

Suicide and Life Expectancy

For this Extract, Transform, and Load process suicide and life expectancy data were reviewed. There was an interest in the relationships between the number of suicides and the life expectancy of people around the world. The suicide data was extracted from the following site: <https://www.kaggle.com/russellyates88/suicide-rates-overview-1985-to-2016>. The suicide data, in the form of a csv file, contained specifics from 1985 to 2016 within 27,821 rows and 12 columns. The life expectancy data was obtained from the following site: <https://www.kaggle.com/kumarajarshi/life-expectancy-who?select=Life+Expectancy+Data.csv>. The life expectancy data, in the form of a csv file, contained figures from 2000 to 2015 for 193 countries within 2939 rows and 22 columns.

The suicide data had to be transformed to ensure clean data for a future analysis. The suicide data was cleaned by dropping the years to include only 2000 to 2015, as those were the only years included in the life expectancy data. The following columns were not needed and were removed from the dataset: ‘country-year’ and ‘HDI for year’. Any null rows were removed, as they can not be included in the SQL database. Column headers were renamed if they contained a space, ‘/’, or ‘$’. Various data frames were created, each with their own table, for the ability to easily analyze on a future basis. In the country data frame the population was aggregated by country, in a new column, as the population was originally dispersed between country, year, age and sex. In other tables the number of suicides were aggregated, in a new column, as the number was also dispersed between country, year, age, and sex. It appeared to be appropriate for any further review to have the data readily aggregated.

The life expectancy data, also, had to be transformed. With regards to this data, there were various columns that were unneeded. The following columns were thus dropped from the dataset: ‘thinness 1-19 years’, ‘thinness 5-9 years’, ‘percentage expenditure’, ‘total expenditure’, and ‘income composition of resources’. Additionally, any null values were also removed.

The data frames for both datasets were used to create various smaller tables that could be used or not used in the production database, depending on relevance to the end user. The final tables, listed below, were loaded into a SQLite database using pandas in a jupyter notebook:

* life\_exp\_clean table contains fully cleaned data without filtering out specific years of infectious disease data
* developing\_all\_yrs table contains fully cleaned data, filtered for Status of country to be equal to ‘developing’ and for years 2000-2015.
* developed\_all\_yrs table contains fully cleaned data, filtered for Status of country to be equal to ‘developed’ and for years 2000-2015.
* all\_countries\_2015 table contains fully cleaned data, filtered for all countries (regardless of status) and for only the year 2015 for most recent data.
* developing\_2015 table contains fully cleaned data, filtered for developing countries and for only the year 2015 for most recent data.
* developed\_2015 table contains fully cleaned data, filtered for developed countries and for only the year 2015 for most recent data.
* suicide\_clean includes suicide data after specific years, columns, and column headers were modified
* suicide\_table contains only the following column headers of the suicide\_clean data: country, year, sex, age, suicides\_no, suicides\_100k\_pop, and generation
* country\_table contains only the following column headers of the suicide\_clean data: country, year, population, gdp\_for\_year, and gdp\_per\_capita
* suicides\_by\_country\_by\_year\_df\_table contains the previously cleaned suicide data and aggregates suicide data by country and year
* Suicides\_by\_country\_table contains the previously cleaned suicide data and aggregates suicide data by country
* suicide\_by\_age\_by\_year\_df\_table contains the previously cleaned suicide data and aggregates suicide deathsdata by age and year
* suicide\_by\_age\_df\_table contains the previously cleaned suicide data and aggregates suicide deaths by their age category.

The decision to review suicide and life expectancy data came about from an interest in the relationship between the two datasets. Originally, there was time spent to identify happiness data to also include. However, the datasets for happiness scores tended to be lacking in time and quantity leading the impression that this has only become of importance in recent years. In reviewing suicides against life expectancy, it brought up many questions that could be answered by this data set. Are the number of suicides higher or lower in developing countries? Are the suicides higher or lower for certain age groups across countries or is a certain age group more likely to be at risk of suicide regardless of country? Over time, has the number of suicides contributed to shortening the life expectancy of a particular age group or country? There were so many questions that would be interesting to see if there was an effect one way or another. Given all of the questions, we thought this would be a good data set to explore.