World Life Expectancy Analysis

Introduction

Global Life expectancy has changed dramatically over the last 150 years or so. Yet some countries have changed more than others. Using the World Development Indicators database from the World bank, I intend to evaluate the factors that affect life expectancy, by country and possibly by geographic region to determine if Life expectancy can be predicted based on the factors most closely correlated. My initial hypothesis is that the various forms of health spending will be one of the most significant factors in determining life expectancy.

The eventual use of this analysis is a potential evaluation of government and health care spending, and determining policy for best allocation of resources to have the greatest impact on life expectancy.

Literature Review

1. Hans Rosling's 200 countries, 200 years, 4 minutes

This Data Scientist explores visually the change in life expectancy vs income over the last 200 years and while he does not delve in to the reasons, he does show interesting trends of growth in life expectancy. Initially the Western countries see rapid growth but by the 2nd half of the 20th century, much of the rest of the world begins to catch up. BY 2009, the last year of his analysis, many of the countries have caught up to the West, yet the gap between longest and shortest is larger than ever before.
http://whatsthebigdata.com/2016/01/23/data-visualization-plotting-life-expectancy-against-income-for-200-countries-over-200-years/

2. Moonshot 1: How Can Big Data Improve Life Expectancy?

This article attempts to explain the potential benefits of applying Big Data techniques to improving health care and life expectancy around the world. The article begins to explore some of the statistics around life expectancy such as geography, wealth, diet, smoking, and other characteristics. It then goes on to extrapolate the use of Big Data into Digital Health – such as the ability to track and personalize health car, which has never been possible before until the advent of Big Data collection and analysis techniques. Additional uses of big data in healthcare is the ability of doctor's to more accurately diagnose, using ever larger databases, etc. http://healthxl.org/what-are-healthcare-moonshots/

3. Why is Big Data so big in Healthcare?

This articles lightly examines how big data is being used to improve healthcare – an industry dying to become more cost effective, while providing better care to it's patients. In short, the article concludes that Big data will allow a more efficient expenditure of R&D in pharmaceutical development, positive patient outcomes, data transparency, and potential for preventive disease methods. These analytics are expected to not only improve life expectancy around the globe, but also reduce costs of this health care. Finally the article concludes that like in healthcare, Big Data techniques can be applied to any area, yielding insights that can both improve output and reduce costs.
http://www.forbes.com/sites/howardbaldwin/2015/05/18/why-is-big-data-is-so-big-in-health-care/#32971f6e53b3

Dataset

https://www.kaggle.com/worldbank/world-development-indicators

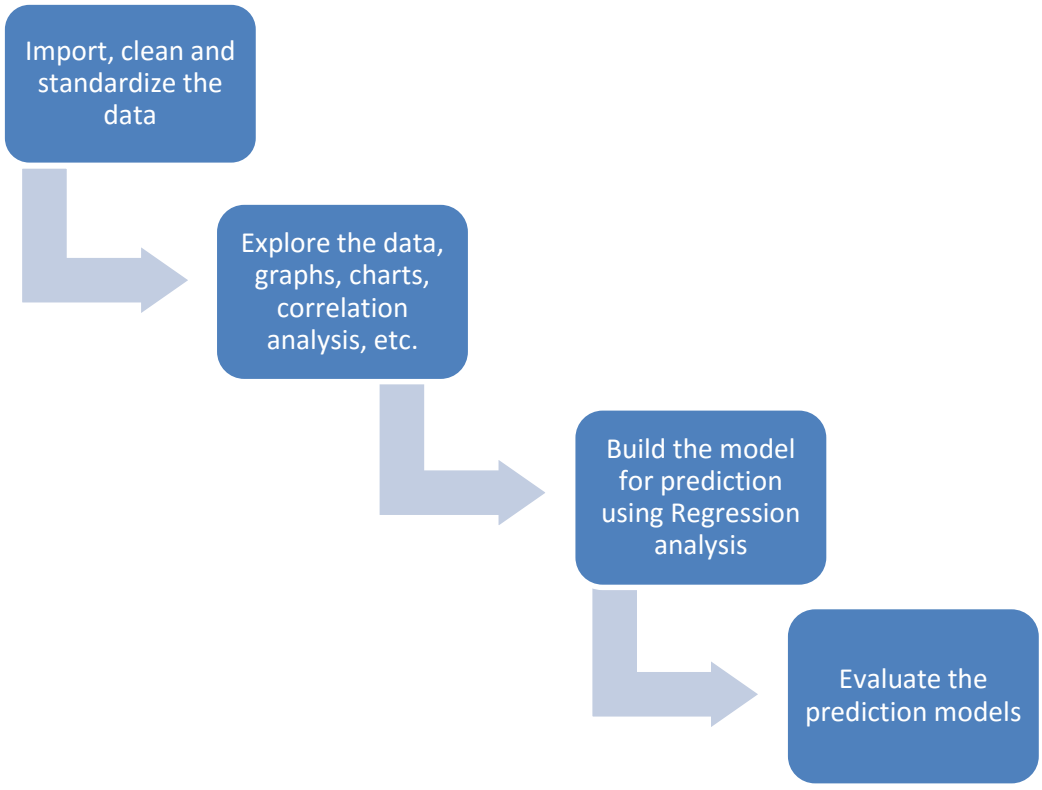
The dataset, called, The World Development Indicators was downloaded from Kaggle and was originally sourced from The World Bank. It contains thousands of annual indicators of economic development, social well-being, democracy, freedom, social evaluations, gender equality, age factors, health and quality of life among others, from 247 countries from 1960 – 2015.

