

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.

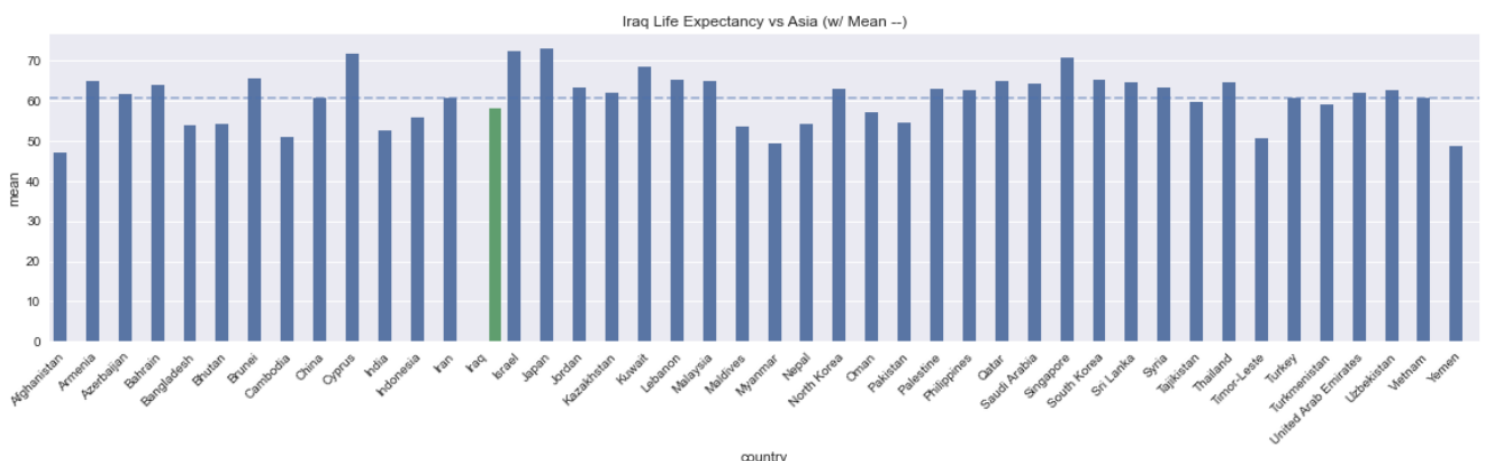
This report begins by focusing on life expectancy of the people of Iraq from 1940 through 2020, how it compares to the rest of Asia and the world and how it changes over time. Also investigated were a few additional variables, and how these correlate to (and are predictors of) life expectancy. The date 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 and the rest of the world?
2. What does Iraqi life expectancy look like over time?
3. Are there any anomalies? Investigate
4. What are the differences in life expectancy by gender?
5. Are there any good predictive factors for life expectancy available from the Gapminder data?
6. Iraqi GDP vs. 10 western nations – 1950 and 2015
7. Income per person – Iraq and 10 Western Nations

Findings and commentary

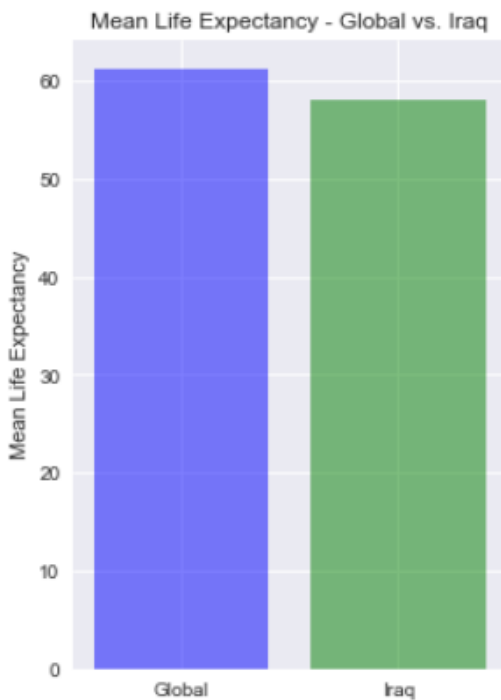
- How does Iraqi life expectancy compare to other countries in Asia?



