Daiyan Chowdhury

The Knowledge House

April 2022

Linear Regression Analysis on Worldwide Life Expectancy

Throughout human history, the average human lifetime has remained between 25 and 40 years old. However, modern advances in technology, health, and medical care, as well as access to food and shelter, have increased the average human longevity globally by about 100 percent. The development of vaccinations and treatments for numerous illnesses, including smallpox and measles, as well as improvements in general health, water availability, and other aspects of everyday living, increased the average human lifetime. However, there is a difference in life expectancy between industrialized countries (first-world countries) and poor ones (3rd-world countries). The World Health Organization's Life Expectancy Dataset records average life expectancy by nation as well as variables that increase or degrade these values.

The initial step was to retrieve the dataset through the Kaggle API key. Because our dataset came from Kaggle, we used KaggleAPI. The next thing that was done was to clean up the dataset. The dataset had over 3000 rows and 22 unique columns that tracked some form of data that affects average human life expectancy. These columns include information such as the percentage of illnesses, a nation's GDP, and even how many individuals in that country were rated "thin" on the BMI that year. The majority of the data was already clean; however, several columns, such as the hepatitis b column, had a high number of NaN values. Fortunately, because that column was not of relevance to us, we did not need to remove the NaN values.

After cleaning our data, the next step was to save it in a local database using SQLite3. However, in order to do so, the easiest approach to input data was to establish a dataframe containing a list of tuples representing all of the values in our dataset. Using the Pandas package, this was simply accomplished, followed by entering raw data and cleaned data as distinct tables in the same local database.Inserting data into tables necessitates the use of SQLite3's SQL commands, which differ significantly from ordinary SQL (string values are TEXT rather than VARCHAR(), and AutoIncrement has no underscore between auto and increment, among other variations). We were able to extract data utilizing helper methods for faster system runtimes after we had a suitable database. Data extraction from csv files consumes more memory than extracting a few rows of data depending on a conditional value. Using the Seaborn library, we were able to create some scatter plot distribution and linear regression lines using the retrieved data.

A few notable discrepancies emerged from our visualization of life expectancy data. Globally, between 2000 and 2014, the global average life expectancy for each nation increased overall, with a positive regression line. On a worldwide scale, most values kept within the marginal range, with only a few outliers. However, when the visuals are divided between first and third world nations, the scale becomes significantly imbalanced. In the 2000s, developed-country life expectancy was greater than in developing-country life expectancy and climbed somewhat slower than in developing-country life expectancy. Although emerging nations have lower life expectancies, the average increased faster than in developed countries throughout time.

Based on this data, we may conclude that, while wealthy nations have considerably greater food, shelter, clothes, and health care, emerging countries are gradually catching up. While this data does not account for potential wars and conflicts inside some of these nations, the advancement of human health is prevalent, allowing individuals to live longer lives than earlier generations.