The original raw data (available here) consists of 333,560 records(rows) and 60 attributes (Columns). The records are broken out by by country, then sub-divided by Indicator. It is worth noting that while most countries have data from 1960 to 2015, not all the indicators have full data. This will pose a challenge for correlation matrix analysis.

The Kaggle curated version of the data at the link above is a slightly transformed version of the data set into a "long form" with 5,656,458 records (rows) only 6 attributes (columns), to facilitate analytics. None of the data has been removed. This is the primary data set I will work with.

Approach

I Intend to analyze the various factors in the database to determine if any can be used to effectively predict Life expectancy. I will begin by cleaning and standardizing the data, followed by correlation analysis to determine superfluous attributes as well as potential candidates for prediction. Then we build the model, run it and evaluate



Step 1: Import and clean the data

i. Many of these steps will be done either in R or Excel or both.
ii. Import the data
iii. Rearrange into a usable format
iv. Clean and standardize the data, ensure all attributes are in the correct class and format
v. Deal with anomalies and outliers
vi. Deal with missing data – either using various averages, prediction or deletion of the relevant records.

See the R Code for these specific actions.

Step 2: Examine, graph, chart, correlation analysis

i. Break the data into smaller usable chunks
ii. Explore interesting graphs, charts and plots using R and Excel.
iii. Full table correlation matrix to determine if any attributes show strong correlations (positive or inverse) and so one or the other can be eliminated.
iv. Correlation analysis of all the remaining "input variables" to the "output variable" (Life expectancy at birth, total) to determine which if any have a strong correlation and can serve a s a good predictor of Life expectancy

Step 3: Build the Prediction Model

i. Divide the data into random training and testing sets
ii. Perform Regression Analysis
    a. Multivariate Linear Regression Analysis
    b. Multinomial Logistics Regression Analysis

