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 healthcare 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 dataset is available here.

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: Disease prevention, Health services, Mortality, Nutrition, Dynamics, Structure, Reproductive 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 machine learning, the dataframe needs to be reshaped 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. First 'country' and 'Indicators' dataframes were merged 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, the dataframe was tranformed into a more workable state by unstacking the dataframe with pivot_table().

Inferential statistics & analysis

Exploratory data analysis

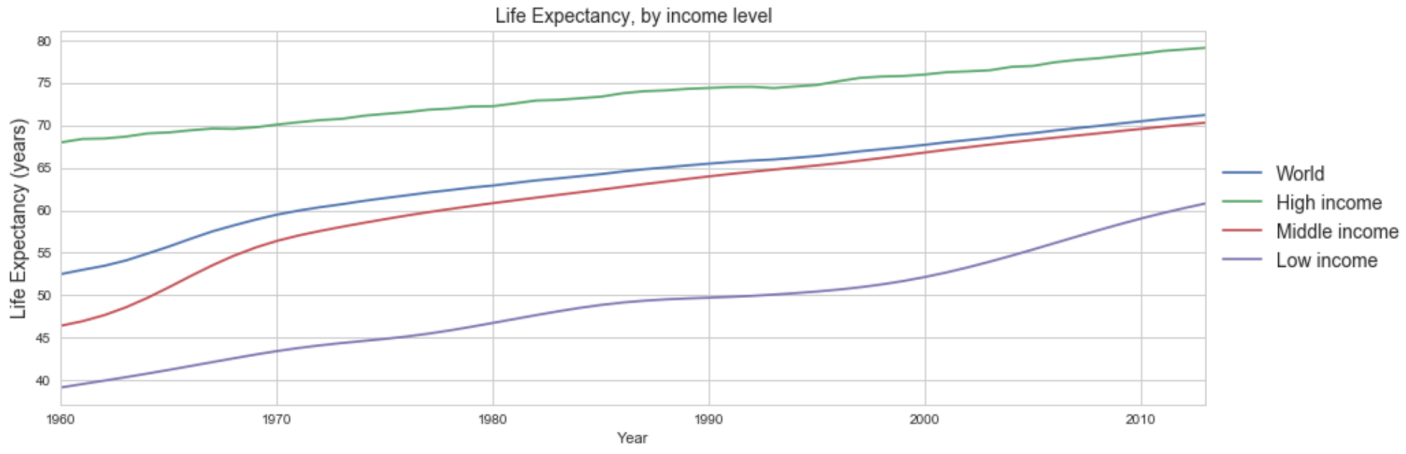
First, let's explore the commonly-assumed relationship between life expectancy and economic prosperity. The World Bank data already includes the following classifications which we can leverage:

Classification	Definition: 2015 GNI per capita
Low income group aggregate	$1,025 or less
Lower middle income group aggregate	between $1,026 and $4,035
Middle income group aggregate	between $1,026 and $12,475
Upper middle income group aggregate	between $4,036 and $12,475
High income group aggregate	$12,476 or more

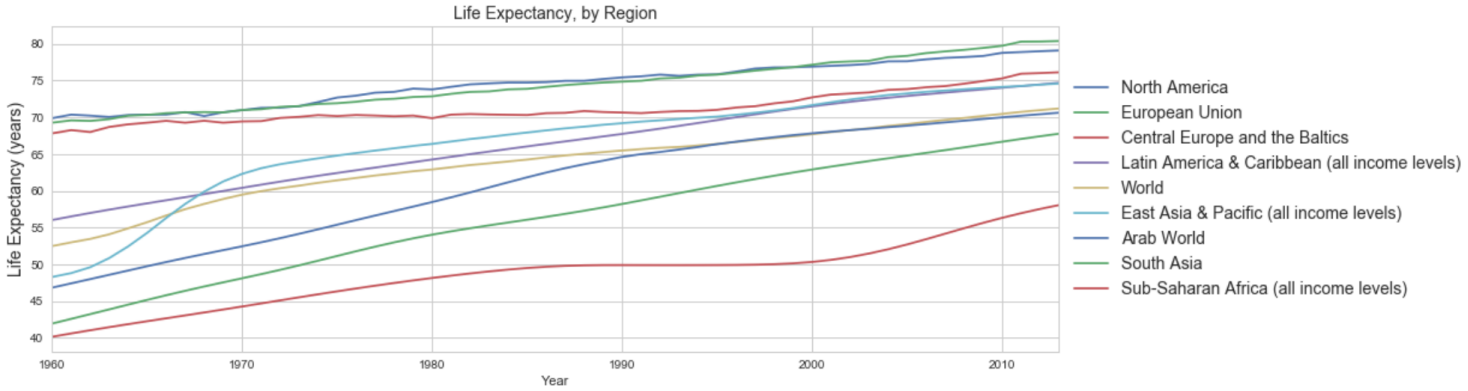
Using the income level classifications in the data, we plot the life expectancy trends from 1960 to 2015 for the high, middle, and low income, as well as the life expectancy trends for the world. In the graph below, we can see a very large and steady different between in life expectancy by income level. 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:

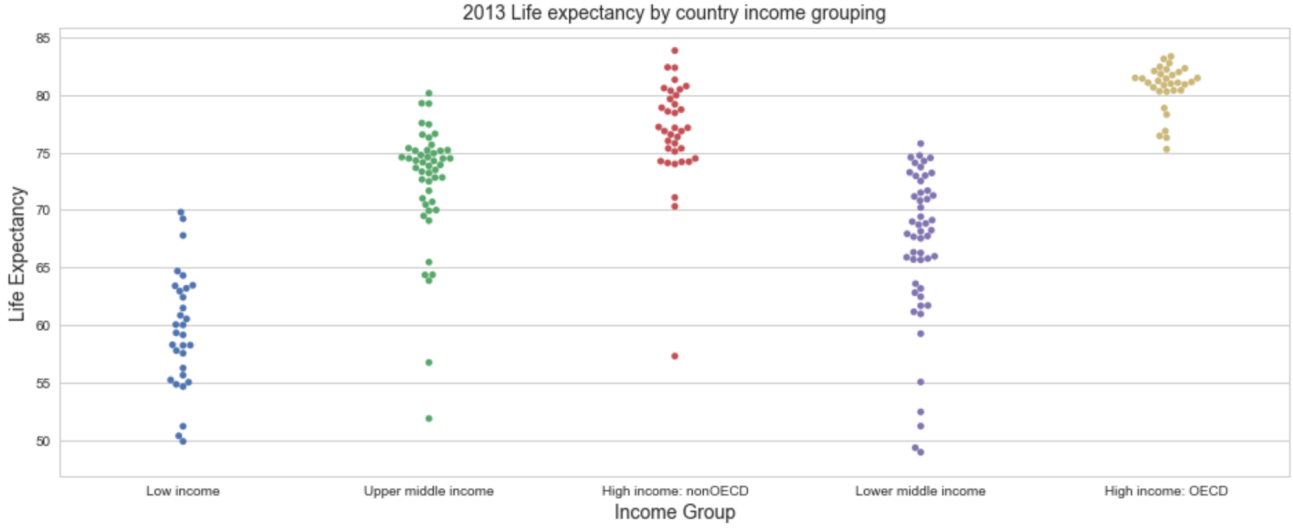
A. The rate of increase in life expectancy for the high income countries has slowed down. This is perhaps not surprisingly, since the high income countries started with a much higher life expectancy.
B. There is a decent uptick in life expectancy for the middle income countries in the mid- to late-60s.
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 plateau 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, life expectancy also increases, and the distribution of life expectancy tights up (lower standard deviation). There is one notable outlier in the "High income: nonOECD" countries, which has a life expectancy of 57.3 years; on closer inspection this data point is Equatorial Guinea.



Below are the summary statistics for life expectancy by income grouping (reflecting the same discussion as above):

IncomeGroup	count	mean	std	min	25%	50%	75%	max
High income: OECD	32.0	80.6	2.0	75.3	80.4	81.1	81.8	83.3
High income: nonOECD	36.0	76.8	4.6	57.3	74.9	77.0	79.7	83.8
Low income	31.0	59.5	5.0	49.9	55.9	59.3	63.1	69.8
Lower middle income	51.0	66.9	6.7	48.9	63.4	68.1	71.6	75.8
Upper middle income	48.0	72.5	5.3	51.9	70.9	74.0	75.1	80.1

Machine Learning: Reshaping the dataframe and EDA

Narrowing down the predictors of 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

Considering these criteria, the following indicators were identified as spanning 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

Additionally, there are a few key mortality-related indicators:

- 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)

However, since the above indicators are tightly coupled with overall life expectancy, they will be excluded from the model. We will reference them again to understand *how* certain features influence life expectancy (e.g. is a lower expectancy lower correlated to a higher risk of maternal death? Higher mortality rate in newborns, infants, or young children?).

