Overview

Can we correctly predict life expectancy? And, if so, which features are most relevant for prediction?

Hypothetical client: Forty years after starting a hugely-successful tech company, Monica E Banks (MoniE Banks), is committed to divesting a large percentage of her stock to create a non-profit. A strong believer that access to health care is a right, MoniE know she wants her nonprofit to help people live longer, healthier lives. The challenge is, however, she's not sure in which of the many worthy causes to invest. She's asked for help to look for strong predictors of increased life expectancy, as a way to narrow down potential areas for investment.

With 1,345 indicators across 20 categories and 56 years, the World Bank Data Bank is an invaluable resource to probe this question.

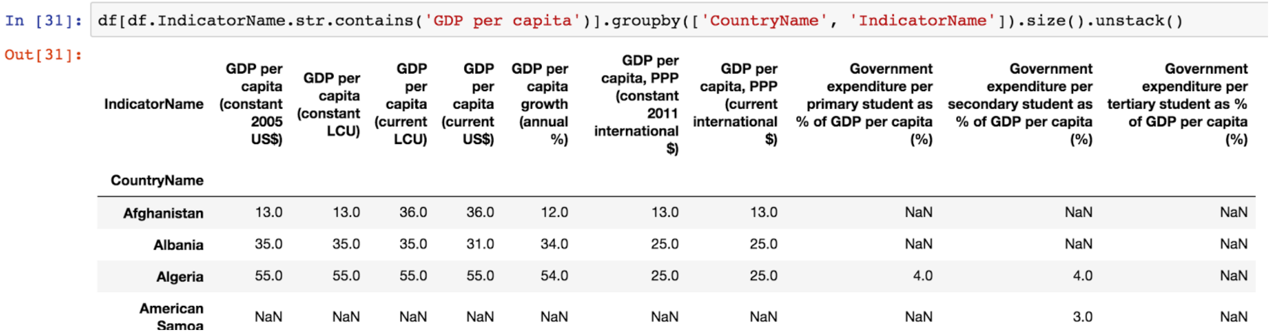
The data

This project leverages the World Data Bank's Development Indicators dataset, which contains development indicators for 214 countries and territories, along with 33 aggregate country groupings (e.g. world, regions, income levels).

Below is a brief description of the three dataframes in the dataset:

1. Country table:ge
    - Size & structure: 247 rows × 31 columns
        - One row for each of 214 countries
        - Columns provide country descriptors including: Various naming conventions, currency units, region, income group, type of government, data sources
    - Aggregate country groupings (e.g. the world, geographical regions, groupings by income level) account for 33 rows, as determined by a null value for the currently unit and "aggregate" as part of the Special Notes
2. Series table:
    - Size & structure: 1345 rows × 20 columns
        - One row per Indicator
        - Columns provide definitions, units, periodicity sources, etc.
    - Indicator hierarchy:
        - Topics: Indicators are categorized as one of 7 topics:
            1. Economic Policy & Debt
            2. Health
            3. Infrastructure
            4. Poverty
            5. Private Sector & Trade
            6. Public Sector
            7. Social Protection & Labor
        - Subtopics: Indicators are further classified with subtopics. "Health" specifically has the following subcategories:
            1. Health: Disease prevention
            2. Health: Health services