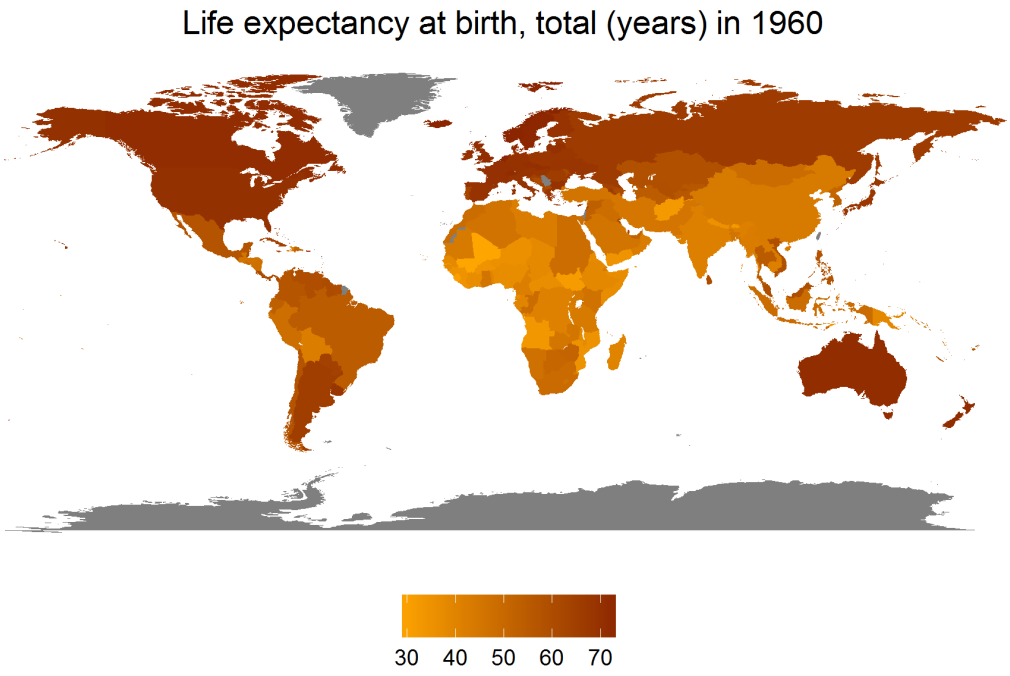
Step 4: Evaluate the Effectiveness of the Model

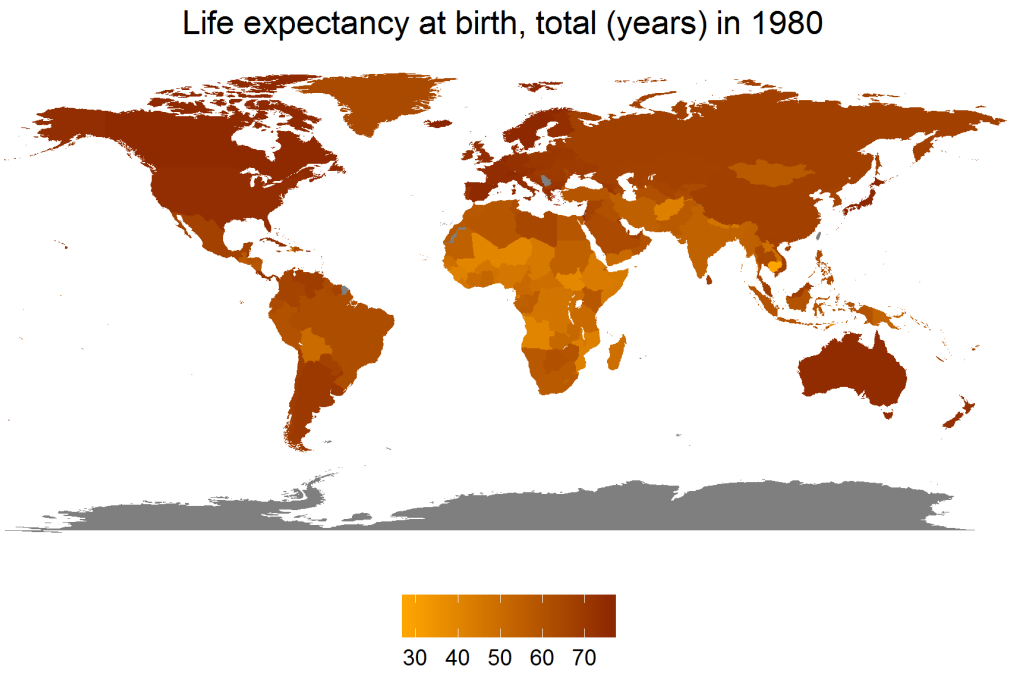
i. Using different forms I will evaluate the success of the Prediction model looking at factors like accuracy, sensitivity, specificity, etc. Specifically we will attempt RMSE, Pred25 and Confusion Matrix.

