Udacity Data Analyst Nanodegree – Investigate a Dataset

Author: Michael Taverner
Date: September 2020
Github Repo: https://github.com/Tavnuh/udacity-DAnano-Iraq-life-expectancy - *contains Jupyter notebook and datasets*

For the Investigate a Dataset project, I selected the Gapminder data as my source. I have had an interest in geopolitical, socioeconomic and population data, and how this can be used to better inform world views for quite some time, so this was a natural choice.

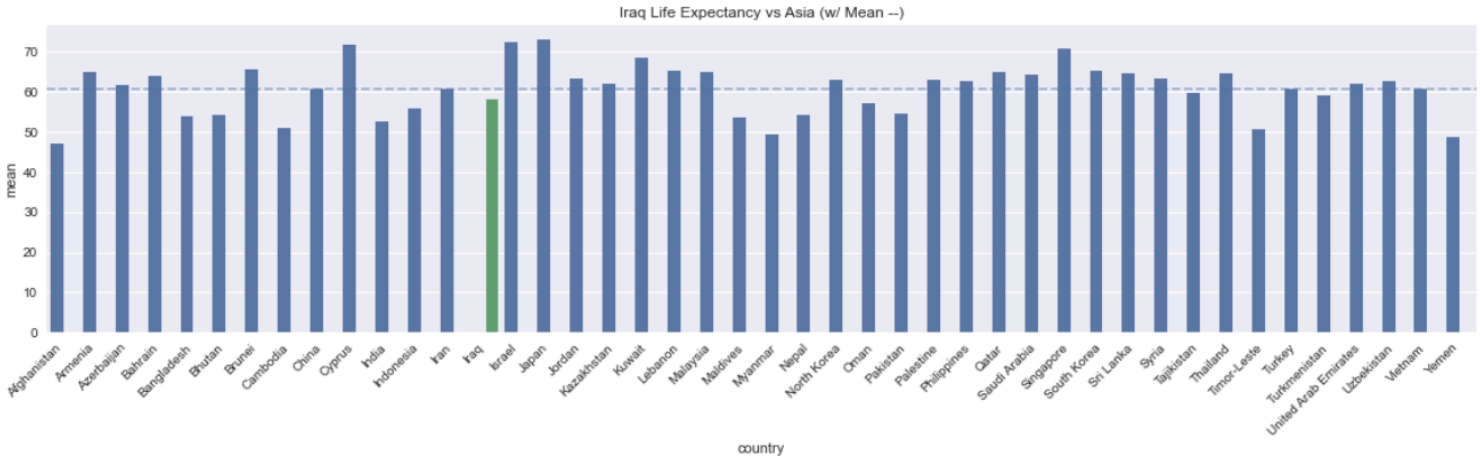
This report focuses on life expectancy of the people of Iraq from 1940 through 2020, how it compares to the rest of Asia and the world and how it changes over time. Also investigated were a few additional variables, and how those correlate to (and are predictors of) life expectancy. The date range was selected based on exploration of the records in the dataset for Iraq – pre-1940 looks to have been populated with estimated values, while post 2020 is forecast data.

After some initial EDA, 6 questions were formulated for exploration;

1. *How does Iraqi life expectancy compare to other countries in Asia?*
2. *How does Iraqi life expectancy compare to the rest of the world?*
3. *What does Iraqi life expectancy look like over time?*
4. *Are there any anomalies? Investigate*
5. *What are the differences in life expectancy by gender?*
6. *Are there any good predictive factors for life expectancy available from the Gapminder data?*

Summary of findings and commentary

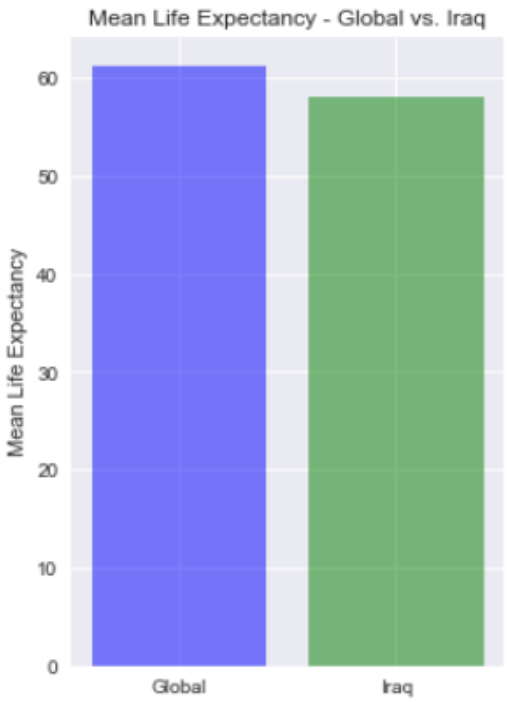
- ### How does Iraqi life expectancy compare to other countries in Asia?