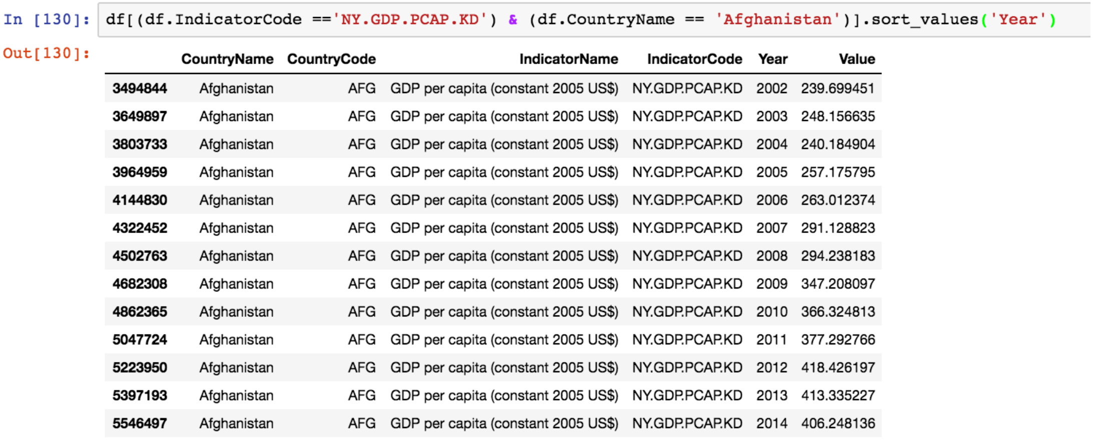
        3. Health: Mortality
        4. Health: Nutrition
        5. Health: Population: Dynamics
        6. Health: Population: Structure
        7. Health: Reproductive health
        8. Health: Risk factors
3. Indicators table:
    - Size & structure: 5656458 rows × 6 columns
        - One row per indicator per country per year (for up to 55 years per indicator)
        - Columns:
            1. Country Name
            2. Country Code (ISO 3-digit standard)
            3. Indicator Name
            4. Indicator Code
            5. Year: from 1960 to 2015
            6. Value
    - Obviously data are not available for all countries and all indicators from 1960 onwards (e.g. countries were founded after 1960, indicators were defined more recently). In cases when values are not available for a particular country-indicator-year, that record is omitted rather than filled with an NA value.
    - For example, Afghanistan only has 13 years of data for the indicator “GDP per capita (constant 2005 US$)” (image 1 below). When we look at all of the records for Afghanistan and this indicator, we can see that the data only starts in 2002 (image 2).



```
In [31]: df[df.IndicatorName.str.contains('GDP per capita')].groupby(['CountryName', 'IndicatorName']).size().unstack()
```

Out[31]:

IndicatorName	GDP per capita (constant 2005 US$)	GDP per capita (constant LCU)	GDP per capita (current LCU)	GDP per capita (current US$)	GDP per capita growth (annual %)	GDP per capita, PPP (constant 2011 international $)	GDP per capita, PPP (current international $)	Government expenditure per primary student as % of GDP per capita (%)	Government expenditure per secondary student as % of GDP per capita (%)	Government expenditure per tertiary student as % of GDP per capita (%)
CountryName										
Afghanistan	13.0	13.0	36.0	36.0	12.0	13.0	13.0	NaN	NaN	NaN
Albania	35.0	35.0	35.0	31.0	34.0	25.0	25.0	NaN	NaN	NaN
Algeria	55.0	55.0	55.0	55.0	54.0	25.0	25.0	4.0	4.0	NaN
American Samoa	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	3.0	NaN



```
In [130]: df[(df.IndicatorCode =='NY.GDP.PCAP.KD') & (df.CountryName == 'Afghanistan')].sort_values('Year')
```

Out[130]:

	CountryName	CountryCode	IndicatorName	IndicatorCode	Year	Value
3494844	Afghanistan	AFG	GDP per capita (constant 2005 US$)	NY.GDP.PCAP.KD	2002	239.699451
3649897	Afghanistan	AFG	GDP per capita (constant 2005 US$)	NY.GDP.PCAP.KD	2003	248.156635
3803733	Afghanistan	AFG	GDP per capita (constant 2005 US$)	NY.GDP.PCAP.KD	2004	240.184904
3964959	Afghanistan	AFG	GDP per capita (constant 2005 US$)	NY.GDP.PCAP.KD	2005	257.175795
4144830	Afghanistan	AFG	GDP per capita (constant 2005 US$)	NY.GDP.PCAP.KD	2006	263.012374
4322452	Afghanistan	AFG	GDP per capita (constant 2005 US$)	NY.GDP.PCAP.KD	2007	291.128823
4502763	Afghanistan	AFG	GDP per capita (constant 2005 US$)	NY.GDP.PCAP.KD	2008	294.238183
4682308	Afghanistan	AFG	GDP per capita (constant 2005 US$)	NY.GDP.PCAP.KD	2009	347.208097
4862365	Afghanistan	AFG	GDP per capita (constant 2005 US$)	NY.GDP.PCAP.KD	2010	366.324813
5047724	Afghanistan	AFG	GDP per capita (constant 2005 US$)	NY.GDP.PCAP.KD	2011	377.292766
5223950	Afghanistan	AFG	GDP per capita (constant 2005 US$)	NY.GDP.PCAP.KD	2012	418.426197
5397193	Afghanistan	AFG	GDP per capita (constant 2005 US$)	NY.GDP.PCAP.KD	2013	413.335227
5546497	Afghanistan	AFG	GDP per capita (constant 2005 US$)	NY.GDP.PCAP.KD	2014	406.248136