Results

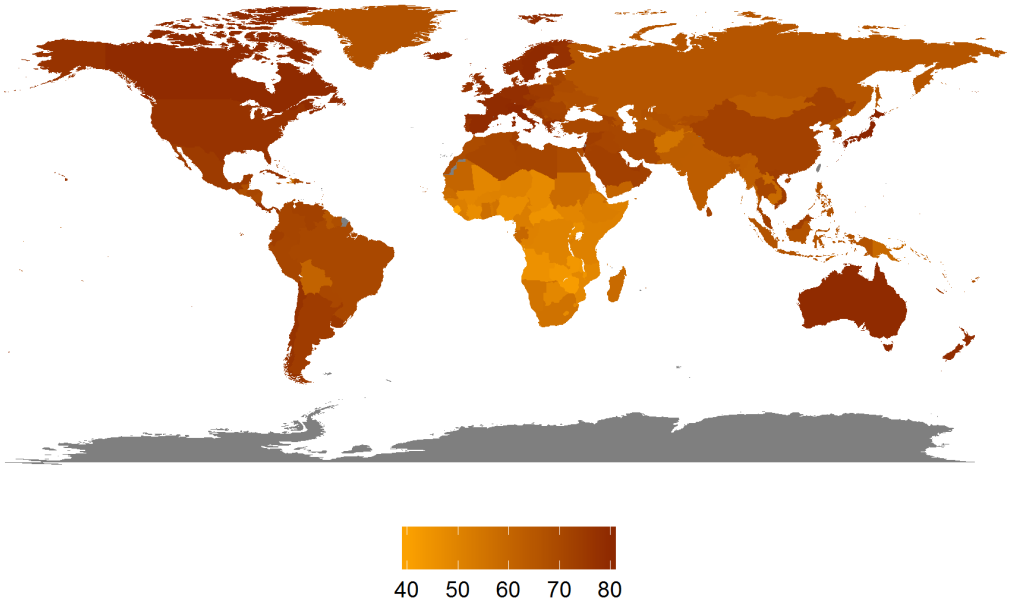
Step 2 – Examining the data, graphs, charts, correlation