Pearson's correlation & heat map

To understand the relationship life expectancy and each of the potential features, Pearson's correlation is used, as shown to the right. We can visualize the strength of the correlation using a heatmap, included below for reference.

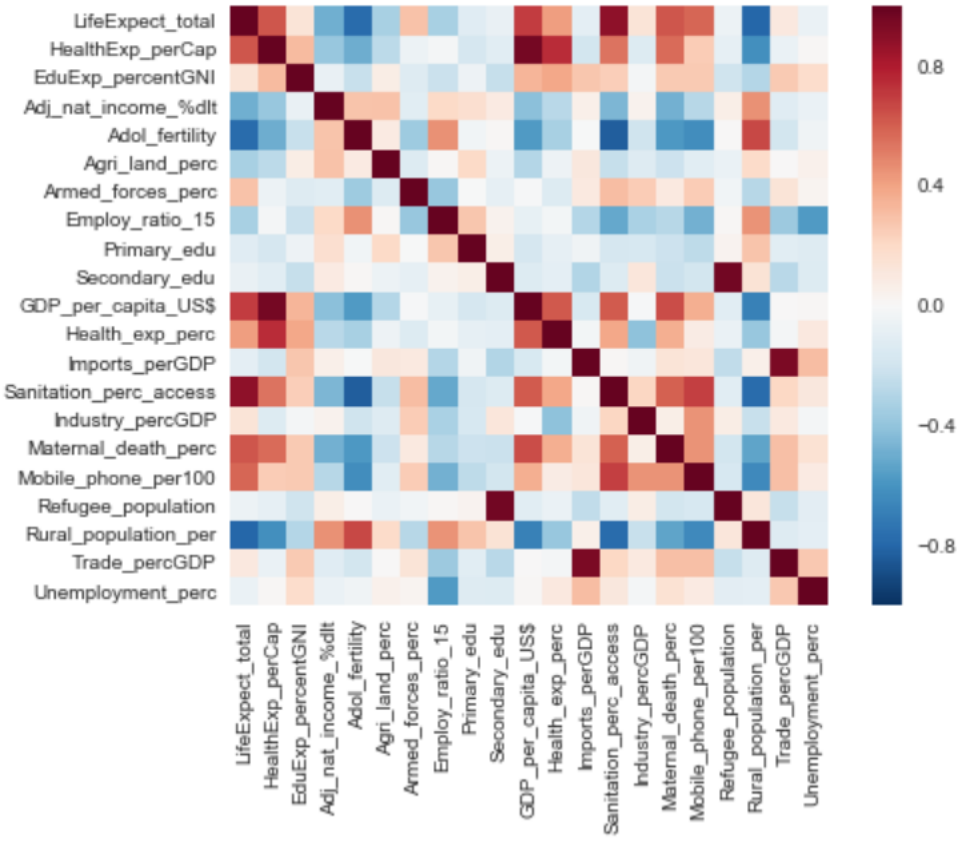
	LifeExpect_total
LifeExpect_total	1.0
HealthExp_perCap	0.63
Adol_fertility	-0.78
GDP_per_capita_US$	0.7
Sanitation_perc_access	0.88
Rural_population_per	-0.8
EduExp_percentGNI	0.14
Adj_nat_income_%dlt	-0.49
Agri_land_%	-0.34
Arable_land_%	-0.018
Armed_forces_%	0.29
Employ_ratio_15	-0.33
Primary_edu	-0.12
Secondary_edu	-0.074
Health_exp_%	0.42
Imports_perGDP	-0.098
Industry_percGDP	0.14
Mobile_phone_per100	0.58
Pop_density	0.083
Refugee_population	-0.06
Trade_percGDP	0.11
Unemployment_%	-0.07

As expected there is a positive correlation between life expectancy and health expenditure per capita (r = 0.63), GDP per capita (r = 0.698), access to sanitation (r = 0.88). There is a strong negative correlation with percentage of the population living in rural areas (r = -0.80) as well as adolescent fertility (r = -0.77).

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.

Using a threshold of |r> 0.5|, we select these five indicators as the initial features for the model:

- Health expenditure per capita
- GDP per capita
- Access to sanitation
- Percentage of the population living in rural areas
- Adolescent fertility



Machine Learning: Predicting life expectancy with linear regression

We'll use linear regression to predict life expectancy (a continuous variable) from a number of independent variables (development indicators).

Linear regression takes the form $y_i = \beta_0 + \beta_1 X_1 + ... + \beta_i X_i$ where

y is the dependent variable (in this case, Life Expectancy) which we will predict

X are the regressors (development indicators)

β_0 is the intercept, and β_i are the regression coefficients.