Data wrangling

While the datasets are fairly clean, there are missing data - for several years for some metrics, and large gaps for specific counties. In some cases, country data is missing because the country is younger than the data set. My discussion of how to deal with missing data is included in the machine learning section.

The original dataset was very tall, with one single row per country-year-indicator combination. For the machine learning and most of the EDA, I needed the dataframe to have features (development indicators) as column headers and the countries as the rows. As a result, the main data wrangling steps were unstacking the data to structure it in a workable state. I first merged the ‘country’ and ‘Indicators’ data frames to pull in additional details on the countries. Specifically, this was to pull in the ‘SpecialNotes’ indicating whether countries are aggregated regions, as well as the IncomeGroup (classification for each country’s income level). Then, I transformed the data frame into a more workable state by unstacking the dataframe with pivot_table().

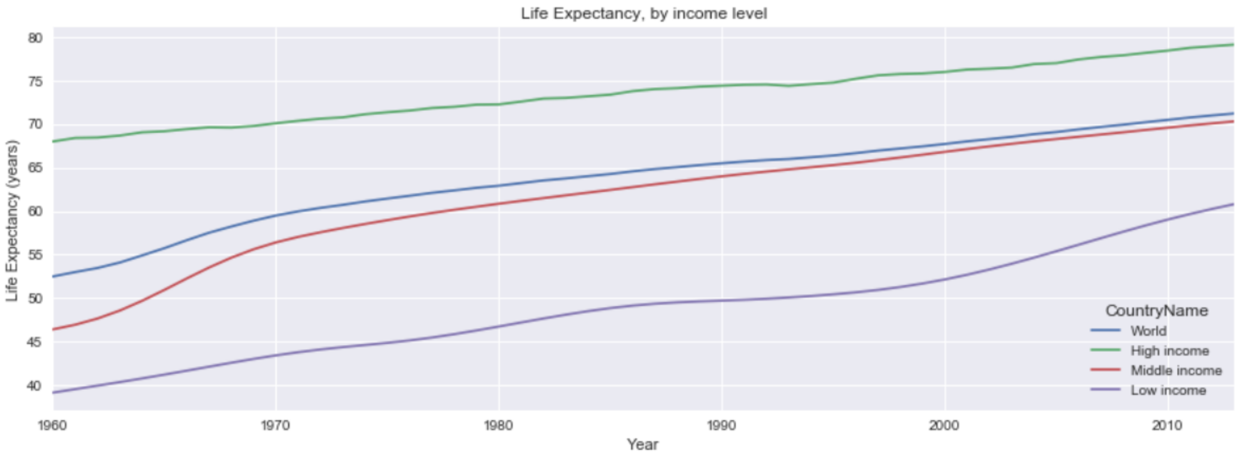
Inferential statistics & analysis

Exploratory data analysis

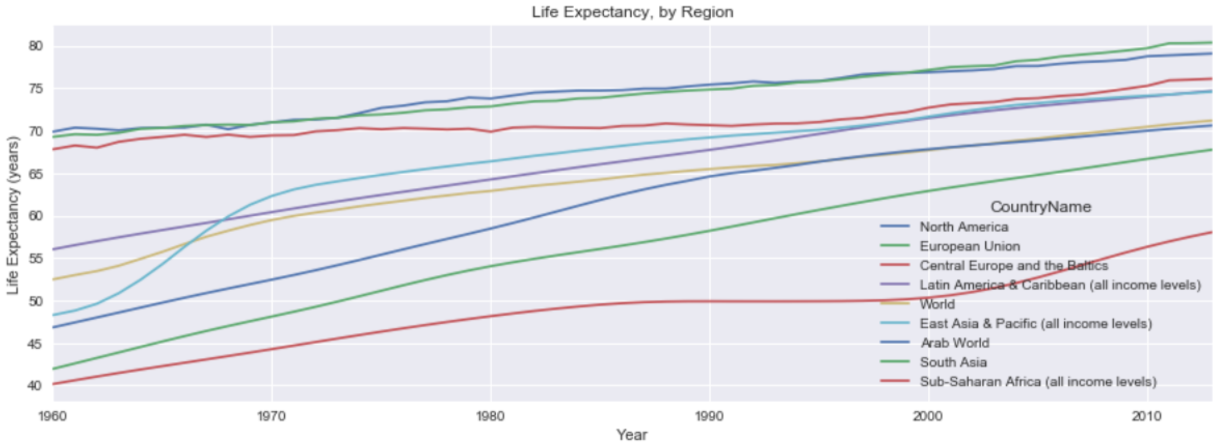