1. Visualizing Global Life Expectancy for 1960, 1980, 2000 and 2013

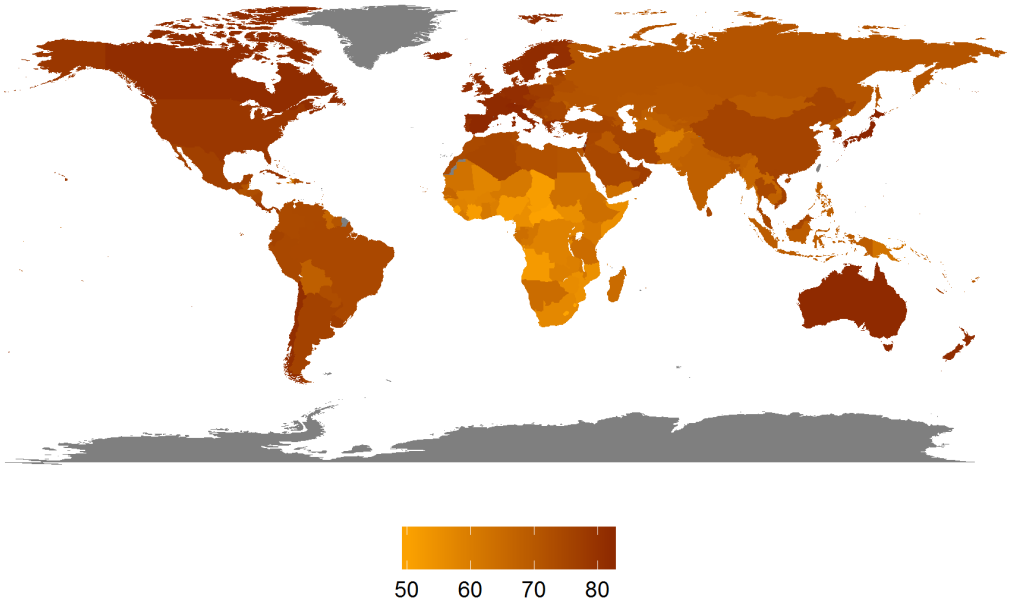




Life expectancy at birth, total (years) in 2000

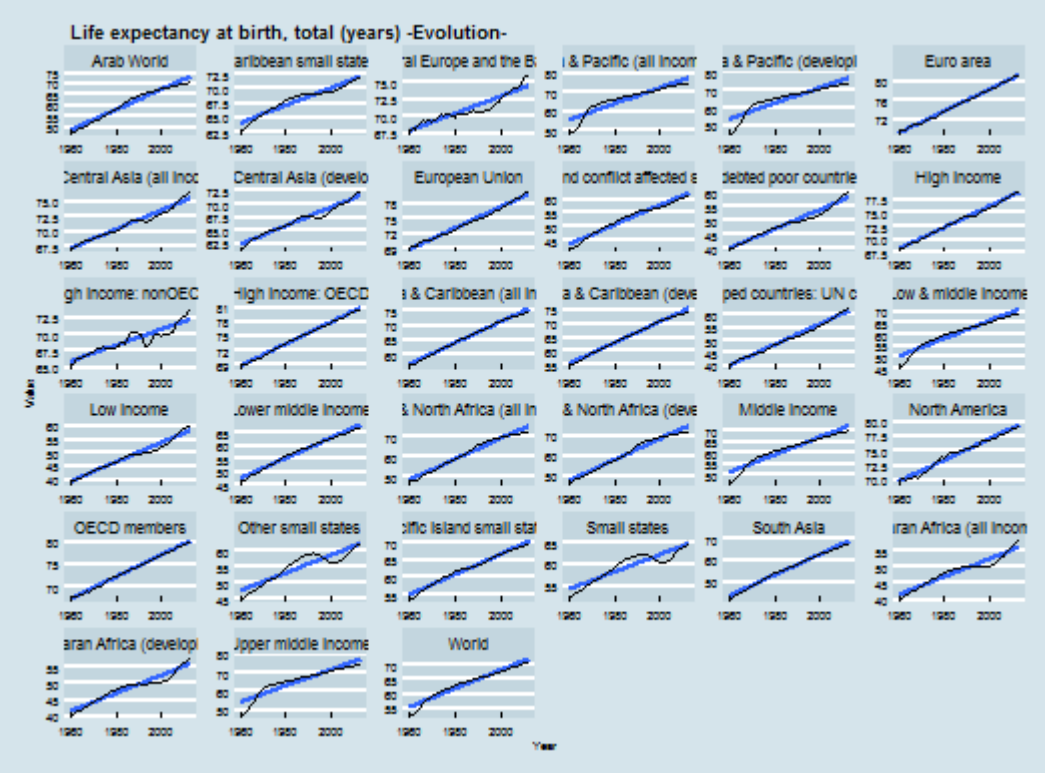


Life expectancy at birth, total (years) in 2013



While most countries seem to have increased Life Expectancy at birth, at first glance it appears that some countries, like Russia have actually decreased (their colour got lighter).

2. Visualizing Life Expectancy by region, using line graphs 1960-2013



With this graph, we can see that every region has increased Life Expectancy steadily since 1960, with a few temporary blips in some regions.

A close inspection of the first set of charts reveals that while the colour of a country like Russia has lightened, it is because the scale has gotten larger. In other words, the maximums are greater and the mean average has increased.

3. Correlation results

Correlation Matrix proved ineffectual since there are far too many results (1073 x 1073). Far too many of the variables are related to each other to decipher anything visually.

Looking at the indicators that correlate strongest to Life Expectancy at birth (total), we see that all the top predictors (strongest correlation to Life Expectancy) are agricultural in nature. Not much of a surprise since food production is a major part of life. However, what I found very interesting is that my original hypothesis, about Health Care spending being a top predictor, didn't even crack the top 20.

The Top 12 strongest indicators (all above 0.75 correlation) are:

Rank	IndicatorName	LE Total
1	Electricity production from oil, gas and coal sources (% of total)	0.875160
2	Crop production index (2004-2006 = 100)	0.824396
3	Food production index (2004-2006 = 100)	0.818857
4	Livestock production index (2004-2006 = 100)	0.802190
5	Cereal production (metric tons)	0.787857
6	Arable land (hectares)	0.785210
7	Cereal yield (kg per hectare)	0.780444
8	Permanent cropland (% of land area)	0.779362
9	Land under cereal production (hectares)	0.774417
10	Agricultural land (sq. km)	0.769273
11	Arable land (% of land area)	0.757985
12	Agricultural land (% of land area)	0.753265

Looking at the linearity and monotonicity

The Pearson correlation shows the top 10 strongest linear correlations to the target variable, "Life Expectancy at birth, (total)".