The above graph shows the mean life expectancy between 1940 and 2020 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

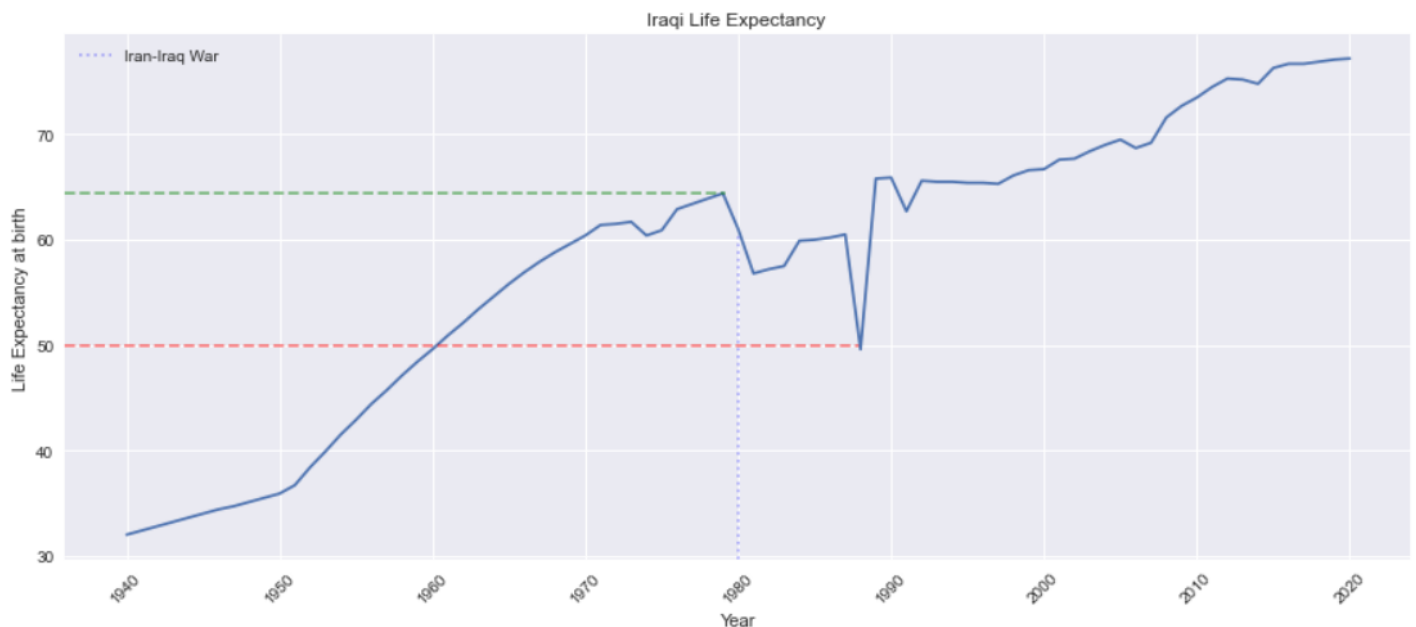
- How does Iraqi life expectancy compare to the rest of the world?



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.