The above graph shows the mean life expectancy for all countries in Asia, as listed in Worldometers, along with an average line plotted across the chart, indicating the mean life expectancy for all countries plotted. Highlighted in green, the chart shows that Iraqi life expectancy fell just short of the Asian average.

Udacity Data Analyst Nanodegree – Investigate a Dataset

- How does Iraqi life expectancy compare to the rest of the world?



On the left is the Iraqi mean life expectancy for 1940-2020 plotted against the global life expectancy for the same period.

In the case of comparisons with both the Asian and global mean life expectancies, the data shows that Iraqi life expectancy falls only marginally short of the average.

- What does Iraqi life expectancy look like over time? And 4. Are there any anomalies? Investigate

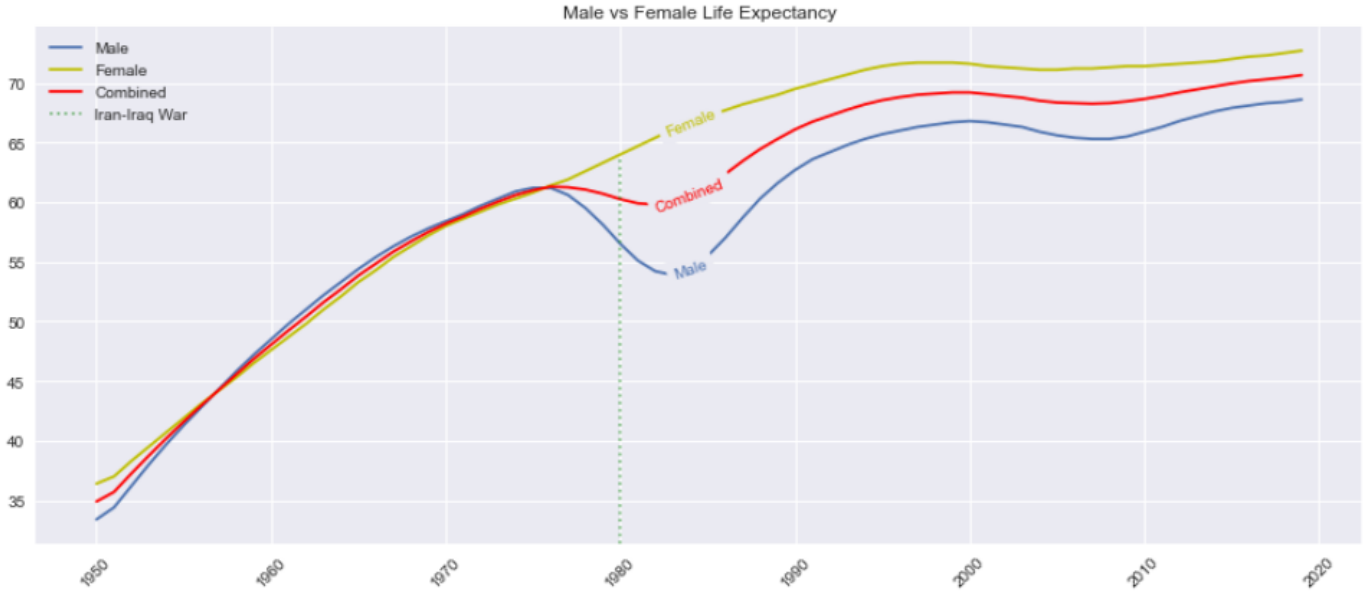
From 1940 through to 1980 the data shows a healthy upward trend with little instability, however from 1979 there is massive deviation. Between 1980 and 1988 the life expectancy drops by as much as 15

years (plotted against the y axis), recovering in 1989. Some investigation showed that 1980 was the beginning of the Iran-Iraq war, plotted with a vertical line above.

To investigate the life expectancy trend at a higher level, a 10-year moving average was calculated and plotted over the original data.

Despite the moving average effectively 'smoothing out' the trend, post-1980 there is still significant deviation from the upward trend in the period prior.

- What are the differences in life expectancy by gender?



To explore this question, two different datasets were used as a breakdown of life expectancy by gender was not available in the original. While it's noted that the observations in the gender-split datasets are much less granular in their detail and may be the result of moving averages, some key information is

available here. Based on this data, the 1980's drop in life expectancy observed in the original data is attributable entirely to the male subset of the population. While both genders recover towards the end of the 1980's, from that point onwards, there remains a 5-7 year gap in life expectancy by gender, with the female subset trending above until the end of the analysis period.