First, let’s explore the commonly-assumed relationship between life expectancy and economic prosperity. Using the income level classifications in the data, we plot the life expectancy trends from 1960 to 2015 for each of the three income groups (High income, medium income, and low income), as well as the life expectancy trends for the world. In the graph below, we can see a very large and steady different between in life expectancy by income level. As expected, higher income is correlated with higher life expectancy.

For all four groupings, the trend is to increase over time. A few other notable observations:

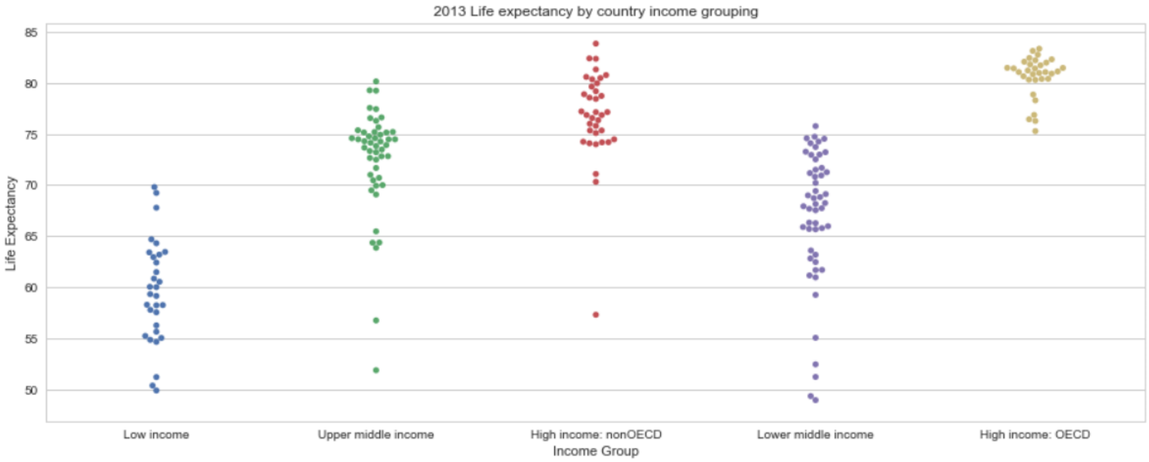
A. The rate of increase in life expectancy for the high income countries has slowed down. This is perhaps not surprisingly, since the high income countries started with a much higher life expectancy.
B. There is a decent uptick in life expectancy for the middle income countries in the mid- to late-60s.
C. In the low income countries, the increase in life expectancy all but stops for nearly a decade between the mid-80s and mid-90s. (Could this be from the HIV/AIDs epidemic?) Following this plateau, low income countries have seen a strong increase in life expectancy.
D. As a result of points A and C, the gap between high and low income countries is closing.



