This project uses four total datasets. However, two of those four are too large to be uploaded to GitHub. Thus, only two are uploaded here. Here are details about all four if you wish to delve further into the raw data:

1. United States Stomach Cancer Mortality Rates by County, Race, and Ethnicity 2000-2019
   * Source: Global Health Data Exchange; <https://ghdx.healthdata.org/record/ihme-data/united-states-stomach-cancer-mortality-by-county-race-ethnicity-2000-2019>
   * Format: 2 csv files (1 for male mortality & 1 for female mortality) that we combine
   * Size: 54 MB each
   * Total size by records and fields: (808920, 19)
2. SDOH data by U.S. county
   * Source: Agency for Healthcare Research and Quality; <https://www.ahrq.gov/sdoh/data-analytics/sdoh-data.html#download>
   * Format: xlsx file but only using the “Data” sheet, so mostly can be treated as csv
   * Size: 13.4 MB
   * Records: (41051, 327)
3. FIPS to ZCTA Crosswalk
   * Source: DataWorld Dataset that is a well-formatted version of US Census Data; <https://data.world/nrippner/fips-to-zip-code-crosswalk> and <https://www.census.gov/programs-surveys/geography/technical-documentation/records-layout/2010-zcta-record-layout.html>
   * Format: comma-delimited txt file
   * Size: 3 MB
   * Records: (44139, 8)
4. Detailed health measures & metrics by county
   * Source: CDC PLACES Local Data for Better Health; <https://data.cdc.gov/500-Cities-Places/PLACES-Local-Data-for-Better-Health-ZCTA-Data-2023/qnzd-25i4/about_data>
   * Format: csv
   * Size: 250.6 MB
   * Records: (1169992, 19)

The combined dataset (*prepped\_dataset.csv*), which is obtained by running through the notebook *1\_Data\_Preparation.ipynb*, is unfortunately also too large to be uploaded to GitHub. To obtain it, you may download all aforementioned 4 datasets, save them in one folder, point to that file path at the beginning of *1\_Data\_Preparation.ipynb*, and run through it.