- Are there any correlative or predictive factors for life expectancy available from the Gapminder data?

For this part of the analysis, a number of additional data points were selected;

- GDP per capita PPP[1]
- GNI per capita PPP[1]
- Primary School Completion Rate (as a measure of education success)
- Government Health Spending as a percentage of total govt. spend
- Military expenditure as a percentage of GDP
- Armed forces personnel as a percentage of the total workforce

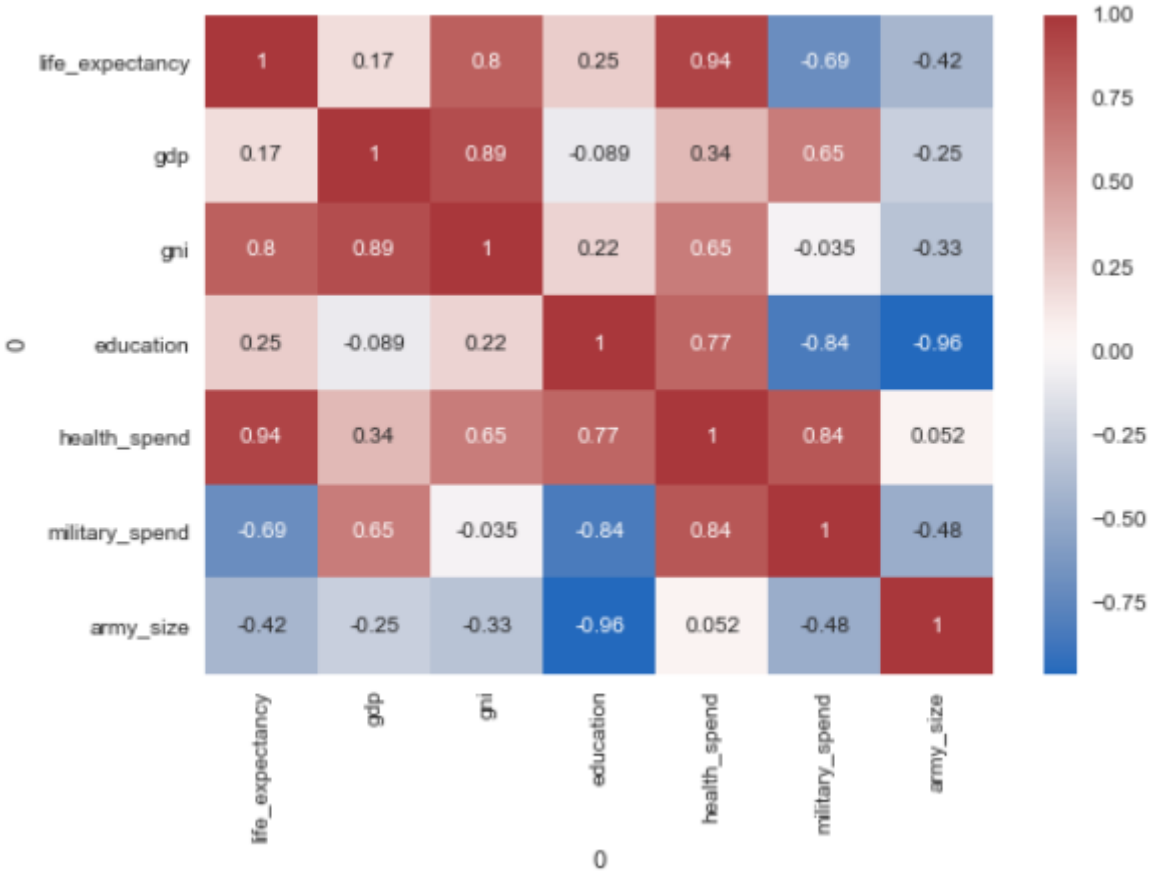
In order to conduct this analysis, I used two separate metrics – Correlation, and the Predictive Power Score[2] – shown visually with a heatmap to demonstrate the relationships between the newly added variables and the dependant variable – life expectancy.

1. PPP – Purchasing Power Parity – An international dollar has the same power over GNI as the US dollar has in the United States.