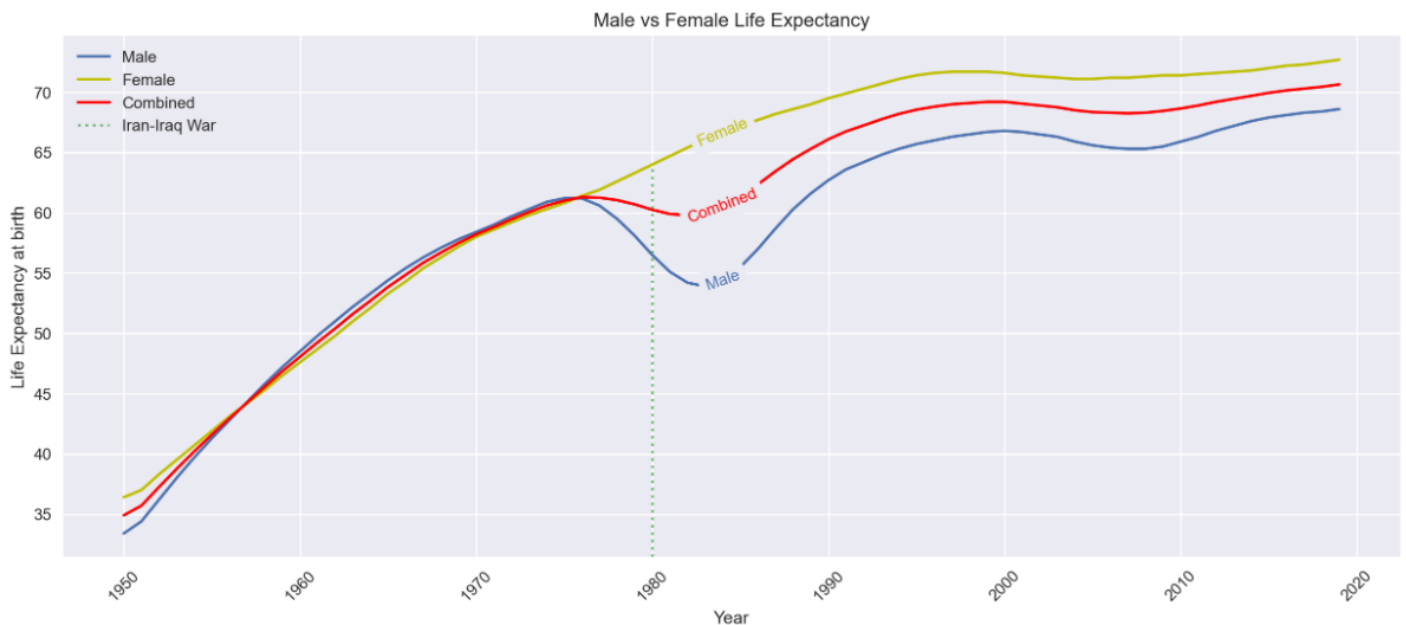
Udacity Data Analyst Nanodegree – Investigate a Dataset

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

Udacity Data Analyst Nanodegree – Investigate a Dataset

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 PPP¹
- GNI per capita PPP¹
- 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

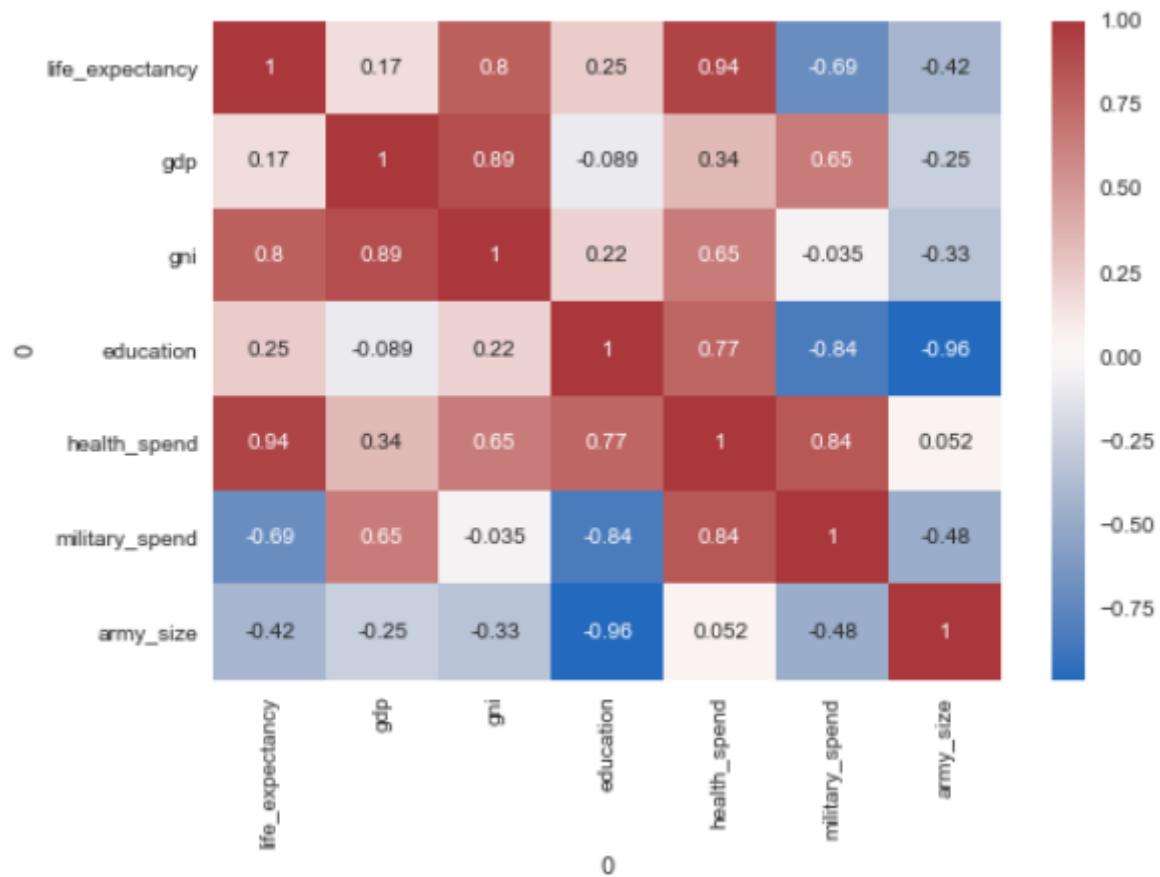
In order to conduct this analysis, I used two separate metrics – Correlation, and the Predictive Power Score² – shown visually with a heatmap to demonstrate the relationships between the newly added variables and the dependant variable – life expectancy.

