**Dataset and Data Wrangling**

The data (birth cohort data) used in this study includes natality (birth data) and fatality (death data) files of infants born in the year 2008, along with a pdf-formatted guide document that describes the data content and type. The birth cohort data for 2008 consists of infant deaths that occurred in 2008 or 2009 linked to births in 2008. The data also includes a separate file that includes infant deaths, and unlinked file, which consists of infant deaths that had not been linked to a corresponding record in the natality file.

The pdf-formatted guide document was converted into usable format in two steps. First, Tabula, a service software, was used to convert the guide document from pdf to tsv (tab separated values file) format. The tsv file was then reformatted using a python code. The original data set did not have column names and it required a second step of writing a python code to extract the field names from the guide document.

Some columns did not have any values and were removed at the start of data wrangling. These include features like county and state of residence of the mother of the infant. The guide document lists the valid values for each feature. I wrote a python code that builds a dictionary of the valid values for each column. Any value outside the list, invalid value, was converted to NaN. All rows that consisted of NaN values were then removed resulting in 1,569,762 rows (records) and 102 columns (features).

**Link to data wrangling notebook:** https://github.com/shaybiz/Capstone\_Final/blob/master/CP\_Data\_Wrangling\_01.ipynb