Pearson

Rank	IndicatorName	V1
1	Electricity production from oil, gas and coal sources (% of total)	0.875160
2	Crop production index (2004-2006 = 100)	0.824396
3	Food production index (2004-2006 = 100)	0.818857
4	Livestock production index (2004-2006 = 100)	0.802190
5	Cereal production (metric tons)	0.787857
6	Arable land (hectares)	0.785210
7	Cereal yield (kg per hectare)	0.780444
8	Permanent cropland (% of land area)	0.779362
9	Land under cereal production (hectares)	0.774417
10	Agricultural land (sq. km)	0.769273

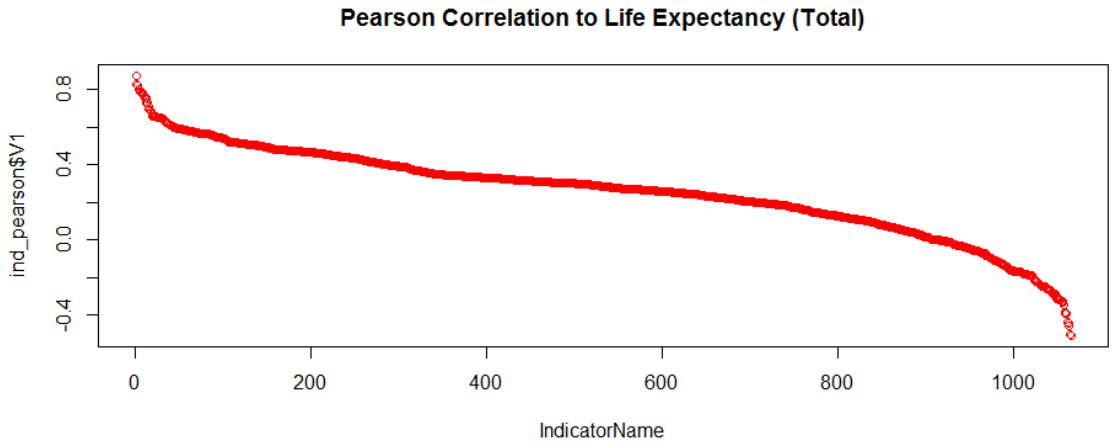
The Spearman correlation on the same data shows the top 10 strongest monotonic correlation.

Spearman

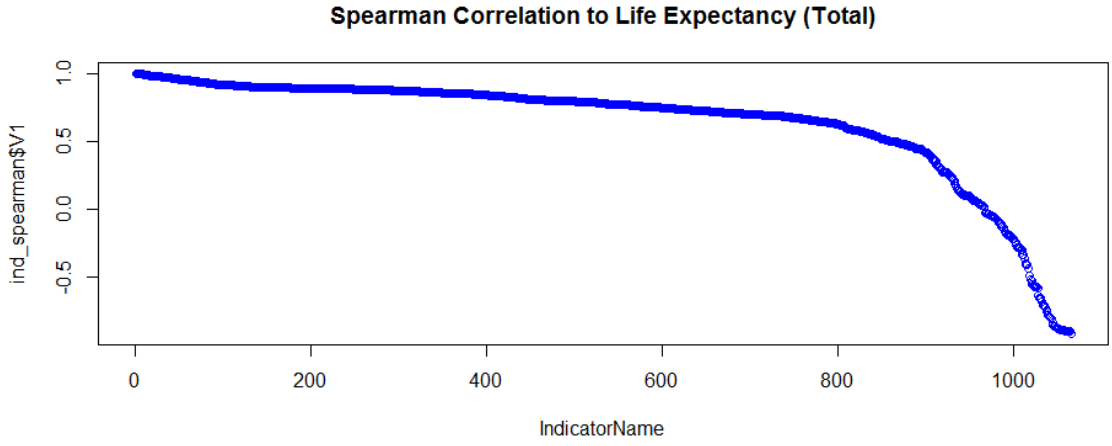
Rank	IndicatorName	V1
1	Livestock production index (2004-2006 = 100)	0.998828
2	Cereal yield (kg per hectare)	0.998539
3	Food production index (2004-2006 = 100)	0.997457
4	Adjusted savings: carbon dioxide damage (current US$)	0.996024
5	Adjusted savings: consumption of fixed capital (current US$)	0.995300
6	Adjusted savings: education expenditure (current US$)	0.994503
7	Crop production index (2004-2006 = 100)	0.994354
8	Adjusted net national income (current US$)	0.994285
9	Cereal production (metric tons)	0.991901
10	Adjusted net national income per capita (current US$)	0.991243

Again, both top 10's are agricultural in nature.