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 Army 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.

Udacity Data Analyst Nanodegree – Investigate a Dataset

- Iraqi GDP vs. 10 western nations – 1950 and 2015

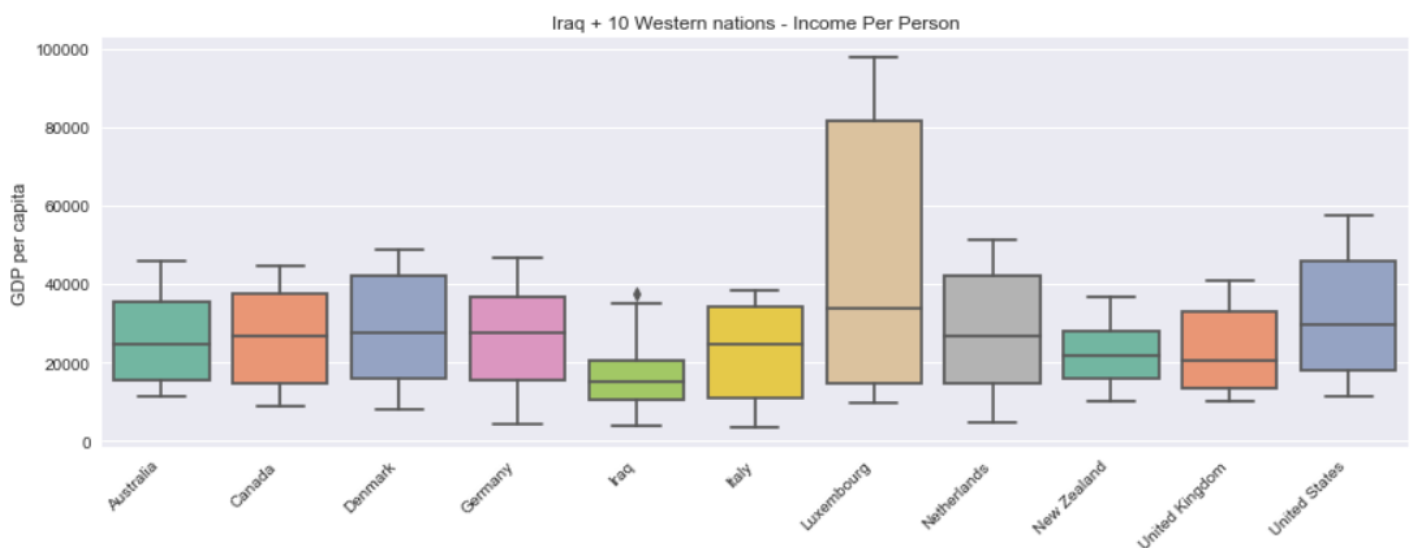
The two graphs below show Iraq's GDP in 1950 and 2015, vs that of 10 other nations representative of the "west".



In 1950, Iraqi GDP was comparable to the selected sample of 10 western nations, however by 2015 it was considerably less than the other sampled nations.