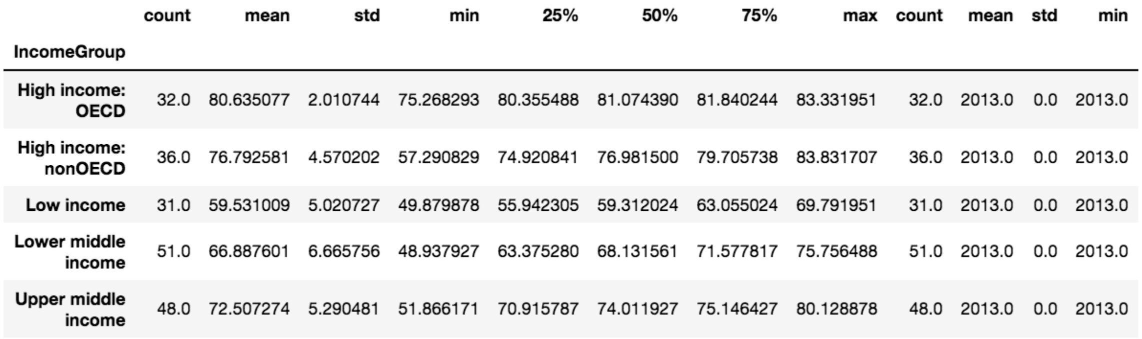
Digging deeper into the life expectancy by region, we can see that the flattening in the mid-80s is also reflected in the Sub-saharan African trend, indicating that it could indeed be the HIV/AIDs epidemic causing this slowdown in low-income countries.



Beyond the longitudinal trends, we can use a bee swarm plot to view the distribution of today's life expectancy for the various income groupings (shown below). We can see a clear relationship between life expectancy and income levels -- as income increases, not only does the life expectancy increase, but the distribution of life expectancy tights up (lower std). A notable outlier in the high income: nonOECD countries is Equatorial Guinea with a life expectancy of 57.29 years.



Summary statistics for life expectancy by income grouping are below:



	count	mean	std	min	25%	50%	75%	max	count	mean	std	min
IncomeGroup												
High income: OECD	32.0	80.635077	2.010744	75.268293	80.355488	81.074390	81.840244	83.331951	32.0	2013.0	0.0	2013.0
High income: nonOECD	36.0	76.792581	4.570202	57.290829	74.920841	76.981500	79.705738	83.831707	36.0	2013.0	0.0	2013.0
Low income	31.0	59.531009	5.020727	49.879878	55.942305	59.312024	63.055024	69.791951	31.0	2013.0	0.0	2013.0
Lower middle income	51.0	66.887601	6.665756	48.937927	63.375280	68.131561	71.577817	75.756488	51.0	2013.0	0.0	2013.0
Upper middle income	48.0	72.507274	5.290481	51.866171	70.915787	74.011927	75.146427	80.128878	48.0	2013.0	0.0	2013.0

Machine Learning

Since life expectancy is a continuous value, a linear regression was used. As mentioned above, we therefore need to structure our data such that we have one sample (country) per row and the features (indicators) as the column headers. The model is taking into account just 2013 data.

Selecting the Indicators to predict life expectancy

To identify appropriate predictors of life expectancy, there were two requirements:

1. Identify predictors that are well-represented in the data set. There are 247 countries in the dataset, but very few indicators have data for all countries. As a lower-bound, n=150 was set.
2. Select a subset of indicators that are representative of the main development categories

Taking these two criteria into account, I identified the following indicators which span healthcare, education, social living conditions, population density, and economics:

Indicator Name	Indicator Code
Adjusted net national income (annual % growth)	NY.ADJ.NNTY.KD.ZG
Adolescent fertility rate (births per 1,000 women ages 15-19)	SP.ADO.TFRT
Agricultural land (% of land area)	AG.LND.AGRI.ZS
Arable land (% of land area)	AG.LND.ARBL.ZS
Armed forces personnel (% of total labor force)	MS.MIL.TOTL.TF.ZS
Adjusted savings: education expenditure (% of GNI)	NY.ADJ.AEDU.GN.ZS
Employment to population ratio, 15+, total (%) (modeled ILO est.)	SL.EMP.1524.SP.ZS
Enrolment in primary education, both sexes (number)	SE.PRM.ENRR
Enrolment in secondary general, both sexes (number)	SE.PRM.ENRL
GDP per capita (current US$)	NY.GDP.PCAP.CD
Health expenditure per capita (current US$)	SH.XPD.PCAP
Health expenditure, total (% of GDP)	SH.XPD.TOTL.ZS
Imports of goods and services (% of GDP)	NE.IMP.GNFS.ZS
Improved sanitation facilities (% of population with access)	SH.STA.ACSN
Industry, value added (% of GDP)	NV.IND.TOTL.ZS
Mobile cellular subscriptions (per 100 people)	IT.CEL.SETS.P2
Population density (people per sq. km of land area)	EN.POP.DNST
Refugee population by country or territory of asylum	SM.POP.REFG
Rural population (% of total population)	SP.RUR.TOTL.ZS
Trade (% of GDP)	NE.TRD.GNFS.ZS
Unemployment, total (% of total labor force)	SL.UEM.TOTL.ZS

