This note is for anyone looking for a topic for their graduate independent project. I have compiled the following references to just a few of the many possible data sets you can use for your project. You are free to choose any suitable public data set for your project.

If you have any questions as you formulate your project, please reach out to me. I am happy to discuss your ideas.

Text Mining Datasets

You can find an annotated list of open source data mining datasets here.

Datasets for Recommender Systems

The Recommender Systems and Personalized Datasets repository from Julian McAuley, of UCSD, contains several dozen datasets for recommenders, ranging from books to social media, entertainment and commerce.

Networks and Social Media Datasets

You can find lists of network, including social media, datasets in the Stanford Large Network Dataset Collection (SNAP).

The Network Data Repository is another large repository containing hundreds of graph datasets covering many subjects from power grids, ecology, transportation, chemical interactions to social graphs.

Entity Resolution

Entity resolution is fundamental and on-going data mining problem. You can find some interesting benchmark data sets from the University of Leipzig database group: https://dbs.uni-leipzig.de/research/projects/object_matching/benchmark_datasets_for_entity_resolution

Human Rights Data

The following list of data sources was recommended by Gloria Ayee of the Harvard University Department of Government.

- Urban Institute maintains a large collection of datasets focused on civil rights and economic development: https://datacatalog.urban.org/search/type/dataset
- European Values Survey: https://europeanvaluesstudy.eu/methodology-data-documentation/
- Harvard Law School International Relations and Human Rights Data Compilation: https://hls.harvard.edu/library/research/find-a-database/international-relations-human-rights-data/
- Human Rights Protection Scores: https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/TADPGE
- The All Minorities at Risk (AMAR) Project: https://cidcm.umd.edu/research/all-minorities-risk-project
- World Values Survey: http://www.worldvaluessurvey.org/wvs.jsp

Global Economic Data

A vast quantity of economic data, both from specific countries and globally, are available for download. Many interesting questions can be explored with these data. These data can be combined with other types of data to address complex real-world problems. A few of the many sources include:

https://www.federalreserve.gov/datadownload/

https://www.bls.gov/bls/data_finder.htm

https://www.gapminder.org/

https://data.worldbank.org/topic/economy-and-growth

https://www.imf.org/en/Data

ASA Bi-Annual Data Exposition

Bi-annually the American Statistical Association (ASA) sponsors a data competition. The data sets selected are usually complex and rich.

As most everyone knows, airline travel can have considerable uncertainty. The **Airline Travel Data Set** was one of the ASA Bi-Annual Data Exposition choices. These data are large and complex and can be applied to several problems: http://stat-computing.org/dataexpo/2009/

Covid 19

The ongoing pandemic has lead to the collection of many complex and challenging data sets. The Johns Hopkins repository has tried to consolidate the most important data set. This site contains links to many other US and global data sets: https://github.com/CSSEGISandData/COVID-19

Kagel Data Sets

Kagel has a large and rapidly growing number of data sets. Many of the Kagel data sets are specialized and might not be of interest for a capstone project. However, there are many interesting data sets.

An example of an interesting Kagel data set is the **Analysis of real and fake news**, a topical problem. This Kagel data set can be used to investigate models which can identify fake news vs. real news: https://www.kaggle.com/clmentbisaillon/fake-and-real-news-dataset