Linear regression models assume that the relationship between the dependent variable (y) and the regressors (X) are linear. The model fits the data such that the error is minimized; in this case, we'll use the least squares method.

Below we'll observe a few variations of the model:

a. Linear regression on all 21 predictors of life expectancy
b. Multivariate regression with a few select predictors, and
c. Regression with a single predictor

For all scenarios, we'll use one of two built-in statistical packages for linear regression: LinearRegression imported from sklearn.linear_model, and ols imported from statsmodels.formula.api. Both models require instantiating a model, then fitting the model to the dataset:

A. Linear regression on all 21 predictors

Initially, we can run a linear regression model with the top 21 features that were identified earlier as possible predictors of life expectancy. As a reminder, these features were selected to present a well-populated and diverse representation of the available development indicators. In the feature-selection step, we sought out indicators that were highly-populated across the countries; we now drop any row that is missing any one of the features, using dropna(how = 'any'). The resulting dataframe contains 96 countries, each of which has a value for all features in X.

Here we'll use statsmodels' OLS regression because of the amount of initial insights this model provides. The OLS Regression Results table below provides tremendous information. Most importantly, let's focus on the p-values for each feature (column 'P>|t|'). The P-value indicates the statistical significance of each feature; a very small P-value (close to zero) indicates statistically significant features. From this, we can see that Access to Sanitation is the most statistically-significant predictor of life expectancy, followed by Percentage of Rural Population and Unemployment.

OLS Regression Results

Dep. Variable:	y	**R-squared:**	0.897
Model:	OLS	**Adj. R-squared:**	0.868
Method:	Least Squares	**F-statistic:**	30.78
Date:	Fri, 01 Dec 2017	**Prob (F-statistic):**	2.40e-28
Time:	13:46:50	**Log-Likelihood:**	-227.84
No. Observations:	96	**AIC:**	499.7
Df Residuals:	74	**BIC:**	556.1
Df Model:	21		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[95.0% Conf. Int.]
Intercept	70.4854	6.609	10.665	0.000	57.317 83.654
X[0]	-0.0009	0.001	-1.202	0.233	-0.002 0.001
X[1]	-0.0164	0.019	-0.871	0.387	-0.054 0.021
X[2]	0.0001	7.72e-05	1.859	0.067	-1.03e-05 0.000
X[3]	0.1549	0.026	6.010	0.000	0.104 0.206
X[4]	-0.1010	0.030	-3.362	0.001	-0.161 -0.041
X[5]	-0.6068	0.264	-2.301	0.024	-1.132 -0.081
X[6]	-0.0621	0.103	-0.605	0.547	-0.266 0.142
X[7]	-0.0316	0.024	-1.332	0.187	-0.079 0.016
X[8]	-0.0392	0.035	-1.132	0.261	-0.108 0.030
X[9]	0.1783	0.586	0.304	0.762	-0.989 1.346
X[10]	-0.0173	0.042	-0.414	0.680	-0.100 0.066
X[11]	0.0293	0.034	0.864	0.390	-0.038 0.097
X[12]	1.143e-08	1.15e-08	0.992	0.324	-1.15e-08 3.44e-08
X[13]	0.2092	0.241	0.867	0.389	-0.272 0.690
X[14]	-0.0131	0.082	-0.160	0.873	-0.176 0.150
X[15]	-0.1093	0.066	-1.648	0.103	-0.241 0.023
X[16]	-0.0087	0.017	-0.521	0.604	-0.042 0.025
X[17]	0.0005	0.004	0.141	0.888	-0.007 0.008
X[18]	-4.56e-07	4.77e-07	-0.956	0.342	-1.41e-06 4.95e-07
X[19]	0.0151	0.044	0.345	0.731	-0.072 0.102
X[20]	-0.2584	0.086	-3.000	0.004	-0.430 -0.087

Omnibus:	2.713	**Durbin-Watson:**	1.967
Prob(Omnibus):	0.258	**Jarque-Bera (JB):**	2.456
Skew:	-0.096	**Prob(JB):**	0.293
Kurtosis:	3.759	**Cond. No.**	2.78e+09