I would also like to include these metrics, but they are too tightly coupled with life expectancy. We'll take a look at them later.

Lifetime risk of maternal death (%)	SH.MMR.RISK
Mortality rate, infant (per 1,000 live births)	SP.DYN.IMRT.IN
Mortality rate, neonatal (per 1,000 live births)	SH.DYN.NMRT
Mortality rate, under-5 (per 1,000)	SH.DYN.MORT

The dataframe is reshaped for machine learning, with one column for y (life expectancy) and the remaining features for X. Because we have already pre-selected highly-populated datasets, we drop any row with missing values. This leaves us with 97 countries with all features populated for the model.

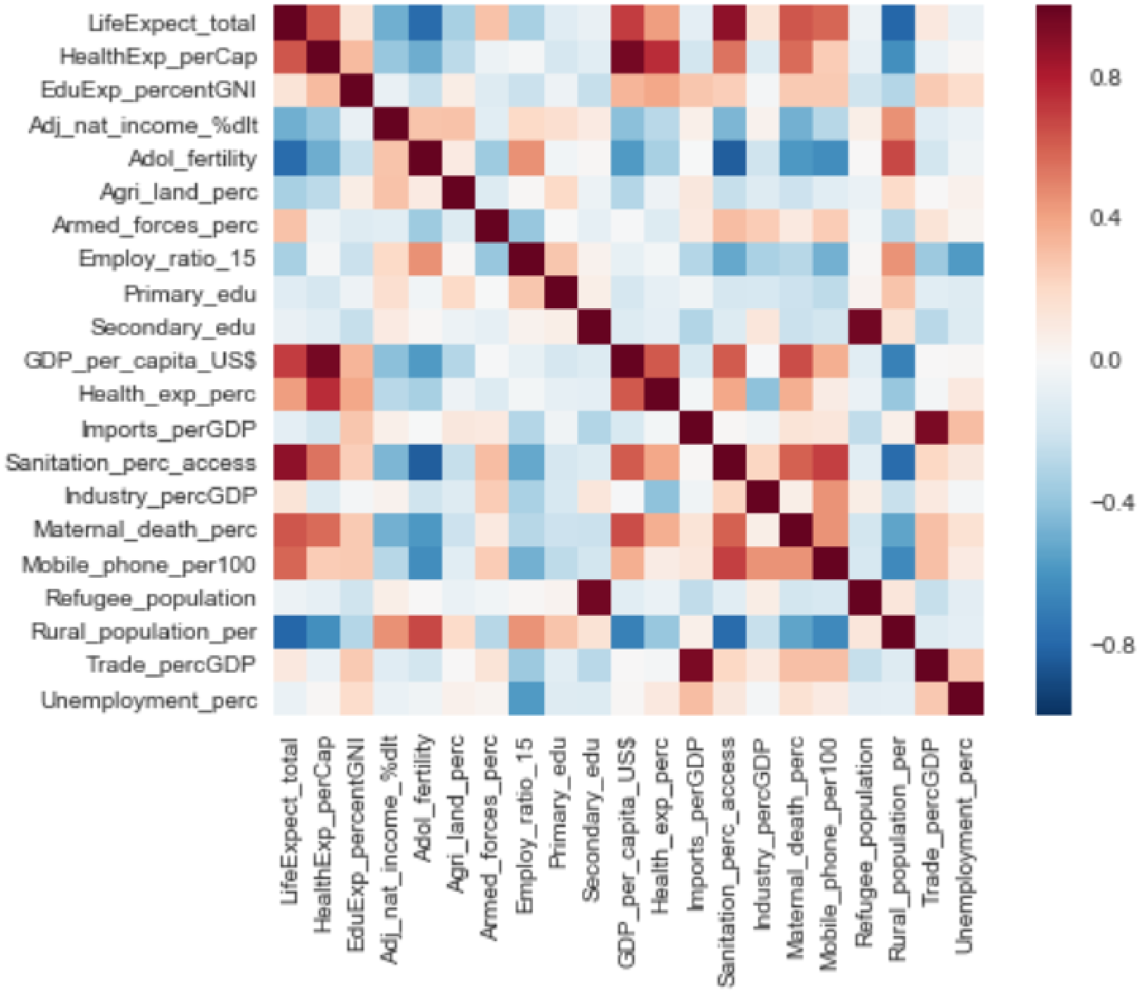
Pearson’s correlation heat map

A preliminary look at the correlation heatmap shows which features (in bold) are highly correlated with life expectancy:

Indicator Name	Indicator Code	Correlation with life exp.
Adjusted net national income (annual % growth)	NY.ADJ.NNTY.KD.ZG	-0.491256348
Adolescent fertility rate (births/1,000 women ages 15-19)	**SP.ADO.TFRT**	**-0.777008787**
Agricultural land (% of land area)	AG.LND.AGRI.ZS	-0.335660885
Arable land (% of land area)	AG.LND.ARBL.ZS	-0.017978765
Armed forces personnel (% of total labor force)	MS.MIL.TOTL.TF.ZS	0.293656078
Adjusted savings: education expenditure (% of GNI)	NY.ADJ.AEDU.GN.ZS	0.139598012
Employment to population ratio, 15+, total (%)	SL.EMP.1524.SP.ZS	-0.331515179
Enrolment in primary education, both sexes (number)	SE.PRM.ENRR	-0.123566415
Enrolment in secondary general, both sexes (number)	SE.PRM.ENRL	-0.073885307
GDP per capita (current US$)	**NY.GDP.PCAP.CD**	**0.69771694**
Health expenditure per capita (current US$)	**SH.XPD.PCAP**	**0.632073168**
Health expenditure, total (% of GDP)	SH.XPD.TOTL.ZS	0.416543232
