## Problem Statement:

The goal of this project is to generate a model to predict country-level maternal mortality rates. Specifically, the goal is to explore the social determinants of maternal mortality to uncover inconsistencies in national rankings in human development as compared to maternal mortality.

Maternal mortality ratios measure all the woman who died during pregnancy or within one year of pregnancy per 100,000 people. Importantly, maternal mortality rates indicate many factors besides a mothers’ health status as it is strongly correlated with social determinants of health including income, education and locality. Regional and sub-regional analysis of maternal mortality has indicated high variability in mortality based on race and income level. It is through this analysis that inequities in intra-country health outcomes as well as global inconsistencies can be found. For example, the United States ranked 55th in the World Health Organization’s latest maternal mortality ranking, and if you limited the list to the ten wealthiest countries in the world, the United States would rank last, despite being the largest economy (Belluz, 2020). Inconsistencies in country ranking among indicators that are highly correlated, such as maternal mortality and gross national income (GNI), can indicate where sub-national analysis needs to undertaken.

## Data Sources:

1. World Health Organization:
   * [The Global Health Observatory Data Repository](https://apps.who.int/gho/data/node.main.15?lang=en) will be scraped for maternal mortality ratios by country for years 2000-2017.
2. World Bank:
   * County-level data will be scraped from the World Bank Data Catalog for GNI, growth, under-five mortality, education attainment and other social determinants of health.
3. United Nations Development Program:
   * The UNDP Human Development Reports will be utilized to scrape a multidimensional index of quality of life such as the Human Development Index.
4. Freedom House:
   * 2019 [Freedom in the World Ranking](https://freedomhouse.org/countries/freedom-world/scores) aggregating the political rights and civil liberties score.
5. Sample of indicators to include:

|  |  |
| --- | --- |
| **Indicator** | **Source** |
| Maternal Mortality Ratio (per 100,000) | World Health Organization |
| GNI | World Bank Open Data |
| Population growth (annual %) | World Bank Open Data |
| Educational attainment (male/female) | World Bank Open Data |
| Mortality Rate, infant (per 1,000 live births) | World Bank Open Data |
| GINI Index | World Bank Open Data |
| Human Development Index | UNDP |
| Freedom in the World Score | Freedom House |

## Methods:

Webscraping will be the primary tool for obtaining data from WHO, the World Bank and UNDP databases. Other datasets may be manually imported as needed for variables that cannot be scraped. Due to the differences in sources there will likely be a larger data wrangling and cleaning portion to this project. The unit of analysis will be country-year but currently all of the sources to be used in this project treat each country as an observation, and year as a variable. During the data cleaning portion of this project, transformations and mutations will be used to change the unit of analysis from country to country-year. Additionally, there is significant cleaning that will need to be undertaken as many data source list confidence intervals within the same vector as the ratio or variable needed. Non-integer characters will need to be removed during cleaning. The k-nearest neighbor method will be used to impute any missing data.

Additionally, data visualization will be a critical component for not only relaying patterns in the data but also discovering patterns in the data. Various facet wraps will be used to analyze trends by region, dominant religion and government regime type. Two-way graphs will be utilized to test correlation and outliers among variables in the cleaned dataset.

Finally, linear regressions, regression trees and random forest models will be run on training data to identify the best model. Once the models are trained, should they need require training, the model with the smallest root-mean-squared-error will be chosen to minimize the prediction errors.

## Success:

By design, publicly accessible data, particularly from highly viewable sources such as the World Bank and World Health Organization, is likely to have been scoured and reanalyzed several times over. While I hope to introduce new detail into the analysis of maternal mortality, I am also aware of how thoroughly this indicator has been explored and thus, will not measure success by the “uniqueness” of my findings. Instead, success will be judged by developing a model that is neither under nor over-fitted to the dataset and can accurately predict mortality ratios.

While intra-country analysis will not be undertaken during this project, finding outlier countries based on relative achievement in other measures, such as the United States, will point to areas of possible further analysis. It would be a success if another country or region be identified with a high likelihood for inequities or adverse health outcomes.