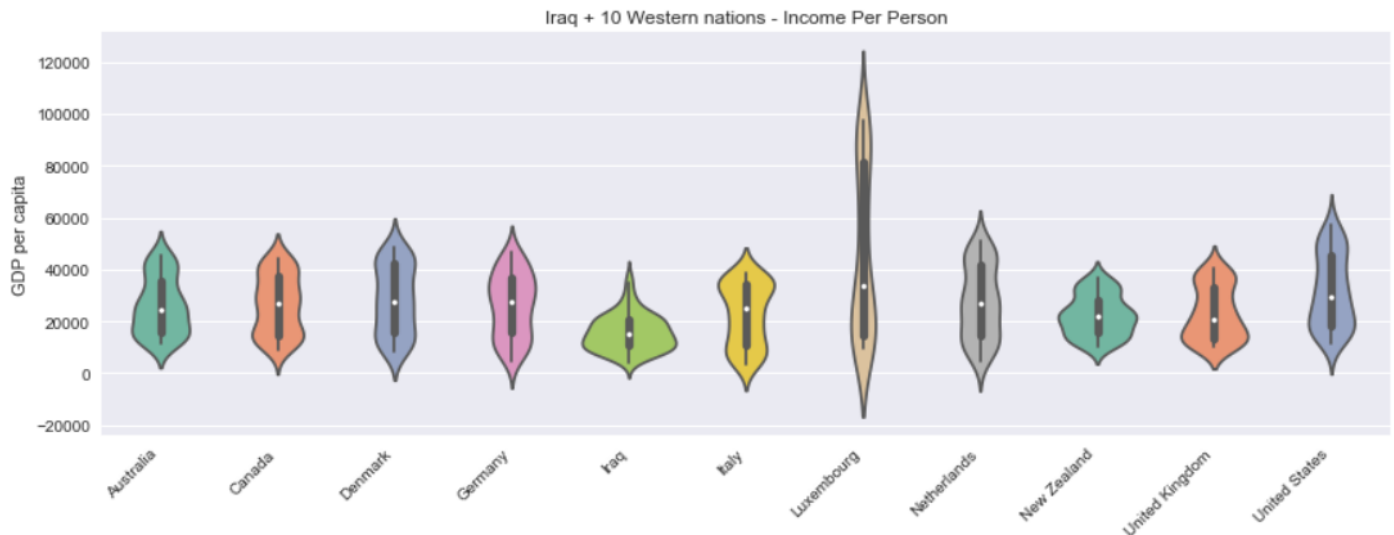
- Income per person – Iraq and 10 Western Nations



Udacity Data Analyst Nanodegree – Investigate a Dataset

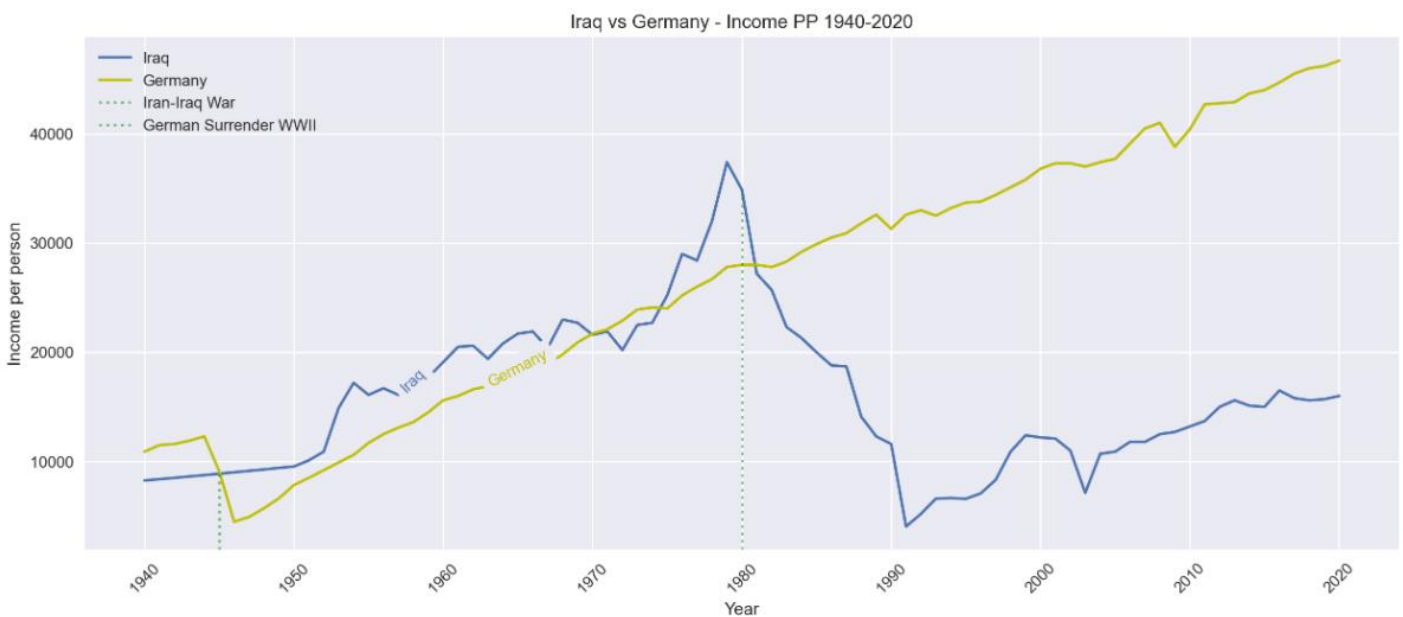
For the period in question (1940 – 2020 inclusive), Income Per Person (GDP/capita) for Iraq was significantly lower. The interquartile range (specifically 50%-75%) does, however, appear to be quite dissimilar to the other sampled countries. Also of note is the fact that outliers were not eliminated, and that Iraq is the only country to have them present in the dataset.