Imports of goods and services (% of GDP)	NE.IMP.GNFS.ZS	-0.097741397
Improved sanitation facilities (% of pop. with access)	**SH.STA.ACSN**	**0.883723624**
Industry, value added (% of GDP)	NV.IND.TOTL.ZS	0.135338206
Mobile cellular subscriptions (per 100 people)	IT.CEL.SETS.P2	0.579734345
Population density (people per sq. km of land area)	EN.POP.DNST	0.083077015
Refugee population by country or territory of asylum	SM.POP.REFG	-0.059503849
Rural population (% of total population)	**SP.RUR.TOTL.ZS**	**-0.803908445**
Trade (% of GDP)	NE.TRD.GNFS.ZS	0.107477203
Unemployment, total (% of total labor force)	SL.UEM.TOTL.ZS	-0.069952485

As expected there is a positive correlation between life expectancy and health expenditure per capita (r = 0.632), GDP per capita (r = 0.697), access to sanitation (r = 0.884). There is a strong negative correlation with percentage of the population living in rural areas as well as adolescent fertility (r = -0.774).

Conversely, we see low correlation between life expectancy and arable land, enrolment in secondary education, imports of good and services as a % of GDP, trade as a % of GDP, and unemployment rates.



Linear regression

The data is split 70:30 into training and test sets. Because the features have vastly different ranges of data, we normalize the data before running it through a linear regressor.

The initial accuracy score with all features is 0.79, but with 20 features and only 150 countries in the dataset, we can assume this is overfit. Let's do some feature engineering then cross-validation to refine this.

Using the five features that have the highest correlation (health expenditure per capita, adolescence fertility, GDP per capita, access to sanitation, percent of population living in rural areas), cross validation shows our model predicts life expectancy with 77% accuracy.

