dataset is about, how it was collected, and any details that are pertinent to utilizing it.
- **Attributes**: The dataset contains information on what each variable is, what it means, and is easy to understand. Sometimes people will list attributes in a manner that isn't easy to understand. Make sure you understand what each variable is and how it contributes to the overall dataset.
- **Size of the dataset**: Make sure the dataset is big enough to perform statistical analyses. This is determined on a case by case basis but we want to have a big dataset when implementing different computational methods.
- **Bias**: Make sure the dataset isn't collected in a biased manner. We want to have equal representation in the dataset. If we have a labeled dataset, make sure that it isn't skewed and all categories have a fair amount of data. Sometimes visualizing the dataset using a histogram can help identify unfavorable datasets.
- **Standardization**: Ask if the dataset is already standardized. If not, make sure to standardize the set if needed.
- **Missing data**: See if the dataset has any missing entries. If you do have missing data, you can still use the set, but you need to find a way to address those missing entries. Sometimes we can delete the entire row of the missing entry or there are methods in replacing the missing entries.
- **Age**: How old is the dataset? Is it too old to use? Is the dataset recent enough to utilize?
- **Pre- vs Post-processed datasets**: Sometimes the datasets will be processed so it is important to know where the dataset is in terms of processing and cleaning. You can use a pre-processed data set, but you will have to clean it yourself.

Where can I find datasets?

A lot of the listed websites are dataset repositories meaning people can post any datasets for others to use. Please look through the list of datasets that others have posted and select the appropriate set. I suggest using a Google dataset search if you have a topic in mind. If not, use any of the other listed sources below to look through the datasets they have.

- **Google Dataset Search**: https://datasetsearch.research.google.com/ Google dataset search will allow you to search for datasets. If you are interested in wine datasets, search "Wine" and look through the results.
- **Kaggle**: https://www.kaggle.com/datasets Kaggle contains datasets in topics such as arts and entertainment, biology, social science, automobiles and vehicles, investing, and social networks.

- **UCI Machine Learning Repository**: https://archive.ics.uci.edu/ UCI ML Repo contains datasets mainly used for machine learning, but it can also be used for other purposes.
- **NASA**: https://data.nasa.gov/browse Earth science and space data.
- **Data.gov**: https://catalog.data.gov/dataset Datasets from various government institutes.
- **NIST**: https://data.nist.gov/sdp/#/ Science datasets from the National Institute of Standards and Technology .
- **NOAA**: https://psl.noaa.gov/data/index.html Atmospheric data.
- **The World Bank**: https://datacatalog.worldbank.org/home International data
- `R` **packages**: many packages in `R` include datasets, and a very helpful person compiled a huge list of these here: https://vincentarelbundock.github.io/Rdatasets/datasets.html