The violin plot below shows more detail about the distributions to give context to this difference.



The plot above indicates that compared to the other sampled nations, the distribution for Iraq appears to much more right skewed, in that the bulk of the observations appear to be at the lower end of the scale. To explore this in further detail, the line plot below shows Iraqi Income Per Person plotted over time against a country with a more uniform observed distribution (Germany).

As anticipated, aside from a decrease in the mid 1940's, Germany's Income per person rose steadily across the majority of the analysis period, while Iraq's shows much greater fluctuation and instability.



Udacity Data Analyst Nanodegree – Investigate a Dataset

Conclusions

Gapminder's core mission is to end ignorance and inform world views through access to data. To that end, the goal of this project was to investigate some key economic and welfare indicators about a country that has fared poorly in western media – Iraq – and expose some interesting and surprising truths through analysis of Life Expectancy, total GDP and GDP/capita (avg income per person). The report looks at the change in these datasets over time, as well as comparatively against other nations and compares them other indicators to investigate correlation. While it should be noted that correlation between these indicators and noteworthy moments in history does not imply causation, the correlative and coincidental relationships are interesting, nonetheless.

Life Expectancy

Iraq's life expectancy was examined for the period of 1940-2020. This period was selected because for Iraq, the data prior to 1940 appeared to be based on estimations or assumptions, whereas the data post-2020 was, of course, forecast data.

When compared to the global average, as well as the average of other countries in Asia, Life Expectancy in Iraq was only marginally lower in both cases. The expectation based on a traditional western perception, was that the data would show a significant difference, however this was pleasingly not the case.

The most significant drop in life expectancy for Iraq coincided with the start of the Iran-Iraq war. While this report does not cover all possibilities for causes of this decline, the coincidence was worthy of note.

One other interesting feature of Iraqi life expectancy was the post-1980 divergence in the male/female datasets. On one hand, it's possible that this is due to errors or shortcomings in the data collection, however it is interesting to note this change in the data coincides, again, with the Iran-Iraq war.

There are some intuitive correlations in the indicators used to compare against life expectancy (government health spending positive correlated with life expectancy for example), however given the limited availability of complete datasets for Iraq across many of the indicators, it's impossible to say if these correlated features are predictive or simply coincidental. Given these limitations, it's also improbable that the same assumptions could be expanded to the entire population.

GDP

In terms of GDP, this report samples 10 nations that would traditionally be thought to be contrasted with Iraq on key economic indicators. One of the most striking observations is the difference in Iraq's GDP compared to a sample of western nations in 1950, and how that compares to the same set of samples in 2015. The difference in the samples of the two different periods shows how far Iraq's economic position has declined compared to that of the sampled western nations over the period in question.

Udacity Data Analyst Nanodegree – Investigate a Dataset

GDP/capita (Income per person)

While it's common and good practice to handle outliers in various ways, given the small number of samples available, and what outliers in this dataset had the potential to represent, the outliers were specifically left in. After it was noted by examining the distributions of the boxplots, it was noted that the only country with outliers present was Iraq. It was then explored over time and compared with a randomly selected western country – Germany - and what it show was the comparatively wild fluctuation present in the Iraq observations