Looking at the plots of the above:



The above plot reinforces the strong linearity of the predictors to the target variable



The above plot reinforces the strong monotnicity of the predictors to the target variable

Now to look at the top 10 negative correlations:

Pearson Negative

Rank	IndicatorName	V1
1	PPG, private creditors (NTR, current US$)	-0.510455
2	PPG, commercial banks (NTR, current US$)	-0.500757
3	External balance on goods and services (current US$)	-0.456499
4	Net transfers on external debt(PPG) (NTR, current US$)	-0.443549
5	PPG, other private creditors (NTR, current US$)	-0.438613
6	PPG, bonds (NTR, current US$)	-0.393239
7	PPG, commercial banks (NFL, current US$)	-0.391545
8	GHG net emissions/removals by LUCF (Mt of CO2 equivalent)	-0.38683
9	Agriculture, value added (% of GDP)	-0.344744
10	Cash surplus/deficit (% of GDP)	-0.334735

Using Pearson we can see that no strong negative correlations exist. (i.e. none below -0.75)

Spearman Negative

Rank	IndicatorName	V1
1	Agriculture, value added (% of GDP)	-0.916378
2	Adolescent fertility rate (births per 1,000 women ages 15-19)	-0.892857
3	Death rate, crude (per 1,000 people)	-0.892857
4	Fertility rate, total (births per woman)	-0.892857
5	Rural population (% of total population)	-0.892857
6	Birth rate, crude (per 1,000 people)	-0.892785
7	Mortality rate, infant (per 1,000 live births)	-0.892785
8	Mortality rate, under-5 (per 1,000)	-0.892641
9	Number of infant deaths	-0.892064
10	Number of under-five deaths	-0.892064

From the Spearman however, we can see that there is a strong monotonic relationship between the strongest negative predictors and the target variable. However, nearly all of them measure death or mortality so they are obviously the opposite of life expectancy.

Step 3 - The Prediction Models

Using a Multivariate Linear Regression model, (See the code for the actual model)

```
Residual standard error: 3.369e-13 on 1016 degrees of freedom
Multiple R-squared:      1, Adjusted R-squared:      1
F-statistic: 9.63e+24 on 54 and 1016 DF,  p-value: < 2.2e-16
```

The Adjusted R Squared value suggests the model is a very strong predictor of the variability in the target variable, “Life.Expectancy”.

Running RMSE on the Multivariate Linear Regression Model to **evaluate the model** yields an exceptionally low value, suggesting the model is an excellent predictor of Life Expectancy.
RMSE = 0.0000000000001485907 (or $1.486e^{-13}$)

Multinomial Logistic Regression

While I was able to create the training model (See the code) , I was unable to perform this regression analysis as the resources required by the analysis are greater than what my PC can provide.

Creating a prediction model using the 12 strongest predictors identified earlier.

Pearson Strongest 12

IndicatorName	V1
Electricity.production.from.oil..gas.and.coal.sources....of.total.	0.8751596
Crop.production.index..2004.2006...100.	0.8243964
Food.production.index..2004.2006...100.	0.8188568
Livestock.production.index..2004.2006...100.	0.8021897
Cereal.production..metric.tons.	0.7878566
Arable.land..hectares.	0.7852101
Cereal.yield..kg.per.hectare.	0.780444
Permanent.cropland....of.land.area.	0.7793618
Land.under.cereal.production..hectares.	0.7744174
Agricultural.land..sq..km.	0.7692734
Arable.land....of.land.area.	0.7579854
Agricultural.land....of.land.area.	0.7532645