2. Predictive Power Score – A measure of the power of a single variable to predict a target variable independent of any other variables. It uses ML algorithms (Decision Tree regressors in this case) to produce a value between 0 (no predictive power) and 1 (perfect predictive power) of variable X as a predictor of variable Y - https://github.com/8080labs/ppscore

Udacity Data Analyst Nanodegree – Investigate a Dataset

Correlation matrix

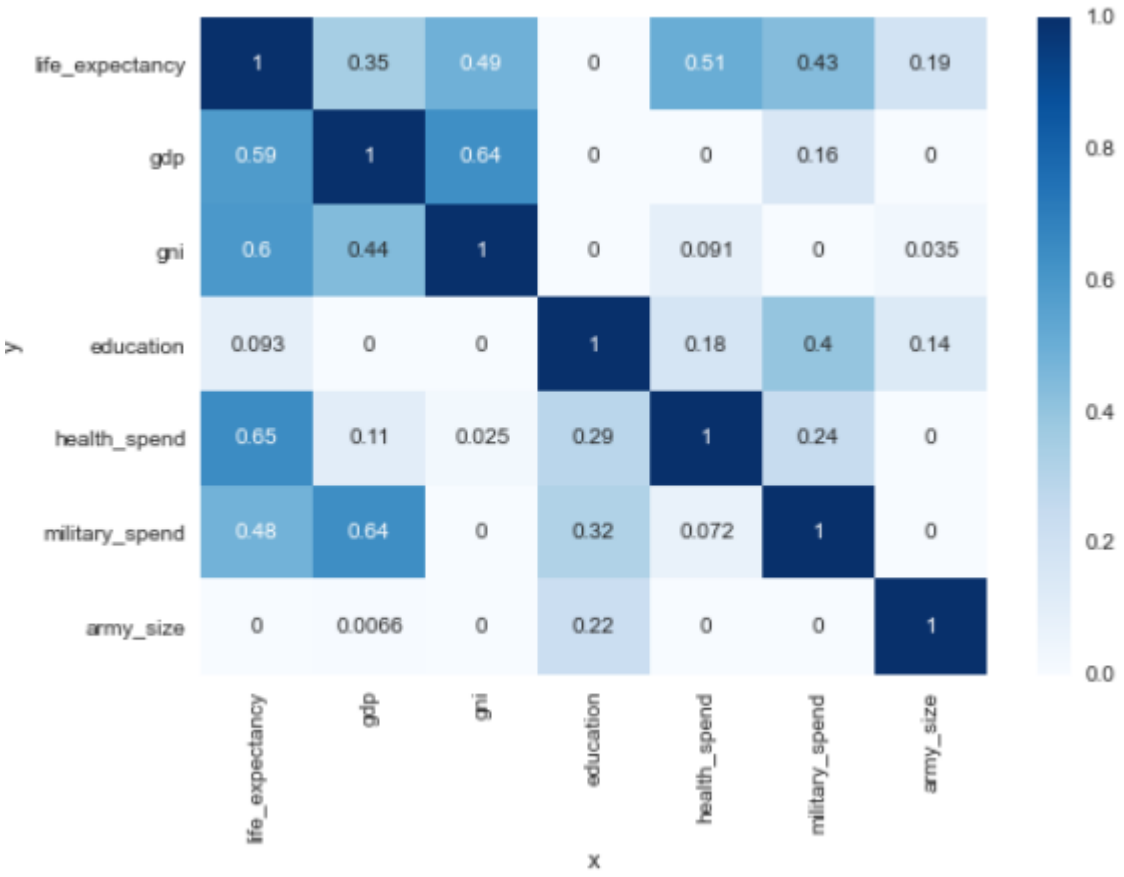


From this, it can be inferred that a high Gross National Income per capita and high Government health spend are strongly correlated with higher life expectancies. Conversely, there is a good negative correlation between Military spend and life expectancy, showing that when Military spend increases, life expectancy tends to decrease somewhat. There is also a weaker negative correlation between Army size vs life expectancy, and a weak positive correlation between GDP and Education vs life expectancy.

One other noteworthy point here is the extremely strong negative correlation of Amy size and Military spend vs education.

Udacity Data Analyst Nanodegree – Investigate a Dataset

Predictive Power Score matrix



Note on PPS vs correlation: while a correlation matrix is symmetrical, the PPS matrix is not. The relationships shown in the PPS matrix are of variable X as a predictor of variable Y.

In the same way that the correlation matrix demonstrated a high degree of positive correlation between GNI and Health spend vs life expectancy, the PPS also shows that these are both reasonably good predictors of life expectancy. While it should be noted that the scale is different (0-1 for PPS vs -1 – +1 for Correlation), the resulting inference for these relationships is similar.

In contrast to the correlation matrix, we do not see the same strength of relationship between Military spend and Army size vs. the education metric. There is a strong correlation, but the PPS has detected only a weak predictive relationship.