B. Multivariate regression with a few select predictors

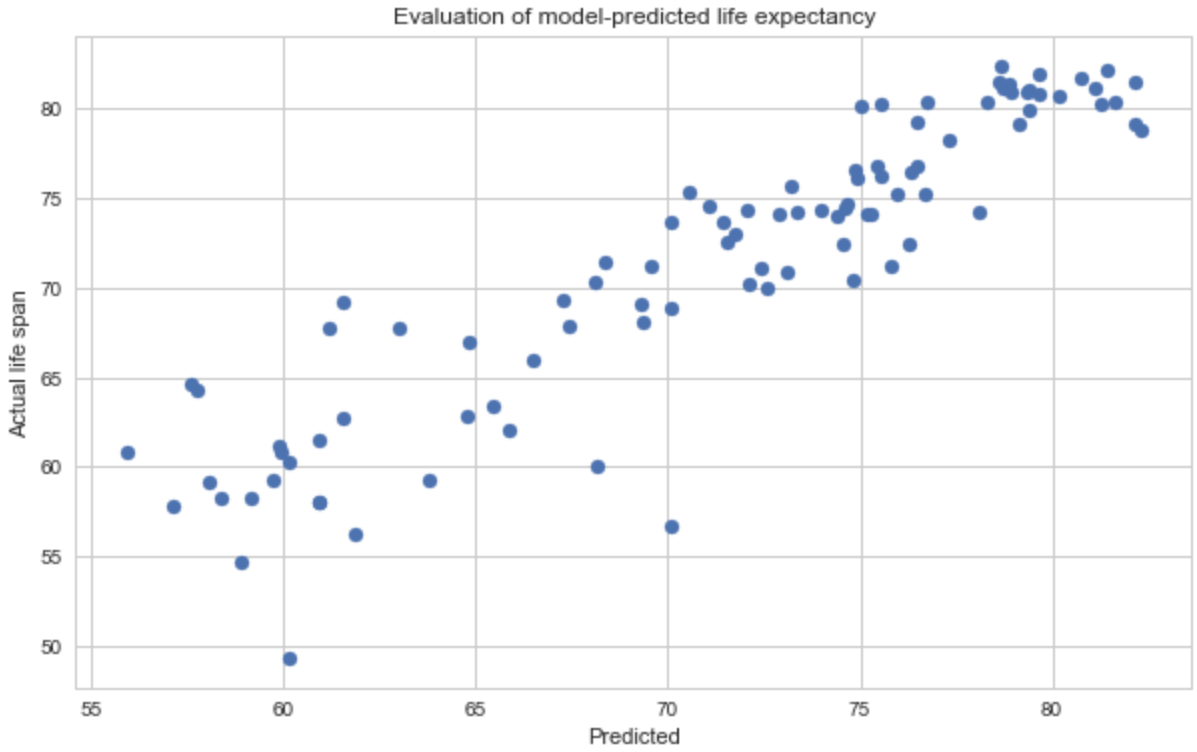
Using the output of the initial model, we can determine which features are strong predictors for a more focused model. Specifically, we'll use the P value as an indicators of the predictor's statistical significance. The lowest P values are for Sanitation (0.000), and Rural population% (0.001).

After fitting a LinearRegression model to the data, we see an intercepts of 62.87 years, and the following coefficients:

	features	estimatedCoefficients
0	Sanitation	0.183838
1	Rural_pop%	-0.117126

This means that on average, life expectancy is 63 years and increases by one year for every 0.18% increase in access to sanitation and 0.11% decrease in percent of population living in rural areas. To train the model, the data is split into training and test sets, with a split of 70:30 (training:test). Cross validation with 5 folds shows our model predicts life expectancy with 80.4% accuracy.

Visualizing the model's predicted values against the actual values (below), we see that the model tends to be better at predicting higher life expectancy. For lower life expectancy, the model does not perform as well, as indicated by the wider spread of data points.



HealthSpend and GDP were also highly correlated with life expectancy. We can check how they impact the model, and because they are very highly correlated with one another (r = 0.96), we should add just one of these features to the model. Adding HealthSpend, we see the accuracy is pretty much unchanged, increasing slightly to 81.32%.

C. Regression with a single predictor

If we look just at the single predictor with the highest statistical significant (Sanitation), we see that the Regression coefficient is 0.2466, meaning on average each 0.25% increase in the percent of the population with access to modern sanitation facilities, is associated with a one-year increase in life expectancy. With Sanitation alone, the model is able to predict life expectancy with a 77% accuracy.

This shows that we are able to predict life expectancy relatively well, and crucially, we know which metric(s) are the key determinants of life expectancy: sanitation.

Just Model analysis

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 predictors 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 high-impact investment areas for MoniE Bank's global non-profit, improving access to sanitation stands out as an especially high-potential area. Not only does increased access to sanitation correlate 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, MoniE Banks should initially focus on investing in sanitation. 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).