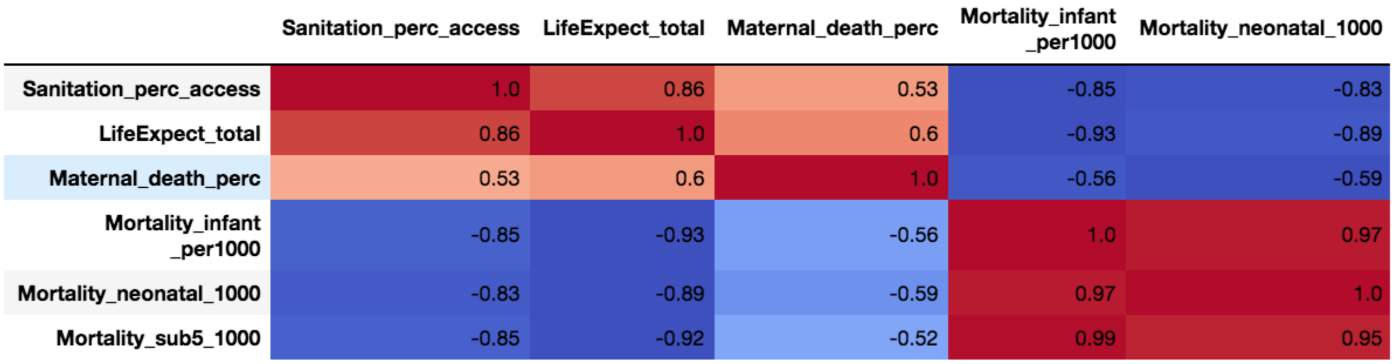
Again dropping the feature with the lowest correlation (health expenditure per capita), the cross-validated model accuracy jumps to 0.784.

This shows that we are able to predict life expectancy relatively well, and most importantly, we know which metrics are the key determinants of life expectancy: sanitation.

Let’s probe deeper into how sanitation impacts life expectancy. Fortunately, in addition to overall life expectancy, we have sufficient cuts of mortality data, including:

- Lifetime risk of maternal death (%)
- Mortality rate, infant (per 1,000 live births)
- Mortality rate, neonatal (per 1,000 live births)
- Mortality rate, under-5 (per 1,000)

Running a correlation on sanitation and these mortality rates by age, we get:



	Sanitation_perc_access	LifeExpect_total	Maternal_death_perc	Mortality_infant_per1000	Mortality_neonatal_1000
Sanitation_perc_access	1.0	0.86	0.53	-0.85	-0.83
LifeExpect_total	0.86	1.0	0.6	-0.93	-0.89
Maternal_death_perc	0.53	0.6	1.0	-0.56	-0.59
Mortality_infant_per1000	-0.85	-0.93	-0.56	1.0	0.97
Mortality_neonatal_1000	-0.83	-0.89	-0.59	0.97	1.0
Mortality_sub5_1000	-0.85	-0.92	-0.52	0.99	0.95

As expected, there is a strong negative correlation between sanitation and mortality (as sanitation improves, mortality rates decrease for neonates, infants, and children under 5). Poor sanitation conditions increased the risk of sharing life-threatening contagions. We also see that sanitation has a relatively weak relationship to maternal death risk (r = 0.53).

Recommendations

The above analysis shows that sanitation is one of the key determinants of life expectancy, with a Pearson’s correlation coefficient of 0.86. Moreover, we see that sanitation has a strong negative correlation with childhood mortality.

When evaluating potential high-impact investment areas for a global non-profit, improving access to sanitation stands out as an especially high-potential area. Not only is increased access to sanitation facilities correlated with increased life expectancy, but it is also associated with decreased mortality in children. According to the World Health Organization, diarrhoeal disease, the leading cause of death for children under five, is largely spread by poor sanitation conditions: *‘Diarrhoea is a symptom of infections caused by a host of bacterial, viral and parasitic organisms, most of which are spread by faeces-contaminated water.’*

Therefore, sanitation issues should likely be addressed first. Once sanitation issues are addressed in communities, issues of secondary importance like education, health care, adolescent fertility (likely linked to education and healthcare) could be addressed. Because sanitation has a high correlation with childhood mortality, an initial focus on sanitation should increase the survival or young children (meaning more children who would be addressed by education initiatives).