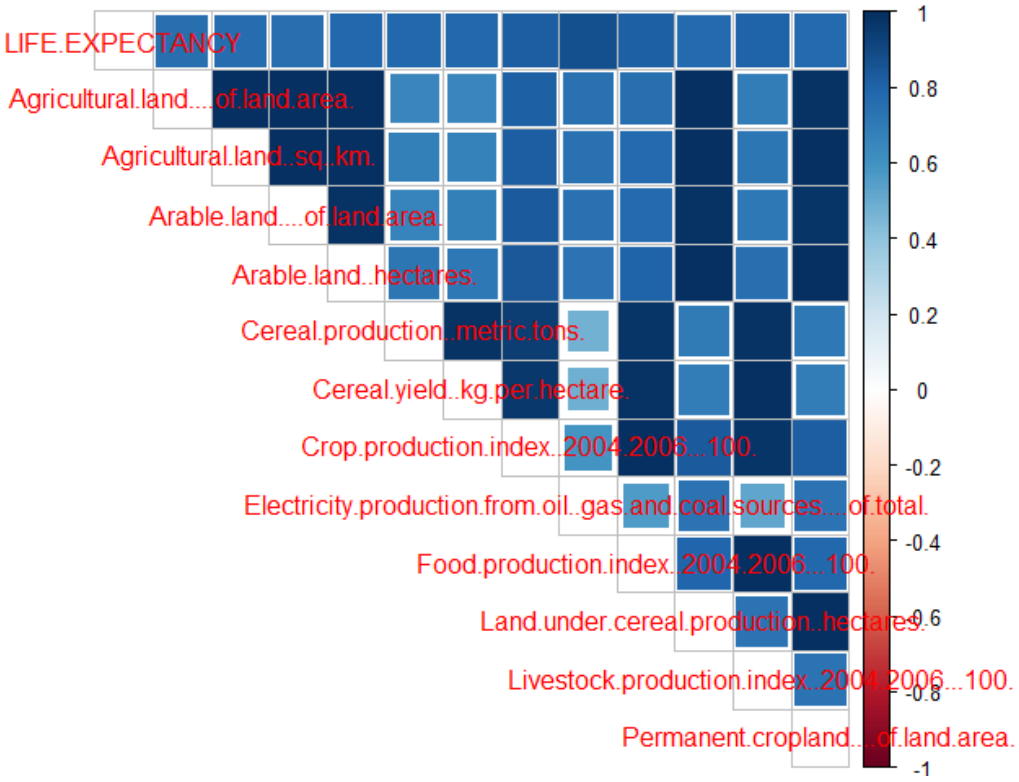
This confirms all 12 have a strong linear correlation to Life Expectancy and as such will make good predictors of Life Expectancy

Spearman Strongest 12

IndicatorName	V1
Livestock.production.index..2004.2006...100.	0.9988275
Cereal.yield..kg.per.hectare.	0.9985389
Food.production.index..2004.2006...100.	0.9974567
Crop.production.index..2004.2006...100.	0.9943542
Cereal.production..metric.tons.	0.991901
Agricultural.land....of.land.area.	0.9020004
Arable.land..hectares.	0.8797778
Agricultural.land..sq..km.	0.8711918
Land.under.cereal.production..hectares.	0.8335287
Arable.land....of.land.area.	0.7281877
Permanent.cropland....of.land.area.	0.7253738
Electricity.production.from.oil..gas.and.coal.sources....of.total.	0.4344156

This confirms all 12 have a strong monotonic correlation to Life Expectancy and will make good predictors of Life Expectancy

Visualizing this correlation



From this plot we can see that all of the 12 indicators have strong correlations to Life Expectancy, although some are clearly stronger (darker) than others.

Conclusions

It is evident that on a global basis, Life Expectancy has been climbing steadily since 1960 (with a few blips in specific regions). What is also evident is that Life expectancy is heavily dependent on agricultural factors. I found this particularly interesting since I had figured health care spending would be a strong predictor yet it did not even crack the top 20.

As a future study, it would be interesting to break this down further by country and do a deeper dive into the specific metrics of that country. Perhaps even comparing the strongest factors in a wealthy Western country like Canada with very high Life Expectancy, with poorer countries with lower ones.

The application of this study is quite simply the better allocation of limited resources – financial, land, physical and otherwise. This can be applied to governments who can invest in wither expanding the amount of agricultural land or improving the efficiency of existing land. R&D can be directed toward improving the efficiency, cost and productivity of existing agricultural resources.

From a corporate perspective, the investment in technologies that can produce agricultural supply more efficiently, less expensively, etc. is a wise investment. That could be new watering techniques, genetically modified foods, machinery, etc.